Schätzen und Testen I

Wintersemester 2014/15

Skript zur Vorlesung von

Sonja Greven
Christian Heumann

nach einer Vorlage von

Christiane Fuchs (geb. Dargatz)
Ludwig Fahrmeir
Volker Schmid

2. Oktober 2014

Verbesserungen und Anregungen ausdrücklich erwünscht
an David.Ruegamer@stat.uni-muenchen.de!

Inhaltsverzeichnis

Kapitel 1

Einführung in statistische Modelle und Inferenzkonzepte

Ziele:

- Statistische Modelle im Überblick, von einfachen hin zu komplexeren Modellen. Auswahl orientiert an Datenstrukturen, Modellklassen und Fragestellungen aus dem Bachelorprogramm und darüber hinaus.
- Problemstellungen der zugehörigen statistischen Inferenz.
- Konzepte statistischer Inferenz im Überblick.

1.1 Statistische Modelle

1.1.1 Einfache Zufallsstichproben

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

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.

Beispiel 1.1.

1. *$X \sim N(\mu, \sigma^2)$; Schätzen/Testen von μ, zum Beispiel Gauß-Test, Student-Test, F-Test für σ^2.*

2. *Analoge Problemstellungen für* $X \sim Bin(n,\pi)$, $X \sim Po(\lambda), \ldots$ *bzw. allgemein*

$$X \sim \text{lineare Exponentialfamilie mit natürlichem (skalarem) Parameter } \theta.$$

3. $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ *mehrdimensional, zum Beispiel* $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

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 $f_0(z)$.

Beispiele:

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

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

- $X \sim DE(a,b)$ *(Laplace- oder Doppelexponentialverteilung) mit Parametern* $a \in \mathbb{R}$ *und* $b > 0$*:*

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

- $X \sim U(a,b)$ *(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.

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

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

5. ***Exponentialfamilien:***

Definition 1.1 (Exponentialfamilien)**.** *Eine Verteilungsfamilie heißt* Exponentialfamilie $\overset{def}{\Leftrightarrow}$

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

mit $h(x) \geq 0$ *und*

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

$\gamma_1, \ldots, \gamma_k$ *heißen die* natürlichen *oder* kanonischen Parameter *der Exponentialfamilie (nach Reparametrisierung von* $\boldsymbol{\theta}$ *mit* $\boldsymbol{\gamma}$*).*

Annahme*:* $1, \gamma_1, \ldots, \gamma_k$ *und* $1, T_1(x), \ldots, T_k(x)$ *sind linear unabhängig, d.h.* f *ist* strikt k-parametrisch*.*

Beispiel 1.2 (Bernoulli-Experiment)**.** $X = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} Bin(1, \pi)$.

$$\begin{aligned} f(x|\pi) &= \pi^{\sum_{i=1}^n x_i}(1-\pi)^{n-\sum_{i=1}^n x_i} \\ &= \exp\left(\sum_{i=1}^n x_i \log(\pi) + \left(n - \sum_{i=1}^n x_i\right)\log(1-\pi)\right) \\ &= \underbrace{1}_{h(x)}\exp\left(\underbrace{n\log(1-\pi)}_{b(\pi)} + \underbrace{\sum_{i=1}^n x_i}_{T_1(x)}\underbrace{\log\left(\frac{\pi}{1-\pi}\right)}_{\gamma_1(\pi)}\right) \\ &= \underbrace{1}_{h(x)}\underbrace{(1-\pi)^n}_{c(\pi)}\exp(\gamma_1(\pi)T_1(x)), \end{aligned}$$

d.h. es liegt eine einparametrische Exponentialfamilie vor mit

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

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

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

bzw. (mit Dispersionsparameter ϕ*) falls*

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

6. ***Mischverteilungen**:*

$$X \ \sim \ \pi_1 f_1(x|\vartheta_1) + \ldots + \pi_k f_k(x|\vartheta_k)$$

mit $\pi_1 + \ldots + \pi_k = 1$*, wobei die* π_i *als* Mischungsanteile *und die* $f_i(x|\vartheta_i)$ *als* Mischungskomponenten *bezeichnet werden. Genauer spricht man von* diskreter Mischung.

