STATISTICAL INFERENCE

Lecture at Ludwig-Maximilians-Universität München
Summer Term 2022

CHRISTIAN HEUMANN CHRISTIAN SCHOLBECK
NURZHAN SAPARGALI

Contents

Preface

What is this course about?

The course is about important statistical concepts needed in science and data science. It will cover

- Statistical inference for estimating data distributions or of parameters characterizing them and the quantification of uncertainty: classical, likelihood and Bayesian inference
- Bootstrap: variance estimation, confidence intervals, prediction error estimation
- Further topics: model selection and model averaging, copulas, statistics of extremes
- Deficient data: missing data, measurement error
- Causality

What is statistical inference?

In **statistical inference**, data is interpreted as observed values of random variables. The goal is to infer the underlying data generating process (DGP) of the random variables from the observations. If we have a dedicated population than the goal is to estimate the distribution or the parameters of the population (e.g. means, variances, quantiles) and to check hypotheses (e.g. in a randomized clinical trial, we want to test whether drug A is better than drug B).
Typical tasks are thus

- **estimating** the density / distribution of the data or of the parameters that characterize them,
- quantification of uncertainty e.g. via a **confidence interval**, to have interval estimates of the parameters,
- **testing of hypotheses** on the parameters of a population.

Different concepts of inference

There are various inference concepts. In this lecture, the focus is on

- Classical inference (chapter 2)
- Likelihood inference (chapter 3)
- Bayesian inference (chapter 4)
- Inference by the Bootstrap (chapter 5)

All other chapters build upon these chapters. Chapter 1 gives an overview of typical examples with respect to data structures and model assumptions where inference is required.

Chapter 1

Introduction to statistical models and inference concepts

1 Introduction to statistical models and inference concepts

Objectives of Chapter 1: Introduction and overview

1.1 – Overview of statistical models, from simple towards more complex models.

– Data structures, model classes and the associated problems of statistical inference.

1.2 Overview and comparison of statistical inference concepts.

1.1 Statistical models

1.1 Statistical models Overview

1.1.1 independently and identically distributed (i.i.d.) case

1.1.2-4 conditionally independent case, especially regression. Generalizations from the simple linear model to

1.1.2 a larger class of distributional assumptions (generalized linear models, GLM)

1.1.3 nonlinear covariable effects (additive models)

1.1.4 distribution-free approaches (quantile regression)

1.1.5 dependent case: longitudinal data as one example

1.1.6 missing and therefore incomplete data

1.1.1 Simple random samples

1.1 Statistical models 1.1.1 Simple random samples

One-sample case: let $x_1, \ldots, x_n$ be the observations as realizations of random (or sampling) variables $X_1, \ldots, X_n$ (where X_i denotes the i-th draw from the population) which are independently and identically distributed (i.i.d.) according to a random variable X with distribution function $F(x) = P(X \leq x)$ and (discrete, continuous or, in general, 'Radon-Nikodym'-) density $f(x)$.

1.1 Statistical models 1.1.1 Simple random samples

Example 1.1

- *continuous normal probability density function (pdf)* $\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right), x \in \mathbb{R}$
- *discrete Bernoulli probability mass function (pmf)* $\pi^x(1-\pi)^{(1-x)}, x \in \{0, 1\}$.
- *general density for* $X \in [0, \infty)$ *with* $P(X = 0) > 0$ *and continuous density function* f^* *for* $X > 0$ *(e.g. plant yield). Note: in this case,* $f(x) = f^*(x)$ *for* $x \in (0, \infty)$, $f(0) = p$ *is a density with respect to a measure which can be written as* $\lambda + \delta_0$, *where* λ *is the Lebesgue measure and* δ_0 *is the Dirac-measure.*

1.1.1. a) Parametric models

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

i.e. the density is completely defined by the distribution assumption (e.g. f is a normal distribution) and the k -dimensional parameter $\boldsymbol{\theta}$.

- As a rule k is fixed and small in relation to n.
- In practice, a model is often an approximation. For model misspecifications see section 3.4.

1.1 Statistical models
1.1.1 Simple random samples

Example 1.2

1. *$X \sim N(\mu, \sigma^2)$; Estimating / testing μ, for example Gauss-Test, t-Test; Estimate / test σ^2. Here, $\boldsymbol{\theta} = (\mu, \sigma)$.*
2. *$\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ multivariate, for example $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Estimating / testing $\boldsymbol{\mu}, \boldsymbol{\Sigma}$. Here, $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$.*
3. *Analogous problems for $X \sim$ simple linear exponential family (see definition 1.1) with natural (scalar) parameter θ, e.g. $X \sim Bin(N, \pi)$, $X \sim Po(\lambda), \ldots$ In these cases, $\boldsymbol{\theta} = \pi$ and $\boldsymbol{\theta} = \lambda$, respectively.*

1.1 Statistical models
1.1.1 Simple random samples

Definition 1.1 (Exponential Family)
A distribution family is called an exponential family $\overset{def}{\Leftrightarrow}$ *the density f has the form (remark: $\boldsymbol{x}$ can be multivariate)*

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

with $h(\boldsymbol{x}) \geq 0$ and

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

$\gamma_1, \ldots, \gamma_r$ *are called the* natural *or* canonical parameters *of the Exponential family (after reparameterization of $\boldsymbol{\theta}$ with $\boldsymbol{\gamma}$).*
Assumption:*: $1, \gamma_1, \ldots, \gamma_r$ and $1, T_1(\boldsymbol{x}), \ldots, T_r(\boldsymbol{x})$ are linearly independent, i.e. f is* strictly r-parametric.

1.1 Statistical models
1.1.1 Simple random samples

- *Remark 1: $T(\boldsymbol{x})$ is called a statistic, see definition 2.1*
- *Remark 2: The density can be factorized in a part which only depends on the data and a part which depends on the parameters and the statistic*
- *Remark 3: Assume a two-parameter exponential family that has two statistics T_1 and T_2 which are linearly dependent, e.g. $T_2(x) = 3 + 5T_1(x) \Rightarrow b(\theta) + \gamma_1(x)T_1(x) + \gamma_2(\theta)T_2(x) = (b(\theta) + 3\gamma_2(\theta)) + (\gamma_1(\theta) + 5\gamma_2(\theta))T_1(x)$. It follows that the exponential family is strictly one-parametric.*

1.1 Statistical models
1.1.1 Simple random samples

Example 1.3 (Bernoulli experiment)
Let π e.g. the percentage of persons with a certain disease.

$$\begin{aligned}
\boldsymbol{X} = (X_1, \ldots, X_n),\ X_i \overset{i.i.d.}{\sim} Bin(1, \pi), \boldsymbol{\theta} = \pi\ . \\
f(\boldsymbol{x}|\pi) &= \prod_{i=1}^{n} \pi^{x_i}(1-\pi)^{(1-x_i)} = \pi^{\sum_{i=1}^n x_i}(1-\pi)^{n-\sum_{i=1}^n x_i} \\
&= \exp\left(\sum_{i=1}^n x_i \log(\pi) + \left(n - \sum_{i=1}^n x_i\right)\log(1-\pi)\right) \\
&= \underbrace{1}_{h(\boldsymbol{x})} \exp\left(\underbrace{n\log(1-\pi)}_{b(\pi)} + \underbrace{\sum_{i=1}^n x_i}_{T_1(x)} \underbrace{\log\left(\frac{\pi}{1-\pi}\right)}_{\gamma_1(\pi)}\right) \\
&= \underbrace{1}_{h(\boldsymbol{x})} \underbrace{(1-\pi)^n}_{c(\pi)} \exp(\gamma_1(\pi)T_1(\boldsymbol{x}))\ .
\end{aligned}$$

1.1 Statistical models
1.1.1 Simple random samples

Example 1.4 (Example 1.3 continued)
The Binomial distribution is therefore a one-parameter exponential family with

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

Remark: *A distribution family is called a* simple linear exponential family*, if*

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

respectively (with dispersion parameter ϕ) if

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

1.1 Statistical models
1.1.1 Simple random samples

Example 1.5 (Example 1.1 continued)

4. ***Location and scale models:*** *A two-parameter family of distributions with*

$$F(x|a,b) = F_0\left(\frac{x-a}{b}\right)$$

with given distribution function $F_0(z)$, $\boldsymbol{\theta} = (a,b)$ with $a \in \mathbb{R}, b > 0$, is called location scale family.

a is the location parameter, b is the scale parameter.

Density in the continuous case:

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

with given density $f_0(z)$.

Examples of location and scale models:

- $X \sim N(a,b^2)$ (normal distribution), $f_0(z) = \phi(z)$ (density of the standard normal distribution N(0,1)).
$$\frac{1}{b} f_0\left(\frac{x-a}{b}\right) = \frac{1}{\sqrt{2\pi}b} \exp\left(-\frac{1}{2}\frac{(x-a)^2}{b^2}\right)$$
- $X \sim \mathrm{DE}(a,b)$ (Laplace or double exponential distribution):
$$\frac{1}{b} f_0\left(\frac{x-a}{b}\right) = \frac{1}{2b} \exp\left(-\frac{|x-a|}{b}\right) \quad , \quad f_0(z) = \frac{1}{2}\exp(-|z|)$$
- $X \sim U(a,b)$ (Uniform distribution) with $f_0(z) = I_{(-\frac{1}{2},\frac{1}{2})}(z)$:
$$\frac{1}{b} f_0\left(\frac{x-a}{b}\right) = \frac{1}{b} I_{(a-\frac{b}{2},a+\frac{b}{2})}(x) = \frac{1}{b} I_{(-\frac{1}{2},\frac{1}{2})}\left(\frac{x-a}{b}\right)$$
The support is a closed interval and depends on the parameters.

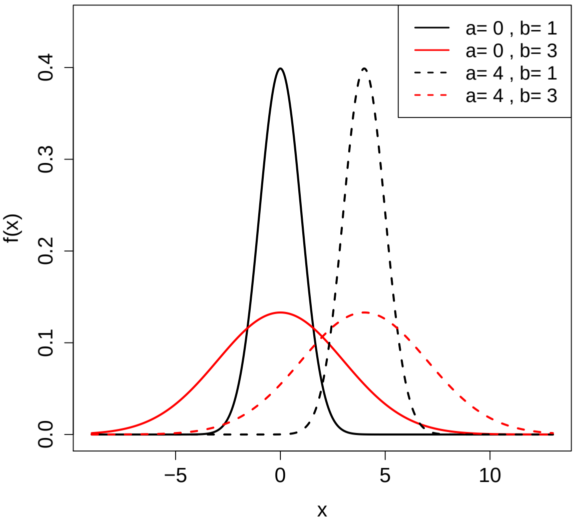


Figure 1: Location and Scale Models: Normal Distribution

Further examples of location and scale models:

- $X \sim C(a,b)$ (Cauchy distribution):
$$\frac{1}{b} f_0\left(\frac{x-a}{b}\right) = \frac{b}{\pi} \cdot \frac{1}{b^2 + (x-a)^2} = \frac{1}{b}\frac{1}{\pi}\frac{1}{1+(\frac{x-a}{b})^2}$$

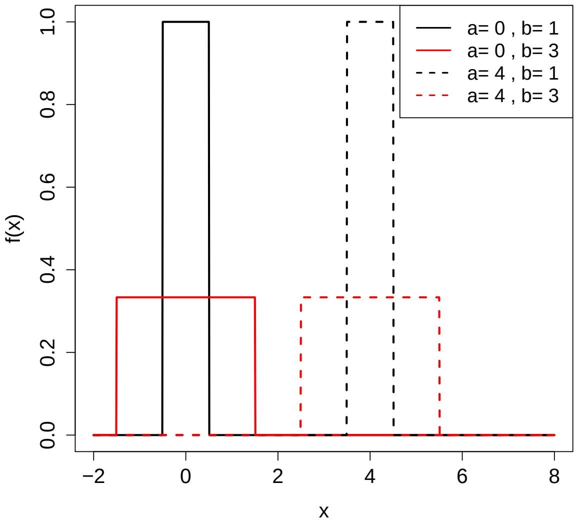


Figure 2: Location and scale models: Uniform Distribution

- $X \sim L(a,b)$ (logistic distribution):

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

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

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

For $a=0$, $\lambda = 1/b$: $f(x) = \lambda \exp\left(-\lambda x\right), x \geq 0$.

1.1 Statistical models
1.1.1 Simple random samples

Example 1.6 (Example 1.1 continued)

5. ***Mixture distributions:***

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

with $\pi_1 + \ldots + \pi_J = 1$, *where the* π_j *are the* mixture weights *and the* $f_j(x|\boldsymbol{\vartheta_j})$ *are called* mixture components. *One speaks more precisely of a* discrete mixture.

Example of a normal distribution mixture:

$$X \;\sim\; f(x|\boldsymbol{\theta}) = \pi_1 \phi(x|\mu_1, \sigma_1^2) + \ldots + \pi_J \phi(x|\mu_J, \sigma_J^2).$$

$\boldsymbol{\vartheta} = (\boldsymbol{\vartheta_1}, \ldots, \boldsymbol{\vartheta_J})$ *and* $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)$ *are unknown.* $\boldsymbol{\theta} = (\boldsymbol{\vartheta}, \boldsymbol{\pi})$ *are estimated by the maximum likelihood principle, often with the help of the EM algorithm, see 3.2.2. Also of interest: testing for the number of the mixture components J.*

1.1 Statistical models
1.1.1 Simple random samples

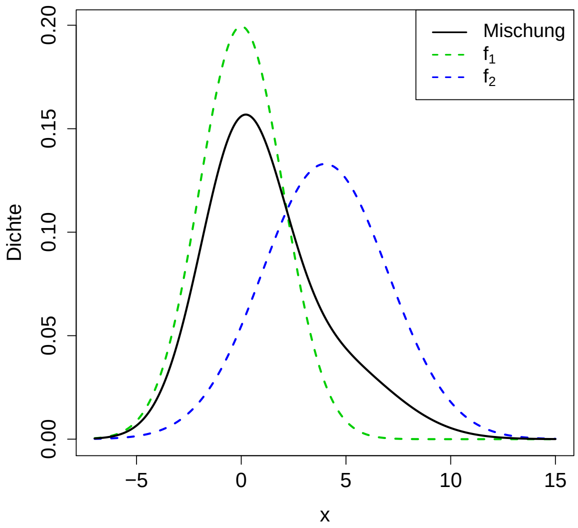


Figure 3: Mixture with $\pi_1 = 0.7, f_1(x) = \phi(x|0, 2^2)$, $\pi_2 = 0.3, f_2(x) = \phi(x|4, 3^2)$.

- *Remark 1: a mixture distribution can be interpreted as a model for heterogeneous populations, e.g. different subtypes of a disease in a population. In that case, the mixture weights correspond to the proportions of this subtypes in the population.*
- *Remark 2: from a sampling perspective, the mixture distribution results if we sample from the mixture components with probalities π_j.*
- *Remark 3: In cluster analysis, a Gaussian Mixture Model (GMM) using multivariate Gaussian mixture components is a popular model based approach, see also section 1.1.6b)*
- *Remark 4: The mixture can be also seen as a complex approximation of a non-standard density, e.g. a density with multiple modes. Then the components themselves have no subject related interpretation.*

In all parametric models in the previous examples, the statistical inference for the parameter $\boldsymbol{\theta}$ is of interest, especially estimating, testing and confidence intervals.

1.1 Statistical models
1.1.1 Simple random samples

1.1.1 b) Non Parametric Model/Inference

1.1 Statistical models
1.1.1 Simple random samples

- $X \sim F(x)$, X is continuous, F continuous distribution $\triangleright$ Kolmogorov-Smirnov-Test for $H_0 : F(x) = F_0(x)$. An empirical distribution is compared to a known reference distribution using the supremum of the differences.
- $X \sim F(x)$, X is discrete or grouped $\triangleright$ χ^2-goodness of fit test on $H_0 : F(x) = F_0(x)$. Comparison of observed numbers and (under F_0) expected numbers. Grouping in intervals may be necessary.
- $X \sim f(x)$, X (f continuous and differentiable except for a finite number of points) $\triangleright$ nonparametric density estimation, for example kernel density estimation.

- *Remark 1: The case of two or more samples can be treated analogously.*

- *Remark 2: Often, inference for the whole distribution or density is of interest, not only for certain quantities, such as expectations.*

1.1.2 Linear and generalized linear parametric models

1.1 Statistical models
1.1.2 Linear and generalized linear parametric models

Data $(y_i, \boldsymbol{x}_i),\ i = 1, \ldots, n$, are given with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$. $y_1|\boldsymbol{x}_1, \ldots, y_n|\boldsymbol{x}_n$ are (conditionally) independent, but not identically distributed.

1.1.2 a) Classical linear model (LM), OLS (ordinary least squares)

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

- *Remark 1: normality assumption is only necessary for tests, confidence intervals. Asymptotically ($n \to \infty$), it can be dropped.*

- *Remark 2: special cases of a linear model are the analysis of variance (ANOVA) and the analysis of covariance (ANCOVA).*

- Assumption: $p = \dim(\boldsymbol{\beta}) < n$ and n fixed, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$.

- Estimate $\boldsymbol{\beta}$ and σ^2, Tests for $\boldsymbol{\beta}$ with or without normality assumption.

- Variable selection and model selection, see 3.3.3.

1.1 Statistical models
1.1.2 Linear and generalized linear parametric models

1.1.2 b) Generalized linear models (GLM)

$y_i|\boldsymbol{x}_i,\ i = 1, \ldots, n$, have a density from a simple linear exponential family, for example normal, binomial, Poisson or gamma distribution, and are conditionally independent.

- *Remark: additional dispersion parameters required for normal and gamma*

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

Here, h is the *inverse link function (or response function).*

The inference problems in the GLM are similar to those in a linear model. Likelihood-based or Bayesian inference are possible.

- *Remark: again, the $y_i|\boldsymbol{x}_i$ are conditionally independent, but not identically distributed.*

1.1 Statistical models
1.1.2 Linear and generalized linear parametric models

Example 1.7

Let $y_i|\boldsymbol{x}_i \in \{0, 1\}$ and

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

Examples of h are the distribution function of the logistic distribution (→ logit model) or the distribution function of the normal distribution (→ probit model).

1.1.3 Non- and semi-parametric regression

1.1 Statistical Model
1.1.3 Non- and semi-parametric regression

1.1.3 a) Nonparametric regression

Data as in the linear model, but x_i only scalar, e.g. a dose-response curve.

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

The regression function $f(x_i) = \mathbb{E}[y_i | x_i]$ is *not* specified parametrically.

- Non- or semi-parametric estimation of f, e.g. LOESS (locally estimated scatterplot smoothing), or penalized splines.
- Testing of
$$\begin{aligned} H_0: &\quad f(x) = \beta_0 + x\beta_1 \quad (\text{or } f(x) = \beta_0) \text{ vs.} \\ H_1: &\quad f \text{ non-linear.} \end{aligned}$$
- *Remark: LOESS allows more than one scalar covariate, the R function* `loess` *from the* `stats` *package e.g. allows up to four predictors*

1.1 Statistical Model
1.1.3 Non- and semi-parametric regression

1.1.3 b) Additive Modell (AM)

Idea: make the LM for flexible in the predictor part.

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

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

with covariate vectors $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})^\top$ and $\boldsymbol{z}_i = (z_{i1}, \dots, z_{iq})^\top$.

- Estimating $f_1, \dots, f_p, \boldsymbol{\beta}$, also testing.
- Variable selection and model selection (e.g. is the influence of a certain covariate linear or nonlinear)

1.1.3 c) Generalized Additive Models (GAM)

Idea: make the GLM for flexible in the predictor part.

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

1.1.4 Quantile regression / robust regression

1.1 Statistical models
1.1.4 Quantile regression / robust regression

Data as in the linear model ($y_i | \boldsymbol{x}_i$ conditionally independent). No assumptions are made about the error distribution.

Idea: Not only $\mathbb{E}[y_i | \boldsymbol{x}_i]$ is of interest, but also the conditional median ($\tau = 0.5$) or, more generally, the (conditional) quantiles $Q_\tau(y_i | \boldsymbol{x}_i)$.

Instead of using the least squares estimate $\widehat{\boldsymbol{\beta}}_{\text{KQ}} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \ \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$ with $\boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}_{\text{KQ}} = \widehat{\mathbb{E}}(y | \boldsymbol{x})$ one optimizes another function (for $\tau = 0.5$ e.g.):

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

The solution provides the conditional median(s) $\boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}_{\text{med}} = \widehat{\text{med}}(y | \boldsymbol{x})$.

One can show that this problem can be embedded in so-called quasi-likelihood methods, see chapter 3.4.2. This enables us to derive also statistical properties of the quantile regression estimates.

1.1.5 Longitudinal data as an example for dependent data

1.1 Statistical Model
1.1.5 Longitudinal data as an example for dependent data

Longitudinal data: $(y_{ij}, \boldsymbol{x}_{ij})$ for $i = 1, \dots, m$ and $j = 1, \dots, n_i$ as observations of target variables y_{ij} and covariables x_{ij} at times $t_{i1} < \dots < t_{ij} < \dots < t_{in_i}$. m statistical units, each with n_i correlated observations. Special case $m = 1$: time series, which are often to be assumed equi-distant.

1.1.5 a) Autoregressive or Markov models:
conditional distribution of $y_{ij}|y_{i,j-1}, y_{i,j-2}, \dots, y_{i1}, \boldsymbol{x}_{ij}$ is e.g. assumed to be a first order Markov model $y_{ij}|y_{i,j-1}, \boldsymbol{x}_{ij}$:

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

Likelihood inference: maximum likelihood estimation is often straightforward, but the asymptotic theory is difficult because the y_{ij} are dependent and occur also on the right hand side in the predictor. Therefore, the theory for standard OLS has to be modified appropriately.

1.1 Statistical Model
1.1.5 Dependent data: Example of longitudinal data

1.1.5 b) Linear mixed model (LMM): e.g. a so-called **random slope** model can be written as

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

- $\beta_0, \beta_1, \boldsymbol{\beta}$: population averaged or so-called **fixed effects**, for example, $\beta_0 + \beta_1 t$ is a linear trend (average trend) in the population modeled by fixed effects.
- b_{0i}, b_{1i}: individual-specific **random effects**. The number of random effects depends on the sample size.
- Typical assumptions:

$$\begin{aligned} b_{0i} &\overset{\text{i.i.d.}}{\sim} N(0, \sigma_{b_0}^2), \\ b_{1i} &\overset{\text{i.i.d.}}{\sim} N(0, \sigma_{b_1}^2) \end{aligned}$$

or

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \overset{\text{i.i.d.}}{\sim} \text{bivariate normal } N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{b_0}^2 & \sigma_{b_0 b_1} \\ \sigma b_0, b_1 & \sigma_{b_1}^2 \end{pmatrix} \right] .$$

1.1 Statistical Model
1.1.5 Dependent data: Example of longitudinal data

- Inference: maximum likelihood, restricted maximum likelihood (REML) inference or Bayesian inference using markov chain Monte Carlo (MCMC) simulation methods. For generalized linear mixed models (GLMMs), inference is much more complex than for LMMs.
- *Remark 1: using random effects reduces the dimensionality of the models and leads to a sparser representation of the data. Effectively, instead of estimating m parameters, one has only to estimate the* ***variance components*** $\sigma_{b_0}^2, \sigma_{b_0 b_1}, \sigma_{b_1}^2$.
- *Remark 2: not every covariate has to be included as fixed and random effect, usually the random effects are only specified for a subset of the covariates.*

1.1 Statistical Model
1.1.5 Dependent data: Example of longitudinal data

1.1.5 c) Marginal models
Marginal models estimate the conditional marginal means, i.e., no previous outcomes are included in the predictor, as for e.g. in autoregressive models.
A brief introduction is in chapter 3.4 (quasi-likelihood inference). A general method are the so-called generalized estimating equations (GEE).

1.1.6 Missing or incompletely observed data

1.1 Statistical Model
1.1.6 Missing or incompletely observed data

- Data: "any" (cross-sectional, survival, longitudinal data)
- Examples:
  - Non-respondents in statistical surveys
  - Drop-outs in clinical trials
  - Censored data (as in survival analysis)
  - Models with latent variables
- *Remark: statistical models and statistical estimation methods often require complete data. E.g. functions* `lm` *and* `glm` *in* R *drop all observations (rows in the data frame) where one or more values of the variables given in the* `formula` *parameter are missing.*

1.1 Statistical Model
1.1.6 a) Life time or event time analysis

Often, the so-called Cox-Model is used in survival analysis or the analysis of time-to-event data T_i. In that case, data may e.g. be right censored.
Right-censored survival data: Right-censored means that one knows only that $T_i \geq t_i$. Observations: $t_1, \dots, t_n$ of independent continuous life times $T_1, \dots, T_n \geq 0$, together with the censoring indicators $\delta_1, \dots, \delta_n$ (1 if t_i is an event time, 0 if it is a censored life time, i.e. the true life time could not be observed). Additionally, covariates $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ may be observed.
Goal: estimate $\lambda(t; \boldsymbol{x})$ or at least the influence of the covariables on the hazard rate

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

Interpretation: $\lambda(t; \boldsymbol{x})\Delta t \approx$ conditional probability of an event ("death") in the interval $[t, t+\Delta t]$ given one had no event ("survived") until time t where Δt is "small".

1.1 Statistical models
1.1.6 a) Life time or event time analysis: Cox model

Cox-Model (also: *Proportional Hazard-Model*)

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

- *Remark:* $\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ *is a multiplicative effect, constant over time!*

Primary interest: Estimation of $\boldsymbol{\beta}$, hypothesis testing.
The "baseline" hazard rate $\lambda_0(t)$ does not depend on i and is considered as nuisance parameter (or nuisance function).
The likelihood is factorized into terms for $T_i = t_i$ and terms for $T_i \geq t_i$):

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

The so-called partial likelihood $L_1(\boldsymbol{\beta})$ is maximized with respect to $\boldsymbol{\beta}$. Surprisingly, the loss of information is minimal when maximizing only L_1 instead of L.

1.1 Statistical models
1.1.6 b) Model-based cluster analysis

- Idea: $\boldsymbol{x} = (x_1, \dots, x_p)^\top$ is sampled from a multivariate mixture distribution as in example 1.1.5

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

  e.g. f_j density of the multivariate normal distribution.

- Task:
  1. Estimates for $\boldsymbol{\vartheta_j}, \pi_j,\ j = 1, \dots, J.$, similar to example 1.6.
  2. Estimates for unknown class membership j of an object with feature vector $\boldsymbol{x}$. Bayes formula provides:
  $$\widehat{\pi}(j|\boldsymbol{x}) = \frac{\widehat{\pi}_j f_j(\boldsymbol{x}|\widehat{\boldsymbol{\vartheta_j}})}{\widehat{f}(\boldsymbol{x})}.$$

- Likelihood maximization: with EM algorithm, see 3.2.2.

  Bayes: with MCMC algorithm, see Chapter 4.

1.2 Concepts of statistical inference

1.2 Concepts of statistical inference

Overview of different inference concepts:

1.2.1 Classical parametric inference

1.2.2 Classical parametric likelihood inference

1.2.3 Advanced likelihood-based inference

1.2.4 Bayesian inference

1.2.5 Statistical decision theory

1.2 Concepts of statistical inference

- $\boldsymbol{x} = (x_1, \dots, x_n)^\top$ or $\boldsymbol{y} = (y_1, \dots, y_n)^\top$ are realizations of random variables $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ or $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$.

  The components $X_1, \dots, X_n$ can again be multivariate.

- Further assumptions:
  - $X_1, \dots, X_n$ i.i.d. as X → simple random sample (see section 1.1.1).
  - $Y_1, \dots, Y_n$ (or $Y_1|X_1, \dots, Y_n|X_n$ in the regression model) are conditionally distributed independently but not identically (refer to sections 1.1.2-1.1.4).
  - $Y_1, \dots, Y_n$ are dependent, e.g. correlated temporally or spatially (see 1.1.5).

1.2 Concepts of statistical inference

- In all cases: $\boldsymbol{x} \in \mathcal{X}$ or $\boldsymbol{y} \in \mathcal{Y}$, where $\mathcal{X}$ or $\mathcal{Y}$ is the corresponding sample space. The sample spaces $\mathcal{X}$ and $\mathcal{Y}$ need a corresponding σ-algebra (well-defined set of possible events) and a probability assigned to each event (probability measure). We will always assume that a suitable σ-algebra exists.

- $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ or $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top$ are distributed on the sample space according to a distribution $\mathbb{P}$ ($\mathbb{P}_X$ or $\mathbb{P}_Y$), $\mathbb{P}(A) = P(\boldsymbol{X} \in A)$ with distribution functions $F(\boldsymbol{x}) = P(X_1 \le x_1, \dots, X_n \le x_n) = \mathbb{P}((-\infty, x_1] \times (-\infty, x_n]) = F(x_1, \dots, x_n)$ and $F(\boldsymbol{y})$.

- $\mathbb{P}$ (or F) belong to a set (or class or family) of distributions $\mathcal{P}_\theta = \{\mathbb{P}_\theta : \boldsymbol{\theta} \in \Theta\}$. Related distribution functions are $F(\boldsymbol{x}|\boldsymbol{\theta}) = P(\boldsymbol{X} \le \boldsymbol{x})$ or (if existing) densities $f(\boldsymbol{x}|\boldsymbol{\theta}) = f(x_1, \dots, x_n|\boldsymbol{\theta})$ (analogously for $\boldsymbol{y}$).

1.2 Concepts of statistical inference

Joint densities $f(\boldsymbol{x}|\boldsymbol{\theta})$:

- i.i.d. case:
$$f(\boldsymbol{x}|\boldsymbol{\theta}) = f(x_1|\boldsymbol{\theta}) \cdot \ldots \cdot f(x_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})$$

- Independent, but not identically distributed random variables $Y_1, \ldots, Y_n$:
$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(y_i|\boldsymbol{\theta}),$$
where the densities depend on the index i, i.e. f_i. If all component densities are from the same family (e.g. all are normal or all are Poisson) than $f_i = f(y_i|\boldsymbol{\theta}_i)$.

- With potentially dependent random variables $Y_1, \ldots, Y_n$, $f(\boldsymbol{y}|\boldsymbol{\theta})$ is not always factorizable and sometimes also difficult or impossible to express analytically.

1.2 Concepts of statistical inference

- Usual **parametric** inference:

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

- **Nonparametric / distribution-free inference**:

  Here, Θ is a function space, θ a certain function. E.g., Θ may be the space of twice continuously differentiable functions.

  Examples of methods:

  - (Kernel) density estimate for density f in 1.1.1 b)
  - Nonparametric regression for nonlinear f in 1.1.3 a).

1.2 Concepts of statistical inference

- **Semi-parametric** inference: Term is used if

  1. Θ has finite-dimensional and one infinite-dimensional component. Example: Cox proportional hazard model, $\boldsymbol{\beta}$ and $\lambda_0(t)$.
     - *Remark 1: term semi-parametric is not always consistently used in the literature.*
     - *Remark 2: frequently, the finite-dimensional part is of interest, while the infinite-dimensional component or function is nuisance.*
  2. the parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top$ is high-dimensional and k grows with n (possibly $k \approx n$), i.e. $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top$ with $k \approx n$. An example is semi-parametric regression with smoothing splines.
     - *Remark: even $k > n$ is possible, e.g. in GLMs with gene expression data as covariates.*

1.2.1 Classical parametric inference

1.2 Concepts of statistical inference
1.2.1 Classical parametric inference

- $\boldsymbol{X} = (X_1, \ldots, X_n)$ has distribution $\mathbb{P} \in \mathcal{P} = \{\mathbb{P}_\theta : \boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top \in \Theta\}$ with $\Theta \subseteq \mathbb{R}^k$ and $k < n$ fixed, often $k \ll n$.

- We assume, that a density function exists for the distribution $\mathbb{P}_\theta$.

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

- *Remark: Generally, this is the Radon-Nikodym-density regarding a dominant measure.*

1.2 Concepts of statistical inference
1.2.1 Classical parametric inference

- **Point estimator**

  $\boldsymbol{\theta}$ should be estimated. An estimator is a function that maps the sample space to the parameter space by assigning each sample a value (statistic) in the parameter space:

  Definition 1.

  $$\boldsymbol{T} : \left\{ \begin{array}{lcl} \mathcal{X} & \longrightarrow & \Theta \\ \boldsymbol{x} & \longmapsto & \boldsymbol{T}(\boldsymbol{x}) =: \widehat{\boldsymbol{\theta}} \end{array} \right.$$

  is called *estimating function* or *estimator* or *estimate.* An estimating function is a statistic (summary of the sample, function of the random variables).

- **Example 1.8**
  $\boldsymbol{\theta}$ is the vector of party proportions to be estimated by an election forecast. A naive estimator in a single election poll are the sample proportions (relative frequencies).

1.2 Concepts of statistical inference
1.2.1 Classical parametric inference

- **Point estimation** continued. The quality of an estimator can be measured by

  - the bias
    $\text{Bias}_{\boldsymbol{\theta}}(\boldsymbol{T}) = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{T}] - \boldsymbol{\theta}$, where $\mathbb{E}_{\boldsymbol{\theta}}$ is calculated with respect to $\mathbb{P}_{\boldsymbol{\theta}}$,
  - the variance
    $\text{Var}_{\boldsymbol{\theta}}(\boldsymbol{T}) = \text{Cov}_{\boldsymbol{\theta}}(\boldsymbol{T}) = \mathbb{E}_{\boldsymbol{\theta}}[(\boldsymbol{T} - \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{T}])(\boldsymbol{T} - \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{T}])^{\top}]$,
  - the mean squared error (MSE)
    $\text{MSE}_{\boldsymbol{\theta}}(\boldsymbol{T}) = \mathbb{E}_{\boldsymbol{\theta}}[\|\boldsymbol{T} - \boldsymbol{\theta}\|^2] = \text{trace}(\text{Var}_{\boldsymbol{\theta}}(\boldsymbol{T})) + \|\text{Bias}_{\boldsymbol{\theta}}(\boldsymbol{T})\|^2$,
  - the matrix valued MSE
    $\text{MSE}_{\boldsymbol{\theta}}(\boldsymbol{T}) = \mathbb{E}_{\boldsymbol{\theta}}[(\boldsymbol{T} - \boldsymbol{\theta})(\boldsymbol{T} - \boldsymbol{\theta})^{\top}] = \text{Cov}_{\boldsymbol{\theta}}(\boldsymbol{T})) + \text{Bias}_{\boldsymbol{\theta}}(\boldsymbol{T})\text{Bias}_{\boldsymbol{\theta}}(\boldsymbol{T})^{\top}$.
  - *Remark: the concept of "quality" is frequentist because it judges how well $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$ works **on average** for "all" possible samples $\boldsymbol{x}$ as realization of $\boldsymbol{X}$. In other words, it is not evaluating the quality with respect to the concrete observed sample, but, in an frequentist interpretation, the "method" $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$.*

1.2 Concepts of statistical inference
1.2.1 Classical parametric inference

- **Confidence sets, confidence intervals:**

  $$C : \left\{ \begin{array}{lcl} \mathcal{X} & \longrightarrow & \mathfrak{P}(\Theta) \\ \boldsymbol{x} & \longmapsto & C(\boldsymbol{x}) \subseteq \Theta \end{array} \right.$$

  with $\mathfrak{P}(\Theta)$ the power set of Θ, such that $\mathbb{P}_{\theta}(C(\boldsymbol{X}) \ni \boldsymbol{\theta}) \geq 1 - \alpha$ for all $\boldsymbol{\theta} \in \Theta$.

  $1 - \alpha$ is the *confidence level* or *coverage probability* of the *confidence set.*

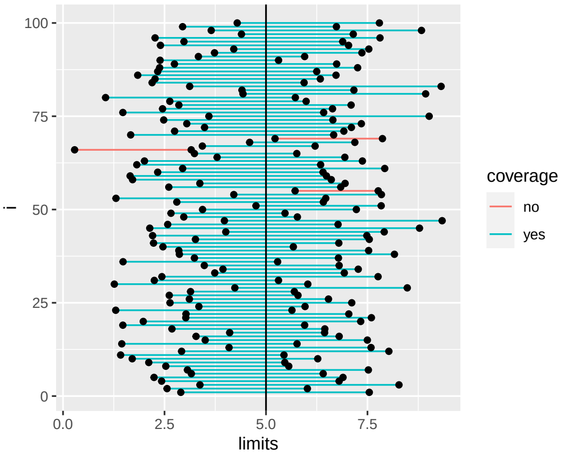
  Note the frequentist interpretation: $C(\boldsymbol{X})$ is a *random* area, but $\boldsymbol{\theta}$ is a fixed but unknown parameter.

  If $\Theta \subseteq \mathbb{R}$ and $C(\boldsymbol{x})$ for all $\boldsymbol{x}$ is an interval, then C is called *confidence interval.*

Example 1.9 (Normal confidence interval for μ, when σ^2 is unknown)
$\bar{X} \pm t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}}$.