Beispiel 1.3.

$$X \ \sim \ \pi_1 \phi(x; \mu_1, \sigma_1^2) + \ldots + \pi_k \phi(x; \mu_k, \sigma_k^2)$$

wird Normalverteilungsmischung *genannt.*

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

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
- $X \sim f(x)$, 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.

1.1.2 Lineare und generalisierte lineare parametrische Modelle

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

Klassisches lineares Modell (LM)

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

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

Generalisierte lineare Modelle (GLM)

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

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

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

Beispiel 1.4. *Sei* $y_i|\boldsymbol{x}_i \in \{0, 1\}$ *und*

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

Beispiele für h *sind die Verteilungsfunktion der logistischen Verteilung (→ Logit-Modell) oder die Verteilungsfunktion der Normalverteilung (→ Probit-Modell).*

Die Inferenzprobleme im GLM sind wie im linearen Modell, jedoch ist likelihoodbasierte oder bayesianische Inferenz möglich.

Beachte: Die $y_i|\boldsymbol{x}_i$ sind nicht identisch verteilt.

1.1.3 Nicht- und semiparametrische Regression

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

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

Additive Modelle (AM)

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

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

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

Generalisierte Additive Modelle (GAM)

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

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

zu.

1.1.4 Quantil-Regression/Robuste Regression

Datenlage wie bei üblicher Regression: $(y_i, \boldsymbol{x}_i),\ i = 1, \ldots, n,\ y_i|\boldsymbol{x}_i$ bedingt unabhängig.

Ziel: Schätze nicht (nur) $\mathbb{E}[y_i|\boldsymbol{x}_i]$, zum Beispiel durch KQ-Schätzer $\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{\text{KQ}}$, sondern den bedingten Median ($\tau = 0.5$) oder allgemeiner die (bedingten) Quantile $Q_\tau(y_i|\boldsymbol{x}_i)$. Statt KQ-Ansatz (ohne Normalverteilungsannahme) und Schätzung von $\widehat{\boldsymbol{\beta}}_{\text{KQ}}$, so dass $\sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$ minimiert wird, suchen wir

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

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

Wichtig dabei: keine Annahme für die Fehlerverteilung, d.h. „verteilungsfreier Ansatz".

Frage: Welche Konzepte zum Schätzen und Testen verwenden? → Quasi-Likelihood-Methoden.

1.1.5 Verweildaueranalyse: Cox-Modell

Grundlegender Begriff: *Hazardrate* $\lambda(t)$ einer stetigen Lebensdauer $T \geq 0$.

Definition 1.2 (Hazardrate)**.**

$$\begin{aligned} \lambda(t) &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} \\ \Leftrightarrow \quad \mathbb{P}(t \leq T \leq t + \Delta t | T \geq t) &= \lambda(t)\Delta t + o(\Delta t) \end{aligned}$$

(Dabei ist $f(x) = o(g(x))$ für $x \to 0$ falls $\lim_{x \to 0} f(x)/g(x) = 0$.)

Interpretation: $\lambda(t)\Delta t \approx$ bedingte Wahrscheinlichkeit für Ausfall in $[t, t + \Delta t]$ gegeben Überleben bis zum Zeitpunkt t bei „kleinem" Δt. Mit Kovariablen $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$:

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

Rechtszensierte Survivaldaten

Verwende $t_1, \ldots, t_n$ für evtl. rechtszensierte Beobachtungen von unabhängigen Lebensdauern $T_1, \ldots, T_n$, $\delta_1, \ldots, \delta_n$ als Zensierungsindikatoren und $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ als zugehörige Kovariablen.

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

Cox-Modell

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

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

Dabei ist $\lambda_0(t)$ die von i bzw. $\boldsymbol{x}_i$ unabhängige „Baseline"-Hazardrate. $\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})$ modifiziert $\lambda_0(t)$ multiplikativ.

Primäres Interesse: Schätzen/Testen von $\boldsymbol{\beta}$ wie im LM oder GLM; $\lambda_0(t)$ wird als Nuisanceparameter (bzw. -funktion) betrachtet.

$\Rightarrow$ Die Likelihood faktorisiert sich in

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

$L_1(\boldsymbol{\beta})$ ist *partielle („partial") Likelihood*, die bezüglich $\boldsymbol{\beta}$ maximiert wird. Erstaunlicherweise ist der Informationsverlust gering. Ferner gibt es einen Zusammenhang zwischen Partial-Likelihood und dem Konzept der Profil-Likelihood.

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.

1.1.7 Konditionale (autoregressive, Markov-) Modelle für Longitudinaldaten

- **Longitudinaldaten:** (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.
- **Autoregressives Modell 1. Ordnung bzw. Markov-Modell 1. Ordnung**: Bedingte Verteilung von $y_{ij}|y_{i,j-1}, y_{i,j-2},\ldots,y_{i1},\boldsymbol{x}_{ij}$ ist $y_{ij}|y_{i,j-1},\boldsymbol{x}_{ij}$, zum Beispiel

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

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

1.1.8 (Generalisierte) Lineare gemischte Modelle für Longitudinaldaten

Lineares gemischtes Modell (LMM)

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \gamma_{0i} + \gamma_{1i} t_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n;$$

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

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

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

1.1.9 Marginale Modelle

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

1.1.10 Modellbasierte Clusteranalyse

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

  $$f(\boldsymbol{x}) = \sum_{k=1}^{g} p(k) f(\boldsymbol{x}|\theta_k) \ ,$$

  zum Beispiel f Dichte der multivariaten Normalverteilung.
- Gesucht:
  1. Schätzungen für $\theta_k, p(k),\ k = 1, \ldots, g$.
  2. Schätzungen für unbekannte Klassenzugehörigkeit k eines Objekts mit beobachtetem Merkmalsvektor $\boldsymbol{x}$. Anwendung der Formel von Bayes liefert:

     $$\widehat{p}(k|\boldsymbol{x}) = \frac{\widehat{p}(k) f(\boldsymbol{x}|\widehat{\theta_k})}{\widehat{f}(\boldsymbol{x})}.$$
- Likelihood: mit EM-Algorithmus
- Bayes: mit MCMC-Algorithmus

1.1.11 Modelle mit latenten Variablen

Beobachtet werden Werte $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ij}, \ldots, y_{ip})^\top$ von p (korrelierten) Variablen, die als Indikatoren für eine latente, unbeobachtete Variable l_i (oder eine kleine Zahl von latenten Variablen) dienen. Primäres Ziel ist die Schätzung der Effekte („Ladungsfaktoren") λ_j von l auf den Vektor $\boldsymbol{y}$ der Indikatoren, die Schätzung der latenten Werte l_i, $i = 1, \ldots, n$, und die Schätzung der festen Effekte $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}$ im folgenden Modell.

1. Beobachtungsmodell:

$$y_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \lambda_j l_i + \varepsilon_{ij} \qquad \text{mit} \quad \varepsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2), \quad j = 1, \ldots, p$$

2. Strukturmodell:

$$l_i = \boldsymbol{u}_i^\top \boldsymbol{\gamma} + \delta_i \qquad \text{mit} \quad \delta_i \overset{i.i.d.}{\sim} N(0, 1)$$

Ohne Kovariablen $\boldsymbol{x}$ und $\boldsymbol{u}$ ergibt sich das klassische Modell der Faktorenanalyse. Erweiterungen entstehen zum Beispiel durch kategoriale Indikatoren oder nichtlineare Effekte von Kovariablen.