1.2 Concepts of statistical inference
1.2.1 Classical parametric inference

Example 1.10 (Frequentist interpretation of confidence intervals)
The following figure shows 100 95%-confidence intervals for μ from 100 samples of size $n = 10$ from a $N(5, 3^2)$.



1.2 Concepts of statistical inference
1.2.1 Classical parametric inference

- With a **statistical tests** ϕ, a hypothesis H_0 against an alternative hypothesis H_1 can be checked.

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

where $\Theta_0 \cap \Theta_1 = \emptyset$. Often, but not necessarily, $\Theta = \Theta_0 \cup \Theta_1$.

I CAN'T BELIEVE SCHOOLS ARE STILL TEACHING KIDS ABOUT THE NULL HYPOTHESIS.
I REMEMBER READING A BIG STUDY THAT CONCLUSIVELY DISPROVED IT YEARS AGO.

Figure 4: Source: https://xkcd.com/892/

1.2 Concepts of statistical inference
1.2.1 Classical parametric inference

Statistical tests continued:

Results / actions:

$A_0 : H_0$ is not rejected,

$A_1 : H_1$ is confirmed, the result "is significant", H_o is rejected.

The test is a mapping

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

1.2 Concepts of statistical inference
1.2.1 Classical parametric inference

- **Testing** continued:

  A non-randomized test takes the form

  $$\phi(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \boldsymbol{x} \in C_\alpha, \\ 0, & \text{if } \boldsymbol{x} \notin C_\alpha. \end{cases}$$

  Here $C_\alpha \subset \mathcal{X}$ is the so-called *critical region*, which depends on the significance level α of the test and has to be understood as a subset of all possible samples.

  The test is often formulated using a (scalar) test statistic $T(\boldsymbol{x})$, i.e. a summary of the sample $\boldsymbol{x}$ as e.g. in the t-test, $T(\boldsymbol{x}) = (\bar{\boldsymbol{x}} - \mu_0)/(s/\sqrt{n})$.

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

  In the (two-sided) t-test case, $C_\alpha = (t_{1-\alpha/2,n-1}, \infty) \cup (-\infty, -t_{1-\alpha/2,n-1})$ (which is a subset of the real numbers).

1.2 Concepts of statistical inference
1.2.1 Classical parametric inference

- **Testing** continued: *Test of level („size") α*, where α "small":

  $$\mathbb{P}_\theta(A_1) \leq \alpha \text{ for all } \boldsymbol{\theta} \in \Theta_0 \text{ , i.e. under } H_0$$

  The probability of an *Type 1 error*, which is the probability of rejecting H_0 under H_0, i.e. if H_0 is true, is smaller than α.

  The function

  $$g_\phi(\boldsymbol{\theta}) = \mathbb{P}_\theta(A_1)$$

  is called the *power function* of ϕ and characterizes the quality of the test.

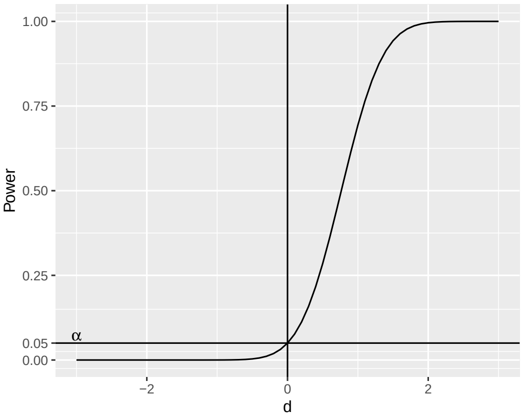
  - Remark: the power function is defined for all $\boldsymbol{\theta} \in \Theta = \Theta_0 \cup \Theta_1$.

  The level condition can also be formulated with the power function:

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

1.2 Concepts of statistical inference
1.2.1 Classical parametric inference

Example 1.11 (Power function of a one-sided t-test for two independent samples)
$d = \frac{\mu_1 - \mu_2}{s}$ is the assumed effect size, s is the pooled standard deviation. $H_0 : d \leq 0$ vs. $H_1 : d > 0$, $\alpha = 0.05$, $n = 10$ per group.



1.2 Concepts of statistical inference
1.2.1 Classical parametric inference

- **Testing** continued: "Program" of classical parametric test theory (see Chapter 2): Find test ϕ for level α with "optimal" power or minimal probability of the *type 2 error*, which is the probability of not rejecting H_0 under H_1 (i.e. if H_1 is true). The type 2 error is

$$1 - g_\phi(\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta_1\ .$$

  - *Remark 1: The concept is again frequentist.*
  - *Remark 2: The "program" is mainly feasible for special distribution families (e.g. for exponential families) and special test problems in the i.i.d. case. It is less or not suitable for more complex models, e.g. for GLMs. Therefore, we need more flexible inference methods:*

    $\rightarrow$ *Likelihood inference, see Chapter 3*

    $\rightarrow$ *Bayesian inference, see Chapter 4*

1.2 Concepts of statistical inference
1.2.1 Classical parametric inference

- **Testing** continued: In the simplest case of two point hypotheses

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

for $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_1$ the "best" test has the same structure as the likelihood ratio test: $H_0$is rejected, if

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

(see Neyman-Pearson theorem, section 2.2).

  - Remark: Indeed, in this simple case, the best test is a likelihood ratio test.

1.2 Concepts of statistical inference
1.2.1 Classical parametric inference

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

Example 1.12 (Gauss test)
$X_1, \dots, X_n$ i.i.d. $N(\mu, \sigma^2)$, σ^2 known. Consider

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

Test statistic is (since $\mathrm{Var}(\bar{X}) = \sigma^2/n$ in the i.i.d. case)

$$\frac{\overline{X} - \mathbb{E}(\overline{X})}{\mathrm{Var}(\overline{X})^{1/2}} \overset{\mu=\mu_0}{=} T(\boldsymbol{X}) = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n} \overset{\mu=\mu_0}{\sim} N(0,1).$$

H_0 is rejected if $T(\boldsymbol{x}) > z_{1-\alpha}$.
The p-value is $p = \mathbb{P}(T(\boldsymbol{X}) > T(\boldsymbol{x})|\mu = \mu_0) = \sup_\mu \mathbb{P}(T(\boldsymbol{X}) > T(\boldsymbol{x})|H_0)$.
Obviously:

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

1.2 Concepts of statistical inference
1.2.1 Classical parametric inference

- ***p*-values** continued:

- Remark: H_0 is rejected, if

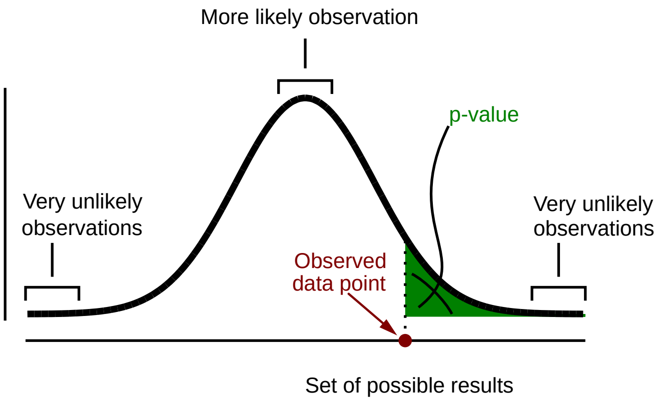
$$\begin{aligned} p &= \mathbb{P}(T(\boldsymbol{X}) > T(\boldsymbol{x}) | \mu = \mu_0) < \mathbb{P}(T(\boldsymbol{X}) > z_{1-\alpha} | \mu = \mu_0) \\ &= \sup_{\mu} \mathbb{P}_{\mu \in \Theta_0}(A_1) \\ &\leq \alpha \end{aligned}$$

- The p-value provides more information (namely how close you are to the decision boundary) than the pure "announcement" of the decision.

1.2 Concepts of statistical inference 1.2.1 Classical parametric inference

Example 1.13 (The p-value)
Credits go to *https://commons.wikimedia.org/wiki/File:P-value_in_statistical_significance_testing.svg*

> p-value
>
> The probability of observing a test statistic T given that some hypothesis is true is not equivalent to the probability that a hypothesis is true given that some test statistic T has been observed.
>
> Using the p-value as a "score" is committing an egregious logical error: the transposed conditional fallacy.



A p-value (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

1.2 Concepts of statistical inference 1.2.1 Classical parametric inference

Definition 1.2 (p-value)
Let $T(\boldsymbol{X})$ a test statistic for H_0 vs. H_1 with the following properties:

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

Then $p(\boldsymbol{x}) = \inf\{\widetilde{\alpha} : T(\boldsymbol{x}) = t \in C_{\widetilde{\alpha}}\}$ and H_0 is rejected, if $p(\boldsymbol{x}) < \alpha$, i.e. p is the smallest α-level at which a rejection is made.

- *Remark 1: p can be used as test statistic as well. The test maintains the level α:* $\sup_{\boldsymbol{\theta}\in\Theta_0}\mathbb{P}_{\boldsymbol{\theta}}(p(\boldsymbol{X}) = \inf\{\tilde\alpha : T(\boldsymbol{X}) \in C_{\tilde\alpha}\} < \alpha) \overset{2.}{\leq} \sup_{\boldsymbol{\theta}\in\Theta_0}\mathbb{P}_{\theta}(T(\boldsymbol{X}) \in C_\alpha) \overset{1.}{\leq} \alpha$

- *Remark 2: since* $\widetilde{\alpha} \leq \alpha \overset{2.}{\Rightarrow} C_{\widetilde{\alpha}} \subseteq C_\alpha \Rightarrow \mathbb{P}_{\boldsymbol{\theta}}(T(\boldsymbol{X}) \in C_{\widetilde{\alpha}}) \leq \mathbb{P}_{\boldsymbol{\theta}}(T(\boldsymbol{X}) \in C_\alpha)$

1.2 Concepts of statistical inference 1.2.1 Classical parametric inference

Example 1.14 (Example 1.12 continued)
Recall that (one-sided case) $C_\alpha = (z_{1-\alpha}, \infty)$.

- *Remark 1.* $\sup_{\mu\leq\mu_0}\mathbb{P}_\mu(T(\boldsymbol{X}) > z_{1-\alpha}) \leq \alpha$

- *Remark 2.* $C_\alpha = (z_{1-\alpha}, \infty) \subseteq (z_{1-\alpha'}, \infty) = C_{\alpha'}$ *for* $\alpha \leq \alpha'$, *since* $z_{1-\alpha'} \leq z_{1-\alpha}$

$$\begin{aligned} p &= \inf\{\widetilde{\alpha} : T(\boldsymbol{x}) \in C_{\widetilde{\alpha}}\} = \inf\{\widetilde{\alpha} : T(\boldsymbol{x}) > z_{1-\widetilde{\alpha}}\} \\ &= \{\widetilde{\alpha} : T(\boldsymbol{x}) = z_{1-\widetilde{\alpha}}\} \\ &= \mathbb{P}_{\mu_0}(Z \geq T(\boldsymbol{x})) = \mathbb{P}_{\mu_0}(T(\boldsymbol{X}) \geq T(\boldsymbol{x})) \\ &= \sup_{\mu\leq\mu_0}\mathbb{P}_\mu(T(\boldsymbol{X}) \geq T(\boldsymbol{x}))\ , \end{aligned}$$

where $Z \sim N(0,1)$.

1.2.2 (Parametric) likelihood inference

1.2 Concepts of statistical inference 1.2.2 (Parametric) likelihood inference

- Let $\mathcal{P} = \{f(\boldsymbol{x}|\boldsymbol{\theta})|\boldsymbol{\theta} \in \Theta\}$, i.e. the densities come from a parametrized distribution family $\mathcal{P}$. After the realized observation of $\boldsymbol{X} = \boldsymbol{x}$,

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

  is called the *likelihood function* of $\boldsymbol{\theta}$ for observation $\boldsymbol{x}$.

  - *Remark 1: The likelihood is defined as a density. But since $\boldsymbol{x}$ is fixed after the observations were made, it can now be seen as a function of $\boldsymbol{\theta}$, while the density is a function of $\boldsymbol{x}$ for fixed $\boldsymbol{\theta}$.*
  - *Remark 2: Since the density is a probability mass function in the discrete case, $L(\boldsymbol{\theta}|\boldsymbol{x}) = \mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x})$, i.e. the likelihood is the probability of the sample $\boldsymbol{x}$ for $\boldsymbol{\theta}$.*

1.2 Concepts of statistical inference 1.2.2 (Parametric) likelihood inference

- Likelihood principle: if two observations $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ have proportional likelihood functions, then this should lead to the same conclusions about the parameter $\boldsymbol{\theta}$, i.e. proportional likelihoods are considered as equivalent.

  Example 1.15 (Normal likelihood)
  $X_1, \dots, X_n$ *i.i.d.* $N(\mu, \sigma^2)$, σ^2 *known.*

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

  Two observations $\boldsymbol{x}$ and $\boldsymbol{y}$ with $\bar{x} = \bar{y}$ lead to the same conclusions about μ according to the likelihood principle, since $f(\boldsymbol{x}|\mu) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp(\sum_{i=1}^{n}\frac{x_i^2}{2\sigma^2} - n\bar{x}\frac{\mu}{\sigma^2} + n\frac{\mu^2}{2\sigma^2}) \propto \frac{1}{(2\pi\sigma^2)^{n/2}}\exp(\sum_{i=1}^{n}\frac{y_i^2}{2\sigma^2} - n\bar{y}\frac{\mu}{\sigma^2} + n\frac{\mu^2}{2\sigma^2}) = f(\boldsymbol{y}|\mu)$. *Note, that the quadratic sums, e.g.* $\sum_{i=1}^{n}x_i^2$, *appear in factors of the likelihood which do not depend on μ and can therefore be ignored.*

- Point estimate: maximum likelihood (ML) estimate

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

  or (in a more suitable notation)

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

- As a rule, there are no finite optimality properties, however asymptotic ones.
- Testing: likelihood ratio test, Wald test, score test (more in chapter 3).

1.2 Concepts of statistical inference 1.2.2 (Parametric) likelihood inference

1.2.3 Advanced likelihood-based inference

- Quasi-likelihood inference
- penalized likelihood inference
- semi-parametric models

1.2 Concepts of statistical inference 1.2.3 Advanced likelihood-based inference

1.2.4 Bayesian inference

We again consider $\mathcal{P} = \{f(\boldsymbol{x}|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, but additionally the uncertainty about $\boldsymbol{\theta}$ is modeled through a *prior density* $p(\boldsymbol{\theta})$ on Θ, i.e. $p(\boldsymbol{\theta})$ is the density of the distribution of $\boldsymbol{\theta}$.
As such, Θ can also be very high-dimensional.

1.2 Concepts of statistical inference 1.2.4 Bayesian inference

- Principle: After observing $\boldsymbol{x}$, all information over $\boldsymbol{\theta}$ is contained in the *posterior distribution.* Using Bayes theorem, we get

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{x}) &= \frac{f(\boldsymbol{x}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{\int f(\boldsymbol{x}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \\ &= \frac{f(\boldsymbol{x}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{f(\boldsymbol{x})} \overset{\text{proportional to}}{\propto} f(\boldsymbol{x}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}) \\ &= L(\boldsymbol{\theta}|\boldsymbol{x}) \cdot p(\boldsymbol{\theta}). \end{aligned}$$

  - *Remark 1: Note, that the denominator $f(\boldsymbol{x})$ is independent of $\boldsymbol{\theta}$, i.e. $\boldsymbol{\theta}$ has been integrated out.*
  - *Remark 2: $f(\boldsymbol{x})$ is called marginal likelihood.*

1.2 Concepts of statistical inference 1.2.4 Bayesian inference

- Although in a Bayesian framework, the posterior distribution contains all information about $\boldsymbol{\theta}$ after observation $\boldsymbol{x}$, Bayesian point estimates are sometimes wanted. Typical examples are:

  - Posterior Mean:

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

  - Posterior median:
$$\boldsymbol{T}_{\text{med}}(\boldsymbol{x}) = \widehat{\boldsymbol{\theta}}_{\text{post-Med}} = \text{med}_{\boldsymbol{\theta}|\boldsymbol{x}}(\boldsymbol{\theta}|\boldsymbol{x})$$
  - Posteriori mode:
$$\boldsymbol{T}_{\text{mod}}(\boldsymbol{x}) = \widehat{\boldsymbol{\theta}}_{\text{post-Mod}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\, p(\boldsymbol{\theta}|\boldsymbol{x}) = \underset{\boldsymbol{\theta}}{\text{argmax}}\, p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{x})$$

1.2 Concepts of statistical inference 1.2.4 Bayesian inference

- *Improper Priors*, i.e. priors which are not integrable, which means that
$$\int_{\Theta} p(\boldsymbol{\theta})d\boldsymbol{\theta} = +\infty,$$
are allowed, as long as the posterior density is integrable (i.e. integral over the whole parameter space is 1).

  A special case is the prior $p(\boldsymbol{\theta}) \propto 1$ ("uniform prior"). In that case
$$\widehat{\boldsymbol{\theta}}_{\text{ML}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\, L(\boldsymbol{\theta}|\boldsymbol{x}) = \underset{\boldsymbol{\theta}}{\text{argmax}} L(\boldsymbol{\theta}|\boldsymbol{x})p(\boldsymbol{\theta}) = \widehat{\boldsymbol{\theta}}_{\text{post-Mod}}$$
  applies, i.e. the ML estimate and the posterior mode are identical.

1.2 Concepts of statistical inference 1.2.4 Bayesian inference

Definition 1.3
A prior distribution $p(\boldsymbol{\theta})$ is called the conjugate distribution *for $f(\boldsymbol{x}|\boldsymbol{\theta})$, if the posterior $p(\boldsymbol{\theta}|\boldsymbol{x})$ and the prior $p(\boldsymbol{\theta})$ have the same form, i.e. if prior and posterior distribution belong to the same distribution family, but with potentially different, updated, parameters.*

- *Remark 1: Conjugate priors are convenient, since they allow a closed-form solution for the posterior, e.g. no (numerical) integration is needed.*
- *Remark 2: Conjugate priors allow to examine the sensitivity of the prior on the posterior and show in a transparent way, how the prior information is updated by the likelihood.*
- *Remark 3: Conjugate priors exist in simple cases such as for exponential families.*
- *Remark 4: The computational power which is nowadays available reduces the benefit of using conjugate priors.*

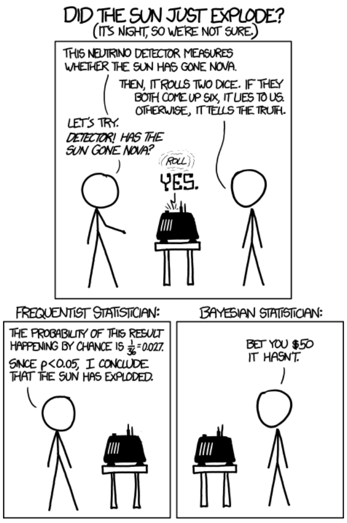
1.2 Concepts of statistical inference 1.2.4 Bayesian inference

- Bayes interval estimate: choose *credibility areas or intervals* $C(\boldsymbol{x})$ such that for fixed $\boldsymbol{x}$
$$\int_{C(\boldsymbol{x})} p(\boldsymbol{\theta}|\boldsymbol{x})\, d\boldsymbol{\theta} = \mathbb{P}_{\boldsymbol{\theta}|\boldsymbol{x}}\Big(\underbrace{\boldsymbol{\theta}}_{\text{randomly}} \in \underbrace{C(\boldsymbol{x})}_{\substack{\text{not by chance,}\\ \text{deterministic}}} \Big) \geq 1 - \alpha.$$
  This is a probability statement for a concrete sample $\boldsymbol{x}$ and no frequentist interpretation is necessary!
  - *Remark 1: See the different views. In Bayesian inference, $C(\boldsymbol{x})$ is fixed and $\boldsymbol{\theta}$ is a random variable, while in the frequentist view, $\boldsymbol{\theta}$ is fixed and $C(\boldsymbol{x})$ (confidence set) is random.*
  - *Remark 2: Nevertheless, a frequentist interpretation can still be made for $C(\boldsymbol{X})$ by introducing asymptotics for Bayes estimators.*

- *Remark 3: Credible intervals (for a scalar component of $\boldsymbol{\theta}$) can be constructed by cutting $\alpha/2$-$\alpha/2$ from the two tails of the posterior. Alternatively, so-called highest posterior density intervals (HPDs) can be constructed.*

1.2 Concepts of statistical inference
1.2.4 Bayesian Inference

Explosion $\theta \in \{0, 1\}$
Detector $X \in \{yes, no\}$
Prior $p(1) = 0.0001$, small
$p(0) = 0.9999$
Frequentist:
$f(\boldsymbol{x}|\boldsymbol{\theta}) =$
$P(X = yes|\theta = 0) = \frac{1}{36} < .05$
Bayes:
$P(X = yes(no)|\theta = 1(0)) = \frac{35}{36}$
$p(\boldsymbol{\theta}|\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})/f(\boldsymbol{x}) =$
$P(\theta = 0|X = yes)$
$= 0.9965119$
$> P(\theta = 1|X = yes)$
$= 0.00348814$


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY. DETECTOR! HAS THE SUN GONE NOVA?
(ROLL)
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027. SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

Source: https://xkcd.com/1132/

1.2.5 Statistical decision theory

1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

- View in decision theory: estimation and testing are decisions under uncertainty. What are the consequences of (wrong) decisions?
- You can learn more in the lecture ”Decision Theory” by Thomas Augustin.
- As before, we consider $\mathbb{P} \in \mathcal{P}_\theta = \{\mathbb{P}_\theta : \boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top \in \Theta\}$ as a statistical model; $\boldsymbol{x}$ the sample, i.e. the observed data.

1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

Additionally, the following functions are considered:

Definition 1.4 (decision function)
A decision function *is a function (mapping from the sample space to space of actions)*

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

D is called the decision *or* action space*.*

1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

Definition 1.5 (loss function)
A loss function *(sometimes a* profit function *instead)*

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

assigns a loss *(”loss”) to a decision* $\mathbf{d}(\boldsymbol{x})$ *(”decision”), depending on $\boldsymbol{\theta}$. In general, L is chosen such that the loss is zero if the decision is correct, i.e. L is a non-negative function.*

1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

Example 1.16

1. **Test***: consider*

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

(e.g. Gauss test).

Let the decision space be $D = \{d_0, d_1\}$ with

d_0*: decision for* H_0*,*

d_1*: decision for* H_1*.*

A possible loss function is:

$$\begin{aligned} L(d_0, \theta) &= \begin{cases} 0, & \textit{if } \theta \leq \theta_0 \quad \textit{(right decision)} \\ a \in \mathbb{R}_+, & \textit{if } \theta > \theta_0 \quad \textit{(type 2 error)} \end{cases} \\ L(d_1, \theta) &= \begin{cases} 0, & \textit{if } \theta > \theta_0 \quad \textit{(right decision)} \\ b \in \mathbb{R}_+, & \textit{if } \theta \leq \theta_0 \quad \textit{(type 1 error)} \end{cases} \end{aligned}$$

- *Remark: $L(d_k, \theta)$ is the loss if one decides for H_k, $k = 1, 2$. a, b can be different weights for an "asymmetric loss", e.g. if it is worse to approve an ineffective drug compared to not discover an effective one.*

1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

Example 1.17 (example 1.16 continued)

2. **Estimator***: "Decision" is a real number:*

$$d(\boldsymbol{x}) = T(\boldsymbol{x}) = \widehat{\theta} \in \Theta, \textit{ i.e. } D = \Theta.$$

Possible loss functions:

$$\begin{aligned} L(d, \theta) &= (d - \theta)^2 & \textit{quadratic loss,} \\ L(d, \theta) &= |d - \theta| & \textit{absolute loss,} \\ L(d, \theta) &= w(\theta)(d - \theta)^2 & \textit{weighted quadratic loss,} \end{aligned}$$

where w is a fixed weight function.

1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

Example 1.18 (example 1.16 continued)

3. **Multiple decision procedure***, for example, choice between three alternatives*

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

4. *Analog: model selection, variable selection*

1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

A loss function depends through $\boldsymbol{d}(\boldsymbol{x})$ on $\boldsymbol{x}$. The risk function considers the expected ("average") loss.

Definition 1.6 (Risk Function)
A risk function *is defined as*

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

The risk function for $\boldsymbol{d}$ as a general rule is independent of $\boldsymbol{x}$. In fact, $\boldsymbol{X}$ is integrated out, i.e., $R(\boldsymbol{d};\boldsymbol{\theta})$ is for a given $\boldsymbol{d}$ only a function of $\boldsymbol{\theta}$.

1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

Example 1.19

1. **Estimate**, *i.e.*

$$\boldsymbol{d}(\boldsymbol{x}) = \boldsymbol{T}(\boldsymbol{x}) \quad \textit{estimated value}, \quad \boldsymbol{d}(\boldsymbol{X}) = \boldsymbol{T}(\boldsymbol{X}) \quad \textit{estimator}.$$

   With a quadratic loss function

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

   the risk function is

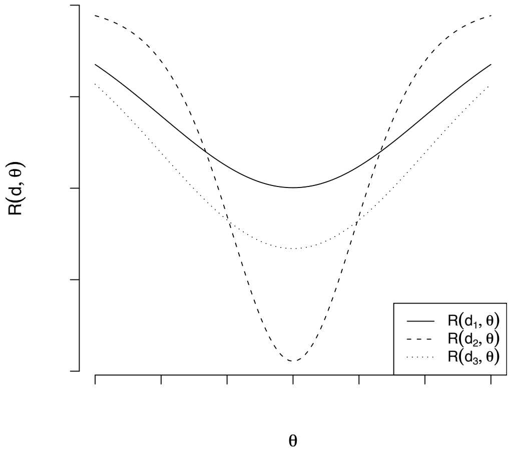
$$R(\boldsymbol{T},\boldsymbol{\theta}) = \mathbb{E}_\theta[\|\boldsymbol{T}(\boldsymbol{X}) - \boldsymbol{\theta}\|^2] = \mathrm{MSE}_\theta(\boldsymbol{T}(\boldsymbol{X})).$$

   Note that the argument $\boldsymbol{T}$ in $R(\boldsymbol{T},\boldsymbol{\theta})$ is the estimator and not the concrete estimated value which depends on the observed sample $\boldsymbol{x}$.

2. **Testing** *: Exercise.*

Comparison of decision rules using the risk function

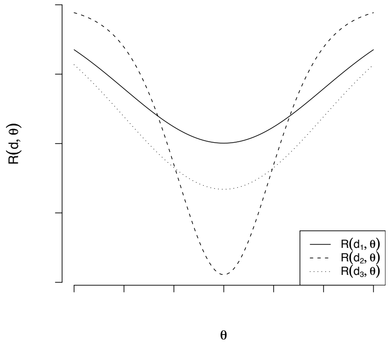
1.2 Concepts of statistical inference
1.2.5 Statistical decision theory



The chart shows that d_3 is better than d_1 for all $\theta \in \Theta$, i.e. d_3 dominates d_1 uniformly.
Goal: Find a rule d^*, which dominates all "competing" rules d.

1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

Problem: This idea doesn't work in general. The risk functions usually cross each other, e.g. in the figure, d_2 is better than d_1 and d_3 only in a certain range.



1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

→ ”Optimal” decision rules only possible through:

- restriction to special classes of loss functions,
- restriction to special classes of decision rules, for example unbiased estimates or unbiased tests (which not only maintain the α-level, but also require that the power is greater than α under H_1),
- additional criteria that summarize $R(\boldsymbol{\theta})$ over $\boldsymbol{\theta}$.

1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

Criteria for optimal decision making rules

1. **Minimax criterion**

   Idea: consider maximum of the risk function, i.e. prefer d_2 in the following figure, because

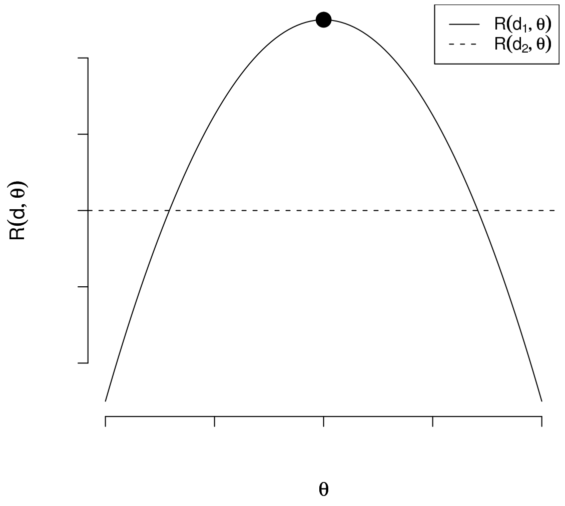
   $$\max_{\theta} R(d_2, \theta) < \max_{\theta} R(d_1, \theta)\ .$$

   Choose the decision rule which minimizes the maximum loss.

- *Remark: Minimax is considered as conservative since it minimizes the worst possible loss*

1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

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



1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

Definition 1.7 (Minimax decision rule)
Let $\boldsymbol{d}^* : \mathcal{X} \to D$ *a decision rule.* $\boldsymbol{d}^*$ *is called* Minimax*, if it minimizes the supremal risk:*

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

- *Remark: In many cases, supremum and Infimum are taken on, so that actually*

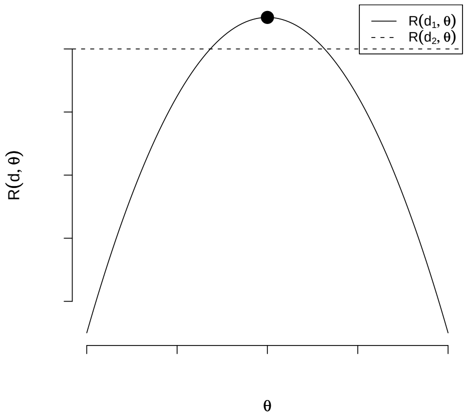
  $$\boldsymbol{d}^* = \operatorname*{argmin}_{\boldsymbol{d}\in D} \max_{\boldsymbol{\theta}\in\Theta} R(\boldsymbol{d}, \boldsymbol{\theta})$$

  holds, hence the name Minimax.

1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

The Minimax criterion protects against the worst case, which is not always reasonable, as the following figure shows:



1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

Here $d^* = d_2$ would only be reasonable if θ values in the middle "are particularly likely". Criteria for optimal decision-making rules continued.

2. **Bayesian criterion**

   As in the Bayesian inference, we assume a prior density $p(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ (from a frequentist point of view, $p(\boldsymbol{\theta})$ is a – not necessarily standardized – weight function).

   Definition 1.8
   The Bayes risk *is defined as*

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

1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

2. **Bayesian criterion** continued:

   The *Bayes risk* is minimized by the *Bayes optimal rule* $\boldsymbol{d}^*$:

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

   Definition 1.9
   Let $p(\boldsymbol{\theta}|\boldsymbol{x})$ *be the posterior density. Then*

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

   is called the Posterior Bayes risk.

- *Remark 1: In the Bayes risk, the integration of the loss function is over $\boldsymbol{x}$ and $\boldsymbol{\theta}$ with respect to the data density or likelihood $f(\boldsymbol{x}|\boldsymbol{\theta})$ and the prior $p(\boldsymbol{\theta})$.*
- *Remark 2: In the posterior Bayes risk, the integration of the loss function is only over $\boldsymbol{\theta}$ with respect to the posterior $p(\boldsymbol{\theta}|\boldsymbol{x})$. In contrast to the Bayes risk, the posterior Bayes risk still depends on the sample $\boldsymbol{x}$.*

1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

The following practical result applies:

Theorem 1.10
A rule $\boldsymbol{d}^*$ is Bayes optimal if and only if $\boldsymbol{d}^*(\boldsymbol{x})$ is minimizing the posterior Bayes risk for each sample $\boldsymbol{x}$.

"Proof": $f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{x})f(\boldsymbol{x})$

$$\begin{aligned}\text{proof" } \Leftarrow \text{ " } r(\boldsymbol{d},p) &= \int_\Theta \int_\mathcal{X} L(\boldsymbol{d}(\boldsymbol{x}),\boldsymbol{\theta}) f(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{x}\; p(\boldsymbol{\theta})\; d\boldsymbol{\theta} \\ &= \int_\Theta \int_\mathcal{X} L(\boldsymbol{d}(\boldsymbol{x}),\boldsymbol{\theta}) f(\boldsymbol{x}) p(\boldsymbol{\theta}|\boldsymbol{x})\; d\boldsymbol{x}\; d\boldsymbol{\theta} \\ &\overset{nonneg.\ Fubini}{=} \int_\mathcal{X} f(\boldsymbol{x}) \left\{ \int_\Theta L(\boldsymbol{d}(\boldsymbol{x}),\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{x})\; d\boldsymbol{\theta} \right\}\; d\boldsymbol{x}\end{aligned}$$

Therefore, $\boldsymbol{d}^*$ minimizes the the Bayes risk $r(\boldsymbol{d},p)$ if it minimizes $\{\dots\}$ for every $\boldsymbol{x}$. For a proof of the reverse direction see Rüger, theorem 2.15.

1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

Remarks:

- Theorem 1.10 facilitates the determination of the Bayes optimal decision rule $d^*(\boldsymbol{x})$ in concrete cases.
- It shows an intuitive quality of the Bayesian procedure: to make a decision in case of observation $\boldsymbol{x}$, it is enough to consider the loss for $\boldsymbol{d}(\boldsymbol{x})$. It is not necessary to consider losses for $\boldsymbol{d}(\boldsymbol{X})$ for other possible samples $\boldsymbol{X}$ which were not observed.
- Bayes-optimal rules $\boldsymbol{d}^*$ are *admissible*, i.e. they are not dominated by any other rule $\boldsymbol{d} \neq \boldsymbol{d}^*$.
- One gets a close relationship to the Minimax rule, if a "least favorable" prior density $p^*(\boldsymbol{\theta})$ is chosen.

1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

Optimality of Bayesian estimators:

- The posterior expected value

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

  is Bayes optimal with the quadratic loss function $L(d,\theta) = (d-\theta)^2$.

  Proof:

  We search $\min_d \int_\Theta (d(\boldsymbol{x}) - \theta)^2 p(\theta|\boldsymbol{x}) d\theta$. Take the derivative with respect to d and set the result to 0 (note, that d is the decision rule and $d\theta$ is the symbol for the integration).

  So $\int_\Theta (d(\boldsymbol{x}) - \theta)\; p(\theta|\boldsymbol{x}) d\theta = 0 \Leftrightarrow d(\boldsymbol{x}) \int_\Theta\; p(\theta|\boldsymbol{x}) d\theta = \int_\Theta \theta\; p(\theta|\boldsymbol{x}) d\theta = \mathbb{E}[\boldsymbol{\theta}|\boldsymbol{x}]$. Since $\int_\Theta\; p(\theta|\boldsymbol{x}) d\theta = 1$, it follows that

  $$d(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{\theta}|\boldsymbol{x}]\ .$$

1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

Optimality of Bayesian estimators:

- The median posterior
$$\widehat{\theta} = \mathrm{med}(\theta|\boldsymbol{x})$$
is Bayes optimal with the absolute loss function $L(d, \theta) = |d - \theta|$.

  Proof (short):

  We search $\min_d \int_{-\infty}^{d} (d - \theta)\ p(\theta|\boldsymbol{x}) d\theta + \int_d^{\infty} (\theta - d)\ p(\theta|\boldsymbol{x}) d\theta$. Using the Leibniz rule (for each integral)
$$\frac{\partial}{\partial d(\boldsymbol{x})} \int_{-\infty}^{d(\boldsymbol{x})} (d(\boldsymbol{x}) - \theta)\ p(\theta|\boldsymbol{x}) d\theta = \int_{-\infty}^{d(\boldsymbol{x})} \frac{\partial}{\partial d(\boldsymbol{x})} (d(\boldsymbol{x}) - \theta) p(\theta|\boldsymbol{x}) d\theta$$
one gets $\int_{-\infty}^{d} p(\theta|\boldsymbol{x}) d\theta - \int_d^{\infty}\ p(\theta|\boldsymbol{x}) d\theta = 0$ or $\int_{-\infty}^{d} p(\theta|\boldsymbol{x}) d\theta = \int_d^{\infty}\ p(\theta|\boldsymbol{x}) d\theta = 1/2$, d posterior median.

1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

Leibniz rule: Let
$$G(x) = \int_{a(x)}^{b(x)} f(x, t) dt\ .$$
Then
$$\frac{\partial G}{\partial x} = \int_{a(x)}^{b(x)} \frac{\partial f}{\partial x} dt + f(x, b(x)) \frac{\partial b}{\partial x} - f(x, a(x)) \frac{\partial a}{\partial x}\ .$$
Proof (long): first substitute the lower and upper infinity bounds by some constants a and b. Then
$$\begin{aligned}
&\frac{\partial}{\partial d(\boldsymbol{x})} \int_a^{d(\boldsymbol{x})} (d(\boldsymbol{x}) - \theta)\ p(\theta|\boldsymbol{x}) d\theta \\
&= \int_a^{d(\boldsymbol{x})} 1 p(\theta|\boldsymbol{x}) d\theta \\
&+ (d(\boldsymbol{x}) - d(\boldsymbol{x})) p(d(\boldsymbol{x})|\boldsymbol{x})(1) - (d(\boldsymbol{x}) - a) p(a|\boldsymbol{x})(0)
\end{aligned}$$
(and similar for the second integral).