1.2 Konzepte der statistischen Inferenz

- $x = (x_1, \ldots, x_n)^\top$ oder $y = (y_1, \ldots, y_n)^\top$ sind Realisierungen von Stichprobenvariablen (Zufallsvariablen) $X = (X_1, \ldots, X_n)^\top$ oder $Y = (Y_1, \ldots, Y_n)^\top$. 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.
  - $Y_1, \ldots, Y_n$ sind abhängig, zum Beispiel zeitlich oder räumlich korreliert.
- In allen Fällen gilt: $x \in \mathcal{X}$ bzw. $y \in \mathcal{Y}$, wobei $\mathcal{X}$ bzw. $\mathcal{Y}$ der entsprechende Stichprobenraum ist. $X = (X_1, \ldots, X_n)^\top$ und $Y = (Y_1, \ldots, Y_n)^\top$ sind auf dem Stichprobenraum nach einer gemeinsamen Verteilung $\mathbb{P}$ bzw. Verteilungsfunktion $F(x) = F(x_1, \ldots, x_n)$ verteilt. $\mathbb{P}$ (bzw. F) gehört einer Menge (oder Klasse oder Familie) von Verteilungen $\mathcal{P}_\theta = \{\mathbb{P}_\theta : \theta \in \Theta\}$ an. Zugehörige Verteilungsfunktionen sind $F(x|\theta)$ bzw. (falls existent) Dichten $f(x|\theta) = f(x_1, \ldots, x_n|\theta)$.
  - I.i.d. Fall:

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

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

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

    die Dichten hängen also vom Index i ab.

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

- (Übliche) **parametrische** Inferenz:

  $\theta = (\theta_1, \ldots, \theta_k)^\top \in \Theta \subseteq \mathbb{R}^k$, 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.

- **Semiparametrische** Inferenz:

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

  Auch: $k > n$, zum Beispiel bei GLMs mit Genexpressionsdaten als Kovariablen: Daten $x_1, \ldots, x_k$ mit $k \sim 1000 - 10000$, bei nur $n \sim 50$ Patienten! Vergleiche multiples Testen in Kapitel 2.

1.2.1 Klassische parametrische Inferenz

$X = (X_1, \ldots, X_n)$ besitzt Verteilung/Dichte $\mathbb{P} \in \mathcal{P} = \{\mathbb{P}_\theta : \theta = (\theta_1, \ldots, \theta_k)^\top \in \Theta\}$ 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

$$f(x|\theta) = f(x_1, \ldots, x_n|\theta).$$

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

- **Punktschätzung**: Geschätzt werden soll θ. Eine messbare Abbildung

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

  heißt *Schätzfunktion* oder *Schätzer*. Eine Beurteilung der Güte/Optimalität kann zum Beispiel durch

  – $\text{Bias}_\theta(T) = \mathbb{E}_\theta[T] - \theta$,

  – $\text{Var}_\theta(T) = \mathbb{E}_\theta[(T - \mathbb{E}_\theta[T])^2]$,

  – $\text{MSE}_\theta(T) = \mathbb{E}_\theta[(T - \theta)^2] = \text{Var}_\theta(T) + (\text{Bias}_\theta(T))^2$

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

- **Bereichsschätzung / Intervallschätzung:**

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

so dass $\mathbb{P}_\theta(C(X) \ni \theta) \geq 1-\alpha$ für alle $\theta \in \Theta$. Dabei ist $1-\alpha$ der *Vertrauensgrad* (auch: Konfidenzniveau oder Überdeckungswahrscheinlichkeit) des *Konfidenzbereiches*. Man beachte die frequentistische/Häufigkeitsinterpretation: $C(X)$ ist ein *zufälliger* Bereich. Ist $\Theta \subseteq \mathbb{R}$ und $C(x)$ für alle x ein Intervall, dann heißt C *Konfidenzintervall*.

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

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

wobei $\Theta_0 \cap \Theta_1 = \emptyset$. Es muss nicht notwendigerweise $\Theta = \Theta_0 \cup \Theta_1$ gelten.