1.2 Concepts of statistical inference
1.2.5 Statistical decision theory

Optimality of Bayesian estimators continued:

- The Posterior mode
$$\widehat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, p(\theta|\boldsymbol{x})$$
is Bayes optimal with 0-1 loss function
$$L_\varepsilon(d, \theta) = \begin{cases} 1, & \text{falls } |d - \theta| \geq \varepsilon, \\ 0, & \text{falls } |d - \theta| < \varepsilon \end{cases}$$
and $\varepsilon \to 0$.

  "Proof:" $d^* = \operatorname{argmin}(\int_{-\infty}^{d-\varepsilon} p(\theta|\boldsymbol{x}) d\theta + \int_{d+\varepsilon}^{\infty} p(\theta|\boldsymbol{x}) d\theta) = \operatorname{argmax} \int_{d-\varepsilon}^{d+\varepsilon} p(\theta|\boldsymbol{x}) d\theta \approx \operatorname{argmax} p(d|\boldsymbol{x}) 2\varepsilon$.

- Note: The ML estimate (posterior mode) is optimal for a flat prior $p(\theta) \propto 1$ for the 0-1 loss function

Chapter 2

Classical Estimation and Testing Theory

2 Classical Estimation and Testing Theory

Basic Model:

Sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ has distribution $\mathbb{P} \in \mathcal{P} = \{\mathbb{P}_\theta : \boldsymbol{\theta} \in \Theta\}, \Theta \subseteq \mathbb{R}^k$, where

- $\boldsymbol{\theta}$: k-dimensional parameter
- Θ: parameter space
- $k < n$, often $k \ll n$, with $\dim(\boldsymbol{\theta}) = k$ fixed for asymptotic $(n \to \infty)$-cases.
- Generally, there is density

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

  such that one can write:

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

2 Classical Estimation and Testing Theory

- Classical estimation and test theory for finite (i.e. for fixed sample size n) i.i.d.-samples is of particular relevance; the following holds:

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

- However, many concepts, especially from estimation theory, are of general significance.
- Literature: Lehmann & Casella (1998), Lehmann & Romano (2005), Rüger (1999, 2002) Volumes I+II

2 Classical Estimation and Testing Theory

Definition 2.1 (Statistic)
A statistic *is a measurable function.*

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

Normally, $l < n$, since with statistic $\boldsymbol{T}$, there is a dimension reduction.

Example 2.1

- $\boldsymbol{T}$ estimator $(l = k)$
- T test statistic $(l = 1)$

2.1 Classical Estimation Theory

2.1 Classical Estimation Theory

Goal: point or interval estimate for $\boldsymbol{\theta}$ or a transformed parameter vector $\boldsymbol{\tau}(\boldsymbol{\theta})$.
Example 2.2
$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ with $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$. Here $\tau(\boldsymbol{\theta}) = \mu$ (i.e. σ^2 is nuisance parameter) or $\tau(\boldsymbol{\theta}) = 1/\sigma^2$ (i.e. the precision is of interest).

2.1 Classical Estimation Theory

Definition 2.2 (Point estimation, Estimator)
Let

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

be a measurable function, the estimator or the estimation function.
$\boldsymbol{T}(\boldsymbol{x})$ *denotes the estimated value or the* point estimate *(for a concrete realization* $\boldsymbol{x}$*) and* $\boldsymbol{T}(\boldsymbol{X})$ *the* point estimator *of* $\boldsymbol{\theta}$*, which is a random variable (also denoted as* $\widehat{\boldsymbol{\theta}}(\boldsymbol{X})$ *or in short* $\widehat{\boldsymbol{\theta}}$*, i.e. in the last expression there is no notational difference between concrete estimated value and estimator).*

2.1 Classical Estimation Theory Overview

2.1.1 Sufficiency

2.1.2 Bias of an estimator, variance and MSE

2.1.3 Fisher information and sufficiency

2.1.4 Unbiased estimators

2.1.5 Asymptotic properties and criteria

2.1.1 Sufficiency

2.1 Classical Estimation Theory 2.1.1 Sufficiency

The notion of sufficiency is of fundamental importance in the classical parametric inference; outside of it, the importance is (strongly) understated, see Probability Theory and Inference II.
Idea: Summarise $\boldsymbol{x}$ in a statistic without information loss; $\boldsymbol{T}(\boldsymbol{x})$ contains all information in $\boldsymbol{x}$ about $\boldsymbol{\theta}$.
Usage: Among other things, we will see in 2.1.3 and 2.1.4 that sufficient statistics play an important role in finding estimators with the smallest possible variance (see theorems 2.22 and 2.23).

2.1 Classical Estimation Theory 2.1.1 Sufficiency

Definition 2.3
A statistic $\boldsymbol{T}$ *is called* sufficient *for* $\boldsymbol{\theta}$ *(or also for* $\mathcal{P}$*)* $\overset{def}{\Leftrightarrow}$ *conditional distribution or density of* $\boldsymbol{X}$ *given* $\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{t}$ *does not depend on* $\boldsymbol{\theta}$ *for all values of* $\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{t}$*, i.e.*

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

does not depend on $\boldsymbol{\theta}$*.*

Idea: Additional information in $\boldsymbol{X}$, that is not contained in $\boldsymbol{T}$ is given by $f_{X|T}$. If $f_{X|T}$ is independent from $\boldsymbol{\theta}$, then sample $\boldsymbol{x}$ does not contain more information about $\boldsymbol{\theta}$ than $\boldsymbol{T}(\boldsymbol{x})$.

2.1 Classical Estimation Theory 2.1.1 Sufficiency

Assume the existence of a density for $\boldsymbol{X}$. The criterion in the following theorem is equivalent and constructive:

Theorem 2.4 (Factorisation theorem, Neyman criterion)
A statistic $\boldsymbol{T}$ is sufficient for $\boldsymbol{\theta}$ if and only if

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

for almost all $\boldsymbol{x}$.

I.e. the density can be factored into two parts, one of which depends on $\boldsymbol{x}$, but not on $\boldsymbol{\theta}$, and the other only on $\boldsymbol{\theta}$ and $\boldsymbol{T}(\boldsymbol{x})$.

2.1 Classical Estimation Theory
2.1.1 Sufficiency

Proof.

„⇒": If $\boldsymbol{T}$ is sufficient, then the following holds:

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

Furthermore,

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

i.e. as a whole $f_{X|T}(\boldsymbol{x}|\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{t}) \cdot f_T(\boldsymbol{t}|\boldsymbol{\theta}) = f_X(\boldsymbol{x}|\boldsymbol{\theta}) \cdot 1_{\boldsymbol{T}(\boldsymbol{x})=\boldsymbol{t}}$ and thus for $\boldsymbol{t} = \boldsymbol{T}(\boldsymbol{x})$

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

2.1 Classical Estimation Theory
2.1.1 Sufficiency

„⇐": One obtains the density of $\boldsymbol{T}$ evaluated at $\boldsymbol{t}$ by summing (or integrating) the above factorization criterion over $\boldsymbol{x}$ for which $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{t}$ holds. In the discrete case:

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

The conditional density $\boldsymbol{X}$ given $\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{t}$,

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

(for $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{t}$, otherwise 0) does not depend on $\boldsymbol{\theta}$, i.e. $\boldsymbol{\theta}$ is sufficient. In the continuous case, sums are replaced with integrals (while considering measurability conditions).

2.1 Classical Estimation Theory
2.1.1 Sufficiency

Example 2.3 (Bernoulli-Experiment)

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bin}(1, \pi)$ and $Z = \sum_{i=1}^n X_i$ be the number of successes. Then Z is sufficient for π, because,

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

(for $\sum_{i=1}^n x_i = z$, else conditional density $= 0$) does not depend on π.

2.1 Classical Estimation Theory
2.1.1 Sufficiency

Example 2.3 (Bernoulli-Experiment) continued

Alternatively: According to the factorisation theorem,

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

Note that this shows that the factorization needs not to be unique.

2.1 Classical Estimation Theory
2.1.1 Sufficiency

Example 2.4 (Normal Distribution)

Let $\boldsymbol{X} = (X_1, \dots, X_n)$, $X_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ and $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$.

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

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

i.e. $\boldsymbol{T}(\boldsymbol{X}) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is sufficient for $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$.

But: the bijective transformation $\widetilde{\boldsymbol{T}}(\boldsymbol{X}) = (\bar{X}, S^2)$ is also sufficient for $\boldsymbol{\theta}$, where S^2 denotes the sample variance.

2.1 Classical Estimation Theory
2.1.1 Sufficiency

Example 2.5 (Exponential Distribution)

Let $\boldsymbol{X} = (X_1, \dots, X_n) \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$, then

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

with $T(\boldsymbol{x}) = \sum_{i=1}^n x_i$. For $t = \sum_{i=1}^n x_i$, $f_{X|T}(\boldsymbol{x}|t, \lambda)$ is

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

which does not depend on λ. Therefore, $T(\boldsymbol{x})$ is sufficient. Here, we used the fact that $\sum_{i=1}^n X_i$ is gamma distributed with parameters n and λ.

2.1 Classical Estimation Theory
2.1.1 Sufficiency

Example 2.6 (Order statistics) Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} f(x|\boldsymbol{\theta})$ (where f is the density of a continuous distribution) and $\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{X}_{(\cdot)} = (X_{(1)}, \dots, X_{(n)})$ is the order statistic, $X_{(1)} \leq ... \leq X_{(n)}$.

Then the following holds,

$$f_{X|T}(\boldsymbol{x}|\boldsymbol{T} = \boldsymbol{x}_{(\cdot)}, \boldsymbol{\theta}) = \frac{1}{n!}\ ,$$

which is independent from $\boldsymbol{\theta}$. The equality follows from the continuity, since $x_i \neq x_j\ \forall i \neq j$ (with probability 1). $\boldsymbol{X}_{(\cdot)}$ is sufficient for $\boldsymbol{\theta}$. Thus, we have no loss of information with iid observations from ordering the data.

2.1 Classical Estimation Theory
2.1.1 Sufficiency

Remarks

- Obviously $\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{X}$, i.e. the sample itself is sufficient.
- Likewise, every one-to-one transformation of $\boldsymbol{X}$ or of a sufficient statistic $\boldsymbol{T}(\boldsymbol{X})$ is sufficient.
- If $\boldsymbol{T}$ is sufficient, then so is $(\boldsymbol{T}, \boldsymbol{T}^*)$, where $\boldsymbol{T}^*$ is any further statistic.

This shows that the dimension of a sufficient statistic should be reduced as far as possible.

2.1 Classical Estimation Theory
2.1.1 Sufficiency

Definition 2.5 (Minimal Sufficiency)
A statistic $\boldsymbol{T}$ *is called* minimal sufficient *for* $\boldsymbol{\theta} \overset{def}{\Leftrightarrow} \boldsymbol{T}$ *is sufficient, and for every other sufficient statistic* $\boldsymbol{V}$ *there exists a function* $\boldsymbol{H}$ *such that*

$$\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{H}(\boldsymbol{V}(\boldsymbol{X})) \ \ \mathcal{P} - \textit{almost everywhere.}$$

Question: Do minimal sufficient statistics exist? If so, are they unique?

2.1 Classical Estimation Theory
2.1.1 Sufficiency

Example 2.7 (Normal Distribution)
Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$.

1. $T(\boldsymbol{X}) = \bar{X}$ is minimal sufficient for μ for a known σ^2.
2. $T(\boldsymbol{X}) = \sum_{i=1}^n (X_i - \mu)^2$ is minimal sufficient for σ^2 for a known μ.
3. $\boldsymbol{T}(\boldsymbol{X}) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is minimal sufficient for μ and σ^2.
4. $T(X) = |X|$ is minimal sufficient for σ^2 if $X \sim N(0, \sigma^2)$ $(n = 1, \mu = 0)$. X is also sufficient, but not minimal sufficient (despite the same dimension). Only the absolute value contributes to the information about σ^2, not the sign. Also: $X^2 = |X|^2$ is minimal sufficient.

2.1 Classical Estimation Theory
2.1.1 Sufficiency

Lemma 2.6
If $\boldsymbol{T}$ *and* $\boldsymbol{S}$ *are minimal sufficient statistics, then there exist injective functions* $\boldsymbol{g}_1$, $\boldsymbol{g}_2$, *such that* $\boldsymbol{T} = \boldsymbol{g}_1(\boldsymbol{S})$ *and* $\boldsymbol{S} = \boldsymbol{g}_2(\boldsymbol{T})$.

Proof. Because of minimal sufficiency, there exist functions $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ such that $\boldsymbol{T} = \boldsymbol{g}_1(\boldsymbol{S})$ and $\boldsymbol{S} = \boldsymbol{g}_2(\boldsymbol{T})$. Let $\boldsymbol{x}, \tilde{\boldsymbol{x}}$ with $\boldsymbol{g}_1(\boldsymbol{S}(\boldsymbol{x})) = \boldsymbol{g}_1(\boldsymbol{S}(\tilde{\boldsymbol{x}}))$. Then follows that $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{T}(\tilde{\boldsymbol{x}})$. But from this we get $\boldsymbol{S}(\boldsymbol{x}) = \boldsymbol{g}_2(\boldsymbol{T}(\boldsymbol{x})) = \boldsymbol{g}_2(\boldsymbol{T}(\tilde{\boldsymbol{x}})) = \boldsymbol{S}(\tilde{\boldsymbol{x}})$. Therefore $\boldsymbol{g}_1$ is injective (and also $\boldsymbol{g}_2$), since equal arguments are mapped to equal function value. □

2.1 Classical Estimation Theory
2.1.1 Sufficiency

Theorem 2.7 (Characterization of minimal sufficiency via likelihood ratios)
Define the likelihood ratio

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

A necessary and sufficient condition for the minimal sufficiency of a statistic $\boldsymbol{T}$ *for* $\boldsymbol{\theta}$ *is:*

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

Proof. "$\Rightarrow$" use factorization theorem.
$\Lambda_x(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \frac{h(\boldsymbol{x})g(\boldsymbol{T}(\boldsymbol{x})|\boldsymbol{\theta}_1)}{h(\boldsymbol{x})g(\boldsymbol{T}(\boldsymbol{x})|\boldsymbol{\theta}_2)} \overset{\boldsymbol{T}(\boldsymbol{x})=\boldsymbol{T}(\boldsymbol{x}')}{=} \frac{h(\boldsymbol{x}')g(\boldsymbol{T}(\boldsymbol{x}')|\boldsymbol{\theta}_1)}{h(\boldsymbol{x}')g(\boldsymbol{T}(\boldsymbol{x}')|\boldsymbol{\theta}_2)} = \Lambda_{x'}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ □

2.1 Classical Estimation Theory 2.1.1 Sufficiency

Example 2.8 (Sufficiency in Exponential Family)
The density of an r-parametric exponential family has the form

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

i.e. $\boldsymbol{T}(\boldsymbol{X}) = (T_1(\boldsymbol{X}), \ldots, T_r(\boldsymbol{X}))^\top$ is sufficient for $\boldsymbol{\theta}$ by the factorisation theorem. From theorem 2.7 it follows that $\boldsymbol{T}(\boldsymbol{X})$ is also minimal sufficient, if $\boldsymbol{\gamma}(\Theta)$ contains an open rectangle in $\mathbb{R}^r$. The minimal sufficiency can be characterised according to Lehmann-Scheffé. For this purpose, the concept of completeness is needed.

2.1 Classical Estimation Theory 2.1.1 Sufficiency

Definition 2.8
A statistic $\boldsymbol{T}$ is complete $\overset{def}{\Leftrightarrow}$ *for every real (measurable) function g the following holds:*

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

2.1 Classical Estimation Theory 2.1.1 Sufficiency

Completeness example
$X_1, \ldots, X_n \overset{i.i.d.}{\sim} Po(\lambda)$ and $T = \sum_{i=1}^n X_i$. From Example 2.8, T is sufficient.

$$f_X(x) = \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} \exp(-\lambda) = \frac{1}{x_1! \cdots x_n!} \lambda^{\sum_i x_i} \exp(-n\lambda)$$

Let g be measurable. Since $T = \sum_i X_i \sim Po(n\lambda)$, the following holds

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

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

(two converging power series). From this, it follows that $c_t = 0$ for all t and because $\frac{n^t}{t!} \neq 0$, $g(t) = 0$ for all t. Thus, T is complete (and by the following theorem is also minimaly sufficient). It is not immediately clear from the definition why "completeness" should be a desirable property for an estimator. The following theorem provides one possible reason.

Classical Estimation Theory 2.1.1 Sufficiency

Theorem 2.9 (Lehmann-Scheffé)
Assume that $\boldsymbol{X}$ has density $f(\boldsymbol{x}|\boldsymbol{\theta})$ and $\boldsymbol{T}(\boldsymbol{X})$ is sufficient and complete for $\boldsymbol{\theta}$. Then $\boldsymbol{T}(\boldsymbol{X})$ is minimal sufficient for $\boldsymbol{\theta}$.

2.1 Classical Estimation Theory 2.1.1 Sufficiency

Proof of Lehmann-Scheffé theorem:
It is assumed that a minimal sufficient statistic exists - this was proven by Lehmann and Scheffé (1950). If this is the case, it is unique except for bijective transformations (Lemma 2.6). Denote with $\boldsymbol{S} = \boldsymbol{g}_1(\boldsymbol{T})$ such minimal sufficient statistic for a function $\boldsymbol{g}_1$. Define $\boldsymbol{g}_2(\boldsymbol{S}) = \mathbb{E}[\boldsymbol{T}|\boldsymbol{S}]$. Since $\boldsymbol{S}$ is sufficient for $\boldsymbol{\theta}$, $\boldsymbol{g}_2(\boldsymbol{S})$ does not depend on $\boldsymbol{\theta}$. Consider now

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

Applying the Law of Iterated Expectation yields:

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

Since $\boldsymbol{T}$ is complete, $\boldsymbol{g}(\boldsymbol{T}) = \boldsymbol{0}$ and $\boldsymbol{g}_2(\boldsymbol{S}) = \boldsymbol{T}$ with probability 1, i.e. $\boldsymbol{T}$ is a function of $\boldsymbol{S}$. Because $\boldsymbol{S}$ is a function of every sufficient statistic, this also applies to $\boldsymbol{T}$ and $\boldsymbol{T}$ is thus minimal sufficient. ($\boldsymbol{S}$ and $\boldsymbol{T}$ are equivalent.) □

Remark (Ancillarity of a statistic)

A statistic $\boldsymbol{V}(\boldsymbol{X})$ is called *ancillary* for $\mathcal{P}$, if its distribution does not depend on $\boldsymbol{\theta}$.
Common situation: $\boldsymbol{T} = (\boldsymbol{U}, \boldsymbol{V})$ is sufficient for $\boldsymbol{\theta}$, $\boldsymbol{V}$ ancillary, $\boldsymbol{U}$ not sufficient.

Example 2.9 (Ancillary Statistic)

$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} Unif\left[\theta - \frac{1}{2}, \theta + \frac{1}{2}\right]$. One can show (Davison, 2004, Ex.12.3) that with

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

$T = (U, V)$ is minimal (but not completely) sufficient for θ . Furthermore U alone is not sufficient and V ancillary.

Addendum: Completeness revisited

The definition was (we only look at the scalar case)
A statistic T is *complete* $\overset{\text{def}}{\Leftrightarrow}$ for every real (measurable) function g the following holds:

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

1. Note, that the condition $\mathbb{E}_\theta[g(T)] = 0$ must hold *for all* θ. Therefore, if it holds, one can also say that it is a property of the family of distributions of T.

2. Example. Let $T = X \sim Uniform(0, \theta)$, $\theta > 0$. Then the condition is

$$\mathbb{E}_\theta[g(T)] = \int_0^\theta \frac{g(t)}{\theta} dt = 0 \quad \forall\theta > 0 \ .$$

   Now, $\forall\theta > 0$, this implies that

$$0 = \int_0^\theta g(t)dt \quad \text{and} \quad 0 = \frac{d}{d\theta} \int_0^\theta g(t)dt = g(\theta) \ .$$

   Therefore $g(t) = 0$ for all $t > 0$ which means that $\mathbb{P}_\theta(g(T) = 0) = 1$, $\forall\theta > 0$. I.e. $Uniform(0, \theta)$ is a complete family of distributions.

   Everything goes similar, if we have a sample $X_1, \ldots, X_n$ and the maximum $X_{(n)}$ plays the role of T (T is sufficient). One can use that the pdf of $X_{(n)}$ is

$$f(x) = \begin{cases} nx^{n-1}\theta^{-n} & \text{if } 0 < x < \theta \\ 0 & otherwise \end{cases}$$

3. Lets revisit the example in the slides: $T = X \sim N(\theta, 1)$, $\theta \in \mathbb{R}$. Then

$$\mathbb{E}_\theta[g(T)] = \int g(t) \frac{1}{\sqrt{2\pi}} \exp(-0.5(t - \theta)^2) dt = 0 \quad \forall\theta \in \mathbb{R}$$

is equivalent to

$$\int g(t)\exp(-0.5t^2)\exp(t\theta)dt = 0 \quad \forall\theta \in \mathbb{R}\ . \tag{1}$$

Now, the so-called two-sided Laplace transform of a function is defined as

$$L_f(s) = \int_{-\infty}^{\infty} f(t)\exp(-st)dt\ .$$

We identify in (1) $s = -\theta$ and

$$f(t) = g(t)\exp(-0.5t^2)\ .$$

Therefore, one can conclude that if the Laplace transform is 0, $\forall\theta \in \mathbb{R}$, then $g(t)$ must be zero (almost everywhere) since the exponential function is always greater zero.

We conclude that the normal family (with known variance) is complete.

4. Counter example: let $X, Y \sim B(1,p)$, X, Y independent, $0 < p < 1$. Then $T = X - Y$ is not complete, since taking $g(T) = T$:

$$\mathbb{E}_p(g(T)) = \mathbb{E}_p(T) = \mathbb{E}_p(X-Y) = 0 \quad \forall p \in (0,1)\ ,$$

but $P_p(T = 0) = P_p(X - Y = 0) = P_p(X = Y = 1) + P_p(X = Y = 0) = p^2 + (1-p)^2 < 1$.

2.1.2 Bias of an estimator, variance and MSE

In this section we consider the question of "good estimators". The notion of "optimality" of an estimator depends on the chosen optimality criterion.

2.1 Classical Estimation Theory 2.1.1 Bias of an estimator, variance and MSE

2.1 Classical Estimation Theory 2.1.2 Bias of an estimator, variance and MSE

- Error of an estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{X})$ is $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$.
- Measurement of the error through loss function, for example

$$\begin{aligned}
L(\widehat{\theta}, \theta) &= |\widehat{\theta} - \theta| && \text{distance (θ scalar, else L1-norm),}\\
L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &= \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 && \text{quadratic error,}\\
L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &= \frac{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2}{\|\boldsymbol{\theta}\|^2} && \text{relative square error,}\\
L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &= (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \boldsymbol{D}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) && \text{weighted square error}\\
& && \text{($\boldsymbol{D}$ is positive definite).}
\end{aligned}$$

- Risk function $R(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \mathbb{E}_\theta[L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})]$.
- Here we (mainly) consider quadratic loss, i.e. the risk corresponds to the MSE

2.1 Classical Estimation Theory 2.1.2 Bias of an estimator, variance and MSE

Definition 2.10 (Unbiasedness, bias, variance of an estimator)

- $\widehat{\boldsymbol{\theta}}$ *is called* unbiased $\overset{def}{\Leftrightarrow}$ $\mathbb{E}_\theta[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$.
- $\text{Bias}_\theta(\widehat{\boldsymbol{\theta}}) = \mathbb{E}_\theta[\widehat{\boldsymbol{\theta}}] - \boldsymbol{\theta}$.
- $\text{Var}_\theta(\widehat{\theta}) = \mathbb{E}_\theta[(\widehat{\theta} - \mathbb{E}_\theta[\widehat{\theta}])^2]$, *$\theta$ scalar.*

  $\text{Cov}_\theta(\widehat{\boldsymbol{\theta}}) = \mathbb{E}_\theta[(\widehat{\boldsymbol{\theta}} - \mathbb{E}_\theta[\widehat{\boldsymbol{\theta}}])(\widehat{\boldsymbol{\theta}} - \mathbb{E}_\theta[\widehat{\boldsymbol{\theta}}])^\top]$

Definition 2.11 (MSE)
The mean squared error (MSE) *is defined as*

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

The total error can thus be divided into a random error (variance) and a systematic one (squared bias).

2.1 Classical Estimation Theory 2.1.2 Bias of an estimator, variance and MSE

When comparing two estimators with respect to their MSE, it is possible that for one subrange of Θ the MSE of one estimator may be smaller, while for other subranges, the MSE of the second estimator may be smaller **Example 2.10** (see exercise 2, problem 3)
$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bin}(1, \pi)$.

1. MSE of $\widehat{\pi} = \bar{X}$:
$$\mathbb{E}_\pi[(\bar{X} - \pi)^2] = \text{Var}_\pi(\bar{X}) = \frac{\pi(1-\pi)}{n}.$$

2.1 Classical Estimation Theory 2.1.2 Bias of an estimator, variance and MSE

Example 2.10 continued

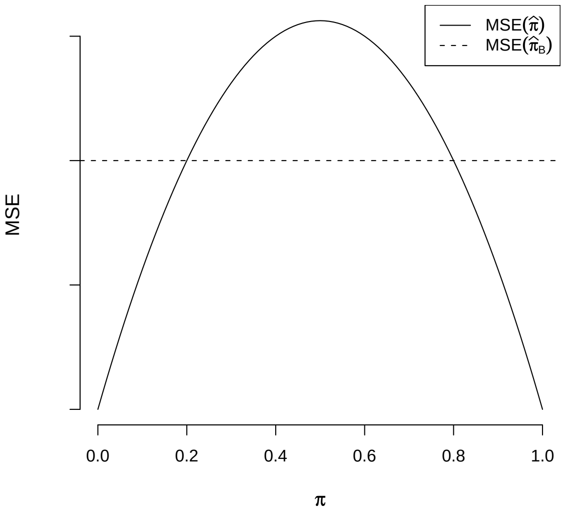
2. MSE of the Bayesian estimator (posterior expectation) with a prior $p(\pi) \sim \text{Be}(a, b)$:
$$\widehat{\pi}_B = \frac{n\bar{X} + a}{a + b + n},$$

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

For $a = b = \sqrt{n}/2$

$$\text{MSE}_\pi(\widehat{\pi}_B) = \frac{1}{4(1+\sqrt{n})^2} = \text{constant with respect to } \pi.$$

2.1 Classical Estimation Theory 2.1.2 Bias of an estimator, variance and MSE



2.1 Classical Estimation Theory 2.1.2 Bias of an estimator, variance and MSE

Bottom line: as a rule, there is no „MSE-optimal" estimator $\widehat{\theta}^{\text{opt}}$, in the sense that $\text{MSE}_\theta(\widehat{\theta}^{\text{opt}}) \leq \text{MSE}_\theta(\widehat{\theta})$ for all θ and all competing $\widehat{\theta}$. However, when restricting ourselves to unbiased estimators, this becomes possible more often. Hence the requirement:

Definition 2.12 ("Admissable" estimator)
An estimator $\widehat{\theta}$ is called admissible $\overset{def}{\Leftrightarrow}$ *there is no estimator $\widetilde{\theta}$ with $\text{MSE}_\theta(\widetilde{\theta}) \leq \text{MSE}_\theta(\widehat{\theta})$ for all θ and $\text{MSE}_\theta(\widetilde{\theta}) < \text{MSE}_\theta(\widehat{\theta})$ for at least one θ, i.e. there is no estimator $\widetilde{\theta}$, $\widehat{\theta}$ that strictly „dominates $\widehat{\theta}$.*

2.1 Classical Estimation Theory 2.1.2 Bias of an estimator, variance and MSE

Definition 2.13 (Generalizations of the MSE on $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k) \in \mathbb{R}^k, k > 1$)
The following two alternatives are common:

1. *MSE (scalar):*

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

2. *MSE-Matrix:*

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

   This variant is often considered in linear models.

2.1 Classical Estimation Theory 2.1.2 Bias of an estimator, variance and MSE

Remark. The j-th diagonal element of the MSE matrix is $\mathrm{MSE}_\theta(\widehat{\theta}_j)$. Comparison of MSE matrices is done according to the *„Löwner"* order:

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

means the difference $\mathrm{MSE}_\theta(\widehat{\boldsymbol{\theta}}) - \mathrm{MSE}_\theta(\widetilde{\boldsymbol{\theta}})$ is positive (semi-)definite. One defines generally for symmetric $(m \times m)$-matrices $\mathbf{A}, \mathbf{B}$:

$$\begin{aligned} \mathbf{A} \leq \mathbf{B} &\overset{\text{def}}{\Leftrightarrow} \mathbf{B} - \mathbf{A} \text{ is positive semi definite,} \\ \mathbf{A} < \mathbf{B} &\overset{\text{def}}{\Leftrightarrow} \mathbf{B} - \mathbf{A} \text{ is positive definite.} \end{aligned}$$

2.1 Classical Estimation Theory 2.1.2 Bias of an estimator, variance and MSE

Example 2.11 (Gaussian experiment)
Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$.

- σ^2 known, μ unknown: MSE comparison of $\bar{X}$ and $T = a\bar{X} + b$.

  $\bar{X}$ is admissable with the smallest MSE among unbiased estimators, but $T = a\bar{X} + b$ has smaller MSE depending μ, a, b

- σ^2 unknown, μ known:
  - One possibility:
  $$S_\mu^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2, \ \mathbb{E}_{\sigma^2}(S_\mu^2) = \sigma^2$$
  - Another possibility:
  $$V_\mu^2 = \frac{1}{n+2}\sum_{i=1}^n (X_i - \mu)^2, \mathbb{E}_{\sigma^2}(V_\mu^2) = \frac{n}{n+2}\sigma^2$$

  It turns out that $\mathrm{MSE}_{\sigma^2}(V_\mu^2) < \mathrm{MSE}_{\sigma^2}(S_\mu^2)$ and so S_μ^2 is not an admissable estimator.

2.1 Classical Estimation Theory 2.1.2 Bias of an estimator, variance and MSE

- μ and σ^2 unknown:
  - One possibility:
  $$S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2,$$
  $$\mathbb{E}_{\sigma^2}(S^2) = \sigma^2, \ \ \mathrm{MSE}_{\sigma^2}(S^2) = \mathrm{Var}_{\sigma^2}(S^2) = \frac{2}{n-1}\,\sigma^4.$$
  - Another possibility:
  $$V^2 = \frac{1}{n+1}\sum_{i=1}^n (X_i - \bar{X})^2,$$
  $$\mathbb{E}_{\sigma^2}(V^2) = \frac{n-1}{n+1}\,\sigma^2, \ \mathrm{MSE}_{\sigma^2}(V^2) = \frac{2}{n+1}\,\sigma^4,$$

    i.e. V^2 dominates S^2.

2.1 Classical Estimation Theory 2.1.2 Bias of an estimator, variance and MSE

- μ and σ^2 (still) unknown:
  - The so called *Stein estimator*

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

dominates V^2 (and thus S^2). Plausibility check: If $\mu = 0$, then $\sum_{i=1}^{n} X_i^2/(n+2)$ is a better estimator than V^2. If $\mu \neq 0$, then V^2 is a better estimator than $\sum_{i=1}^{n} X_i^2/(n+2)$. In the case of Stein estimator, the better estimator is used with higher probability.

2.1 Classical Estimation Theory 2.1.2 Bias of an estimator, variance and MSE

Example 2.12 (Stein's Paradox)
Let $(X_1, \dots, X_m)^\top \sim N_m(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_m)$ be a multivariate normally distributed variable with known $\boldsymbol{\mu} = (\mu_1, \dots, \mu_m)^\top$ and σ^2. The expected values $\mu_1, \dots, \mu_m$ are estimated simultaneously. Note here that the individual components are assumed to be independent. The sample has the form

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

(i.i.d. samples from „groups" $1, \dots, m$). Usual estimators:

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

The (scalar) MSE is:

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

2.1 Classical Estimation Theory 2.1.2 Bias of an estimator, variance and MSE

Paradoxically, the following applies:

1. For $m \leq 2$, $\boldsymbol{T}$ is admissable.
2. For $m \geq 3$, $\boldsymbol{T}$ is *not* admissible and is dominated by the (James-)Stein's estimator

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

It can be shown that $\boldsymbol{T}^*$ is inadmissible itself. The Stein estimator is a so-called *shrinkage estimator.*

2.1 Classical Estimation Theory 2.1.2 Bias of an estimator, variance and MSE

Example 2.13 (Linear model)

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

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

where $\boldsymbol{D}$ is a diagonal matrix with positive diagonal elements. The ridge estimator is also a shrinkage estimator. For comparison of MSE, see exercise.

2.1 Classical Estimation Theory 2.1.2 Bias of an estimator, variance and MSE

Conclusion:: In the simple example of estimating π in $B(1, \pi)$ (see Example 2.10) it can be seen already that in general there are no MSE optimal estimators. Ways out:

1. Restriction to subclass of estimators, for example unbiased (and linear) estimators.
2. Change MSE criterion:
   - Replace $\text{MSE}_\theta(\widehat{\boldsymbol{\theta}})$ by minimizing $\max_{\boldsymbol{\theta}\in\Theta} \text{MSE}_\theta(\widehat{\boldsymbol{\theta}})$ (Minimax criterion)
   - or replace $\text{MSE}_\theta(\widehat{\boldsymbol{\theta}})$ by $\mathbb{E}_{p(\boldsymbol{\theta})}[\text{MSE}_\theta(\widehat{\boldsymbol{\theta}})]$ with a prior distribution $p(\boldsymbol{\theta})$ (Bayes estimator).

Here: strategy 1 with unbiased estimators, see 2.1.4, but first, some definitions and results in 2.1.3.

2.1.3 Fisher Information and Sufficiency

2.1 Classical Estimation Theory 2.1.3 Fisher information and sufficiency

Definition 2.14 (Fisher-regular distribution families)
A family of distributions $\mathcal{P}$ with density $f(\boldsymbol{x}|\boldsymbol{\theta}) = f(x_1, \dots, x_n|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, is called Fisher-regular, *if the following applies:*

1. *The support $\{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}|\boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta}$*
2. *Θ is open in $\mathbb{R}^k$*
3. *The first and second derivatives of $f(\boldsymbol{x}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ exist and are finite functions of $\boldsymbol{\theta}$ for every $\boldsymbol{x}$.*
4. *Interchangeability: for both $f(\boldsymbol{x}|\boldsymbol{\theta})$ and $\log(f(\boldsymbol{x}|\boldsymbol{\theta}))$ first and second differentiation with respect to $\boldsymbol{\theta}$ and integration over $\boldsymbol{x}$ can be interchanged.*

Condition 1. is violated, for example, with $X_1, \dots, X_n \overset{i.i.d.}{\sim} Unif[0;\theta]$ or for Pareto distribution. 2., for example, when $\sigma^2 \geq 0$).

2.1 Classical Estimation Theory 2.1.3 Fisher information and sufficiency

Definition 2.15 (Log-Likelihood, score function and information)

$$
\begin{aligned}
\ell(\boldsymbol{\theta}|\boldsymbol{x}) &= \log f(\boldsymbol{x}|\boldsymbol{\theta}) \quad \textit{(}\text{log-likelihood } \textit{of } \boldsymbol{\theta} \textit{ with respect to sample } \boldsymbol{x}\textit{)} \\
\boldsymbol{s}(\boldsymbol{\theta}|\boldsymbol{x}) &= \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}|\boldsymbol{x}) = \left(\frac{\partial}{\partial \theta_1} \ell(\boldsymbol{\theta}|\boldsymbol{x}), \dots, \frac{\partial}{\partial \theta_k} \ell(\boldsymbol{\theta}|\boldsymbol{x}) \right)^\top \\
&\qquad\qquad\qquad\qquad\qquad \textit{(}\text{score function}\textit{)} \\
\boldsymbol{J}(\boldsymbol{\theta}|\boldsymbol{x}) &= -\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \quad \textit{(}\text{observed information matrix } \textit{of the} \\
&\qquad \textit{sample with entries} \quad (\boldsymbol{J}(\boldsymbol{\theta}|\boldsymbol{x}))_{ij} = -\frac{\partial^2 \log f(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\textit{)} \\
\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) &= \text{E}_\theta[\boldsymbol{J}(\boldsymbol{\theta}|\boldsymbol{X})] \quad \textit{(}\text{expected } \textit{or } \text{Fisher information matrix}\textit{)}
\end{aligned}
$$

2.1 Classical Estimation Theory 2.1.3 Fisher information and sufficiency

Theorem 2.16
If $\mathcal{P}$ is Fisher-regular, then:

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

2.1 Classical Estimation Theory
2.1.3 Fisher information and sufficiency

Proof For 1.:

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

2.1 Classical Estimation Theory
2.1.3 Fisher information and sufficiency

For 2.:

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

using the quotient rule of differentiation

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

2.1 Classical Estimation Theory
2.1.3 Fisher information and sufficiency

For 2. (continued): Upon interchanging differentiation and integration, the first term of the sum is equal to 0. For the second term, we use result from 1.

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

□

2.1 Classical Estimation Theory
2.1.3 Fisher information and sufficiency

Further properties:

- If $X_1, \ldots, X_n$ are independent and distributed by $X_i \sim f_i(x|\boldsymbol{\theta}),\ i = 1, \ldots, n$, then the following holds:

$$\begin{aligned}
\ell(\boldsymbol{\theta}|\boldsymbol{x}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}|x_i) \quad &, \quad \ell_i(\boldsymbol{\theta}|x_i) = \log f_i(x_i|\boldsymbol{\theta}) \\
\boldsymbol{s}(\boldsymbol{\theta}|\boldsymbol{x}) = \sum_{i=1}^{n} \boldsymbol{s}_i(\boldsymbol{\theta}|x_i) \quad &, \quad s_i(\boldsymbol{\theta}|x_i) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f_i(x_i|\boldsymbol{\theta}) \\
\boldsymbol{J}(\boldsymbol{\theta}|\boldsymbol{x}) = -\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \quad &= \quad \sum_{i=1}^{n} -\left(\frac{\partial^2 \log f_i(x_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right)
\end{aligned}$$

2.1 Classical Estimation Theory
2.1.3 Fisher information and sufficiency

- For $X_1, \dots, X_n$ i.i.d. as $X_1 \sim f_1(x|\boldsymbol{\theta})$ it follows that,

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

  where

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

  is the expected information of an individual observation, i.e. the expected information matrix of the sample $X_1, \dots, X_n$ is n-times the expected information of a (typical) sample variable X_1.

2.1 Classical Estimation Theory 2.1.3 Fisher information and sufficiency

- For a statistic $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X}), \boldsymbol{X} = (X_1, \dots, X_n)^\top$ with $\boldsymbol{T} \sim f_T(\boldsymbol{t}|\boldsymbol{\theta})$ one can analagously define score function and Fisher information. In particular

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

2.1 Classical Estimation Theory 2.1.3 Fisher information and sufficiency

Theorem 2.17 (Sufficiency and Fisher information)
Let $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ be the Fisher information for $\boldsymbol{X}$. Then for every statistic $\boldsymbol{T}$, the following holds under Fisher-regularity:

1. $\boldsymbol{\mathcal{I}}_T(\boldsymbol{\theta}) \leq \boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$.
2. $\boldsymbol{\mathcal{I}}_T(\boldsymbol{\theta}) = \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) \;\Leftrightarrow\; \boldsymbol{T}$ *is sufficient for* $\boldsymbol{\theta}$.

So, if a statistic $\boldsymbol{T}$ is sufficient, no (expected) information is „given away".

2.1 Classical Estimation Theory 2.1.3 Fisher information and sufficiency

Sufficiency and Fisher information example
Let $X_1, \dots, X_n$ be i.i.d. with $X \sim N(\mu, \sigma^2)$, known σ^2 and $\boldsymbol{\theta} = \mu$. Then, for information matrices based on $\boldsymbol{X}$ and sufficient statistic $T = \frac{1}{n}\sum_{i=1}^n X_i \sim N(\mu, \sigma^2/n)$, the following applies:

$$\begin{aligned}
\ell(\boldsymbol{\theta}|\boldsymbol{x}) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2 \\
\Rightarrow \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) &= -\mathbb{E}[-\frac{\partial^2}{\partial\mu^2}\ell(\boldsymbol{\theta}|\boldsymbol{x})] = \frac{1}{2\sigma^2}2n = \frac{n}{\sigma^2} \\
\ell(\boldsymbol{\theta}|\boldsymbol{x}) &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\frac{\sigma^2}{n}) - \frac{n}{2\sigma^2}(t - \mu)^2 \\
\Rightarrow \boldsymbol{\mathcal{I}}_{\boldsymbol{T}}(\boldsymbol{\theta}) &= -\mathbb{E}[-\frac{\partial^2}{\partial\mu^2}\ell(\boldsymbol{\theta}|\boldsymbol{t})] = \frac{n}{2\sigma^2}2 = \frac{n}{\sigma^2}
\end{aligned}$$

so $\boldsymbol{\mathcal{I}}_{\boldsymbol{T}}(\boldsymbol{\theta}) = \boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$

2.1.4 Unbiased estimators

2.1 Classical Estimation Theory 2.1.4 Unbiased estimators

- In this section we restrict oursevles to unbiased estimators.
- „Nice" results for finite n, but for comparatively simple statistical models.
- Problem: there are no „reasonable" unbiased estimators for more complex models.

- But: a few results can be translated to asymptotic case for $n \to \infty$ (cf. 2.1.5).

2.1 Classical Estimation Theory
2.1.4 Unbiased estimators

Information inequalities

I. $\theta \in \mathbb{R}$ (scalar). In addition to θ, transformed parameters $\tau(\theta)$ are also considered. Whenever derivatives are needed, we implicitly assume they exist.

2.1 Classical Estimation Theory
2.1.4 Unbiased estimators

Theorem 2.18
Let $f(\boldsymbol{x}|\theta)$ be Fisher-regular.

1. If $\widehat{\theta}$ is unbiased with respect to θ, then:

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

2. If $T = T(\boldsymbol{X})$ is unbiased with respect to $\tau(\theta)$, then:

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

$\frac{(\tau'(\theta))^2}{\mathcal{I}(\theta)}$ *is called* Cramer-Rao bound.

3. If $\widehat{\theta}$ has bias $B(\theta) = \mathbb{E}_\theta[\widehat{\theta}] - \theta$, then:

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

2.1 Classical Estimation Theory
2.1.4 Unbiased estimators

Proof.
We show 2. From there follows 1. for $\tau(\theta) = \theta$, and 3. for $\tau(\theta) = \theta + B(\theta)$
Differentiation of

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

with respect to θ and using Fisher's regularity yields:

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

2.1 Classical Estimation Theory
2.1.4 Unbiased estimators

Using the Cauchy-Schwarz inequality

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

yields

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

Therefore,

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

□

2.1 Classical Estimation Theory 2.1.4 Unbiased estimators

Remark.
Equality is assumed only for one-parametric exponential family $f(\boldsymbol{x}|\theta) = h(\boldsymbol{x})\exp(b(\theta)+\gamma(\theta)T(\boldsymbol{x}))$. In this case $T(\boldsymbol{x})$ is an efficient estimator for its expected value $\tau(\theta) = -b'(\theta)/\gamma'(\theta)$, i.e. unbiased estimator with the smallest variance. So, rather a small class of models.

Sketch of proof: $s(\theta|\boldsymbol{x}) = b'(\theta) + \gamma'(\theta)T(\boldsymbol{x})$ is linear in $T(\boldsymbol{x})$. In the proof of theorem 2.18, the equality $\mathrm{Cov}_\theta(T(\boldsymbol{X}), s(\theta|\boldsymbol{X})) = \sqrt{\mathrm{Var}(T(\boldsymbol{X})}\sqrt{\mathrm{Var}(s(\theta|\boldsymbol{X}))}$ is assumed only if $T(\boldsymbol{x})$ and $s(\theta|\boldsymbol{X})$ are linearly dependent.

2.1 Classical Estimation Theory 2.1.4 Unbiased estimators

II. **Information inequalities** for multidimensional $\boldsymbol{\theta}$ or $\boldsymbol{\tau}(\boldsymbol{\theta})$.

Theorem 2.19
Let $f(\boldsymbol{x}|\boldsymbol{\theta})$ be Fisher-regular.

1. *If $\widehat{\boldsymbol{\theta}}$ is unbiased with respect to $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, then:*

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

   where „≥" is according to the the Löwner order (see. slide 117). It follows in particular that $\mathrm{Var}_\theta(\widehat{\theta}_j) \geq f_{jj}$, $j = 1, \ldots, k$, where f_{jj} is the j-th diagonal element of $\boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta})$.
2. *If $\boldsymbol{T}$ unbiased with respect to $\boldsymbol{\tau}(\boldsymbol{\theta})$, then we have*

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

   with the derivative matrix $(\boldsymbol{H}(\boldsymbol{\theta}))_{ij} = \frac{\partial}{\partial\theta_j}\tau_i(\boldsymbol{\theta})$. The matrix $\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})^\top$ is the Cramer-Rao bound*.*

2.1 Classical Estimation Theory 2.1.4 Unbiased estimator

Remark
The previous remark on scalar θ applies similarly

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

for multi-parametric exponential family.

2.1 Classical Estimation Theory 2.1.4 Unbiased estimators

Example 2.14 (Cramer-Rao bound for $X \sim N(\mu, \sigma^2)$)
$X_1, \ldots, X_n$ i.i.d. with $X \sim N(\mu, \sigma^2)$, $\boldsymbol{\theta} = (\mu, \sigma^2)$. Then, for the information matrix,

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

is the lower bound for the covariance of all unbiased estimators of $\boldsymbol{\theta}$.

2.1 Classical Estimation Theory 2.1.4 Unbiased estimators

Best unbiased estimators
Unbiased estimators with minimum variance within a given class of estimators are called *efficient.* The information inequalities motivate the definition:

Definition 2.20 (Uniformly Minimum Variance Unbiased (UMVU) estimator)

1. *Scalar θ:*

   The estimator $\widehat{\theta}_{\mathit{eff}}$ of θ is called a uniformly minimum variance unbiased *(or* UMVU*) estimator $\overset{def}{\Leftrightarrow}$ $\widehat{\theta}_{\mathit{eff}}$ is unbiased, and $\operatorname{Var}_\theta(\widehat{\theta}_{\mathit{eff}}) \leq \operatorname{Var}_\theta(\widehat{\theta})$ for all θ and any unbiased estimator $\widehat{\theta}$.*

2. *Multidimensional $\boldsymbol{\theta}$:*

   Replace $\operatorname{Var}_\theta(\widehat{\theta}_{\mathit{eff}}) \leq \operatorname{Var}_\theta(\widehat{\theta})$ with $\operatorname{Cov}_\theta(\widehat{\boldsymbol{\theta}}_{\mathit{eff}}) \leq \operatorname{Cov}_\theta(\widehat{\boldsymbol{\theta}})$.

2.1 Classical Estimation Theory 2.1.4 Unbiased estimators

Theorem 2.21 (Efficiency and information inequalities)
Let $f(\boldsymbol{x}|\boldsymbol{\theta})$ be Fisher-regular and $\widehat{\boldsymbol{\theta}}$ be an unbiased estimator of $\boldsymbol{\theta}$. If $\operatorname{Cov}_\theta(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$, then $\widehat{\boldsymbol{\theta}}$ is a UMVU-estimator.

Proof. The statement follows directly from the information inequality and the above definition. □

2.1 Classical Estimation Theory 2.1.4 Unbiased estimators

Example 2.15 (Gaussian experiment)
Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ with unknown μ, σ^2. From example 2.14 we know that $\mathcal{I}(\mu) = n/\sigma^2$ and thus $\mathcal{I}^{-1}(\mu) = \sigma^2/n = \operatorname{Var}(\bar{X})$. Then $\bar{X}$ is UMVU for μ. But

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

therefore the Cramer-Rao bound is not achieved, so it cannot be concluded that S^2 is UMVU for σ^2. (This follows from theorem 2.23.)

2.1 Classical Estimation Theory 2.1.4 Unbiased estimators

Example 2.16 (Linear Model)

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

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{\text{LS}} &= \widehat{\boldsymbol{\beta}}_{\text{ML}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} \text{ is efficient (UMVU)} \\ &\quad \text{and satisfies Cramer-Rao bound} \\ \widehat{\sigma}^2 &= \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \text{ is a UMVU-estimator for } \sigma^2, \\ &\quad \text{that does not achieve Cramer-Rao bound (theorem 2.23).} \end{aligned}$$

2.1 Classical Estimation Theory 2.1.4 Unbiased estimators

Comment. A distinction must be made between the following situations:

1. There exists a UMVU estimator, the variance of which is equal to the Cramer-Rao bound.
2. There is a UMVU estimator, the variance of which is greater than the Cramer-Rao bound (can be shown by the Lehmann-Scheffé theorem, see theorem 2.23).
3. Most common case: there exists (for finite sample size) no UMVU estimator.

2.1 Classical Estimation Theory 2.1.4 Unbiased estimators

Conclusion: finite theory unbiased estimator is of limited practical use But: there is an analogous asymptotic theory with a broader practical use based on the finite theory (see Section 2.1.5). To construct UMVU estimators, the following two statements are useful:

2.1 Classical Estimation Theory 2.1.4 Unbiased estimators

Theorem 2.22 (Rao-Blackwell)
Let $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$ be sufficient for $\boldsymbol{\theta}$ or $\mathcal{P}$ and $\widehat{\boldsymbol{\theta}}$ be unbiased for $\boldsymbol{\theta}$. For estimator

$$\widehat{\boldsymbol{\theta}}_{RB} = \mathbb{E}_\theta[\widehat{\boldsymbol{\theta}}|\boldsymbol{T}] \quad \text{(„Rao-Blackwellization”)}$$

the following holds:

1. $\widehat{\boldsymbol{\theta}}_{RB}$ *is unbiased for* $\boldsymbol{\theta}$*.*
2. $\mathrm{Cov}_\theta(\widehat{\boldsymbol{\theta}}_{RB}) \leq \mathrm{Cov}_\theta(\widehat{\boldsymbol{\theta}})$
3. *In 2. the equality holds iff* $\widehat{\boldsymbol{\theta}}$ *depends only on* $\boldsymbol{T}$ *, i.e.* $\widehat{\boldsymbol{\theta}}_{RB} = \widehat{\boldsymbol{\theta}}$ *with probability 1.*

2.1 Classical Estimation Theory
2.1.4 Unbiased estimators

Theorem 2.23 (Lehmann-Scheffé)
If $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$ *is sufficient and complete (i.e. minimal sufficient) and* $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ *is an unbiased estimator, then*

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

is a UMVU estimator uniquely determined with probability 1.

Specifically: an unbiased estimator $\boldsymbol{U} = \boldsymbol{g}(\boldsymbol{T})$ that depends only on a complete and sufficient statistic $\boldsymbol{T}$ is the UMVU estimator.

2.1 Classical Estimation Theory
2.1.4 Unbiased estimators

Examples of Lehmann-Scheffé's theorem

- Normal distribution: let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. $\boldsymbol{\theta} = (\mu, \sigma^2)$.

  $\boldsymbol{T} = (\sum X_i, \sum X_i^2)$ is sufficient and complete. Equivalently $\widetilde{\boldsymbol{T}} = (\bar{X}, S^2)$ is sufficient and complete (examples 2.7, 2.8). Thus, the following holds:

  - $\bar{X}$ as an estimator for μ is UMVU.
  - S^2 as an estimator for σ^2 is UMVU, even if the Cramer-Rao bound is not achieved.
  - $\frac{\bar{X}}{S^2}$ as an estimator for $\mathbb{E}(\frac{\bar{X}}{S^2})$ is UMVU.

- In the example 16.2 (linear model) $\hat{\sigma}^2$ is unbiased for σ^2. Moreover it is possible to show that $\hat{\sigma}^2$ is a function of a complete sufficient statistic.

  Therefore, $\hat{\sigma}^2$ is a UMVU estimator for σ^2.

2.1.5 Asymptotic properties and criteria

2.1 Classical Estimation Theory
2.1.5 Asymptotic properties and criteria

Important estimators (moment estimators, shrinkage estimators, ML- and quasi-ML estimators etc.) are generally not unbiased, but have favorable asymptotic ($n \to \infty$) properties. Let the following

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

be an estimator for $\boldsymbol{\theta}$.

Definition 2.24 (Asymptotical unbiasedness)
$\widehat{\boldsymbol{\theta}}_n$ *is called* asymptotically unbiased $\overset{def}{\Leftrightarrow}$

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

2.1 Classical Estimation Theory
2.1.5 Asymptotic properties and criteria

Definition 2.25 (Consistency)

1. $\widehat{\boldsymbol{\theta}}_n$ *is* (weakly) consistent *for* $\boldsymbol{\theta}$ *(in notation:* $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathbb{P}} \boldsymbol{\theta}$ *(for all* $\boldsymbol{\theta}$*))* $\overset{def}{\Leftrightarrow}$

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

2. $\widehat{\boldsymbol{\theta}}_n$ *is called* MSE-consistent *for* $\boldsymbol{\theta} \overset{def}{\Leftrightarrow}$

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

3. $\widehat{\boldsymbol{\theta}}_n$ *is* strongly consistent *for* $\boldsymbol{\theta} \overset{def}{\Leftrightarrow}$

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

Remark.

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

1. From the (generalized) Chebyshev inequality, it follows that

$$\widehat{\boldsymbol{\theta}}_n \text{ MSE-consistent } \Rightarrow \widehat{\boldsymbol{\theta}}_n \text{ weakly consistent.}$$

2. Since $\mathrm{MSE}_\theta(\widehat{\theta}_n) = \mathrm{Var}_\theta(\widehat{\theta}_n) + (\mathrm{Bias}_\theta(\widehat{\theta}_n))^2$, it follows that:

$$\widehat{\theta}_n \text{ is MSE-consistent}$$

$$\Leftrightarrow \mathrm{Var}_\theta(\widehat{\theta}_n) \to 0 \text{ and } \mathrm{Bias}_\theta(\widehat{\theta}_n) \to 0 \text{ for all } \theta$$

(or component-wise for vectors $\boldsymbol{\theta}$).

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

3. If $\widehat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}$ and g is a continuous mapping, then $g(\widehat{\boldsymbol{\theta}}_n)$ is also consistent for $g(\boldsymbol{\theta})$ (Continuous Mapping Theorem).

4. Proof of consistency usually exists in the application(weak) laws of large numbers (for iid variables; inid Variables; dependent variables, e.g. martingales,Markov processes,, ...).

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

Example 2.17 Let $\mu = \mathbb{E}(X_i), \sigma^2 = Var(X_i)$

1. $\bar{X}_n = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$ is (MSE-)consistent for $n \to \infty$ because $\mathbb{E}(\bar{X}_n) = \mu$ and $\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n} \to 0$.

2. $S_n^2 = \frac{1}{n-1}\sum\limits_{i=1}^{n}(X_i - \bar{X}_n)^2$ and $\widetilde{S}_n^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \bar{X}_n)^2$ are (MSE-)consistent for σ^2.

3. With $g(x) = \sqrt{x}$ it follows that

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

are consistent for σ.

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

4. $S_n^2/\bar{X}_n$ is consistent for σ^2/μ for $\mu > 0$, because we can apply the continuity theorem again for $\boldsymbol{\theta} = (\mu, \sigma)$ and $g(\boldsymbol{\theta}) = \sigma^2/\mu$.

5. $\widehat{\pi}_n$ is consistent for π (in Bernoulli experiment, example 2.10).

6. $\widehat{\boldsymbol{\beta}}_{LS}, \widehat{\boldsymbol{\beta}}_{Ridge}$ are consistent for $\boldsymbol{\beta}$ in the linear model under certain weak assumptions on $\boldsymbol{X}$, see example 2.19.

2.1 Classical Estimation Theory
2.1.5 Asymptotic properties and criteria

Asymptotic Normality

Many estimators (LS, moment based, ML-, quasi-ML, Bayesian estimators) are asymptotically normally distributed under regularity assumptions. Informally, this means that for large n, $\widehat{\boldsymbol{\theta}}_n$ is not only approximately unbiased, but also approximately normally distributed, in short

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

with (approximate) covariance matrix

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

which is estimated by

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

2.1 Classical Estimation Theory
2.1.5 Asymptotic properties and criteria

The (estimated) variances form the diagonal of $\boldsymbol{V}(\widehat{\boldsymbol{\theta}}_n)$

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

of the components $\widehat{\theta}_j, j = 1, \ldots, k$, of $\widehat{\boldsymbol{\theta}}_n$. $\Rightarrow$ "The usual" output of a statistical software is

$$\underbrace{\widehat{\theta}_j}_{\text{Estimator}} \quad \underbrace{\sqrt{v_{jj}(\widehat{\boldsymbol{\theta}})}}_{\text{Standard error}} \quad \underbrace{\text{t}}_{\text{t-statistic}} \quad \underbrace{\text{p}}_{\text{p-value}}$$

2.1 Classical Estimation Theory
2.1.5 Asymptotic properties and criteria

Example 2.18.

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F(x|\boldsymbol{\theta})$ with $\mathbb{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. But let F *not* be equal to Φ (normal distribution), but rather, for example, distribution functions of $B(\pi) = Bin(1, \pi)$ or $Po(\lambda)$. For $\bar{X}_n$ we have

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

From the central limit theorem, it follows

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

for example

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

2.1 Classical Estimation Theory
2.1.5 Asymptotic properties and criteria

More precise formulation:

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

so in our example

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

or

$$\left.\begin{array}{rcl} \frac{\bar{X}_n - \mu}{\sigma}\sqrt{n} & \overset{d}{\to} & N(0,1), \\ \frac{\bar{X}_n - \pi}{\sqrt{\pi(1-\pi)}}\sqrt{n} & \overset{d}{\to} & N(0,1). \end{array}\right\} \text{central limit theorem}$$

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

The $\sqrt{n}$-normalization is particularly suitable for iid sample variables.

For sample variables that are not identically distributed, such as $y_1|\boldsymbol{x}_1, \ldots, y_n|\boldsymbol{x}_n$ in regression, to normalise by $\sqrt{n}$, one needs to impose requirements that are (in part) unnecessarily restrictive .

It is better thus to perform „matrix normalization" with the help of a „root" $\boldsymbol{\mathcal{I}}^{\frac{1}{2}}(\boldsymbol{\theta})$ of the information matrix.

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

Sidenote: root of a positive definite matrix

- $\boldsymbol{A} \in \mathbb{R}^{m \times m}$ is positive definite, if $\boldsymbol{A}$ is symmetric and $\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} > 0$ holds for all $\boldsymbol{x} \neq \boldsymbol{0}$.
- The matrix $\boldsymbol{A}^{\frac{1}{2}}$ is then called *(left) root* of $\boldsymbol{A}$ $\overset{\text{def}}{\Leftrightarrow}$

$$\boldsymbol{A}^{\frac{1}{2}} \quad \underbrace{(\boldsymbol{A}^{\frac{1}{2}})^\top}_{=\boldsymbol{A}^{\frac{\top}{2}}\text{, right root}} \quad = \boldsymbol{A}.$$

However, $\boldsymbol{A}^{\frac{1}{2}}$ is not unique, as for any orthogonal matrix $\boldsymbol{Q}$, $\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{Q}$ is also a left root:

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

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

- Two common types of roots:
  1. **Symmetric root:** consider the spectral decomposition of $\boldsymbol{A}$ with matrix of orthonormal eigenvectors $\boldsymbol{P} \in \mathbb{R}^{m \times m}$ as columns and eigenvalues $\lambda_i > 0$.

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

     Then, the following applies

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

     $\boldsymbol{A}^{\frac{1}{2}}$ is called a *symmetric root of* $\boldsymbol{A}$. (This decomposition is numerically complex!)

     2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

  2. **Cholesky root:** Let $\boldsymbol{A}^{\frac{1}{2}} := \boldsymbol{C}$ be the lower triangular matrix with positive diagonal elements and $\boldsymbol{C}\boldsymbol{C}^\top = \boldsymbol{A}$. Then $\boldsymbol{C}$ is the *unique Cholesky-root* of $\boldsymbol{A}$. (This is comparatively easy to compute numerically!)

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

- **Applications in statistics**
  1. Generation of $N_m(\boldsymbol{0}, \boldsymbol{\Sigma})$ distributed random numbers ($\boldsymbol{\Sigma}$ is specified):
     Draw m independent $N(0,1)$ distributed random variables $\boldsymbol{Z} = (Z_1, ..., Z_m)^T$.
     Compute Cholesky root of $\boldsymbol{\Sigma}$ and $\boldsymbol{Y} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}$. The following then applies

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

2.1 Classical Estimation Theory
2.1.5 Asymptotic properties and criteria

2. Matrix normalization with asymptotic normal distribution:
Example 2.19 (Asymptotic normality of the LS estimator in linear model)
Let $y_1|\boldsymbol{x}_1, \dots, y_n|\boldsymbol{x}_n$ be independent. Then,

$$\begin{aligned} \mathbb{E}[y_i|\boldsymbol{x}_i] = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \quad \mathrm{Var}(y_i|\boldsymbol{x}_i) = \sigma^2, \quad i = 1, \dots, n, \\ \Leftrightarrow\ \boldsymbol{y}_n = \boldsymbol{X}_n \boldsymbol{\beta} + \boldsymbol{\varepsilon}_n,\ \ \mathbb{E}[\boldsymbol{\varepsilon}_n] = \boldsymbol{0},\ \mathrm{Cov}(\boldsymbol{\varepsilon}_n) = \sigma^2 \boldsymbol{I}_n, \end{aligned}$$

where $\boldsymbol{X}_n$ here is the design matrix with rows denoted as $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$. The LS estimator is

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

The information matrix under the assumption of normal distribution is

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

2.1 Classical Estimation Theory
2.1.5 Asymptotic properties and criteria

Central limit theorems (for independent, not identically distributed random variables, short: inid) yield under suitable requirements (informally):

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

More precise formulations assume that

$$\underset{n \to \infty}{\text{(p-)lim}} \frac{1}{n} \boldsymbol{X}_n^\top \boldsymbol{X}_n =: \boldsymbol{A} > 0 \tag{2}$$

exists (so: $\boldsymbol{X}_n^\top \boldsymbol{X}_n \approx n\boldsymbol{A} \ \Leftrightarrow\ (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \approx \boldsymbol{A}^{-1}/n$ for large n).

2.1 Classical Estimation Theory
2.1.5 Asymptotic properties and criteria

Application of the (multivariate) central limit theorem then yields:

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

or

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

Assumption (2) is fulfilled, if, for example, $\boldsymbol{x}_i,\ i = 1, \dots, n$, are iid. realizations of stochastic covariates $\boldsymbol{X} = (X_1, \dots, X_p)^\top$.

2.1 Classical Estimation Theory
2.1.5 Asymptotic properties and criteria

According to the law of large numbers:

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

Typically assumption (2) is *not* fulfilled for deterministic regressors with trend. The simplest example for this is a linear trend: $x_i = i$ for $i = 1, \dots, n$ and $y_i = \beta_1 i + \varepsilon_i$. Then

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

and therefore

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

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

In this case a different normalisation is necessary instead of $\sqrt{n}$, for example a matrix normalization

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

Then the asymptotic normality of the LS estimator

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

or

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

can be shown under following very weak conditions:

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

(D) Divergence condition: For $n \to \infty$:

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

An equivalent requirement is:

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

where $\lambda_{\min}$ is the smallest eigenvalue of $\boldsymbol{X}_n^\top \boldsymbol{X}_n$. The divergence condition ensures that the „information matrix"

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

for $n \to \infty$ diverges against $\boldsymbol{\infty}$, the information with $n \to \infty$ thus grows continuously. The following is true: (D) is sufficient and necessary for the (weak and strong) consistency of the least squares estimator $\widehat{\boldsymbol{\beta}}_n$.

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

(N) Normality condition:

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

(N) ensures that the information of each observation i compared to the total information $\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top$ is negligible asymptotically.

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

The following holds under (D) and (N)

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

(Proof with limit theorems for independent, not identically distributed random variables), , i.e. for practical purposes:

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

for a sufficiently large sample size n. In addition σ^2 by a consistent estimator $\widehat{\sigma}^2$ to be replaced.

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

Definition 2.26 (Asymptotic Normality)

$\widehat{\boldsymbol{\theta}}_n$ *is called* asymptotically normal *for* $\boldsymbol{\theta}$ $\overset{def}{\Leftrightarrow}$ *there exists one sequence of matrices* $\boldsymbol{A}_n$ *with* $\lambda_{\min}(\boldsymbol{A}_n) \overset{(\mathbb{P})}{\longrightarrow} \infty$*, such that*

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

with non-negative definite (usually positive definite) matrix $\boldsymbol{V}(\boldsymbol{\theta})$*.*

A special case where $\boldsymbol{A}_n = n\boldsymbol{I}_k$ *corresponds to* $\sqrt{n}$*-normalization*

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

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

Remark.

1. Practical formulation:
$$\widehat{\boldsymbol{\theta}}_n \overset{a}{\sim} N_k(\boldsymbol{\theta}, \boldsymbol{V}(\boldsymbol{\theta})/n)$$
or
$$\widehat{\boldsymbol{\theta}}_n \overset{a}{\sim} N_k(\boldsymbol{\theta}, (\boldsymbol{A}_n^{1/2})^{-1} V(\boldsymbol{\theta})(\boldsymbol{A}_n^{1/2})^{-\top}).$$
In this case $\boldsymbol{\theta}$ in $\boldsymbol{V}(\boldsymbol{\theta})$ may be replaced with $\widehat{\boldsymbol{\theta}}_n$.
2. Often $\boldsymbol{V}(\boldsymbol{\theta}) = \boldsymbol{I}_k$ is possible, if normalized appropriately, for example in ML estimation.

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

Example 2.20.
Let $X_1, \ldots, X_n$ be i.i.d. random variables with (known) expected value μ and variance σ^2. Then

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

is unbiased and asymptotically normally distributed (for σ^2) with $V(\sigma^2) = \mu_4 - \sigma^4$, $\mu_4 = \mathbb{E}[(X_i - \mu)^4] < \infty$.
For the variance one gets:

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

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

The requirements for using the central limit theorem are satisfied Thus it follows that:

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

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

Delta method
Let $\widehat{\boldsymbol{\theta}}_n$ be an asymptotically normally distributed estimator for $\boldsymbol{\theta}$. Question: For a given mapping

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

how is the estimator $\boldsymbol{h}(\widehat{\boldsymbol{\theta}})$ for $\boldsymbol{h}(\boldsymbol{\theta})$ distributed?

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

Theorem 2.27 (Delta Method)
Let h be defined like above.

1. *θ scalar: for all θ, for which h is continuously differentiable with $h'(\theta) \neq 0$, it holds that:*

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

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

2.1 Classical Estimation Theory
2.1.5 Asymptotic properties and criteria

2. $\boldsymbol{\theta}$ vector: let $\boldsymbol{h}$ be given by

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

with Jacobian

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

with full rank r. For all $\boldsymbol{\theta}$, for which $\boldsymbol{h}(\boldsymbol{\theta})$ is component-wise continuously partially differentiable and every row of $H(\boldsymbol{\theta})$ is not equal to the zero vector, the following holds:

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

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

2.1 Classical Estimation Theory
2.1.5 Asymptotic properties and criteria

Sketch of proof for scalar θ Taylor expansion of $h(\widehat{\theta}_n)$ at θ yields:

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

For a sequence of random variables Z_n

$$Z_n = o(a_n) \quad \text{if } Z_n/a_n \overset{\mathbb{P}}{\rightarrow} 0 \text{ for } n \rightarrow \infty.$$

Thus:

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

or

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

From $\sqrt{n}(\widehat{\theta}_n - \theta) \overset{d}{\rightarrow} N(0, V(\theta))$ it follows that $\sqrt{n}(h(\widehat{\theta}_n) - h(\theta)) \overset{d}{\rightarrow} N(0, h'(\theta)^2 V(\theta))$. □

2.1 Classical Estimation Theory
2.1.5 Asymptotic properties and criteria

Delta method example
Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} B(\pi)$. For the estimator $\widehat{\pi} = \frac{1}{n}\sum_{i=1}^n X_i$, it holds that:

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

(see example 2.18). Now consider the odds $h(\pi) = \frac{\pi}{1-\pi}$ and the estimator $h(\widehat{\pi}) = \frac{\widehat{\pi}}{1-\widehat{\pi}}$. According to the delta method, the following holds

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

and therefore $\frac{\widehat{\pi}}{1-\widehat{\pi}} \overset{a}{\sim} N(\frac{\pi}{1-\pi}, \frac{\pi}{n(1-\pi)^3})$, since

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

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

Asymptotic Cramer-Rao bound and asymptotic efficiency
Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} f(x|\boldsymbol{\theta})$ and

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

be the expected Fisher information of observation X_i. The information of the entire sample is then $X_1, \dots, X_n$

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

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

Theorem 2.27 (Asymptotic Cramer-Rao inequality)
Under Fisher regularity and slight additional assumptions, the following applies:

1. *From* $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \overset{d}{\to} N_k(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\theta}))$ *follows* $\boldsymbol{V}(\boldsymbol{\theta}) \geq \boldsymbol{i}^{-1}(\boldsymbol{\theta})$.
2. *From* $\sqrt{n}(\boldsymbol{h}(\widehat{\boldsymbol{\theta}}_n) - \boldsymbol{h}(\boldsymbol{\theta})) \overset{d}{\to} N_r(\boldsymbol{0}, \boldsymbol{D}(\boldsymbol{\theta}))$ *follows*

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

with "$\geq$" according to the Löwner-order (and the notation from the delta method rule, theorem 2.27).

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

Definition 2.28 (Best asymptotically normal (BAN) estimator)
$\widehat{\boldsymbol{\theta}}_n$ *is called* BAN-estimator*, if in 1. it holds that:*

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

From the delta method definition it follows immediately that:

Theorem 2.29 (Transformation of BAN estimators)
If $\widehat{\boldsymbol{\theta}}_n$ *is a BAN-estimator for* $\boldsymbol{\theta}$*, then* $\boldsymbol{h}(\widehat{\boldsymbol{\theta}}_n)$ *is BAN-estimator für* $\boldsymbol{h}(\boldsymbol{\theta})$.

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

Remark. The concept of asymptotic efficiency can be transferred to the matrix normalisation: $\widehat{\boldsymbol{\theta}}$ is BAN-estimator for $\boldsymbol{\theta}$ exactly when

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

or $\widehat{\boldsymbol{\theta}}_n \overset{a}{\sim} N_k(\boldsymbol{\theta}, \boldsymbol{\mathcal{I}}^{-1}(\widehat{\boldsymbol{\theta}}_n))$, with $\boldsymbol{\mathcal{I}}^{1/2}(\boldsymbol{\theta})$ being the root of Fisher information $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ of the sample $X_1, \dots, X_n$. Instead of the expected Fisher information, observed $\boldsymbol{J}(\boldsymbol{\theta})$ can also be used.

2.1 Classical Estimation Theory 2.1.5 Asymptotic properties and criteria

Example of BAN estimator

Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} B(\pi)$ be as in slide 191. For the estimator $\widehat{\pi} = \bar{X}$, it was true that $V(\pi) = \pi(1-\pi)$.

$$
\begin{aligned}
i(\pi) &= -\mathbb{E}\left[\frac{\partial^2[x\log(\pi) + (1-x)\log(1-\pi)]}{\partial\pi\partial\pi}\right] \\
&= \mathbb{E}\left[\frac{x}{\pi^2} + \frac{1-x}{(1-\pi)^2}\right] = \frac{1}{\pi} + \frac{1}{1-\pi} = \frac{1}{\pi(1-\pi)},
\end{aligned}
$$

so $V(\pi) = i(\pi)^{-1}$. Thus $\widehat{\pi}$ is a BAN-estimator. From theorem 2.30, it immediately follows that $h(\widehat{\pi}) = \frac{\widehat{\pi}}{1-\widehat{\pi}}$ is a BAN-estimator for the odds $h(\pi) = \frac{\pi}{1-\pi}$.

2.2 Classical Test Theory

2.2.1 Problem statement

2.2 Classical Test Theory 2.2.1 Problem statement

Aim: find test at level α with optimal power for $\boldsymbol{\theta} \in \Theta_1$. Here n is finite.

Problem statement:

- Let Θ be the parameter space; the hypotheses are

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

  with $\Theta_0 \cap \Theta_1 = \emptyset$, i.e. Θ_0 and Θ_1 are disjoint. Possibly, but not necessarily, it can be that $\Theta_0 \cup \Theta_1 = \Theta$.