Ergebnisse/Aktionen:

$$A_0 : H_0 \text{ wird nicht abgelehnt,}$$
$$A_1 : H_1 \text{ wird bestätigt, „ist signifikant".}$$

Der Test ist eine Abbildung

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

Ein nicht-randomisierter Test hat die Form

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

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 $T(x)$:

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

Test zum Niveau („size") α, wobei α „klein":

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

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

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

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

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

„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*, $1 - g_\phi(\theta)$, $\theta \in \Theta_1$. 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:

→ Likelihood-Inferenz

→ Bayes-Inferenz

→ Nicht- und semiparametrische Inferenz

Im einfachsten Fall von zwei Punkthypothesen

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta = \theta_1$$

für $\theta_0 \neq \theta_1$ hat der „beste" Test Likelihood-Quotienten-Struktur: H_0 wird abgelehnt, falls

$$\frac{f(x|\theta_1)}{f(x|\theta_0)} > k_\alpha$$

(vgl. Neyman-Pearson Theorem).

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

  Beispiel 1.5 (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

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

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

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

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

  Definition 1.3 (p-Wert)**.** *Gegeben sei ein Test bzw. eine Teststatistik $T(X)$ für H_0 vs. H_1 mit*

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

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

1.2.2 (Parametrische) Likelihood-Inferenz

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

  $$L(\theta|x) := f(x|\theta)$$

  Likelihoodfunktion von θ zur Beobachtung x.

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

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

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

bzw.

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

- In der Regel existieren keine finiten Optimalitätseigenschaften, jedoch asymptotische.
- Testen: Likelihood-Quotienten-Test, Wald-Test, Score-Test.

1.2.3 Likelihoodbasierte Inferenz

Quasi-Likelihood-Inferenz, penalisierte Likelihood-Inferenz, semiparametrische Modelle.

1.2.4 Bayes-Inferenz

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

- Prinzip: Nach Beobachtung von $\boldsymbol{x}$ ist sämtliche Information über θ in der *Posteriori-dichte*

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

- Bayes-Schätzung:
  - Posteriori-Erwartungswert:

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

  - Posteriori-Median:

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

  - Posteriori-Modus:

$$T_{\mathrm{mod}}(x) = \widehat{\theta}_{\text{post-Mod}} = \operatorname*{argmax}_{\theta} p(\theta|\boldsymbol{x}) = \operatorname*{argmax}_{\theta} p(\theta) L(\theta|\boldsymbol{x})$$

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

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

die sich somit nicht normieren lassen. Allerdings muss die Posterioridichte eigentlich sein! Ein Spezialfall ist $p(\theta) \propto 1$ („Gleichverteilungs-Priori"), bei deren Verwendung

$$\widehat{\theta}_{\mathrm{ML}} = \operatorname*{argmax}_{\theta} L(\theta|\boldsymbol{x}) = \widehat{\theta}_{\text{post-Mod}}$$

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

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

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

  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. (→ Asymptotik der Bayes-Schätzer)

1.2.5 Statistische Entscheidungstheorie

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

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

Definition 1.4 (Entscheidungsfunktion)**.** *Als* Entscheidungsfunktion *bezeichnet man eine Funktion*

$$d : \left\{ \begin{array}{lcl} \mathcal{X} & \longrightarrow & D \\ x & \longmapsto & d(x). \end{array} \right.$$

Mit D wird der Entscheidungs- *oder* Aktionenraum *bezeichnet.*

Definition 1.5 (Verlustfunktion)**.** *Eine* Verlustfunktion *(oft auch stattdessen* Gewinnfunktion*)*

$$L : \left\{ \begin{array}{lcl} D \times \Theta & \longrightarrow & \mathbb{R} \\ (d, \theta) & \longmapsto & L(d, \theta) \end{array} \right.$$

ordnet einer Entscheidung $d(x)$ („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 1.6.

1. **Test***: Betrachte*

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

   (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:

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

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

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

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

Mögliche Verlustfunktionen:

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

wobei w eine feste Gewichtsfunktion ist.

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 1.6 (Risikofunktion)**.** *Eine* Risikofunktion *ist definiert als*

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

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

Beispiel 1.7.