2.2 Classical Test Theory 2.2.1 Problem statement

- A null hypothesis is called *simple*, if it consists of a single element of Θ, i.e. $\Theta_0 = \{\theta_0\}$. Otherwise one speaks of *composite* hypotheses. There are many null hypotheses that are only seemingly simple, but are actually composite, especially in the case of nuisance parameters

  Example: let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ with unknown μ and σ^2. The null hypothesis $H_0 : \mu = 0$ is a composite hypothesis, since

$$\Theta = \{(\mu, \sigma^2) : -\infty < \mu \leq \infty, 0 < \sigma^2 < \infty\}$$

  and

$$\Theta_0 = \{(\mu, \sigma^2) : \mu = 0, 0 < \sigma^2 < \infty\}.$$

2.2 Classical Test Theory 2.2.1 Problem statement

- Results / Actions:

$$
\begin{aligned}
A_0 : &\quad H_0 \text{ is not rejected} \\
A_1 : &\quad H_0 \text{ is rejected}
\end{aligned}
$$

- Test at level α:

$$\mathbb{P}_\theta(A_1) \leq \alpha, \quad \text{for all } \boldsymbol{\theta} \in \Theta_0$$

2.2 Classical Test Theory 2.2.1 Problem statement

- **Test functions** (see section 1.2.1): tests are often carried out as follows: Choose a test statistic $T(\boldsymbol{X})$, a sample $\boldsymbol{X}$ and a critical region C_α. Then, the test is performed by

$$\phi(\boldsymbol{x}) = \begin{cases} 1 & \text{, if } \ T(\boldsymbol{x}) \in C_\alpha \qquad \text{(reject } H_0 \text{),} \\ 0 & \text{, if } T(\boldsymbol{x}) \notin C_\alpha \qquad \text{(do not reject } H_0 \text{).} \end{cases}$$

2.2 Classical Test Theory 2.2.1 Problem statement

- For the test theory in this section, test functions $\phi(\boldsymbol{x}) \in \{0, 1\}$ are extended to *randomized test functions* $\phi(\boldsymbol{x}) \in [0, 1]$:

  1. $\phi(\boldsymbol{x}) \in [0, 1]$ for given data $\boldsymbol{X} = \boldsymbol{x}$.
  2. Draw a (independent) Bernoulli variable $W \sim \text{Bin}(1, \phi(\boldsymbol{x}))$.
  3. Reject H_0 if and only if $W = 1$.

  Interpretation: $\phi(\boldsymbol{x})$ is the probability of the rejection of H_0 given the observation $\boldsymbol{X} = \boldsymbol{x}$. In the special case $\phi(\boldsymbol{x}) \in \{0, 1\}$, a randomized test is reduced to a standard, non-randomized test. Randomized tests are (for theory) mostly relevant for discrete test statistics.

2.2 Classical Test Theory 2.2.1 Problem statement

Example 2.21 (Randomized binomial test)
Let $X \sim \text{Bin}(10, \pi)$ and

$$H_0 : \pi \leq \frac{1}{2}, \quad H_1 : \pi > \frac{1}{2}.$$

Test: reject $H_0 \Leftrightarrow X \geq k_\alpha$, where k_α is such that

$$\mathbb{P}_\pi(X \geq k_\alpha) \leq \alpha \quad \text{for } \pi = \frac{1}{2}.$$

It is

$$\mathbb{P}_{0.5}(X \geq k) = \begin{cases} 0.00098 & , k = 10 \\ 0.01074 & , k = 9 \\ 0.05469 & , k = 8 \\ \ldots \end{cases}$$

2.2 Classical Test Theory 2.2.1 Problem statement

For $\alpha = 0.05$, the value $k_\alpha = 8$ is not possible because $0.054 > 0.05$. For $k_\alpha = 9$, $\alpha = 0.05$ is nowhere near enough, i.e. the test is very conservative. The solution is a randomized test

$$\phi(x) = \begin{cases} 1 & , x \in \{9, 10\} \\ 67/75 & , x = 8 \\ 0 & , x \leq 7, \end{cases}$$

i.e. if $x = 8$, take a Bernoulli distributed random variable with probability 67/75. If 1 is drawn, then H_0 is rejected.

2.2 Classical Test Theory 2.2.1 Problem statement

Randomization is an artificial process to exploit the significance level α, i.e. to reach

$$\mathbb{P}_\theta(A_1) = \alpha$$

for that $\boldsymbol{\theta}$ on the margin between Θ_0 and Θ_1. A randomized test usually has the following structure:

$$\phi(\boldsymbol{x}) = \begin{cases} 1 & , \boldsymbol{x} \in B_1 \\ \gamma(\boldsymbol{x}) & , \boldsymbol{x} \in B_{10} \\ 0 & , \boldsymbol{x} \in B_0. \end{cases}$$

2.2 Classical Test Theory 2.2.1 Problem statement

The sample space is thus divided into **three** parts:

B_1 strict *rejection region* of H_0, i.e. $\boldsymbol{x} \in B_1 \Rightarrow$ action A_1.

- B_0 strict *acceptance region*, i.e. $\boldsymbol{x} \in B_0 \;\Rightarrow$ Action A_0.
- B_{10} *randomization region*, i.e. $\boldsymbol{x} \in B_{10}$ leads to rejection with probability $\gamma(\boldsymbol{x})$ and no rejection of H_0 with probability $1 - \gamma(\boldsymbol{x})$. B_{10} can be interpreted as an *indifference region.*

2.2 Classical Test Theory 2.2.1 Problem statement

In general, a test is formulated with a test statistic $T = T(\boldsymbol{X})$. Then, randomized tests often take the form:

$$\phi(\boldsymbol{x}) = \begin{cases} 1, & T(\boldsymbol{x}) > c \\ \gamma, & T(\boldsymbol{x}) = c \\ 0, & T(\boldsymbol{x}) < c\,. \end{cases}$$

If $T(\boldsymbol{X})$ is a continuous random variable, then $\mathbb{P}(T(\boldsymbol{X}) = c) = 0$, i.e. for continuous T, $\phi(\boldsymbol{x})$ is reduced to

$$\phi(\boldsymbol{x}) = \begin{cases} 1, & T(\boldsymbol{x}) \geq c \\ 0, & T(\boldsymbol{x}) < c\,. \end{cases}$$

2.2 Classical Test Theory 2.2.1 Problem statement

For discrete test statistics T usually $\gamma > 0$, as in the exact binomial test, since $\mathbb{P}(T(\boldsymbol{X}) = c) > 0$. The value c is at the „decision boundary" between A_1 and A_0. The fact that the decision is made by a random procedure is met with reservations in practice.. The (frequentist) theory shows that the prior probability

$$\mathbb{P}_\theta(A_1) = \int_{\mathcal{X}} \underbrace{\mathbb{P}(A_1|\boldsymbol{x})}_{\phi(\boldsymbol{x})}\; \underbrace{f(\boldsymbol{x}|\boldsymbol{\theta})d\boldsymbol{x}}_{d\mathbb{P}_\theta} = \mathbb{E}_\theta[\phi(\boldsymbol{X})], \quad \boldsymbol{\theta} \in \Theta_1$$

can be maximized via randomization ($\phi(\boldsymbol{x})$ is the conditional probability, a posteriori, i.e. for a given sample, to decide in favour of A_1). „Maximum " refers to the „average " optimality of the test when performed repeatedly.

2.2 Classical Test Theory 2.2.1 Problem statement

Subjective view: at $T(\boldsymbol{x}) = c$ and $\boldsymbol{x} \in B_{10}$ one is more likely not to make a decision („indifference region"). For $n \to \infty$ (as a rule) $\mathbb{P}(T(\boldsymbol{X}) = c)$ approaches 0, i.e. for large n the randomization area B_{10} becomes smaller and smaller.

Idea: collect additional data at $T(\boldsymbol{x}) = c$.

2.2 Classical Test Theory 2.2.1 Problem statement

Power, power function

When making a test decision, there are following options:

	A_0: H_0 not rejected	A_1: H_1 is significant
H_0 is true	correct statement	Type 1 error
H_1 is true	Type 2 error	correct statement

2.2 Classical Test Theory 2.2.1 Problem statement

$\phi(\boldsymbol{x}) = \mathbb{P}(A_1|\boldsymbol{x})$ is the conditional probability of A_1 given the sample $\boldsymbol{x}$. If $\mathbb{P}_\theta(A_1)$ is the unconditional probability / prior probability, then (as in the above)

$$\mathbb{P}_\theta(A_1) = \int_{\mathcal{X}} \mathbb{P}(A_1|\boldsymbol{x}) f(\boldsymbol{x}|\boldsymbol{\theta})\; d\boldsymbol{x} = \int \phi(\boldsymbol{x}) f(\boldsymbol{x}|\boldsymbol{\theta})\; d\boldsymbol{x} = \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X})]$$

and thus also $\mathbb{P}_\theta(A_0) = \mathbb{E}_\theta(1 - \phi(\boldsymbol{X}))$ for $\boldsymbol{\theta} \in \Theta$.

2.2 Classical Test Theory 2.2.1 Problem statement

Definition 2.30 (Power function of a test ϕ)

2.2 Classical Test Theory 2.2.1 Problem statement

1. *The mapping* $g_\phi(\boldsymbol{\theta}) = \mathbb{E}_\theta[\phi(\boldsymbol{X})] = \mathbb{P}_\theta(A_1)$, $\boldsymbol{\theta} \in \Theta$, *is called* power function *of the test* ϕ.

$$\begin{aligned} g_\phi(\boldsymbol{\theta}) = \mathbb{P}_\theta(A_1) \qquad & \textit{probability of Type 1 error,} \\ & \boldsymbol{\theta} \in \Theta_0 \\ 1 - g_\phi(\boldsymbol{\theta}) = \mathbb{P}_\theta(A_0) \qquad & \textit{probability of Type 2 error,} \\ & \boldsymbol{\theta} \in \Theta_1 \end{aligned}$$

Furthermore:

$$g_\phi(\boldsymbol{\theta}) = \mathbb{P}_\theta(A_1) \qquad \textit{power of the tests},\ \boldsymbol{\theta} \in \Theta_1$$

2.2 Classical Test Theory
2.2.1 Problem statement

Definition 2.30 (continued)

2. *The value*

$$\alpha(\phi) = \sup_{\boldsymbol{\theta}\in\Theta_0} \mathbb{P}_\theta(A_1) = \sup_{\boldsymbol{\theta}\in\Theta_0} g_\phi(\boldsymbol{\theta})$$

is called the (actual) level (size) *of* ϕ *and is the* supremal probability of Type 1 error*.*

$$\beta(\phi) = \sup_{\boldsymbol{\theta}\in\Theta_1} \mathbb{P}_\theta(A_0) = \sup_{\boldsymbol{\theta}\in\Theta_1} (1 - g_\phi(\boldsymbol{\theta})) = 1 - \inf_{\boldsymbol{\theta}\in\Theta_1} g_\phi(\boldsymbol{\theta})$$

is the supremal probability of Type 2 error*.*

2.2 Classical Test Theory
2.2.1 Problem statement

- For „common” tests (e.g. one-sided Z-test), the following holds due to the monotony and continuity of $g_\phi(\theta)$

$$\alpha(\phi) + \beta(\phi) = 1,$$

i.e. $\alpha(\phi)$ can only be kept small at the expense of $\beta(\phi)$ (and vice versa). In general,

$$\alpha(\phi) + \beta(\phi) \leq 1$$

(for unbiased tests, see definition 2.37).

2.2 Classical Test Theory
2.2.1 Problem statement

- *Classical test theory approach*: maximise the power for $\boldsymbol{\theta} \in \Theta_1$ under constraint

$$g_\phi(\boldsymbol{\theta}) \leq \alpha \text{ for all } \boldsymbol{\theta} \in \Theta_0$$

with a given $\alpha > 0$, i.e.

$$g_\phi(\boldsymbol{\theta}) \geq \max_{\widetilde{\phi}} g_{\widetilde{\phi}}(\boldsymbol{\theta}) \quad \text{for } \boldsymbol{\theta} \in \Theta_1$$

over „competing” tests $\widetilde{\phi}$. H_0 and H_1 are thus viewed asymmetrically.

2.2 Classical Test Theory
2.2.1 Problem statement

- Because of the relationship $\alpha(\phi)+\beta(\phi) = 1$, the given significance level α must be exploited fully, i.e.

$$\alpha(\phi) = \alpha$$

must hold. If $\alpha(\phi) < \alpha$,

$$\beta(\phi) = 1 - \inf_{\theta\in\Theta_1} g_\theta(\phi)$$

becomes larger than needed for $\boldsymbol{\theta} \in \Theta_1$, i.e. the power of the test is worse.

2.2 Classical Test Theory 2.2.1 Problem statement

- The following problems that we consider are based on this concept:

1. *simple* H_0 *vs. simple* H_1: Neyman-Pearson theorem shows how to construct the best test.
2. *simple* H_0 *vs. composite* H_1: for certain cases a „uniformly most powerful" (UMP) test can be constructed based on the Neyman-Pearson theorem. In other cases — without any further restrictions — there exists no UMP test.
3. *composite* H_0 *vs. composite* H_1: search for a UMP test is even more difficult.

2.2.2 Neyman-Pearson's theorem

2.2 Classical Test Theory 2.2.2 Neyman-Pearson's theorem

Problem: simple null hypothesis vs. simple alternative hypothesis, so

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0, \quad \text{vs.} \quad H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$$

with $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_1$. Let $f_0(\boldsymbol{x}) = f(\boldsymbol{x}|\theta_0), f_1(\boldsymbol{x}) = f(\boldsymbol{x}|\theta_1)$. Then

$$\Lambda(\boldsymbol{x}) = \frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})}$$

is called *likelihood ratio.* According to Neyman-Pearson, there is a (best) test of the form:

$$H_0 \text{ reject } \Leftrightarrow \Lambda(\boldsymbol{x}) > k_\alpha$$

with k_α chosen such that the test complies with level α . But: if $\Lambda(\boldsymbol{x})$ is discrete, there is a theoretical problem

2.2 Classical Test Theory 2.2.2 Neyman-Pearson's theorem

Definition 2.31 (Randomized LR test)
A test $\phi^*(\boldsymbol{x})$ *is called* randomized likelihood ratio test*, in short* LR-Test *(LRT)* $\overset{def}{\Leftrightarrow}$ $\phi^*(\boldsymbol{x})$ *has the structure*

$$\phi^*(\boldsymbol{x}) = \begin{cases} 1 & , f_1(\boldsymbol{x}) > kf_0(\boldsymbol{x}) \;\Leftrightarrow\; \Lambda(\boldsymbol{x}) > k \\ \gamma(\boldsymbol{x}) & , f_1(\boldsymbol{x}) = kf_0(\boldsymbol{x}) \;\Leftrightarrow\; \Lambda(\boldsymbol{x}) = k \\ 0 & , f_1(\boldsymbol{x}) < kf_0(\boldsymbol{x}) \;\Leftrightarrow\; \Lambda(\boldsymbol{x}) < k \end{cases}$$

with constant $k > 0$ *and* $0 < \gamma(\boldsymbol{x}) < 1$*. If* $\Lambda(X)$ *is continuously distributed, then* $\mathbb{P}_\theta(\Lambda(X) = k) = 0$*. Then a non-randomized test is enough*

$$\phi^*(\boldsymbol{x}) = \begin{cases} 1, \; f_1(\boldsymbol{x}) > kf_0(\boldsymbol{x}) \;\Leftrightarrow\; \Lambda(\boldsymbol{x}) > k \\ 0, \; \textit{otherwise.} \end{cases}$$

2.2 Classical Test Theory 2.2.2 Neyman-Pearson's theorem

Theorem 2.32 (Neyman-Pearson lemma)

1. *Optimality: for every* k *and* $\gamma(\boldsymbol{x})$*, the test* ϕ^* *has maximal power among all tests whose level is at most equal to that level of* ϕ^**.*
2. *Existence: for a given* $\alpha \in (0, 1)$*, there are constants* k^* *and* γ^**, such that the LR test* ϕ^* *with* k^* *and* $\gamma(\boldsymbol{x}) = \gamma^*$ *has exactly the level* α *for all* $\boldsymbol{x}$*.*

3. *Uniqueness: if a test ϕ with level α has the maximum power (= smallest Type 2 error) among all other tests with level α, then ϕ is an LR test (possibly with the exception of a null set $\mathcal{X}_0 \subset \mathcal{X}$ of samples $\boldsymbol{x}$, i.e. $\mathbb{P}_{\theta_0}(\mathcal{X}_0) = \mathbb{P}_{\theta_1}(\mathcal{X}_0) = 0$).*

2.2 Classical Test Theory 2.2.2 Neyman-Pearson's theorem

Proof

1. Optimality:

   Let ϕ be a test with
   $$\mathbb{E}_{\theta_0}[\phi(\boldsymbol{X})] \leq \mathbb{E}_{\theta_0}[\phi^*(\boldsymbol{X})] \tag{2.2}$$
   and define
   $$U(\boldsymbol{x}) = (\phi^*(\boldsymbol{x}) - \phi(\boldsymbol{x}))(f_1(\boldsymbol{x}) - kf_0(\boldsymbol{x})).$$

   - $\phi^*(\boldsymbol{x}) = 1$ for $f_1(\boldsymbol{x}) - kf_0(\boldsymbol{x}) > 0$, so $U(\boldsymbol{x}) \geq 0$.
   - $\phi^*(\boldsymbol{x}) = 0$ for $f_1(\boldsymbol{x}) - kf_0(\boldsymbol{x}) < 0$, so $U(\boldsymbol{x}) \geq 0$.
   - $U(\boldsymbol{x}) = 0$ for $f_1(\boldsymbol{x}) - kf_0(\boldsymbol{x}) = 0$.

   So: $U(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x}$.

2.2 Classical Test Theory 2.2.2 Neyman-Pearson's theorem

Since $U(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x}$:

$$\begin{aligned}
0 &\leq \int U(\boldsymbol{x}) d\boldsymbol{x} \\
&= \int (\phi^*(\boldsymbol{x}) - \phi(\boldsymbol{x}))(f_1(\boldsymbol{x}) - kf_0(\boldsymbol{x}))\, d\boldsymbol{x} \\
&= \int \phi^*(\boldsymbol{x}) f_1(\boldsymbol{x})\, d\boldsymbol{x} - \int \phi(\boldsymbol{x}) f_1(\boldsymbol{x})\, d\boldsymbol{x} \\
&\qquad + k\left(\int \phi(\boldsymbol{x}) f_0(\boldsymbol{x})\, d\boldsymbol{x} - \int \phi^*(\boldsymbol{x}) f_0(\boldsymbol{x})\, d\boldsymbol{x}\right) \\
&= \mathbb{E}_{\theta_1}[\phi^*(\boldsymbol{X})] - \mathbb{E}_{\theta_1}[\phi(\boldsymbol{X})] + \underbrace{k(\mathbb{E}_{\theta_0}[\phi(\boldsymbol{X})] - \mathbb{E}_{\theta_0}[\phi^*(\boldsymbol{X})])}_{\leq 0 \text{ due to } (2.2)}
\end{aligned}$$

$\Rightarrow \mathbb{E}_{\theta_1}[\phi^*(\boldsymbol{X})] \geq \mathbb{E}_{\theta_1}[\phi(\boldsymbol{X})]$, i.e. the power of ϕ^* is greater than the power of ϕ.

2.2 Classical Test Theory 2.2.2 Neyman-Pearson's theorem

2. The distribution function $G(k) = \mathbb{P}_{\theta_0}(\Lambda(\boldsymbol{x}) \leq k)$ is increasing monotonically in k. It is also continuous to the right, i.e.
   $$G(k) = \lim_{y \downarrow k} G(y) \quad \text{for all } k.$$
   Considering the equation
   $$G(k^*) = 1 - \alpha$$
   and trying to solve with respect to k^*, there are two possibilities:

   (i) Either such k^* exists,

   (ii) or the equation cannot be solved exactly, but there exists k^*, such that
   $$G_-(k^*) = \mathbb{P}_{\theta_0}(\Lambda(\boldsymbol{X}) < k^*) \leq 1 - \alpha < G(k^*)$$
   (which corresponds to the „level condition").

2.2 Classical Test Theory 2.2.2 Neyman-Pearson's theorem

In the first case you set $\gamma^* = 0$, while in the second

$$\gamma^* = \frac{G(k^*) - (1-\alpha)}{G(k^*) - G_-(k^*)}.$$

In this case the test has exactly the level α, as claimed, because:

$$\begin{aligned}
\mathbb{E}_{\theta_0}[\phi(\boldsymbol{X})] &= \mathbb{P}_{\theta_0}\left(\frac{f_1(\boldsymbol{X})}{f_0(\boldsymbol{X})} > k^*\right) + \frac{G(k^*) - 1 + \alpha}{G(k^*) - G_-(k^*)}\mathbb{P}_{\theta_0}\left(\frac{f_1(\boldsymbol{X})}{f_0(\boldsymbol{X})} = k^*\right) \\
&= (1 - G(k^*)) + \frac{G(k^*) - 1 + \alpha}{G(k^*) - G_-(k^*)}(G(k^*) - G_-(k^*)) \\
&= \alpha.
\end{aligned}$$

2.2 Classical Test Theory 2.2.2 Neyman-Pearson's Theorem

3. For a given α let ϕ^* be the existing LR test according to 2, defined by a constant k and a function $\gamma(\boldsymbol{x})$. Assume, ϕ is a different test with the same level α and the same (after 1st maximum) power as ϕ^*. If one defines $U(\boldsymbol{x})$ as in 1., then $U(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x}$ and $\int U(\boldsymbol{x})\, d\boldsymbol{x} = 0$, since $\mathbb{E}_{\theta_1}[\phi^*(\boldsymbol{X})] - \mathbb{E}_{\theta_1}[\phi(\boldsymbol{X})] = 0$ and $\mathbb{E}_{\theta_0}[\phi^*(\boldsymbol{X})] - \mathbb{E}_{\theta_0}[\phi(\boldsymbol{X})] = 0$ after assumption. Following the fact that U is non-negative with an integral 0, it holds that $U(\boldsymbol{x}) = 0$ for almost all $\boldsymbol{x}$. This in turn means that $\phi(\boldsymbol{x}) = \phi^*(\boldsymbol{x})$ or $f_1(\boldsymbol{x}) = kf_0(\boldsymbol{x})$, i.e. $\phi(\boldsymbol{x})$ is one LR test (for almost all $\boldsymbol{x}$). □

2.2 Classical Test Theory 2.2.2 Neyman-Pearson's theorem

Remark. For simple hypotheses H_0 and H_1 classical test theory and likelihood ratio test are still identical. For composite hypotheses (a more practical case) the concepts diverge:

- Classical test theory continues to search for optimal tests (for finite samples).
- Likelihood-based tests are either quadratic approximations of $\Lambda(\boldsymbol{x})$, the distribution function of which (under H_0) holds only asymptotically ($n \to \infty$), or they try to generalise $\Lambda(\boldsymbol{x})$.

2.2 Classical Test Theory 2.2.2 Neyman-Pearson's theorem

Example 2.22 (Binomial test)
Consider

$$H_0 : \pi = \pi_0 \quad \text{vs.} \quad H_1 : \pi = \pi_1$$

with $0 < \pi_0 < \pi_1 < 1$. The density (probability function) of the i.i.d. Bernoulli distributed sample $X = (X_1, \dots, X_n)^\top$ is given by

$$f(\boldsymbol{x}|\pi) = \prod_{i=1}^{n} \pi^{x_i}(1-\pi)^{1-x_i} = \pi^z(1-\pi)^{n-z} \quad \text{mit} \quad z = \sum_{i=1}^{n} x_i,$$

The likelihood ratio is then

$$\Lambda(\boldsymbol{x}) = \frac{\pi_1^z(1-\pi_1)^{n-z}}{\pi_0^z(1-\pi_0)^{n-z}} = \left(\frac{1-\pi_1}{1-\pi_0}\right)^n \cdot \left(\frac{\pi_1(1-\pi_0)}{\pi_0(1-\pi_1)}\right)^z := \Lambda(z).$$

2.2 Classical Test Theory 2.2.2 Neyman-Pearson's theorem

Since $\Lambda(\boldsymbol{x}) = \Lambda(z)$ is strictly monotonic in z, $\Lambda(z) > k$ can equivalently be transformed into $z > \Lambda^{-1}(k) =: c$.
The likelihood ratio test ϕ^* with critical number k and (constant) randomization γ^* then has the form

$$\phi^*(\boldsymbol{x}) = \begin{cases} 1 & , Z = Z(\boldsymbol{x}) > c \\ \gamma^* & , Z = Z(\boldsymbol{x}) = c \\ 0 & , Z = Z(\boldsymbol{x}) < c \end{cases}$$

with „test statistics" Z. We can (because of the range of values of Z) restrict ourselves to $c \in \{0, 1, \ldots, n\}$.
γ^* is to be determined from the level condition

$$\mathbb{P}_{\pi_0}(Z > c) + \gamma^* \mathbb{P}_{\pi_0}(Z = c) \overset{!}{=} \alpha$$

The test ϕ^* depends on π_0 but not on π_1!

2.2 Classical Test Theory 2.2.2 Neyman-Pearson's theorem

Remark. If H_1 is true, then π_1 determines the probability for „realized " Type 2 error $\mathbb{P}_{\pi_1}(A_0)$. The further away π_1 is from π_0, the lower the probability of Type 2 error and the greater the power at the point $\pi = \pi_1$.

2.2.3 Uniformly Most Powerful Test

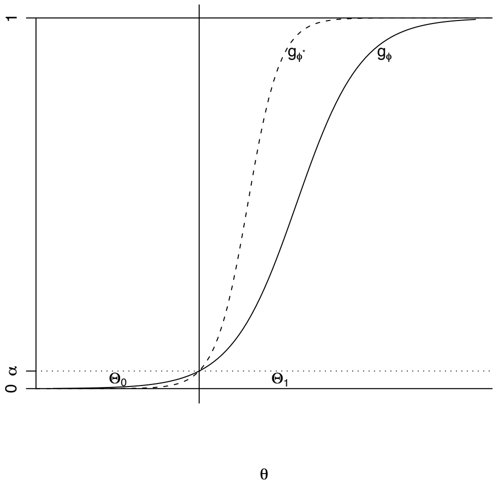
2.2 Classical Test Theory 2.2.3 Uniformly Most Powerful Test

Definition 2.33 (Uniformly Most Powerful (UMP) Test)
An α-level test ϕ^* is called Uniformly Most Powerful *or* UMP test *at level α* $\overset{def}{\Leftrightarrow}$

1. $\mathbb{E}_\theta[\phi^*(\boldsymbol{X})] \leq \alpha$ *for all* $\boldsymbol{\theta} \in \Theta_0$.
2. *For every other α-level test ϕ with $\mathbb{E}_\theta[\phi(\boldsymbol{X})] \leq \alpha$ for all $\boldsymbol{\theta} \in \Theta_0$ the following applies:*

$$\mathbb{E}_\theta[\phi^*(\boldsymbol{X})] \geq \mathbb{E}_\theta[\phi(\boldsymbol{X})] \textit{ for all } \boldsymbol{\theta} \in \Theta_1.$$

2.2 Classical Test Theory 2.2.3 Uniformly Most Powerful Test



Comment: The term „uniformly" in the above definition refers to the uniformity of the property $g_{\phi^*} \geq g_\phi$ on Θ_1 for every other test ϕ.

2.2 Classical Test Theory 2.2.3 Uniformly Most Powerful Test

Best one-sided tests with scalar θ
In example 2.22 (binomial test for simple hypotheses) the test ϕ^* did not depend on specific values of $\pi_1(\equiv H_1) > \pi_0(\equiv H_0)$. It follows that ϕ^* is better than other tests ϕ for all $\pi_1 > \pi_0$. It is crucial that the density or likelihood ratio is monotonic in z. This holds for a more general case and leads to the following definition.

2.2 Classical Test Theory 2.2.3 Uniformly Most Powerful Test

Definition 2.34 (Distributions with monotonic density ratio)
The distribution family $\{f(\boldsymbol{x}|\theta),\ \theta \in \Theta \subseteq \mathbb{R}\}$ with scalar parameter θ has monotonic density *or, respectively,* likelihood ratio *(in short: MLR)* $\overset{def}{\Leftrightarrow}$ *there exists a statistic T, such that*

$$\Lambda(\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta_1)}{f(\boldsymbol{x}|\theta_0)}$$

is monotonically increasing in $T(\boldsymbol{x})$ for every two $\theta_0, \theta_1 \in \Theta$ with $\theta_0 \leq \theta_1$.

2.2 Classical Test Theory 2.2.3 Uniformly Most Powerful Test

Remark.

1. Monotonously *increasing* is not a real limitation; if $\Lambda(\boldsymbol{x})$ decreases monotonically in $\widetilde{T}(\boldsymbol{x})$, then one defines $T(\boldsymbol{x}) = -\widetilde{T}(\boldsymbol{x})$.
2. Every *single-parametric* exponential family in $T(\boldsymbol{x})$ and $\gamma(\theta)$ has monotonic density ratios if $\gamma(\theta)$ is monotonic in θ. The latter is especially true for the *natural* parameterisation $\gamma(\theta) = \theta$.

2.2 Classical Test Theory 2.2.3 Uniformly Most Powerful Test

Theorem 2.35 (UMP Test with MLR)
Given is $\mathcal{P}_\theta = \{f(x|\theta) : \theta \in \Theta \subseteq \mathbb{R}\}$ with MLR in $T(\boldsymbol{x})$ and the hypotheses

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

1. *Existence: there is an UMP test ϕ^* at level α, namely*

$$\phi^*(x) = \begin{cases} 1, & T(\boldsymbol{x}) > c \\ \gamma, & T(\boldsymbol{x}) = c \\ 0, & T(\boldsymbol{x}) < c. \end{cases}$$

Here c and γ are uniquely determined by the level condition

$$\mathbb{E}_{\theta_0}[\phi^*(\boldsymbol{X})] = \mathbb{P}_{\theta_0}(T(\boldsymbol{X}) > c) + \gamma \mathbb{P}_{\theta_0}(T(\boldsymbol{X}) = c) = \alpha.$$

2.2 Classical Test Theory 2.2.3 Uniformly Most Powerful Test

Theorem 2.35 (continued)

2. *The power function $g_{\phi^*}(\theta)$ is monotonically increasing in θ and strictly monotonically increasing for all θ with $0 < g_{\phi^*}(\theta) < 1$. The maximum probability of Type 1 error is $g_{\phi^*}(\theta_0) = \alpha$.*
3. *ϕ^* has also uniformly minimal probabilities for the Type 1 error among all tests ϕ for H_0 vs. H_1 with $g_\phi(\theta_0) = \alpha$.*
4. *ϕ^* is uniquely determined (with probability 1) .*

2.2 Classical Test Theory 2.2.3 Uniformly Most Powerful Test

Remark. The following further holds: if ϕ^* is the best test for the simple alternative problem

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta = \theta_1,$$

then ϕ^* is also the UMP test at level α for composite hypotheses

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

if ϕ^* does not depend on specific values of $\theta_1 \in H_1$ and for all $\theta \in H_0$ complies with the level α.

2.2 Classical Test Theory 2.2.3 Uniformly Most Powerful Test

Example 2.23

1. Binomial test with $H_0 : \pi \leq \pi_0$ against $H_1 : \pi > \pi_0$ has MLR in $Z(x) =$ "number of successes" (see example 2.22 and remark 2.). The binomial test is therefore a UMP test.
2. Uniform distribution
3. Gauss Z-test with known σ^2

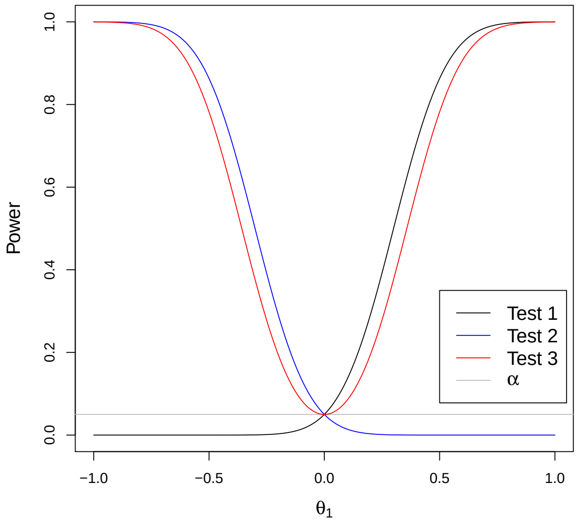


Figure 5: Power functions for Tests 1 and 2, both one-sided Gauss Z-tests, and Test 3, a two-sided Gauss Z-test

4. Exponential distribution
5. Poisson distribution

2.2 Classical Test Theory 2.2.3 Uniformly Most Powerful Test

Remark Often there is no UMP test, but a locally best (one-sided) test: ϕ_{lok} is called *locally best test* at level $\alpha \overset{\text{def}}{\Leftrightarrow}$

$$g'_{\phi_{\text{lok}}}(\theta_0) = \frac{d}{d\theta} g_{\phi_{\text{lok}}}(\theta_0) \geq \frac{d}{d\theta} g_\phi(\theta_0),$$

in which $g_{\phi_{\text{lok}}}(\theta_0) = g_\phi(\theta_0) = \alpha$ follows.

2.2 Classical Test Theory 2.2.3 Uniformly Most Powerful Test

Best unbiased two-tailed tests with scalar θ
For two-sided test problems of the form

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0$$

there is usually no UMP test, even if there is MLR. Therefore, a restriction to a smaller class of competing tests is necessary.

2.2 Classical Test Theory 2.2.3 Uniformly Most Powerful Test

Example: let $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$ with known σ^2. Consider $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ (in the plot $\theta_0 = 0$).

2.2 Classical Test Theory 2.2.3 Uniformly Most Powerful Test

Definition 2.36 (Unbiased α-level test)
A test ϕ for H_0 vs. H_1 is called unbiased α-level test $\overset{def}{\Leftrightarrow}$

$$g_\phi(\theta) \leq \alpha \textit{ for all } \theta \in \Theta_0, \quad g_\phi(\theta) \geq \alpha \textit{ for all } \theta \in \Theta_1.$$

2.2 Classical Test Theory 2.2.3 Uniformly Most Powerful Test

Theorem 2.37 (Two-sided Uniformly Most Powerful Unbiased (UMPU) tests)
Let

$$f(\boldsymbol{x}|\theta) = h(\boldsymbol{x})c(\theta)\exp(\theta T(\boldsymbol{x}))$$

be a one-parametric exponential family with a natural parameter $\theta \in \Theta$ (let Θ be an open interval) and statistic $T(\boldsymbol{x})$. Then

$$\phi^*(\boldsymbol{x}) = \begin{cases} 1 & , T(\boldsymbol{x}) < c_1 \\ \gamma_1 & , T(\boldsymbol{x}) = c_1 \\ 0 & , c_1 < T(\boldsymbol{x}) < c_2 \\ \gamma_2 & , T(\boldsymbol{x}) = c_2 \\ 1 & , T(\boldsymbol{x}) > c_2 \end{cases}$$

is a UMPU test at level α among all unbiased tests ϕ at level α for the test problem $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.

2.2 Classical Test Theory 2.2.3 Uniformly Most Powerful Test

Here $c_1, c_2, \gamma_1, \gamma_2$ are determined from

$$\mathbb{E}_{\theta_0}[\phi^*(X)] = \alpha, \quad \mathbb{E}_{\theta_0}[\phi^*(X)T(X)] = \alpha \mathbb{E}_{\theta_0}[T(X)]$$

Addendum: Theorem 2.37 revisited

The conditions are

$$\mathbb{E}_{\theta_0}[\phi^*(X)] = \alpha, \quad \mathbb{E}_{\theta_0}[\phi^*(X)T(X)] = \alpha \mathbb{E}_{\theta_0}[T(X)] \ .$$

We show that the second condition follows from the fact the first derivative wrt θ of the power function must be zero in θ_0 (minimum, unbiasedness). Note, that $g_\phi(\theta) = \mathbb{E}_\theta(\phi(X))$ Now

$$g'_\phi(\theta) \quad = \quad \frac{d}{d\theta}\mathbb{E}_\theta(\phi(X))|\theta = \theta_0 = 0$$

The theorem explicitly addresses a one parameter exponential family. We have

$$\begin{aligned} \int c(\theta)h(x)\exp(\theta T(x))dx &= 1 \\ c(\theta)\int h(x)\exp(\theta T(x))dx &= 1 \\ c(\theta) &= [\int h(x)\exp(\theta T(x))dx]^{-1} \\ c'(\theta) &= -c(\theta)^2 \frac{d}{d\theta}\int h(x)\exp(\theta T(x))dx \\ &= -c(\theta)^2 \int h(x)\left(\frac{d}{d\theta}\exp(\theta T(x))\right)dx \\ &= -c(\theta)^2 \int h(x)\exp(\theta T(x))T(x)dx \end{aligned}$$

Therefore

$$\begin{aligned}
g'_\phi(\theta) &= \frac{d}{d\theta}\left[\int \phi(x)c(\theta)h(x)\exp(\theta T(x))dx\right] \\
&= c'(\theta)\int \phi(x)h(x)\exp(\theta T(x))dx + c(\theta)\int \phi(x)h(x)\exp(\theta T(x))T(x)dx \\
&= -c(\theta)^2\int h(x)\exp(\theta T(x))T(x)dx\int \phi(x)h(x)\exp(\theta T(x))dx \\
&\quad + c(\theta)\int \phi(x)h(x)\exp(\theta T(x))T(x)dx \\
&= -\int T(x)c(\theta)h(x)\exp(\theta T(x))dx\int \phi(x)c(\theta)h(x)\exp(\theta T(x))dx \\
&\quad + \int [\phi(x)T(x)]c(\theta)h(x)\exp(\theta T(x))dx \\
&= \mathbb{E}_\theta[\phi(X)T(X)] - \mathbb{E}_\theta T(X)\mathbb{E}_\theta\phi(X)\ .
\end{aligned}$$

Now, since we have a α-level test, $\mathbb{E}_{\theta_0}\phi(X) = \alpha$. Thus, we have the condition

$$\begin{aligned}
g'_\phi(\theta_0) &= 0 \\
\mathbb{E}_{\theta_0}[\phi(X)T(X)] - \mathbb{E}_{\theta_0}T(X)\mathbb{E}_{\theta_0}\phi(X) &= 0 \\
\mathbb{E}_{\theta_0}[\phi(X)T(X)] &= \mathbb{E}_{\theta_0}T(X)\mathbb{E}_{\theta_0}\phi(X) \\
\mathbb{E}_{\theta_0}[\phi(X)T(X)] &= \alpha\mathbb{E}_{\theta_0}T(X)\ .
\end{aligned}$$

2.2 Classical Test Theory 2.2.3 Uniformly Most Powerful Test

Example 2.24

1. Two-sided binomial test for

$$H_0:\ \pi = \pi_0 \quad \text{vs.} \quad H_1:\ \pi \neq \pi_0$$

.