1. **Schätzen**, *d.h.*

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

Bei quadratischer Verlustfunktion ist

$$L(T(X),\theta) = (T(X)-\theta)^2$$

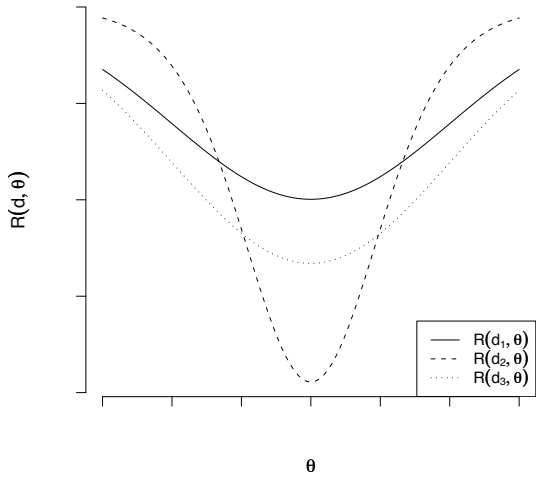
mit Risikofunktion

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

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

2. **Testen**: *vgl. Übung.*

Vergleich von Entscheidungsregeln/-strategien 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.

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.

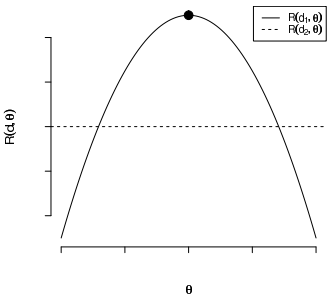
→ „Optimale" Entscheidungsregeln nur möglich durch:

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

1. **Minimax-Kriterium**

   Idee: Betrachte Maximum der Risikofunktion, d.h. präferiere in der folgenden Abbildung d_2, da

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



Definition 1.7 (Minimax-Entscheidungsregel)**.** *Sei $d^* : \mathcal{X} \to D$ eine Entscheidungsregel. d^* heißt* Minimax*, falls es das supremale Risiko minimiert:*

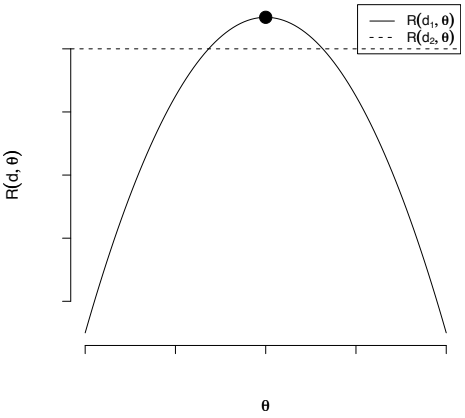
$$\sup_{\theta\in\Theta} R(d^*, \theta) \leq \sup_{\theta\in\Theta} R(d, \theta) \;\; \forall d \in D \;\Leftrightarrow\; d^* = \operatorname*{arginf}_{d\in D} \sup_{\theta\in\Theta} R(d, \theta).$$

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

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

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:



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

2. **Bayes-Kriterium**

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

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

und wird durch den *Bayes-optimalen Schätzer* d^* minimiert:

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

Sei $p(\theta|x)$ (eigentliche) Posterioridichte. Dann heißt

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

das *Posteriori-Bayes-Risiko*. Es gilt folgendes praktische Resultat:

Satz 1.8. *Eine Regel d^* ist genau dann Bayes-optimal, wenn d^* für jede Beobachtung/Stichprobe x das Posteriori-Bayes-Risiko minimiert.*

Anmerkungen:

- Bayes-optimale Regeln d^* sind *zulässig*, d.h. sie werden von keiner anderen Regel $d \neq d^*$ dominiert.
- Eine enge Beziehung zur Minimax-Regel ist durch die Wahl einer „ungünstigsten" Prioridichte $p^*(\theta)$ herstellbar.

Optimalität von Bayes-Schätzern:

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

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

$$\widehat{\theta} = \mathrm{med}(\theta|x)$$

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

$$\widehat{\theta} = \operatorname*{argmax}_{\theta\in\Theta} p(\theta|x)$$

ist Bayes-optimal bei 0-1 Verlustfunktion

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

Der Grenzübergang $\varepsilon \to 0$ liefert den Posteriori-Modus.

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

1.2.6 Weitere Inferenzkonzepte

- Struktur-Inferenz
- Fiduzial-Inferenz