2. Two-sided Gauss Z-test with $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, known σ^2 for

$$H_0:\ \mu = \mu_0 \quad \text{vs.} \quad H_1:\ \mu \neq \mu_0$$

3. Two-sided Poisson test with $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Po}(\lambda)$ for

$$H_0:\ \lambda = \lambda_0 \quad \text{vs.} \quad H_1:\ \lambda \neq \lambda_0$$

2.2 Classical Test Theory 2.2.3 Uniformly Most Powerful Test

3. (continued) since it is a one-parametric exponential family with natural parameter $\theta = \log\lambda$ an equivalent hypotheses in terms of θ is

$$H_0:\ \theta = \theta_0 \quad \text{vs.} \quad H_1:\ \theta \neq \theta_0.$$

Determining the test variable

$$\begin{aligned}
f(x_i|\theta) &= h(x_i)c(\theta)\exp(\theta x_i) \\
f(x|\theta) &= f(x_1|\theta)\cdot\ldots\cdot f(x_n|\theta) \propto \exp\Big(\theta\underbrace{\sum_{i=1}^n x_i}_{T(x)}\Big)
\end{aligned}$$

2.2 Classical Test Theory 2.2.3 Uniformly Most Powerful Test

3. (continued) and thus the test is

$$\phi^*(x) = \begin{cases} 1 & , \sum_{i=1}^n x_i < c_1 \\ \gamma_1 & , \sum_{i=1}^n x_i = c_1 \\ 0 & , c_1 < \sum_{i=1}^n x_i < c_2 \\ \gamma_2 & , \sum_{i=1}^n x_i = c_2 \\ 1 & , \sum_{i=1}^n x_i > c_2 \; . \end{cases}$$

Cases 1-3 all belong to one-parametric exponential family, therefore the UMPU test has the form as in theorem 2.38 with $T(\boldsymbol{x}) = \sum_{i=1}^n x_i$.

2.2 Classical Test Theory 2.2.3 Uniformly Most Powerful Test

4. Two-sided χ^2-test for variance: Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, with known μ. We test for

$$H_0 : \; \sigma^2 = \sigma_0^2 \quad \text{vs.} \quad H_1 : \; \sigma^2 \neq \sigma_0^2.$$

The resulting density is a one-parametric exponential family

$$f(\boldsymbol{x}|\sigma^2) \propto exp\left(- \sum_{i=1}^n \frac{(x_i - \mu)^2}{2} \frac{1}{\sigma^2} \right) = exp(T(\boldsymbol{x})\theta).$$

with $\theta = \frac{1}{\sigma^2}$. Test as in theorem 2.38. Better: test statistic $\sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma_0} \right)^2$ is $\chi^2(n)$-distributed under H_0. The test is constructed as

$$\phi^*(x) = \begin{cases} 1, \; \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma_0} \right)^2 > \chi_{1-\alpha/2}(n) \text{ or } < \chi_{\alpha/2}(n) \\ 0, \; \text{otherwise} \end{cases}$$

The test constructed in this way is UMPU.

2.2 Classical Test Theory 2.2.3 Uniformly Most Powerful Test

Multi-parametric distribution assumption

- So far: θ is scalar.
  - ⇒ For $N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$. This distribution is outside the scope of the theory of optimal tests.
  - ⇒ t-test on μ (with unknown σ^2) and others are not included.
- Idea: „optimal" tests can (still) be done for a scalar component η from $\theta = (\eta, \xi)$, where ξ may be multidimensional. ξ is to be regarded as a disturbance/ nuisance parameter.

2.2 Classical Test Theory 2.2.3 Uniformly Most Powerful Test

- $\{f(\boldsymbol{x}|\boldsymbol{\theta}), \; \theta \in \Theta \subseteq \mathbb{R}^k\}$ must be a (strictly) k-parameter exponential family with natural parameters $\boldsymbol{\theta} = (\eta, \boldsymbol{\xi})$ and $\boldsymbol{T} = (U, \boldsymbol{V})$, η end U scalar. This leads to the theory of conditional tests.
- Suitable, for example, for
  - t-test: comparison of μ_1, μ_2 for independent samples, only if $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

  - Test for significance of β_1 in simple linear regression.

2.2 Classical Test Theory
2.2.3 Uniformly Most Powerful Test

- Already no longer applicable for
  - Comparison of μ_1, μ_2 when $\sigma_1^2 \neq \sigma_2^2$ (Behrens-Fisher-Problem).
  - Test for significance of β_1 in the logit or Poisson regression model.

$\Rightarrow$ (Asymptotic) likelihood theory, Bayesian inference.

2.3 Interval Estimation and Confidence Intervals

2.3.1 Definition and assessment of the quality

2.3 Interval Estimation and Confidence Intervals
2.3.1 Definition and assessment of the quality

Definition 2.37 (Interval estimate)
An interval estimate *(*confidence region*) C for $\tau(\boldsymbol{\theta})$, $\tau : \Theta \to \Sigma$, for a (given)* degree of confidence *(*confidence level*) $1-\alpha$ is a mapping of the sample space $\mathcal{X}$ onto the power set $\mathfrak{P}(\Sigma)$, i.e. $\boldsymbol{x} \to C(\boldsymbol{x}) \in \mathfrak{P}(\Sigma)$, with measurable $\{\tau(\boldsymbol{\theta}) \in C(\boldsymbol{X})\}$ and*

$$\mathbb{P}_\theta(\tau(\boldsymbol{\theta}) \in C(\boldsymbol{X})) \geq 1-\alpha \quad \textit{for all } \boldsymbol{\theta} \in \Theta.$$

2.3 Interval Estimation and Confidence Intervals
2.3.1 Definition and assessment of the quality

$C(\boldsymbol{X})$ is a random set in $\mathfrak{P}(\Sigma)$. After observing the sample $\boldsymbol{X} = \boldsymbol{x}$, $C(\boldsymbol{x})$ is given. The statement

$$\tau(\boldsymbol{\theta}) \in C(\boldsymbol{x}) \quad (\text{true} \quad \overset{!}{\text{or}} \quad \text{false})$$

is assigned the confidence level $1-\alpha$. The is the well-known frequency interpretation. If $C(\boldsymbol{x})$ is an interval for every $\boldsymbol{x}$, then $C(\boldsymbol{x})$ is called *confidence interval* and C an *interval estimate.*
A probability statement about

$$\tau(\boldsymbol{\theta}) \in C(\boldsymbol{x})$$

for a given $\boldsymbol{x}$ is possible within the framework of Bayesian inference (without logical problems).
The „precision" of $C(\boldsymbol{X})$ is measured by the expected size of the area or the length of the confidence interval.

2.3 Interval Estimation and Confidence Intervals
2.3.1 Definition and assessment of the quality

Example 2.25
Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ and

$$C(\boldsymbol{X}) = \left[\bar{X} - t_{n-1}\Big(\frac{\alpha}{2}\Big)\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1}\Big(\frac{\alpha}{2}\Big)\frac{S}{\sqrt{n}}\right]$$

be a confidence interval for μ (where $t_{n-1}\Big(\frac{\alpha}{2}\Big)$ is the $\Big(1-\frac{\alpha}{2}\Big)$-th quantile of the t_{n-1} distribution).
The length of $C(\boldsymbol{X})$

$$L = 2\, t_{n-1}\Big(\frac{\alpha}{2}\Big)\frac{S}{\sqrt{n}}$$

is random with expected value

$$\mathbb{E}(L) = 2\, t_{n-1}\Big(\frac{\alpha}{2}\Big)\frac{1}{\sqrt{n}}\,\mathbb{E}(S) = 2\, t_{n-1}\Big(\frac{\alpha}{2}\Big)\frac{\sigma}{\sqrt{n}}\sqrt{\frac{2}{n-1}}\frac{\Gamma(n/2)}{\Gamma((n-1)/2)}.$$

2.3 Interval Estimation and Confidence Intervals
2.3.1 Definition and assessment of the quality

The following applies:

$$\begin{aligned} 1-\alpha \ \text{larger} \quad &\to \quad \mathbb{E}(L) \ \text{larger}, \\ n \ \text{larger} \quad &\to \quad \mathbb{E}(L) \ \text{smaller}. \end{aligned}$$

2.3 Interval Estimation and Confidence Intervals 2.3.1 Definition and assessment of the quality

When assessing the precision of a confidence interval by its length, the shorter the expected length, the better is the confidence interval. In general, confidence *region C* is assessed by its average „size ". Let π be a measure on Σ or Θ. Then

$$\pi(C(\boldsymbol{x}))$$

is the size of $C(\boldsymbol{x})$. If π is a Lebesgue measure, then, in case of confidence intervals, this results in length. Then

$$\mathbb{E}_\theta(\pi(C(\boldsymbol{X})))$$

is the expected size. To assess the quality, the expected length or size alone is not sufficient.

2.3 Interval Estimation and Confidence Intervals 2.3.1 Definition and assessment of the quality

Definition 2.38 (Characteristic function of a confidence region)
A characteristic function *is a function*

$$k_C(\boldsymbol{\theta},\boldsymbol{\theta}^{'}) := \mathbb{P}_\theta(C(\boldsymbol{X}) \ni \tau(\boldsymbol{\theta}^{'})).$$

where $\boldsymbol{\theta}$ is the „true" value and $\boldsymbol{\theta}'$ is any value in Θ.

- For $\boldsymbol{\theta} = \boldsymbol{\theta}^{'}$, the probability of „$C(\boldsymbol{X}) \ni \tau(\boldsymbol{\theta}^{'})$" should be as large as possible.
- For $\boldsymbol{\theta} \neq \boldsymbol{\theta}^{'}$ with $\tau(\boldsymbol{\theta}^{'}) \neq \tau(\boldsymbol{\theta})$, the probability of „$C(\boldsymbol{X}) \ni \tau(\boldsymbol{\theta}^{'})$" should be kept as small as possible.

In the following we consider the special case $\tau(\theta) = \theta$ with a scalar θ. Then

$$k_C(\theta,\theta^{'}) = \mathbb{P}_\theta(C(\boldsymbol{X}) \ni \theta^{'}).$$

2.3 Interval Estimation and Confidence Intervals 2.3.1 Definition and assessment of the quality

Definition 2.39

1. *A confidence region has the* confidence level $1-\alpha$ $:\overset{def}{\Leftrightarrow}$

$$k_C(\boldsymbol{\theta},\boldsymbol{\theta}') \geq 1-\alpha \quad \textit{for all } \boldsymbol{\theta}^{'} = \boldsymbol{\theta}.$$

2. *A confidence region at confidence level $1-\alpha$ is called* unbiased $:\overset{def}{\Leftrightarrow}$

$$k_C(\boldsymbol{\theta},\boldsymbol{\theta}') \leq 1-\alpha \quad \textit{for all } \boldsymbol{\theta}^{'} \neq \boldsymbol{\theta}.$$

3. *An [unbiased] confidence region C_0 at confidence level $1-\alpha$ is called* uniformly best (selective) [unbiased] confidence region $:\overset{def}{\Leftrightarrow}$ *for all $\boldsymbol{\theta}^{'} \neq \boldsymbol{\theta}$ and all [unbiased] confidence regions C at confidence level $1-\alpha$, the following applies*

$$k_{C_0}(\boldsymbol{\theta},\boldsymbol{\theta}^{'}) \leq k_C(\boldsymbol{\theta},\boldsymbol{\theta}^{'}).$$

2.3 Interval Estimation and Confidence Intervals 2.3.1 Definition and assessment of the quality

Lemma 2.40
For a given confidence level $1-\alpha$, every uniformly best confidence region also has the smallest expected size with respect to any measure π (but not vice versa).

Proof

$$\begin{aligned}
\mathbb{E}_\theta(\pi(C(\boldsymbol{X}))) &= \int\limits_{\mathcal{X}} \pi(C(\boldsymbol{x}))d\mathbb{P}_\theta(\boldsymbol{x}) \\
&= \int\limits_{\mathcal{X}} \int\limits_{\Theta} I_{C(\boldsymbol{x})}(\boldsymbol{\theta}^{'})d\pi(\boldsymbol{\theta}^{'})d\mathbb{P}_\theta(\boldsymbol{x}) \\
&= \int\limits_{\Theta} \int\limits_{\mathcal{X}} I_{C(\boldsymbol{x})}(\theta^{'})d\mathbb{P}_\theta(\boldsymbol{x})d\pi(\boldsymbol{\theta}^{'}) \qquad \text{(Fubini)} \\
&= \int\limits_{\Theta} \underbrace{\mathbb{P}_\theta(C(\boldsymbol{x}) \ni \boldsymbol{\theta}^{'})}_{k_C(\boldsymbol{\theta},\boldsymbol{\theta}^{'})} d\pi(\boldsymbol{\theta}^{'}).
\end{aligned}$$

For each „true“ $\boldsymbol{\theta}$ therefore applies

2.3 Interval Estimation and Confidence Intervals 2.3.1 Definition and assessment of the quality

$$\underbrace{\int\limits_{\mathcal{X}} \pi(C(\boldsymbol{x}))d\mathbb{P}_\theta(\boldsymbol{x})}_{\text{expected size}} = \int\limits_{\Theta} \underbrace{k_C(\boldsymbol{\theta},\boldsymbol{\theta}^{'})}_{\text{Characteristic function of the confidence region}} d\pi(\boldsymbol{\theta}^{'}).$$

□

2.3.2 Duality between confidence intervals and tests

2.3 Interval Estimation and Confidence Intervals 2.3.2 Duality between confidence intervals and tests

We assume the special case described above: $\tau(\theta) = \theta$ with scalar θ. For every fixed θ we consider an α-level test $\phi_\theta(\boldsymbol{x})$ for the null hypothesis $H_0 = \{\theta\}$ against the alternative $H_1 = \Theta \backslash H_0$. The tests are not intended to be randomized, so that they can be determined by specifying a test variable $T_\theta = T_\theta(\boldsymbol{x})$ and a critical region (rejection region) K_θ:

$$\phi_\theta(\boldsymbol{x}) = \begin{cases} 1 & \text{for } T_\theta(\boldsymbol{x}) \in K_\theta, \\ 0 & \text{otherwise.} \end{cases}$$

2.3 Interval Estimation and Confidence Intervals 2.3.2 Duality between confidence intervals and tests

After observing $\boldsymbol{X} = \boldsymbol{x}$, the null hypothesis „The unknown parameter has the value of θ“ is not rejected if and only if

$$T_\theta\ (\boldsymbol{x})\ \in\ \bar{K}_\theta\ =\ \text{rejection region of the test } \phi_\theta$$

Therefore, after observing $\boldsymbol{X} = \boldsymbol{x}$, it is obvious to define the set

$$C(\boldsymbol{x}) := \{\theta \in \Theta : T_\theta(\boldsymbol{x}) \in \bar{K}_\theta\}$$

as a confidence region; this corresponds to the random set before the observation

$$C(\boldsymbol{X}) = \{\theta \in \Theta : T_\theta(\boldsymbol{X}) \in \bar{K}_\theta\}$$

or.

$$C(\boldsymbol{X}) = \{\theta \in \Theta : \phi_\theta(\boldsymbol{X}) = 0\}$$

The following theorem serves as a confirmation of the above procedure.

2.3 Interval Estimation and Confidence Intervals 2.3.2 Duality between confidence intervals and tests

Theorem 2.41 (Correspondence theorem)

1. *If $\{\phi_\theta\}$ is a set of tests ϕ_θ for $H_0 = \{\theta\}$ vs $H_1 = \Theta\backslash\{\theta\}$ at level α, then $C(\boldsymbol{X}) := \{\theta \in \Theta : \phi_\theta(\boldsymbol{X}) = 0\}$ is a confidence set for confidence level $\gamma = 1 - \alpha$.*
2. *If $\{\phi_\theta\}$ is a set of uniformly most powerful [unbiased] tests, then $C(\boldsymbol{X})$ is also a uniformly best [unbiased] confidence region.*

2.3 Interval Estimation and Confidence Intervals 2.3.2 Duality between confidence interval and tests

Proof. The proof for 1. stems from

$$\mathbb{P}_\theta(C(\boldsymbol{X}) \ni \theta) = \mathbb{P}_\theta(\phi_\theta(\boldsymbol{X}) = 0) \geq 1 - \alpha \quad \text{for all } \ \theta \in \Theta,$$

the one for 2. – out of the relationship

$$\begin{aligned} k_C(\theta, \theta^{'}) &= \mathbb{P}_\theta(C(\boldsymbol{X}) \ni \theta^{'}) = \mathbb{P}_\theta(\phi_{\theta'}(\boldsymbol{X}) = 0) \\ &= 1 - \mathbb{P}_\theta(\phi_{\theta'}(\boldsymbol{X}) = 1) = 1 - g_{\phi_{\theta'}}(\theta) \end{aligned}$$

for all $\theta, \theta^{'} \in \Theta$. Here $g_{\phi_{\theta'}}$ denotes the power function of the test $\phi_{\theta'}$. □

2.3 Interval Estimation and Confidence Intervals 2.3.2 Duality between confidence intervals and tests

The correspondence theorem can be generalized to

- one-sided tests for one-sided confidence intervals
- randomised tests with corresponding randomised confidence regions: $C(\boldsymbol{x})$ is the set of all θ that were not rejected at observation $\boldsymbol{x}$ by test ϕ_θ (also after randomization).

2.3 Interval Estimation and Confidence Intervals 2.3.2 Duality between confidence intervals and tests

In this way, the theory of the interval estimates can be traced back to the test theory except for the following problem: In order for a "reasonable" confidence region (in the topological sense, i.e., for example, a confidence interval) to be constructible from the test family,

- the test function $\phi_\theta(\boldsymbol{x})$, or better, the test variable $T_\theta(\boldsymbol{x})$ *as a function in* θ (for each fixed $\boldsymbol{x}$) must be „good-natured" (ideally monotonic in θ).
- furthermore, the distribution of $T_\theta(\boldsymbol{X})$ must not depend on θ, i.e. $T_\theta(\boldsymbol{X})$ must be a *pivot quantity.*

2.3 Interval Estimation and Confidence Intervals 2.3.2 Duality between confidence interval and tests

Example 2.25 continued:
Let $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. $T_\mu(\boldsymbol{X}) = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$ is a pivot quantity, since the distribution is independent of μ and σ^2.
For every $\mu_0 \in \mathbb{R}$, the test ϕ_{μ_0} is α-level test for

$$H_0 : \mu = \mu_0 \quad \text{gegen} \quad H_1 : \mu \neq \mu_0,$$

which rejects H_1 for $|T_{\mu_0}(\boldsymbol{X})| > t_{n-1}(\alpha/2)$.
A confidence interval for level $1 - \alpha$ is as before

$$\begin{aligned} C(\boldsymbol{x}) &= \{\mu_0 \in \mathbb{R} : \phi_{\mu_0}(\boldsymbol{x}) = 0\} \\ &= \{\mu_0 \in \mathbb{R} : |T_{\mu_0}(\boldsymbol{x})| \leq t_{n-1}(\alpha/2)\} \\ &= \left[\bar{x} - t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}; \bar{x} + t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}\right]. \end{aligned}$$

Addendum: Sufficiency and pivot statistics

Pivots are useful since they allow an easy derivation of confidence intervals.
In general, there are some hints for pivots. For location families (location parameter μ), $X - \mu$, or $\bar{X} - \mu$ are pivots (differences). For scale families with scale parameter σ, X/σ or $\bar{X}/\sigma$ are pivotal (ratios).
From that its clear that a pivot may not even be a statistic (since it may depend on unknown parameters) while a sufficient statistics is only dependent on the data.

2.4 Multiple Testing

2.4 Multiple Testing

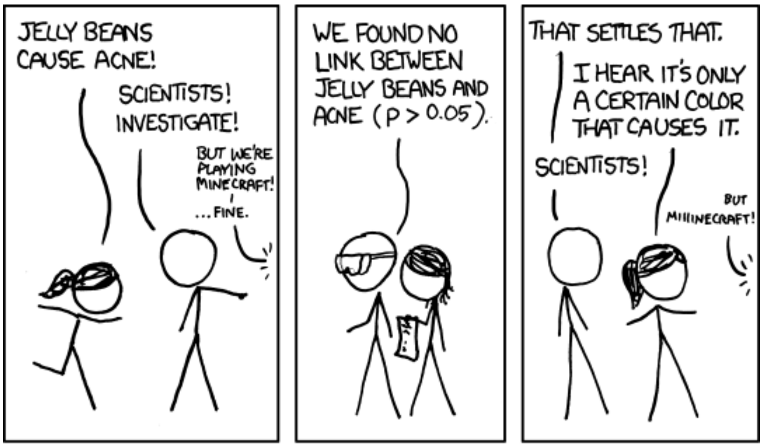

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!

Figure 6: Source: https://xkcd.com/882/

Literature:

2.4 Multiple Testing

- Lehmann & Romano, Chapter 9
- Dudoit, Shaffer & Boldrick (2003): *Multiple Hypothesis Testing in Microarray Experiments*, Statistical Science (**18**), Page 71-103

2.4 Multiple Testing

2.4 Multiple Testing

2.4 Multiple Testing

Problem: a finite set of (null) hypotheses $H_1, ..., H_m$ should be tested simultaneously with the help of *single* data set.
Examples:

2.4 Multiple Testing

- *Analysis of Variance:* comparison of several treatments with control (e.g. placebo, „common” therapy). A simultaneous test of the form

$$H_0 : \theta_1 = \ldots = \theta_m = 0 \quad \text{vs.} \quad H_{\text{alt}} : \quad \text{at least one} \;\; \theta_j \neq 0$$

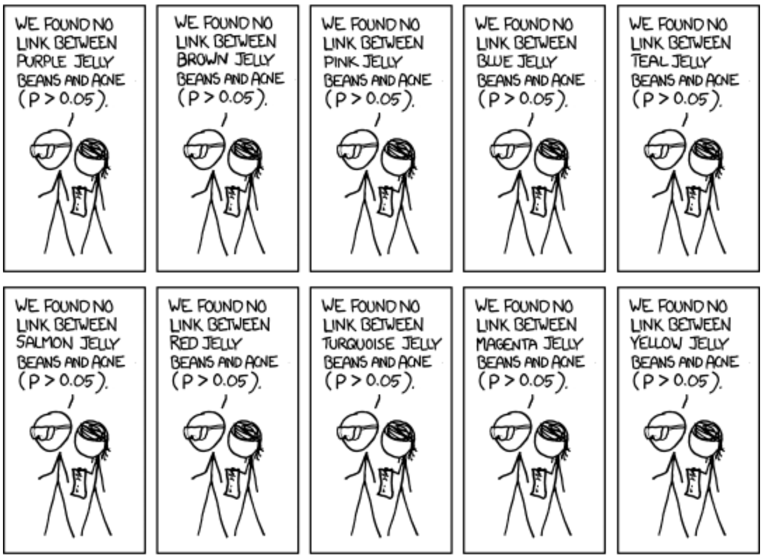

WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).

Figure 7: Source: https://xkcd.com/882/

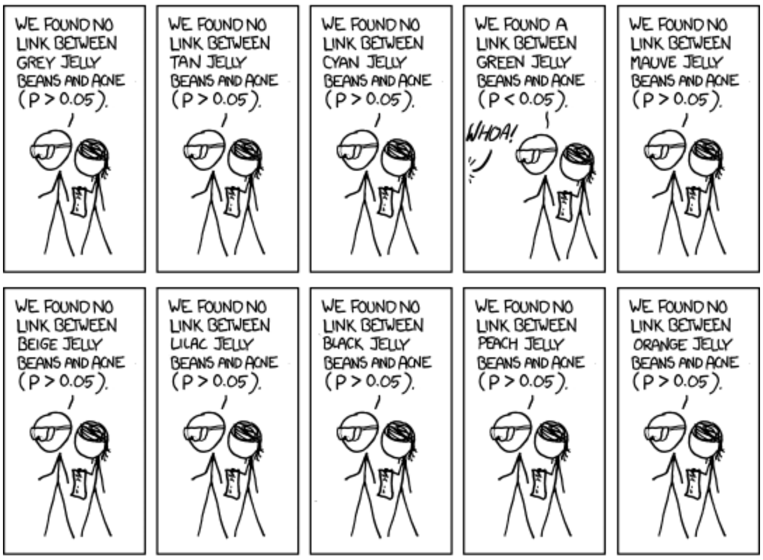

WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

Figure 8: Source: https://xkcd.com/882/

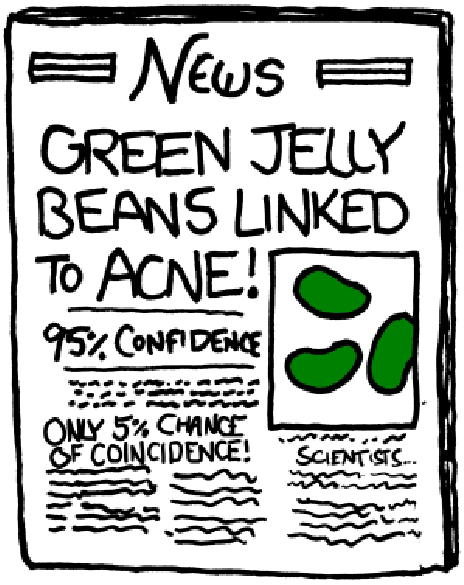

News
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Figure 9: Source: https://xkcd.com/882/

is often not sufficient: if H_0 is rejected, one would like to know which θ_j's are significantly different from 0 . For this purpose, the individual hypotheses

$$H_j := H_{0j} : \theta_j = 0$$

can (simultaneously) be tested for $j = 1, \dots, m$. Usually, m is comparatively small; „classic" multiple testing procedures can be used.

2.4 Multiple Testing

- *Microarray experiments:* Let $X_1, \dots, X_m$ be (normalized log-) expressions of genes $1, \dots, m$ on microarrays, $X_j \overset{a}{\sim} N(\mu_j, \sigma_j)$ for $j = 1, \dots, m$ and m is of the order of several thousands. . The aim is to investigate which genes have a significant influence on a phenotype Y, for example a certain disease. In a naive approach, this might be examined by simultaneous tests as in the above. In microarray analysis, rejecting H_j, means that the gene j is „deferentially expressed". However, if m and the number correct hypotheses m_0 are large, there is a high probability that one or more hypotheses will be falsely rejected.

2.4 Multiple Testing

- For example, for independent test statistics $T_1, \dots, T_m$ and $\alpha = 0.05$, the following table illustrates the problem.

$m = m_0$	1	2	5	10	50
$\mathbb{P}$(at least one false rejection)	0.05	0.10	0.23	0.40	0.92

A „new" multiple testing procedure must be utilised to control the error rates.

2.4.1 Error rates

2.4 Multiple Testing
2.4.1 Error rates

The setting with m given hypotheses can be described as follows:

	Number of not-rejected null hypotheses	Number of-rejected null hypotheses	
Number of true null hypotheses	U	V	m_0
Number of false null hypotheses	T	S	m_1
	$m-R$	R	m

There are

- m_0 the (unknown) number of true null hypotheses,
- $m_1 = m - m_0$ the (unknown) number of false null hypotheses,
- R is an observable random variable,
- S, T, U, V unobservable random variables.

2.4 Multiple Testing
2.4.1 Error rates

Ideally: minimize

- Number of Type 1 errors, V (false positives),
- Number of Type 2 errors, T (false negatives).

Classical test theory ($m = 1$):

$$\begin{aligned} \mathbb{P}(\text{Type 1 error}) &\leq \alpha \\ \mathbb{P}(\text{Type 2 error}) &\to \min \end{aligned}$$

Various generalizations for controlling error rates are possible in case of multiple testing.

2.4 Multiple Testing
2.4.1 Error rates

Type 1 error rates

- PCER (per-comparison error rate):

  $$\text{PCER} = \frac{\mathbb{E}(V)}{m}$$

  This is the relative number of expected Type 1 errors.

- PFER (per-family error rate):

  $$\text{PFER} = \mathbb{E}(V)$$

  This is the absolute number of expected Type 1 errors.

2.4 Multiple Testing
2.4.1 Error rates

- FWER (family-wise error rate):

  $$\text{FWER} = \mathbb{P}(V \geq 1)$$

  This is the probability of at least one Type 1 error.

- FDR (false discovery rate; Benjamini & Hochberg, 1995):

  $$\text{FDR} = \mathbb{E}(Q) \qquad \text{mit} \qquad Q = \begin{cases} \frac{V}{R} & \text{für } R > 0, \\ 0 & \text{für } R = 0. \end{cases}$$

  This is the expected relative frequency of Type 1 errors for R rejected hypotheses.

It is true that PCER $\leq$ FDR $\leq$ FWER $\leq$ PFER (FDR = FWER at $m = m_0$).

2.4 Multiple Testing
2.4.1 Error rates

Strong and weak controls

Typically, the following is true: for an *unknown* subset

$$\Lambda_0 \subseteq \{1, \ldots, m\}$$

the hypotheses $H_j, j \in \Lambda_0$ are true, while being false for the rest.

Strong control arises when an error rate for *each* subset Λ_0 is bounded above by α for example, when

$$\text{FWER} \leq \alpha$$

is satisfied.

Weak control is present when the error rate is controlled if *all* null hypotheses are true.

Classic approaches (e.g. Bonferroni and Holm procedures, see following subsection) control *strongly.* The FDR approach of Benjamini and Hochberg controls the FDR *weakly* and is (therefore) less conservative.

2.4 Multiple Testing
2.4.2 Multiple testing procedures

Bonferroni procedure

For $j = 1, \ldots, m$ reject hypotheses H_j, if the p-value satisfies: $p_j \leq \frac{\alpha}{m}$The following holds:

$$\text{FWER} \leq \alpha \quad \textit{strong},$$

i.e..

$$\mathbb{P}\left(V \geq 1 \,\middle|\, \bigcap_{j \in \Lambda_0} H_j\right) \leq \alpha.$$

Drawback: the level α/m of the individual tests becomes extremely small with large m and common α. For the microarrays example, therefore, there is a high probability that relevant genes will pass undetected.

2.4 Multiple Testing
2.4.2 Multiple testing procedures

Step-down procedures

Order the p-values $p_1, \ldots, p_m$ of tests $H_1, \ldots, H_m$ by size in ascending order. Then

$$p_{(1)} \leq \ldots \leq p_{(m)}$$

correspond to sorted hypotheses $H_{(1)}, \ldots, H_{(m)}$. Let

$$\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_m.$$

If $p_{(1)} \geq \alpha_1$, accept all hypotheses. Otherwise reject hypotheses $H_{(1)}, \ldots, H_{(r)}$ *step by step* and accept $H_{(r+1)}, \ldots, H_{(m)}$, if

$$p_{(1)} < \alpha_1, \ldots, p_{(r)} < \alpha_r, \text{ but } p_{(r+1)} \geq \alpha_{r+1}.$$

The **Holm procedure** specifically uses $\alpha_j = \alpha/(m - j + 1)$.
For the Holm procedure:

2.4 Multiple Testing
2.4.2 Multiple testing procedures

$$\text{FWER} \leq \alpha \quad \textit{strong.}$$

Proof:

Let j^* be the smallest (random) index with $p_{(j^*)} = \min_{j \in \Lambda_0} p_j$.
A false rejection arises if

$$p_{(1)} \leq \alpha/m, p_{(2)} \leq \alpha/(m-1), \ldots, p_{(j^*)} \leq \alpha/(m - j^* + 1).$$

Since $j^* \leq m - m_0 + 1$

$$\min_{j \in \Lambda_0} p_j = p_{(j^*)} \leq \alpha/(m - j^* + 1) \leq \alpha/m_0.$$

Thus, the probability of a false rejection ($V \geq 1$) is bounded above by

$$FWER \leq \mathbb{P}(\min_{j \in \Lambda_0} p_j \leq \alpha/m_0) \leq \sum_{j \in \Lambda_0} \mathbb{P}(p_j \leq \alpha/m_0) \leq \alpha.$$

□

2.4 Multiple Testing 2.4.2 Multiple testing procedures

An alternative are: **Step-up procedures**
If $p_{(m)} < \alpha_m$, reject all hypotheses. Otherwise reject hypotheses $H_{(1)}, \ldots, H_{(r)}$ for $r = 1, \ldots, m$, if

$$p_{(m)} \geq \alpha_m, \ldots, p_{(r+1)} \geq \alpha_{r+1},$$

but $p_{(r)} < \alpha_r$.

2.4 Multiple Testing 2.4.2 Multiple testing procedures

Remark.

- Statements about *strong* control can be found, for example, in Lehmann & Romano, Chapter 9.
- For $m \sim 100, 1000$ and larger: there will still be low power, significantly less than for the individual tests. Benjamini & Hochberg (1995) recomment controlling the FDR. The properties of multiple testing procedures are still subject of current research.
- Some of the various procedures can be formulated conveniently with the help of *adjusted* p-values $\widetilde{p}_j$,see Dudoit, Shaffer & Boldrick (2003)
- Resampling methods (bootstrap , permutations , ...) are necessary to calculate (adjusted) p-values.
- Software: `www.bioconductor.org`.

2.5 Tests, p-values and Reproducibility

There is currently a heated discussion in science about the lack of reproducibility of many published „significant“ scientific results.

Some of the possible reasons are

- absence of correction for (often implicit) multiple testing
- absence of correction for model selection (see 3.3.3)
- publication bias: often only significant results are released
- the points on the following slides.

Literature: Colquhoun (2014): An investigation of the false discovery rate and the misinterpretation of p-values. R. Soc. open sci. 1: 140216.

2.5 Tests, p-values and Reproducibility

Digression: the Screening Problem
Consider a medical test that

- tests sick people positive with 80% sensitivity
- tests healthy people negative with 95% specificity.

For a disease with 1% prevalence in the population, the proportion of false-positive tests using Bayes' theorem is

$$\begin{aligned}
&\mathbb{P}(\text{Healthy}|\text{Test positive}) \\
=\;& \frac{\mathbb{P}(\text{Test positive}|\text{Healthy})\mathbb{P}(\text{Healthy})}{\mathbb{P}(\text{Test pos.}|\text{Healthy})\mathbb{P}(\text{Healthy}) + \mathbb{P}(\text{Test pos.}|\text{Sick})\mathbb{P}(\text{Sick})} \\
=\;& \frac{0.05 \cdot 0.99}{0.05 \cdot 0.99 + 0.8 \cdot 0.01} \approx 0.86.
\end{aligned}$$

In a population-wide preventive screening, 86% of tests would be false alarms.

2.5 Tests, p-values and Reproducibility

Statistical tests and power

Now consider a statistical test with

- 80% power to reject H_0 under the alternative H_1
- 5% α level, Type 1 error rate under H_0.

If in a scientific study 10% [1%, 50%] of tested null hypotheses are not true (e.g. 10% of the new tested drugs are effective), the proportion of false positive tests is

$$\begin{aligned}
&\mathbb{P}(H_0 \text{ is true}|\text{Test significant}) \\
=\;& \frac{\mathbb{P}(\text{Test sign.}|H_0)\mathbb{P}(H_0)}{\mathbb{P}(\text{Test sign.}|H_0)\mathbb{P}(H_0) + \mathbb{P}(\text{Test sign.}|H_1)\mathbb{P}(H_1)} \\
=\;& \frac{0.05 \cdot 0.9}{0.05 \cdot 0.9 + 0.8 \cdot 0.1} = 0.36.
\end{aligned}$$

Of the published „significant"results, 36% [86%, 6%] would be false positive. The numbers increase with decreasing power!

2.5 Tests, p-values and Reproducibility

Some points on p-values:

- In 1.2.1 we saw that
$$\sup_{\boldsymbol{\theta} \in \Theta_0} \mathbb{P}_{\theta}(p(\boldsymbol{X}) \leq \alpha) \leq \alpha.$$
For $\Theta_0 = \{\boldsymbol{\theta}_0\}$ we usually have
$$\mathbb{P}_{\theta_0}(p(\boldsymbol{X}) \leq \alpha) = \alpha$$
and under H_0 the p-value is thus evenly distributed $Unif[0,1]$, i.e. all values are equally likely.
- It can be shown that for $p(\boldsymbol{x}) \approx 0.05$ the share of false-positive tests is at least 28.9% and for $p(\boldsymbol{x}) \approx 0.001$ is at least 1.8% .

2.5 Tests, p-values and Reproducibility

The Inflation Effect

- Estimates 'far away' from Θ_0 tend to be associated with rejection of H_0.
- Therefore, the exclusive consideration of significant results even for an unbiased estimator often lead to *overestimation of effect strength.* Decreasing power only exacerbates problem.
- **Example:** $X_1, \ldots, X_{30}$ i. i. d. $N(\mu = 0.5, \sigma^2)$ t-test for $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$.

σ	Power	Mean $\hat{\mu}$ on significance
0.5	99.9 %	0.50
1.0	75.4 %	0.57
1.5	41.6 %	0.75
2.0	26.1 %	0.93

General notes

- Since lower power increases both the proportion of false-positive tests and the inflation effect, calculations of power (and necessary n) can provide important information.

- It is important to correctly interpret p-values as well as test results!

- American Statistical Association opinion from 2016 on the subject „Statistical Significance and P-Values" at http://amstat.tandfonline.com/doi/pdf/10.1080/00031305.2016.1154108?needAccess=true.

Addendum: Weak and strong control in multiple testing problems

In the edge case $R = 0$, also $V = 0$ (and $Q = 0$ by definition), and we cannot make a wrong rejection of a null hypothesis since we do not reject any of the null hypotheses.
Assume first that $m = m_0$ (i.e. all null hypotheses are true). Than $V = R$. If $R \geq 1$ than $Q = V/R = 1$ (otherwise $Q = 0$). Then
$FDR = \mathbb{E}(Q) = 1 \cdot P(Q = 1) + 0 \cdot P(Q = 0) = P(Q = 1) = P(V \geq 1) = FWER.$
Now assume that $m_0 < m$.
If $V > 0$ (that means at least one false rejection of a null hypothesis), $Q = V/R \leq 1$. Now the indicator function $1_{V\geq1}$ (which is 1 if $V \geq 1$) is never lower than Q (note, if $V = 0$, $Q = 0$ and $1_{V\geq1} = 0$, so that both are equal to 0). I.e. $Q \leq 1_{V\geq1}$. Taking expectations on both sides and using the monotony of the expectation, one gets

$$FDR = \mathbb{E}(Q) \leq E 1_{V\geq1} = P(V \geq 1) = FWER \ ,$$

since the expectation of an indicator variable is the probability that it takes the value 1.
So if $FWER$ is controlled (strong control), i.e. $FWER \leq \alpha$ than also $FDR \leq \alpha$.
But if $FDR \leq \alpha$ (weak control), $FWER$ may exceed α.

Chapter 3

Likelihood inference

Likelihood inference

Objectives of Chapter 3: Introduction to likelihood inference

3.1 Parametric Likelihood-Inference

3.1 Parametric Likelihood-Inference

Assumption: $\mathcal{P}_{\boldsymbol{\theta}} = \{f(\boldsymbol{x}|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, $\Theta \subseteq \mathbb{R}^p$, $p \ll n$, p is kept constant for $n \to \infty$. $f(\boldsymbol{x}|\boldsymbol{\theta})$ is a discrete or continuous or more generally a Radon-Nikodym density.

Definition 3.1 (Likelihood function)
The likelihood function *of* $\boldsymbol{\theta} \in \Theta$,

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

is defined as the density of the observed data $\boldsymbol{X} = (X_1, \ldots, X_n) = \boldsymbol{x} = (x_1, \ldots, x_n)$*, considered as a function of* $\boldsymbol{\theta}$*. With* $L(\boldsymbol{\theta})$, $\widetilde{L}(\boldsymbol{\theta}) = const \times L(\boldsymbol{\theta})$ *is also a likelihood function.*

- *Remark 1: p constant is required for standard asymptotic theory for likelihood inference*
- *Remark 2: The likelihood is unique except for the proportionality constant*
- *Remark 3: Proportional likelihoods should lead to the same conclusions (i.e. inference) about* $\boldsymbol{\theta}$

3.1 Parametric likelihood inference

- An alternative definition defines the likelihood as the *probability* of the observed data as a function of $\boldsymbol{\theta}$, that parametrizes the statistical model, i.e. the likelihood is a probability.
- For discrete observations, the definitions agree.
- For continuous observations and scalar x, let us consider for a small ε

$$\mathbb{P}_{\boldsymbol{\theta}}(x - \frac{\varepsilon}{2} \leq X \leq x + \frac{\varepsilon}{2}) = \int_{x-\frac{\varepsilon}{2}}^{x+\frac{\varepsilon}{2}} f(x|\boldsymbol{\theta})dx \approx \varepsilon f(x|\boldsymbol{\theta}).$$

  Since ε does not depend on $\boldsymbol{\theta}$, the constant factor in the likelihood $L(\boldsymbol{\theta}|x)$ can be ignored.

3.1 Parametric likelihood inference

Common situations:

1. The random variables $X_1, \ldots, X_n$ are i.i.d. as random variable $X_1 \sim f_1(x|\boldsymbol{\theta})$. Then the factorization

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{n} f_1(x_i|\boldsymbol{\theta})$$

   holds. Note that the density f_1 is the same for all observations.

2. $X_1, \ldots, X_n$ are independent, but no longer distributed identically. Then the factorization

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{n} f_i(x_i|\boldsymbol{\theta})$$

applies. Note that the densities f now have an observation specific index i. *Remark: Regression is a special case, where $y_1|\boldsymbol{z}_1, \ldots, y_n|\boldsymbol{z}_n$ – for a random outcome variable y and covariates $\boldsymbol{z}$ – are assumed to be (conditionally) independent but not identically distributed.*

3. $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent, the individual components within a vector $\boldsymbol{X}_i$, however, may be dependent.

3.1 Parametric likelihood inference

4. Temporally correlated observations $X_1, \ldots, X_t, \ldots, X_n$ with density function

$$\begin{aligned} f(x_1, \ldots, x_t, \ldots, x_n|\boldsymbol{\theta}) = \\ f(x_n|x_{n-1}, \ldots, x_t, \ldots, x_1; \boldsymbol{\theta}) \cdot f(x_{n-1}|x_{n-2}, \ldots, x_1; \boldsymbol{\theta}) \cdot \ldots \\ \cdot f(x_2|x_1; \boldsymbol{\theta}) f(x_1|\boldsymbol{\theta}). \end{aligned}$$

For first order Markov chains with the property

$$f(x_n|x_{n-1}, \ldots, x_1; \boldsymbol{\theta}) = f(x_n|x_{n-1}; \boldsymbol{\theta}) \ ,$$

the likelihood is simplified to

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \left(\prod_{i=2}^{n} f(x_i|x_{i-1}; \boldsymbol{\theta}) \right) f(x_1|\boldsymbol{\theta}).$$

3.1 Parametric likelihood inference

Example 3.1

1. *Recall the examples in section 1.1.1*

2. *Regression situations (cross-sectional data) with independent outcome variables $y_1|\boldsymbol{z}_1, \ldots, y_n|\boldsymbol{z}_n$ and covariates $\boldsymbol{z}_i$:*
   - *classical linear model: $y_i|\boldsymbol{z}_i \sim N(\boldsymbol{z}_i^\top \boldsymbol{\beta}, \sigma^2)$,*
   - *Logit or probit model: $y_i|\boldsymbol{z}_i \sim Bin(1, \pi_i = h(\boldsymbol{z}_i^\top \boldsymbol{\beta}))$,*
   - *Poisson regression: $y_i|\boldsymbol{z}_i \sim Po(\lambda_i = h(\boldsymbol{z}_i^\top \boldsymbol{\beta}))$.*

3. - *Multivariate data*
   - *Survival data $\boldsymbol{X}_i = (T_i, \delta_i)$ with constant lifetime T_i und discrete censoring indicator δ_i*

4. *Markov chains or autoregressive models for time series and longitudinal data.*

3.1 Parametric likelihood inference

Autoregressive process of order one: Let

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \varepsilon_t$$

with $\varepsilon_t \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ or, with additional (time-dependent) covariate vectors $\boldsymbol{z}_t$,

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \boldsymbol{z}_t^\top \boldsymbol{\beta} + \varepsilon_t.$$

Then the likelihood is

$$L(\boldsymbol{\theta}) = \left(\prod_{t=2}^{n} f_t(y_t|y_{t-1};\boldsymbol{\theta}) \right) f_1(y_1)$$

with

$$f_t(y_t|y_{t-1};\boldsymbol{\theta}) = \phi(y_t|\mu_t = \alpha_0 + \alpha_1 y_{t-1} + \boldsymbol{z}_t^\top \boldsymbol{\beta}, \sigma^2),$$

where $\phi(y_t|\mu_t, \tau^2)$ is the value of the normal density with expected value μ_t und variance τ^2 evaluated at y_t.

3.1 Parametric likelihood inference

Example 3.2

Let us consider independent, but partly incomplete observations from a normal distribution $N(\theta, 1)$ with known variance.

1. *Let the first observation $x_1 = 2.45$. Then*

$$\begin{aligned} L_1(\theta) &= L(\theta|X_1 = 2.45) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(2.45 - \theta)^2 \right). \end{aligned}$$

3.1 Parametric likelihood inference

2. Assume that for the second observation, we only know that $0.9 < x_2 < 4$ (we call that an interval censored observation). Then

$$\begin{aligned} L_2(\theta) = L(\theta|0.9 < X_2 < 4) &= \mathbb{P}_\theta(0.9 < X_2 < 4) \\ &= \Phi(4 - \theta) - \Phi(0.9 - \theta) \ , \end{aligned}$$

where Φ denotes the PDF of the standard normal distribution.
Remark: $X_2 < 4 \Leftrightarrow 4 - \theta > X_2 - \theta \sim N(0, 1)$.
Formally, one could also define a binary variable

$$X_2^d = \begin{cases} 1, \ 0.9 < X_2 < 4, \\ 0, \ \text{sonst} \end{cases}$$

with density (probability) function

$$f_2^d(1) = \mathbb{P}_\theta(X_2^d = 1) = \Phi(4 - \theta) - \Phi(0.9 - \theta).$$

3.1 Parametric likelihood inference

3. Let $z_1, \ldots, z_n$ i.i.d. realizations from $N(\theta, 1)$. But we only observe the maximum

$$x_3 = \max_{1 \le i \le n} z_i = z_{(n)}.$$

All other observations are missing. The distribution function of the maximum $X_3 = Z_{(n)}$ is

$$\begin{aligned} F_\theta(z_{(n)}) &= \mathbb{P}_\theta(Z_{(n)} \le z_{(n)}) = \mathbb{P}_\theta(Z_1 \le z_{(n)} \wedge \ldots \wedge Z_n \le z_{(n)}) \\ &= \mathbb{P}_\theta(Z_i \le z_{(n)} \ \forall \ i) \overset{\text{indep}}{=} [\Phi(z_{(n)} - \theta)]^n. \end{aligned}$$

The density results from differentiation with respect to $z_{(n)}$:

$$f_\theta(z_{(n)}) = n[\Phi(z_{(n)} - \theta)]^{n-1} \phi(z_{(n)} - \theta) \ ,$$

where ϕ is the density of the normal distribution. For example, with $n = 5$ and $z_{(n)} = x_3 = 3.5$ one gets

$$L_3(\theta) = L(\theta|X_3 = 3.5) = 5[\Phi(3.5 - \theta)]^4 \phi(3.5 - \theta).$$

3.1 Parametric likelihood inference

The total likelihood of the three observations is therefore

$$L(\theta|x_1, 0.9 < X_2 < 4, x_3) \overset{\text{indep.}}{=} L_1(\theta) \cdot L_2(\theta) \cdot L_3(\theta),$$

i.e. the product of the likelihood functions L_1, L_2 und L_3.
Conclusion: Likelihood is a very generally construct.
Remark: The described situations are not only academic but occur in practice: a blood marker may be censored, i.e. only observed if the concentration is higher than a detection limit; the income of a person may be not known exactly but only that it lies in an interval; a certificate may contain only the best grade from several attempts.

3.1 Parametric likelihood inference

Relationship to Bayesian Inference

In Bayesian inference, the parameters *and* the data are considered as random variables.

- Let $p(\boldsymbol{\theta})$ the prior density for θ.
- $L(\boldsymbol{\theta}|\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\theta})$ is the likelihood.
- Then ($\propto$ means proportional to)

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})/f(\boldsymbol{x}) \quad &\propto \quad p(\boldsymbol{\theta}) \cdot L(\boldsymbol{\theta}|\boldsymbol{x}) \\ \text{"posterior"} \quad &\propto \quad \text{"prior"} \times \text{likelihood}. \end{aligned}$$

3.1 Parametric likelihood inference

Likelihood ratio

Question: How do you compare two likelihoods $L(\boldsymbol{\theta}_1|\boldsymbol{x})$ and $L(\boldsymbol{\theta}_2|\boldsymbol{x})$ for $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$?
Answer: One looks at the ratio (not the difference), since the ratio is invariant to one-to-one transformations, e.g. it doesn't matter whether we measure the temperature in $^\circ C$ or $^\circ F$:

$$\boldsymbol{x} \mapsto \boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}) \;\Leftrightarrow\; \boldsymbol{y} \mapsto \boldsymbol{x}(\boldsymbol{y})$$

For continuous $\boldsymbol{x}, \boldsymbol{y}$, this can be shown using the density transformation formula:

$$f_Y(\boldsymbol{y}|\boldsymbol{\theta}) = f_X(\boldsymbol{x}(\boldsymbol{y})|\boldsymbol{\theta}) \left| \det\left(\frac{\partial \boldsymbol{x}(\boldsymbol{y})}{\partial \boldsymbol{y}}\right) \right|$$

and therefore (factors independent of $\boldsymbol{\theta}$ cancel out)

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = L(\boldsymbol{\theta}|\boldsymbol{x}) \left| \det\left(\frac{\partial \boldsymbol{x}(\boldsymbol{y})}{\partial \boldsymbol{y}}\right) \right| \;\Rightarrow\; \frac{L(\boldsymbol{\theta}_2|\boldsymbol{y})}{L(\boldsymbol{\theta}_1|\boldsymbol{y})} = \frac{L(\boldsymbol{\theta}_2|\boldsymbol{x})}{L(\boldsymbol{\theta}_1|\boldsymbol{x})}.$$

Remark: in the discrete case, $P(\boldsymbol{X} = \boldsymbol{x}) = P(\boldsymbol{Y} = \boldsymbol{y})$ holds for the PMF.

3.1 Parametric likelihood inference

Theorem 3.2

1. *Let $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$ be sufficient for $\boldsymbol{\theta}$. Then $L(\boldsymbol{\theta}|\boldsymbol{x}) = const \times L(\boldsymbol{\theta}|\boldsymbol{t})$ with $\boldsymbol{t} = \boldsymbol{T}(\boldsymbol{x})$, i.e. $L(\boldsymbol{\theta}|\boldsymbol{x})$ and $L(\boldsymbol{\theta}|\boldsymbol{t})$ are equivalent.*
2. *$L(\boldsymbol{\theta}|\boldsymbol{x})$ is minimally sufficient.*

Proofs: Immediately follow from the results from Section 2.

□

- *Remark 1: Sufficiency follows from the factorization theorem 2.4.*
- *Remark 2: Minimal sufficiency follows from definition 2.5, since the likelihood is a function of every sufficient statistics.*

3.2 Maximum likelihood estimation

3.2 Maximum likelihood estimation (MLE)

Maximum likelihood estimation (MLE) is the most popular method for the construction of point estimators for purely parametric estimation problems.

Definition 3.3 (Maximum likelihood estimator)
The maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ is defined as the value of $\boldsymbol{\theta}$ for which

$$L(\widehat{\boldsymbol{\theta}}|\boldsymbol{x}) \geq L(\boldsymbol{\theta}|\boldsymbol{x}) \textit{ for all } \boldsymbol{\theta} \in \Theta \ ,$$

i.e.

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}} \, L(\boldsymbol{\theta}|\boldsymbol{x}) \ .$$

Interpretation: Find that value $\hat{\boldsymbol{\theta}}$ such that the "probability" of the observed data is maximized. This is then considered as the most plausible estimate.

3.2 Maximum likelihood estimation (MLE)

Definition 3.4 (Maximum likelihood estimator)
Equivalently to the maximization in definition 3.3 is to find the value $\boldsymbol{\theta}$ such that

$$\ell(\widehat{\boldsymbol{\theta}}|\boldsymbol{x}) \geq \ell(\boldsymbol{\theta}|\boldsymbol{x}) \textit{ for all } \boldsymbol{\theta} \in \Theta \ .$$

ℓ is called log-likelihood, i.e. $\ell(\boldsymbol{\theta}|\boldsymbol{x}) = \log L(\boldsymbol{\theta}|\boldsymbol{x})$.

- *Remark: with* log*, we always consider the natural logarithm, i.e. the inverse function of the exponential function* exp*.*
- *Remark: The equivalency holds because* log *is a strictly monotonic function.*

3.2 Maximum likelihood estimation
3.2.1 Estimation concept

Theorem 3.5 (Invariance of the MLE)
If $\widehat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ and $\boldsymbol{h}(\cdot)$ is a one-to-one function, then $\boldsymbol{h}(\widehat{\boldsymbol{\theta}})$ is the ML estimator of $\boldsymbol{h}(\boldsymbol{\theta})$.

Example 3.3
$N(\mu, \sigma^2)$. If S^2 is the MLE of σ^2, S is the (possibly biased) MLE of σ.

Example 3.4
$B(n, \pi)$ (Binomial distribution). Let $\hat{\pi}$ the MLE of π. Then

$$\hat{\theta} = \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) \quad \textit{is the MLE of the log-odds} \quad \log\left(\frac{\pi}{1-\pi}\right) \ .$$

3.2.1 Estimation concept

3.2 Maximum likelihood estimation
3.2.1 Estimation concept

Usually, one searches for (local) maxima of $\ell(\boldsymbol{\theta}|\boldsymbol{x})$. Setting the score function to zero

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

(as far as the first derivation of the log-likelihood exists) provides a solution of the so-called *ML or likelihood equation*

$$\boldsymbol{s}(\widehat{\boldsymbol{\theta}}|\boldsymbol{x}) = \mathbf{0}.$$

- *Remark 1: This works (mostly) assuming Fisher regularity and if a unique maximum exists which is not at the boundary of Θ.*

- *Remark 2: The solution is only in simple cases analytically available.*

3.2 Maximum likelihood estimation 3.2.1 Estimation concept

Example 3.5 (Linear model)
Consider

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \textit{with} \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n).$$

- *Likelihood ($\boldsymbol{y} \sim N(\boldsymbol{Z}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$, constants are ignored)*

$$L(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|^2\right)$$

- *Log-Likelihood:*

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2} \underbrace{\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|^2}_{\textit{Least-Squares criterion}}$$

3.2 Maximum likelihood estimation 3.2.1 Estimation concept

- Score function:

$$\begin{aligned} \boldsymbol{s}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma^2) &= \frac{\partial \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}\boldsymbol{Z}^\top(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}) \\ s_{\sigma^2}(\boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|^2 \end{aligned}$$

It is easy to verify that $\mathbb{E}[\boldsymbol{s}_{\boldsymbol{\beta}}] = \boldsymbol{0}, \mathbb{E}[s_{\sigma^2}] = 0$. From the ML equations (setting the score functions to zero), it follows:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{\text{ML}} &= (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{y}, \\ \widehat{\sigma}^2_{\text{ML}} &= \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{Z}\beta\|^2. \end{aligned}$$

$\widehat{\boldsymbol{\beta}}_{\text{ML}}$ therefore corresponds to the least squares estimator. $\widehat{\sigma}^2_{\text{ML}}$ is biased, but asymptotically unbiased.

3.2 Maximum likelihood estimation 3.2.1 Estimation concept

- Information matrix:

$$\begin{aligned} -\frac{\partial^2 \ell(\beta, \sigma^2)}{\partial\beta\partial\beta^\top} &= -\frac{\partial s_{\boldsymbol{\beta}}(\beta, \sigma^2)}{\partial\beta^\top} \\ &= \frac{1}{\sigma^2}\boldsymbol{Z}^\top\boldsymbol{Z} = \left(\text{Cov}(\widehat{\beta})\right)^{-1} \text{(independent of } y.) \end{aligned}$$

$$-\frac{\partial^2 \ell(\beta, \sigma^2)}{\partial\beta\partial\sigma^2} = \frac{1}{\sigma^4}\boldsymbol{Z}^\top(\boldsymbol{y} - \boldsymbol{Z}\beta) \qquad \Rightarrow \mathbb{E}\left[-\frac{\partial^2 \ell(\beta, \sigma^2)}{\partial\beta\partial\sigma^2}\right] = \boldsymbol{0}$$

$$-\frac{\partial^2 \ell(\beta, \sigma^2)}{\partial\sigma^2\partial\sigma^2} = -\frac{n}{2(\sigma^2)^2} + \frac{\|\boldsymbol{y} - \boldsymbol{Z}\beta\|^2}{(\sigma^2)^3} \qquad \Rightarrow \mathbb{E}\left[-\frac{\partial^2 \ell(\beta, \sigma^2)}{\partial\sigma^2\partial\sigma^2}\right] = \frac{n}{2\sigma^4}$$

The second expected value follows from

$$\|\boldsymbol{y} - \boldsymbol{Z}\beta\|^2 = \sum_{i=1}^{n} \varepsilon_i^2 \sim \sigma^2\chi^2(n).$$

3.2 Maximum likelihood estimation 3.2.1 Estimation concept

Example 3.5 (Generalized linear model (GLM))
Let $y_i \overset{ind.}{\sim} f(y_i|\mu_i)$ for $i = 1, \dots, n$ with $\mu_i = h(\boldsymbol{x}_i^\top\boldsymbol{\beta})$, e.g. $y_i \sim Po(\lambda_i)$ and $\lambda_i = \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})$ (log linear Poisson-Model).

3.2 Maximum likelihood estimation
3.2.1 Estimation concept

Example 3.6 (Generalized linear mixed model (GLMM) for longitudinal data)
Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ij}, \ldots, y_{in_i})$ *with conditionally independent* $y_{ij} \sim f(y_{ij}|\mu_{ij})$ *and* $\mu_{ij} = h(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{b}_i)$.
The $\boldsymbol{b}_i$ *are, for example, individual-specific intercepts* b_i *(*$\boldsymbol{z}_{ij} \equiv 1$*) with distributional assumption* $b_i \overset{i.i.d.}{\sim} N(0, \tau^2)$. *The (marginal) likelihood of the parameter* $\boldsymbol{\theta} = (\boldsymbol{\beta}, \tau^2)$ *is then*

$$L(\boldsymbol{\beta}, \tau^2) = \int \prod_{i=1}^{m} \prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\beta}, b_i)\phi(b_i|0, \tau^2)\, db_i.$$

where the random effects b_i *have been integrated out.*
Remark: Approaches to maximize the likelihood are the expectation maximization (EM) algorithm or numerical integration or Bayesian inference.

3.2.2 Iterative numerical methods for calculating the ML estimator

3.2 Maximum likelihood estimation
3.2.2 Iterative numerical methods for calculating the ML estimator

If no analytical solution exists, one solves the ML equations using methods such as Newton-Raphson, Quasi-Newton, Fisher scoring or the EM algorithm.
The first three methods work with the (negative) Hessian matrix of log-likelihood, the observed information matrix

$$\boldsymbol{J}(\boldsymbol{\theta}|\boldsymbol{x}) = -\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \left(-\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{x})}{\partial \theta_i \partial \theta_j} \right)_{ij}$$

or approximations to this or the expected Information matrix

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \mathbb{E}_\theta[\boldsymbol{J}(\boldsymbol{\theta}|\boldsymbol{X})].$$

Under Fisher regularity:

$$\mathbb{E}_\theta[\boldsymbol{s}(\boldsymbol{\theta}|\boldsymbol{X})] = \boldsymbol{0} \quad \text{and} \quad \mathrm{Cov}_\theta(\boldsymbol{s}(\boldsymbol{\theta}|\boldsymbol{X})) = \mathbb{E}_\theta[\boldsymbol{s}(\boldsymbol{\theta}|\boldsymbol{X})\boldsymbol{s}(\boldsymbol{\theta}|\boldsymbol{X})^\top] = \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}).$$

3.2 Maximum likelihood estimation
3.2.2 Iterative numerical methods for the computation of the ML estimator

Newton-Raphson
The Newton-Raphson procedure is a general method for searching the roots of a continuously differentiable function $\boldsymbol{g}(\boldsymbol{\theta})$.
In the scalar case one makes a Taylor series expansion around the current iteration value $\theta^{(t)}$:

$$g(\theta) \approx g(\theta^{(t)}) + g'(\theta^{(t)})(\theta - \theta^{(t)}) \overset{!}{=} 0$$

The next iteration value $\theta^{(t+1)}$ is then a new approximate solution for the root $\widehat{\theta}$:

$$\theta^{(t+1)} = \theta^{(t)} - \frac{g(\theta^{(t)})}{g'(\theta^{(t)})}.$$

3.2 Maximum likelihood estimation
3.2.2 Iterative numerical methods for calculating the ML estimator

For the multidimensional analog we get with $\boldsymbol{s}(\widehat{\boldsymbol{\theta}}|\boldsymbol{x})(\hat{=}g(\hat{\theta})) = \boldsymbol{0}$ and $\boldsymbol{J}(\boldsymbol{\theta}^{(t)}|\boldsymbol{x})(\hat{=}g'(\theta^{(t)}))$:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + [\boldsymbol{J}(\boldsymbol{\theta}^{(t)}|\boldsymbol{x})]^{-1}\boldsymbol{s}(\boldsymbol{\theta}^{(t)}|\boldsymbol{x}),$$

as the observed information matrix

$$\boldsymbol{J}(\boldsymbol{\theta}|\boldsymbol{x}) = -\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{x})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} = -\frac{\partial \boldsymbol{s}(\boldsymbol{\theta}|\boldsymbol{x})}{\partial\boldsymbol{\theta}^\top}$$

corresponds to the negative derivative of $\boldsymbol{s}(\boldsymbol{\theta}|\boldsymbol{x})$.

Quasi-Newton works with approximations to $\boldsymbol{J}(\boldsymbol{\theta}|\boldsymbol{x})$,
Fisher-Scoring with the expected information matrix $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \mathbb{E}_\theta[\boldsymbol{J}(\boldsymbol{\theta}|\boldsymbol{X})]$.

3.2 Maximum likelihood estimation 3.2.2 Iterative numerical methods for calculating the ML estimator

EM (Expectation-Maximization)-Algorithms
The EM algorithm is an alternative to Newton-Raphson and Fisher scoring, especially in models with incomplete data or so-called latent (not directly observable) variables, e.g. random effects or factors.

3.2 Maximum likelihood estimation 3.2.2 Iterative numerical methods for calculating the ML estimator

Examples: (see 1.1.6)

- Missing values, e.g. due to missing answers / laboratory values
- As in 1.1.6 a) lifetimes may be only partially observed. For the right-censored observations, the exact life time is not known.
- As in 1.1.6 b), the data may come from a mixture distribution, where the component of the mixture from which each observation originates, is unknown (not observable).
- Mixed models (see 1.1.5. b)). The random effects are not observed.

3.2 Maximum likelihood estimation 3.2.2 Iterative numerical methods for calculating the ML estimator

Notation and setup:

- $\boldsymbol{x}$ observed ("incomplete") data
- $\boldsymbol{z}$ unobserved data/latent variables
- $(\boldsymbol{x}, \boldsymbol{z})$ complete data
- $L(\boldsymbol{\theta}|\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\theta})$ likelihood of the observed data
- $L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{z}) = f(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})$ likelihood of the complete data

Goal: Maximize $L(\boldsymbol{\theta}|\boldsymbol{x})$ using $L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{z})$ wrt $\boldsymbol{\theta}$.

The EM algorithm is particularly useful when $L(\boldsymbol{\theta}|\boldsymbol{x})$ is difficult to calculate and $L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{z})$ is easier to handle.

3.2 Maximum likelihood estimation 3.2.2 Iterative numerical methods for calculating the ML estimator

Intuition:

$$\begin{aligned}
\log f(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) &= \log f(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}) + \log f(\boldsymbol{x}|\boldsymbol{\theta}) \\
\Rightarrow \underbrace{\mathbb{E}_{\boldsymbol{\theta}^{(m)}}[\ell(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{Z})]}_{Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)})} &= \underbrace{\mathbb{E}_{\boldsymbol{\theta}^{(m)}}[\ell(\boldsymbol{\theta}\big|(\boldsymbol{Z}|\boldsymbol{x}))]}_{C(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)})} + \ell(\boldsymbol{\theta}|\boldsymbol{x}),
\end{aligned}$$

where the expected values $\mathbb{E}_{\boldsymbol{\theta}^{(m)}}$ are calculated with respect to $f(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(m)})$ at the current iteration value $\boldsymbol{\theta}^{(m)}$ to eliminate $\boldsymbol{z}$.

We have $C(\boldsymbol{\theta}^{(m)};\boldsymbol{\theta}^{(m)}) \geq C(\boldsymbol{\theta}^{(m+1)};\boldsymbol{\theta}^{(m)})$ because of the information inequality: generally it holds, that $\mathbb{E}_{\boldsymbol{\theta}}(\log f(\boldsymbol{Y}|\boldsymbol{\theta})) \geq \mathbb{E}_{\boldsymbol{\theta}}(\log f(\boldsymbol{Y}|\boldsymbol{\theta}'))$, since

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\log \frac{f(\boldsymbol{Y}|\boldsymbol{\theta}')}{f(\boldsymbol{Y}|\boldsymbol{\theta})}\right] \leq \log \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{f(\boldsymbol{Y}|\boldsymbol{\theta}')}{f(\boldsymbol{Y}|\boldsymbol{\theta})}\right] = \log \int f(\boldsymbol{y}|\boldsymbol{\theta}')d\boldsymbol{y} = \log 1 = 0$$

(Jensen's inequality), and since the log function is concave.

3.2 Maximum likelihood estimation
3.2.2 Iterative numerical methods for calculating the ML estimator

$$\underbrace{\mathbb{E}_{\boldsymbol{\theta}^{(m)}}[\ell(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{Z})]}_{Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)})} \quad = \quad \underbrace{\mathbb{E}_{\boldsymbol{\theta}^{(m)}}[\ell(\boldsymbol{\theta}\big|(\boldsymbol{Z}|\boldsymbol{x}))]}_{C(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)})} + \ell(\boldsymbol{\theta}|\boldsymbol{x}),$$

Therefore it follows, as long as $Q(\boldsymbol{\theta}^{(m+1)};\boldsymbol{\theta}^{(m)}) \geq Q(\boldsymbol{\theta}^{(m)};\boldsymbol{\theta}^{(m)})$, that

$$\begin{aligned} & Q(\boldsymbol{\theta}^{(m+1)};\boldsymbol{\theta}^{(m)}) - C(\boldsymbol{\theta}^{(m+1)};\boldsymbol{\theta}^{(m)}) \\ & \quad \geq Q(\boldsymbol{\theta}^{(m)};\boldsymbol{\theta}^{(m)}) - C(\boldsymbol{\theta}^{(m)};\boldsymbol{\theta}^{(m)}) \\ \Leftrightarrow \quad & \ell(\boldsymbol{\theta}^{(m+1)}|\boldsymbol{x}) \geq \ell(\boldsymbol{\theta}^{(m)}|\boldsymbol{x}) \ . \end{aligned}$$

Algorithm: For given $\boldsymbol{\theta}^{(m)}$ maximize $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)})$ over $\boldsymbol{\theta}$ and get $\boldsymbol{\theta}^{(m+1)}$. The new $\boldsymbol{\theta}^{(m+1)}$ increases the likelihood. Iterate.

3.2 Maximum likelihood estimation
3.2.2 Iterative numerical methods for calculating the ML estimator

Algorithm 1: EM-Algorithm

Start value: $\boldsymbol{\theta}^{(0)}$. Iterate for $m = 0, 1, 2, \ldots$ **E**/**M**-steps until convergence of $\boldsymbol{\theta}^{(m)}$, e.g. until the absolute difference $|\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}| < \epsilon$.

- **E**-Step (expectation step): Calculate

$$Q(\boldsymbol{\theta}) = Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)}) = \underbrace{\mathbb{E}_{\boldsymbol{\theta}^{(m)}}[\ \ell(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{Z})]}_{\text{regarding the distribution of } \boldsymbol{Z}|\boldsymbol{x} \text{ with parameter } \boldsymbol{\theta}^{(m)}} \ .$$

- **M**-Step (maximization step): Calculate $\boldsymbol{\theta}^{(m+1)}$, such that $Q(\boldsymbol{\theta})$ is maximized:

$$\boldsymbol{\theta}^{(m+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}).$$

3.2 Maximum likelihood estimation
3.2.2 Iterative numerical methods for calculating the ML estimator

Theorem 3.6
Under relatively general assumptions $\boldsymbol{\theta}^{(m)} \rightarrow \widehat{\boldsymbol{\theta}}_{ML}$ *for* $m \rightarrow \infty$.

Remark: The ML estimator maximizes the likelihood of the observed data.
Properties of the EM algorithm:

- Monotony: $\ell(\boldsymbol{\theta}^{(m+1)}|\boldsymbol{x}) \geq \ell(\boldsymbol{\theta}^{(m)}|\boldsymbol{x})$.
- Slow convergence (linear, not quadratic as e.g. Newton-Raphson).
- The standard error of the resulting estimator is difficult to determine since the information matrix is not directly accessible as with Fisher scoring.

3.2 Maximum likelihood estimation
3.2.2 Iterative numerical methods for calculating the ML estimator

Example 3.7 (Mixture distribution)
Let $X_1, \dots, X_n$ i.i.d. as $X \sim f(x|\boldsymbol{\theta})$, where X is also allowed to be multivariate (note that we have in general reserved the symbol $\boldsymbol{X}$ for $\boldsymbol{X} = (X_1, \dots, X_n)$). Consider the mixture distribution

$$f(x|\boldsymbol{\theta}) = \sum_{j=1}^{J} \pi_j f_j(x|\boldsymbol{\vartheta}_j) \quad \textit{with} \quad \boldsymbol{\theta} = (\boldsymbol{\vartheta}_1, ..., \boldsymbol{\vartheta}_J,\ \pi_1, ..., \pi_J). \tag{3.2}$$

- *π_j unknown mixture proportions, $\sum_{j=1}^{J} \pi_j = 1$,*
- *$f_j(x|\boldsymbol{\vartheta}_j)$ is the j-th mixture component,*
- *$\boldsymbol{\vartheta}_j$ is the unknown parameter (vector) for component j.*

3.2 Maximum likelihood estimation
3.2.2 Iterative numerical methods for calculating the ML estimator

With a mixture of multivariate normal distributions, we get especially

$$\begin{aligned} f_j(x|\boldsymbol{\vartheta}_j) &\propto |\boldsymbol{\Sigma}_j|^{-1/2} \exp\left(-\frac{1}{2}(x - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1}(x - \boldsymbol{\mu}_j)\right) \\ X &\sim \pi_1 N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \pi_2 N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + \ldots + \pi_J N(\boldsymbol{\mu}_J, \boldsymbol{\Sigma}_J). \end{aligned}$$

In the univariate case with two mixture components:

$$X \sim \pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2).$$

Interpretation of the mixture model (3.2): x_i comes from one of J subpopulations, where, in each subpopulation j,

$$X_i|j \sim f_j(x_i|\boldsymbol{\vartheta}_j).$$

3.2 Maximum likelihood estimation
3.2.2 Iterative numerical methods for calculating the ML estimator

Define the unobserved (latent) indicator variable Z_i for $j = 1, \dots, J$ as

$$Z_i = j \ \Leftrightarrow x_i \text{ is from Population } j.$$

The marginal distribution is $\mathbb{P}(Z_i = j) = \pi_j,\ j = 1, \dots, J$. Then the conditional distribution of $X_i|Z_i$ is

$$X_i|Z_i = j \sim f_j(x_i|\boldsymbol{\vartheta}_j)\ .$$

3.2 Maximum likelihood estimation
3.2.2 Iterative numerical methods for calculating the ML estimator

The log-likelihood of the observed data $\boldsymbol{x}$ is (log of sums!)

$$\ell(\boldsymbol{\theta}|\boldsymbol{x}) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{J} \pi_j f_j(x_i|\boldsymbol{\vartheta}_j)\right), (\quad \text{ugly!})$$

that of the complete data $(\boldsymbol{x}, \boldsymbol{z})$ is

$$\begin{aligned} \ell(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{z}) &= \sum_{i=1}^{n} \log f(x_i, z_i|\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \log\left(f(x_i|z_i, \boldsymbol{\theta}) \cdot f(z_i|\boldsymbol{\theta})\right) \\ &= \sum_{i=1}^{n} (\log f_{z_i}(x_i|\boldsymbol{\vartheta}_{z_i}) + \log \pi_{z_i}). \end{aligned}$$

3.2 Maximum likelihood estimation 3.2.2 Iterative numerical methods for calculating the ML estimator

Note: the EM algorithm uses $\ell(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{z})$ to maximize $\ell(\boldsymbol{\theta}|\boldsymbol{x})$.

E-Step:

$$\begin{aligned} Q(\boldsymbol{\theta}) &= \mathbb{E}_{\theta^{(m)}}[\ \ell(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{Z})] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{J} p_{ij}^{(m)}\{\log\pi_j - \frac{1}{2}\log|\boldsymbol{\Sigma}_j| - \frac{1}{2}(x_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(x_i-\boldsymbol{\mu}_j)\} \end{aligned}$$

where

$$p_{ij}^{(m)} = \mathbb{P}(Z_i = j|x_i;\boldsymbol{\theta}^{(m)}) \overset{\text{Bayes}}{=} \frac{\pi_j^{(m)} f_j(x_i|\boldsymbol{\vartheta}_j^{(m)})}{\sum_{s=1}^{J}\pi_s^{(m)} f_s(x_i|\boldsymbol{\vartheta}_s^{(m)})}.$$

for $i = 1,\dots,n$, $j = 1,\dots,J$.

Note: can be calculated, as $\boldsymbol{\theta}^{(m)}$ is known.

3.2 Maximum likelihood estimation 3.2.2 Iterative numerical methods for calculating the ML estimator

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{J} p_{ij}^{(m)}\{\log\pi_j - \frac{1}{2}\log|\boldsymbol{\Sigma}_j| - \frac{1}{2}(x_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(x_i-\boldsymbol{\mu}_j)\}$$

M-Step: Calculate (weighted with actual prob. that x_i is in j)

$$\begin{aligned} \pi_j^{(m+1)} &= \underset{\pi_j}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}) \overset{1.}{=} \frac{1}{n}\sum_{i=1}^{n} p_{ij}^{(m)} \text{ (actual prob. to belong to j)} \\ \boldsymbol{\mu}_j^{(m+1)} &= \underset{\mu_j}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}) \overset{2.}{=} \sum_{i=1}^{n} w_{ij}^{(m)} x_i \text{ (weighted means and covariances)} \\ \boldsymbol{\Sigma}_j^{(m+1)} &= \underset{\Sigma_j}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}) \overset{3.}{=} \sum_{i=1}^{n} w_{ij}^{(m)}(x_i-\boldsymbol{\mu}_j^{(m+1)})(x_i-\boldsymbol{\mu}_j^{(m+1)})^T \end{aligned}$$

with $w_{ij}^{(m)} = \frac{p_{ij}^{(m)}}{\sum_{i=1}^{n} p_{ij}^{(m)}}$.

3.2 Maximum likelihood estimation 3.2.2 Iterative numerical methods for calculating the ML estimator

Example 3.6 (mixed models)

A derivation for E and M step for linear mixed models can be found in Pawitan, Chapter 12.8.

Addendum: EM example in R

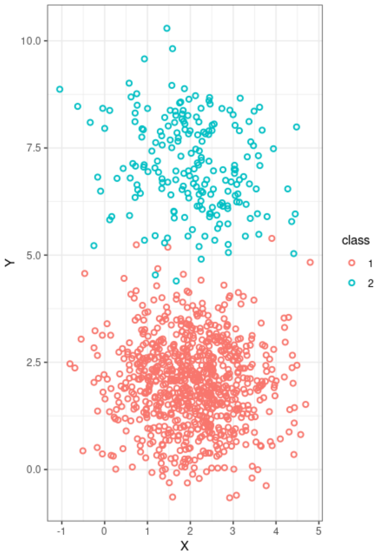
We first generate 2 clusters from a bivariate normal distribution with identity covariance matrix where 20% of points are drawn from the first cluster and 80% from the second:

```
library(mvtnorm)
set.seed(123)
n = 1000
p1 = 0.8
mu1 = c(2, 2)
Sigma = matrix(nrow = 2, ncol = 2, c(1, 0, 0, 1))
rnd1 = rmvnorm(n*p1, mu1, Sigma)
rnd1 = as.data.frame(rnd1)
rnd1$class = 1
```

```
mu2 = c(2, 7)
p2 = 0.2
rnd2 = rmvnorm(n*p2, mu2, Sigma)
rnd2 = as.data.frame(rnd2)
rnd2$class = 2
df = rbind(rnd1, rnd2)
colnames(df) = c("X", "Y", "class")
df$class = as.factor(df$class)
```

We plot the data:

```
library(ggplot2)
ggplot() +
geom_point(data = df, aes(x = X, y = Y, group = class, color = class),
pch = 21, stroke = 1) +
theme_bw()
```



EM algorithm:

```
computeJoint = function(x, priors, mean, cov) {
marginals = priors[1] * dmvnorm(x, mean[[1]], cov[[1]]) + priors[2] *
dmvnorm(x, mean[[2]], cov[[2]])
return(marginals)
}
expectationMaximization = function(iter, priors, data, mu, sigma) {
posterior1 = rep(NA, nrow(data))
posterior2 = rep(NA, nrow(data))
posterior_mat = data.frame(posterior1, posterior2)
cat("Initiate EM | Mean1:" , mu[[1]], "| Mean2:", mu[[2]], "| Sigma1:",
sigma[[1]], "| Sigma2:", sigma[[2]], "| Prior1:", priors[1], "| Prior2:",
priors[2])
for (i in 1:iter) {

# expectation step
joint = computeJoint(data, priors = priors, mean = mu, cov = sigma)
posteriors1 = (priors[1] * dmvnorm(data, mean = mu[[1]], sigma[[1]])) /
joint
posteriors2 = (priors[2] * dmvnorm(data, mean = mu[[2]], sigma[[2]])) /
joint
posterior_mat$posterior1 = posteriors1
posterior_mat$posterior2 = posteriors2

# maximization step
prior1 = mean(posteriors1)
prior2 = mean(posteriors2)
priors = c(prior1, prior2)

mu1 = (1 / (n * priors[1])) * apply(data * posteriors1, MARGIN = 2, sum)
mu2 = (1 / (n * priors[2])) * apply(data * posteriors2, MARGIN = 2, sum)
mu = list(mu1, mu2)
cat("Iteration:", i, "| Mean1:" , round(mu1, 2), "| Mean2:", round(mu2,
2), "| Sigma1:", round(sigma1, 2), "| Sigma2:", round(sigma2, 2),"|
Prior1:", round(prior1, 2) , "| Prior2:", round(prior2, 2)) }

return(list(means = mu, priors = priors, posteriors = posterior_mat,
sigma = sigma))
}
```

Test implementation:

```
mu1 = c(-20, -20)
mu2 = c(20, 20)
mu = list(mu1, mu2)
sigma1 = diag(c(1, 1))
sigma2 = sigma1
sigma = list(sigma1, sigma2)
n = nrow(df)
c = 2
priors = c(0.5, 0.5)

res = expectationMaximization(iter = 20, priors = priors, data = df[,
1:2], mu = mu, sigma)
```

Output:

```
Initiate EM | Mean1: -20 -20 | Mean2: 20 20 | Sigma1: 1 0 0 1 | Sigma2: 1
0 0 1 | Prior1: 0.5 | Prior2: 0.5

Iteration: 1 | Mean1: -0.51 0.43 | Mean2: 2.04 3.02 | Sigma1: 1 0 0 1 |
Sigma2: 1 0 0 1 | Prior1: 0 | Prior2: 1
```

```
Iteration: 2 | Mean1: 0.31 0.42 | Mean2: 2.04 3.03 | Sigma1: 1 0 0 1 |
Sigma2: 1 0 0 1 | Prior1: 0 | Prior2: 1

Iteration: 3 | Mean1: 0.86 0.36 | Mean2: 2.05 3.04 | Sigma1: 1 0 0 1 |
Sigma2: 1 0 0 1 | Prior1: 0.01 | Prior2: 0.99

Iteration: 4 | Mean1: 1.34 0.37 | Mean2: 2.06 3.09 | Sigma1: 1 0 0 1 |
Sigma2: 1 0 0 1 | Prior1: 0.02 | Prior2: 0.98

Iteration: 5 | Mean1: 1.71 0.51 | Mean2: 2.06 3.2 | Sigma1: 1 0 0 1 |
Sigma2: 1 0 0 1 | Prior1: 0.07 | Prior2: 0.93

Iteration: 6 | Mean1: 1.93 0.75 | Mean2: 2.06 3.38 | Sigma1: 1 0 0 1 |
Sigma2: 1 0 0 1 | Prior1: 0.14 | Prior2: 0.86

Iteration: 7 | Mean1: 2.03 1.02 | Mean2: 2.04 3.65 | Sigma1: 1 0 0 1 |
Sigma2: 1 0 0 1 | Prior1: 0.24 | Prior2: 0.76

Iteration: 8 | Mean1: 2.06 1.28 | Mean2: 2.02 4.04 | Sigma1: 1 0 0 1 |
Sigma2: 1 0 0 1 | Prior1: 0.37 | Prior2: 0.63

Iteration: 9 | Mean1: 2.06 1.52 | Mean2: 2.01 4.66 | Sigma1: 1 0 0 1 |
Sigma2: 1 0 0 1 | Prior1: 0.52 | Prior2: 0.48

Iteration: 10 | Mean1: 2.05 1.74 | Mean2: 2.01 5.6 | Sigma1: 1 0 0 1 |
Sigma2: 1 0 0 1 | Prior1: 0.67 | Prior2: 0.33

Iteration: 11 | Mean1: 2.05 1.9 | Mean2: 2.02 6.6 | Sigma1: 1 0 0 1 |
Sigma2: 1 0 0 1 | Prior1: 0.76 | Prior2: 0.24

Iteration: 12 | Mean1: 2.04 1.97 | Mean2: 2.03 7 | Sigma1: 1 0 0 1 |
Sigma2: 1 0 0 1 | Prior1: 0.79 | Prior2: 0.21

Iteration: 13 | Mean1: 2.04 1.99 | Mean2: 2.03 7.07 | Sigma1: 1 0 0 1 |
Sigma2: 1 0 0 1 | Prior1: 0.8 | Prior2: 0.2

Iteration: 14 | Mean1: 2.04 1.99 | Mean2: 2.03 7.08 | Sigma1: 1 0 0 1 |
Sigma2: 1 0 0 1 | Prior1: 0.8 | Prior2: 0.2

Iteration: 15 | Mean1: 2.04 1.99 | Mean2: 2.03 7.08 | Sigma1: 1 0 0 1 |
Sigma2: 1 0 0 1 | Prior1: 0.8 | Prior2: 0.2

Iteration: 16 | Mean1: 2.04 1.99 | Mean2: 2.03 7.08 | Sigma1: 1 0 0 1 |
Sigma2: 1 0 0 1 | Prior1: 0.8 | Prior2: 0.2

Iteration: 17 | Mean1: 2.04 1.99 | Mean2: 2.03 7.08 | Sigma1: 1 0 0 1 |
Sigma2: 1 0 0 1 | Prior1: 0.8 | Prior2: 0.2
```

```
Iteration: 18 | Mean1: 2.04 1.99 | Mean2: 2.03 7.08 | Sigma1: 1 0 0 1 |
Sigma2: 1 0 0 1 | Prior1: 0.8 | Prior2: 0.2

Iteration: 19 | Mean1: 2.04 1.99 | Mean2: 2.03 7.08 | Sigma1: 1 0 0 1 |
Sigma2: 1 0 0 1 | Prior1: 0.8 | Prior2: 0.2

Iteration: 20 | Mean1: 2.04 1.99 | Mean2: 2.03 7.08 | Sigma1: 1 0 0 1 |
Sigma2: 1 0 0 1 | Prior1: 0.8 | Prior2: 0.2
```

Check classification of instances:

```
res$posteriors = as.data.frame(res$posteriors)
table(apply(res$posteriors, MARGIN = 1, FUN = function(x) which.max(x)))
```

Output:

```
1 2
798 202
```

3.2.3 Asymptotic properties

3.2 Maximum likelihood estimation
3.2.3 Asymptotic properties

Theorem 3.7
Let $X_1, \dots, X_n$ i.i.d. from a density $f(x|\boldsymbol{\theta})$ which satisfies the following assumptions:

- *$f(x|\boldsymbol{\theta})$ is Fisher-regular.*
- *The information matrix $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ is positive definite inside Θ.*
- *There are functions M_{jkl} such that*

  $$\left| \frac{\partial^3}{\partial\theta_j \partial\theta_k \partial\theta_l} \log f(x|\boldsymbol{\theta}) \right| \leq M_{jkl}(x)$$

  and

  $$\mathbb{E}_{\boldsymbol{\theta}_0}[M_{jkl}(X)] < \infty$$

  for all j, k and l, where $\boldsymbol{\theta}_0$ is the true value of the parameter.

3.2 Maximum likelihood estimation
3.2.3 Asymptotic properties

Then under further, relatively weak regularity assumptions (e.g. identifiability, i.e. $f(\boldsymbol{x}|\boldsymbol{\theta}_1) \neq f(\boldsymbol{x}|\boldsymbol{\theta}_2)$ for $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$), the following holds for the MLE $\widehat{\boldsymbol{\theta}}_n$:

- The likelihood (ML) equations have with probability of 1 for $n \to \infty$ a solution $\widehat{\boldsymbol{\theta}}_n$ (i.e. $\mathbb{P}(\widehat{\boldsymbol{\theta}}_n$ exists) $\to 1$) with $\widehat{\boldsymbol{\theta}}_n \overset{\mathbb{P}}{\to} \boldsymbol{\theta}_0$; the consistent solution $\widehat{\boldsymbol{\theta}}_n$ is unique and $\mathbb{P}(\widehat{\boldsymbol{\theta}}_n$ is a (local) maximum) $\to 1$.
- $\widehat{\boldsymbol{\theta}}_n \overset{a}{\sim} N(\boldsymbol{\theta}_0, \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta}_0))$ or $\boldsymbol{\mathcal{I}}^{1/2}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \overset{d}{\to} N(\boldsymbol{0}, \boldsymbol{I}_k)$,
- $\widehat{\boldsymbol{\theta}}_n \overset{a}{\sim} N(\boldsymbol{\theta}_0, \boldsymbol{J}^{-1}(\boldsymbol{\theta}_0|\boldsymbol{X}))$ or $\boldsymbol{J}^{1/2}(\boldsymbol{\theta}_0|\boldsymbol{X})(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \overset{d}{\to} N(\boldsymbol{0}, \boldsymbol{I}_k)$,

that is, MLE's are asymptotically unbiased BAN (best asymptotically normal) estimators. Proof: see Lehmann & Casella, Theorems 3.7, 3.10, 5.1.

3.2 Maximum likelihood estimation
3.2.3 Asymptotic properties

- *Remark 1: There are also other variants of regularity assumptions possible.*
- *Remark 2: Under stronger regularity assumptions, the theorem also applies to inid (independent non identically distributed) and to dependent random variables $X_1, \dots, X_n$.*
- *Remark 3: $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)$ and $\boldsymbol{J}(\boldsymbol{\theta}_0|\boldsymbol{x})$ can also be replaced by $\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\theta}}_n)$ and $\boldsymbol{J}(\widehat{\boldsymbol{\theta}}_n|\boldsymbol{x})$, respectively. This is important to have operational and well defined estimates (at least asymptotically) for the standard errors of the estimates.*

3.3 Testing linear hypotheses and confidence intervals

3.3.1 Testing hypotheses

3.3 Testing linear hypotheses and confidence intervals
3.3.1 Testing hypotheses

Consider linear hypotheses

$$H_0 : \boldsymbol{C\theta} = \boldsymbol{d} \quad \text{vs.} \quad H_1 : \boldsymbol{C\theta} \neq \boldsymbol{d},$$

where $\boldsymbol{C}$ has full row rank $s \leq p = \dim(\boldsymbol{\theta})$ besitze. *Important special case:*

$$H_0 : \boldsymbol{\theta}_s = \boldsymbol{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\theta}_s \neq \boldsymbol{0},$$

where $\boldsymbol{\theta}_s$ an arbitrary s-dimensional subvector of $\boldsymbol{\theta}$, e.g. in a GLM, where $\boldsymbol{\beta}_s = \boldsymbol{0}$ means that the associated covariates have no influence on the mean of the response.

3.3 Testing linear hypotheses and confidence intervals
3.3.1 Testing hypotheses

Likelihood ratio Statistic

The Likelihood ratio Statistic

$$\lambda = 2\left(\ell(\widehat{\boldsymbol{\theta}}|\boldsymbol{x}) - \ell(\widetilde{\boldsymbol{\theta}}|\boldsymbol{x})\right) = 2\log\left[\frac{L(\widehat{\boldsymbol{\theta}}|\boldsymbol{x})}{L(\widetilde{\boldsymbol{\theta}}|\boldsymbol{x})}\right] \geq 0$$

compares the unrestricted maximum of the log-likelihood $\ell(\widehat{\boldsymbol{\theta}}|\boldsymbol{x})$ (i.e. $\boldsymbol{\theta} \in \Theta$) with the maximum of the log-likelihood, if the H_0 restriction holds, i.e. $\widetilde{\boldsymbol{\theta}}$ maximizes $\ell(\boldsymbol{\theta}|\boldsymbol{x})$ under the constraint $\boldsymbol{C\theta} = \boldsymbol{d}$.

The structure of an associated test is:

$$\lambda \text{ too large} \Rightarrow \text{reject } H_0 \text{ (since } L(\widehat{\boldsymbol{\theta}}|\boldsymbol{x}) \text{ is then "much" larger than } L(\widetilde{\boldsymbol{\theta}}|\boldsymbol{x})$$

Disadvantage: A numerical maximization of $\ell(\boldsymbol{\theta}|\boldsymbol{x})$ under a linear constraint is necessary to get $\widetilde{\boldsymbol{\theta}}$.

The maximization is easy if $H_0 : \boldsymbol{\theta}_s = \boldsymbol{0}$.

3.3 Testing linear hypotheses and confidence intervals
3.3.1 Testing hypotheses

Wald statistic

The Wald statistic

$$w = (\boldsymbol{C}\widehat{\boldsymbol{\theta}} - \boldsymbol{d})^\top (\boldsymbol{C}\boldsymbol{\mathcal{I}}^{-1}(\widehat{\boldsymbol{\theta}})\boldsymbol{C}^\top)^{-1}(\boldsymbol{C}\widehat{\boldsymbol{\theta}} - \boldsymbol{d}) \ ,$$

where $(\boldsymbol{C}\boldsymbol{\mathcal{I}}^{-1}(\widehat{\boldsymbol{\theta}})\boldsymbol{C}^\top)^{-1}$ is the inverse of the estimated asymptotic covariance of $\boldsymbol{C}\widehat{\boldsymbol{\theta}}$, measures the (weighted, squared) distance between the unrestricted estimate $\boldsymbol{C}\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{C\theta}$ and the hypothetical value $\boldsymbol{d}$ under H_0.

A test is constructed in such a way that

$$w \text{ too large} \Rightarrow \text{reject } H_0.$$

Advantage compared to λ: No calculation of $\widetilde{\boldsymbol{\theta}}$ is necessary (only estimation of $\widehat{\boldsymbol{\theta}}$ under H_1).

3.3 Testing linear hypotheses and confidence intervals
3.3.1 Testing hypotheses

Score (or Rao) statistic

The score statistic is

$$u = \boldsymbol{s}(\widetilde{\boldsymbol{\theta}}|\boldsymbol{x})^\top \boldsymbol{\mathcal{I}}^{-1}(\widetilde{\boldsymbol{\theta}})\boldsymbol{s}(\widetilde{\boldsymbol{\theta}}|\boldsymbol{x}),$$

where $\boldsymbol{s}(\boldsymbol{\theta}|\boldsymbol{x})$ is the score function of the full model under H_1, evaluated at $\widetilde{\boldsymbol{\theta}}$.

Idea: For $\widehat{\boldsymbol{\theta}}$, $\boldsymbol{s}(\widehat{\boldsymbol{\theta}}|\boldsymbol{x}) = \mathbf{0}$. If H_1 is correct, then $\boldsymbol{s}(\widetilde{\boldsymbol{\theta}}|\boldsymbol{x})$ is significant different from $\mathbf{0} = \boldsymbol{s}(\widehat{\boldsymbol{\theta}}|\boldsymbol{x})$, i.e.

$$u \text{ is large} \Rightarrow \text{reject } H_0.$$

The statistic thus calculates the distance $\boldsymbol{s}(\widetilde{\boldsymbol{\theta}}|\boldsymbol{x})$ from the origin, weighted with $\boldsymbol{\mathcal{I}}^{-1}(\widetilde{\boldsymbol{\theta}})$.
Remark: only the "smaller" model under H_0 must be fitted.

3.3 Testing linear hypotheses and confidence intervals
3.3.1 Testing hypotheses

Example 3.7 (Test for a subvector)
Consider

- $H_1 : \eta = \boldsymbol{x}^\top\boldsymbol{\beta} = \boldsymbol{x}_{-s}^\top\boldsymbol{\beta}_{-s} + \boldsymbol{x}_s^\top\boldsymbol{\beta}_s$ *predictor in the full GLM,*
- $H_0 : \eta_s = \boldsymbol{x}_{-s}^\top\boldsymbol{\beta}_{-s}$ *predictor in reduced GLM (after omission of covariates, $H_0 : \boldsymbol{\beta}_s = \mathbf{0}$, i.e. $\boldsymbol{C}$ selects the corresponding components $\boldsymbol{\beta}_s$ of $\boldsymbol{\beta}$.*

Let $\widetilde{\boldsymbol{\beta}}_{-s}$ maximize the log-likelihood of the reduced sub-model. With $\widetilde{\boldsymbol{\beta}}_{-s}$ and $\widehat{\boldsymbol{\beta}}$ we can determine the likelihood ratio statistic.

For the Wald statistic

$$w = \widehat{\boldsymbol{\beta}}_s^\top [\widehat{\boldsymbol{A}}_s]^{-1}\widehat{\boldsymbol{\beta}}_s,$$

where $\widehat{\boldsymbol{\beta}}_s$ are the elements of the subvector $\boldsymbol{\beta}_s$ in $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{A}}_s$ is the sub-matrix of $\widehat{\boldsymbol{A}} = \boldsymbol{\mathcal{I}}^{-1}(\widehat{\boldsymbol{\beta}})$, that these elements correspond to.

3.3 Testing linear hypotheses and confidence intervals
3.3.1 Testing hypotheses

Theorem 3.8
Under H_0 and the same regularity assumptions as in theorem 3.7:

$$\lambda,\ w,\ u\ \overset{a}{\sim}\ \chi^2(s)\ ,$$

i.e. one rejects H_0, if $\lambda, w, u > \chi^2_{1-\alpha}(s)$.

- *Remark 1: For finite samples λ, w and u may have different values; λ is then often considered to be more reliable.*
- *Remark 2: $s = rank(\boldsymbol{C})$.*

3.3 Testing linear hypotheses and confidence intervals
3.3.1 Testing hypotheses

- *Proof for w:* It holds

  $$\widehat{\boldsymbol{\theta}} \overset{a}{\sim} N(\boldsymbol{\theta}, \boldsymbol{\mathcal{I}}^{-1}(\widehat{\boldsymbol{\theta}}))$$

  and thus

  $$\boldsymbol{C}\widehat{\boldsymbol{\theta}} \overset{a}{\sim} N(\boldsymbol{C}\boldsymbol{\theta}, \boldsymbol{C}\boldsymbol{\mathcal{I}}^{-1}(\widehat{\boldsymbol{\theta}})\boldsymbol{C}^\top).$$

  Under H_0 follows

  $$\boldsymbol{C}\widehat{\boldsymbol{\theta}} - \underbrace{\boldsymbol{C}\boldsymbol{\theta}}_{\boldsymbol{d}} \overset{a}{\sim} N(\mathbf{0}, \underbrace{\boldsymbol{C}\boldsymbol{\mathcal{I}}^{-1}(\widehat{\boldsymbol{\theta}})\boldsymbol{C}^\top}_{\boldsymbol{A}}),$$

so that

$$\boldsymbol{A}^{-1/2}(\boldsymbol{C}\widehat{\boldsymbol{\theta}} - \boldsymbol{d}) \overset{a}{\sim} N(\boldsymbol{0}, \boldsymbol{I}_s)$$

and thus (dimension of $\boldsymbol{I}_s$ is s)

$$w = (\boldsymbol{C}\widehat{\boldsymbol{\theta}} - \boldsymbol{d})^T \boldsymbol{A}^{-1}(\boldsymbol{C}\widehat{\boldsymbol{\theta}} - \boldsymbol{d}) \overset{a}{\sim} \chi^2(s).$$

3.3 Testing linear hypotheses and confidence intervals
3.3.1 Testing hypotheses

- *Proof for λ:* By a Taylor expansion it can be shown that $w \overset{a}{\sim} \lambda$ and thus $\lambda \overset{a}{\sim} \chi^2(s)$. We sketch only the special case

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{vs.} \quad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \ .$$

This corresponds to $\boldsymbol{C} = \boldsymbol{I}_p, \boldsymbol{d} = \boldsymbol{\theta}_0$, $s = \text{rank}(\boldsymbol{C}) = p = \dim(\boldsymbol{\theta})$. A second order Taylor expansion of $\ell(\boldsymbol{\theta}_0|\boldsymbol{x})$ around the unrestricted maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ gives

$$\ell(\boldsymbol{\theta}_0|\boldsymbol{x}) \approx \ell(\widehat{\boldsymbol{\theta}}|\boldsymbol{x}) + \boldsymbol{s}(\widehat{\boldsymbol{\theta}}|\boldsymbol{x})^\top(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}})^\top \boldsymbol{J}(\widehat{\boldsymbol{\theta}}|\boldsymbol{x})(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}) \ .$$

Now $\boldsymbol{s}(\widehat{\boldsymbol{\theta}}|\boldsymbol{x}) = \boldsymbol{0}$ and under H_0 ($\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\theta}}) \mathrel{\widehat{=}} \boldsymbol{A}^{-1}$)

$$\begin{aligned}\lambda = 2\left(\ell(\widehat{\boldsymbol{\theta}}|\boldsymbol{x}) - \ell(\boldsymbol{\theta}_0|\boldsymbol{x})\right) &\approx (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top \boldsymbol{J}(\widehat{\boldsymbol{\theta}}|\boldsymbol{x})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ &\approx (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top \boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = w \overset{a}{\sim} \chi^2(p).\end{aligned}$$

3.3 Testing linear hypotheses and confidence intervals
3.3.1 Testing hypotheses

- *Proof for u:* We take the same special case as in the proof for λ , i.e. $\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$. Under H_0

$$\boldsymbol{s}(\boldsymbol{\theta}_0|\boldsymbol{X}) \overset{a}{\sim} N(\boldsymbol{0}, \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)) \ .$$

$$\boldsymbol{\mathcal{I}}^{-1/2}(\boldsymbol{\theta}_0)\boldsymbol{s}(\boldsymbol{\theta}_0|\boldsymbol{X}) \overset{a}{\sim} N(\boldsymbol{0}, \boldsymbol{I}_p),$$

so

$$u = \boldsymbol{s}(\boldsymbol{\theta}_0|\boldsymbol{X})^\top \underbrace{\boldsymbol{\mathcal{I}}^{-\top/2}(\boldsymbol{\theta}_0)\boldsymbol{\mathcal{I}}^{-1/2}(\boldsymbol{\theta}_0)}_{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)^{-1}} \boldsymbol{s}(\boldsymbol{\theta}_0|\boldsymbol{X}) \overset{a}{\sim} \chi^2(p).$$

□

3.3.2 Confidence Intervals

3.3 Testing linear hypotheses and confidence intervals
3.3.2 Confidence Intervals

- Joint Confidence Intervals: ($\widehat{=} w$ in the special case)

$$(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \overset{a}{\sim} \chi^2(p)$$

$$\Rightarrow \mathbb{P}_\theta\left((\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \le \chi^2_{1-\alpha}(p)\right) \overset{a}{\approx} 1 - \alpha.$$

An $(1 - \alpha)$-confidence ellipsoid can be constructed from this expression.

3.3 Testing linear hypotheses and confidence intervals
3.3.2 Confidence Intervals

- Component wise confidence intervals for θ_j, $j = 1, \ldots, p$:

  $$\frac{\widehat{\theta}_j - \theta_j}{\widehat{\sigma}_j} \overset{a}{\sim} N(0,1),$$

  where $\widehat{\sigma}_j^2$ is the j-th diagonal element of $\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\mathcal{I}}^{-1}(\widehat{\boldsymbol{\theta}})$.

  The corresponding approximate $(1-\alpha)$-confidence interval is

  $$\widehat{\theta}_j \ \pm \ z_{1-\alpha/2}\, \widehat{\sigma}_j \ .$$

  This can be derived from the probability expression

  $$\mathbb{P}_\theta(-z_{1-\alpha/2} \leq \frac{\theta_j - \widehat{\theta}_j}{\widehat{\sigma}_j} \leq z_{1-\alpha/2}) \overset{a}{\approx} 1 - \alpha \ .$$

3.3.3 Model selection

3.3 Testing linear hypotheses and confidence intervals 3.3.3 Model selection

To compare two (or more) models, i.e. different distributional assumptions, existing model selection criteria can sometimes not be used, e.g. transformations of the data may change the distribution family, which results in a non-nested situation. A likelihood ratio test is not valid in such a case.
The goodness-of-fit, as measured by $\ell(\widehat{\boldsymbol{\theta}}|\boldsymbol{x})$, and the model complexity $p = \dim(\boldsymbol{\theta})$ can be combined to a penalized criterion $-2\ell(\widehat{\boldsymbol{\theta}}|\boldsymbol{x}) + \mathrm{pen}(p)$. Here $-2\ell(\widehat{\boldsymbol{\theta}}|\boldsymbol{x})$ is small if the model fits well to the data, while $\mathrm{pen}(p)$ is constructed to be large for models with a large number of parameters.

- *Remark 1: Models with more parameters have usually a better fit to the data and therefore a higher likelihood. This may lead to a large overfit of the models to the data.*
- *Remark 2: Software packages often ignore the constants in the output of the likelihood. This may lead to wrong conclusions, when, e.g., different distribution families should be compared.*

3.3 Testing linear hypotheses and confidence intervals 3.3.3 Model selection

A well-known likelihood based criterion is *Akaikes Information criteria*

$$\mathrm{AIC} = -2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{x}) + 2p$$

with $\mathrm{pen}(p) = 2p$.

Motivation: $\{f_\theta(\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ parametrizes the considered family of models and $g(\boldsymbol{x})$ is the true density of $\boldsymbol{X}$.
Goal: Minimize the Kullback-Leibler distance

$$D(g, f_\theta) = \mathbb{E}_Z \left[\log \frac{g(\boldsymbol{Z})}{f(\boldsymbol{Z}|\boldsymbol{\theta})}\right] \geq 0 \ ,$$

or

$$\mathbb{E}_X[D(g, f_{\hat{\boldsymbol{\theta}}(\boldsymbol{X})})] = \mathbb{E}_X\mathbb{E}_Z[\log g(\boldsymbol{Z}) - \log f(\boldsymbol{Z}|\hat{\boldsymbol{\theta}}(\boldsymbol{X}))] \ ,$$

where $\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ is the MLE of $\boldsymbol{\theta}$ based on sample $\boldsymbol{X}$ and $\boldsymbol{X}, \boldsymbol{Z} \overset{i.i.d}{\sim} g$.

Remark: Expectations are taken with respect to the true density g, $\mathbb{E}_X\mathbb{E}_Z[\log g(\boldsymbol{Z})]$ does not depend on the data $\boldsymbol{X}$ and can be dropped.
The Akaike information (without constants) $\mathbb{E}_X\mathbb{E}_Z[-\log f(\boldsymbol{Z}|\hat{\boldsymbol{\theta}}(\boldsymbol{X}))]$ is a predictive measure of two independent realizations $\boldsymbol{X}$ and $\boldsymbol{Z}$ from g.
The negative maximized log-likelihood $(-\log f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}(\boldsymbol{x})))$ is a biased estimate of $\mathbb{E}_X\mathbb{E}_Z[-\log f(\boldsymbol{Z}|\hat{\boldsymbol{\theta}}(\boldsymbol{X}))]$ since $\boldsymbol{X}$ is doubly used (only $\boldsymbol{X} = \boldsymbol{x}$ is observed, $\boldsymbol{Z}$ is unobserved). Therefore an over-optimistic adaption to the data happens.
It can be shown that under regularity conditions as in theorem 3.7 and if g is in the model family $\{f_{\boldsymbol{\theta}}(\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$, i.e. a $\tilde{\boldsymbol{\theta}}$ exists such that $g(\boldsymbol{x}) = f(\boldsymbol{x}|\tilde{\boldsymbol{\theta}})$, the expression $2p$ corrects (asymptotically) the bias.
An alternative is the *Schwartz- (Bayes-) information criteria*

3.3 Testing linear hypotheses and confidence intervals
3.3.3 Model selection

$$\mathrm{BIC} = -2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{x}) + p\log n,$$

where n is the sample size. For $n \geq 8$, the BIC "punishes" model complexity more than the AIC.

It can be shown that the model choice based on the BIC is asymptotically equivalent to the model choice based on so-called Bayes factors, see Held, Chapter 7.2, for a derivation. The Bayesian factors compare the posterior model probabilities with the prior model probabilities.

3.3 Testing linear hypotheses and confidence intervals
3.3.3 Model selection

Inference by model choice

In practice, model selection is often the first step based on an information criteria such as AIC, e.g. covariates are selected, transformations are processed.
Inference is afterwards performed in the second step (estimating, testing, confidence intervals).

When model selection and inference are carried out on separate data sets (e.g. data record splitting), this leads to more valid inference in the second step.

When model selection and inference are carried out on the same data set, tests and confidence intervals based on the final model may no longer have the nominal properties. Confidence intervals e.g. are usually too narrow since they do not reflect the (random) selection process.

3.3 Testing linear hypotheses and confidence intervals
3.3.3 Model selection

Simple example: Consider the maximal model

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2),\ i = 1, \ldots, n.$$

1. Test in the full model e.g. $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. Then the p-value under H_0 has a uniform distribution on $[0, 1]$ and the test complies with the nominal probability α for a type I error.

3.3 Testing linear hypotheses and confidence intervals
3.3.3 Model selection

2. Hypothesis dependent on the selection: Select model $M_0 : y_i = \beta_0 + \varepsilon_i$ or $M_1 : y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i$ or $M_2 : y_i = \beta_0 + \beta_2 x_2 + \varepsilon_i$ or $M_3 : y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$

   with the smallest AIC. Then test in the selected model:

   $H_0 : \beta_1 = 0$ in M_1 resp. $H_0 : \beta_2 = 0$ in M_2 resp. $H_{0,1} : \beta_1 = 0$ und $H_{0,2} : \beta_2 = 0$ in M_3.

   In the selected model, the p-values do not follow an $\mathrm{Unif}[0, 1]$.

   The type I errors are larger than the nominal α.

A valid inference must taken into account, that *the question was posed*, e.g. that the tested variable was selected into the model, since otherwise it will not be tested. (see e.g. Fithian, Sun, Taylor, 2015, Optimal Inference After Model Selection, https://arxiv.org/pdf/1410.2597v2.pdf).

3.3 Testing linear hypotheses and confidence intervals
3.3.3 Model selection

Intuitively: use the information in the data already used for model selection not again for inference. Otherwise this leads to over-optimism (p-values that are too small).

3.3 Testing linear hypotheses and confidence intervals
3.3.3 Model selection

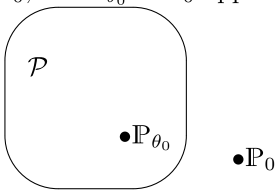
Selective inference / Post selection inference

- Idea: Instead of splitting the data, split the *information in the data* into the one used for the selection and the other used for inference.
- Control $P_{M,H_0}(\text{reject } H_0|\text{ select } M, H_0) \leq \alpha$, i.e. condition the inference on the selection of the model . → characterize the selection as condition $\mathbb{1}_M(\boldsymbol{y})$ and derive the conditional distribution of the test statistic $T(\boldsymbol{y})|\mathbb{1}_M(\boldsymbol{y})$.
  - *Remark:* Data splitting: $f(\boldsymbol{y}) = f(\boldsymbol{y}_2|\boldsymbol{y}_1)f(\boldsymbol{y}_1) = f(\boldsymbol{y}_2)f(\boldsymbol{y}_1)$ if $\boldsymbol{y}_1, \boldsymbol{y}_2$ are independent. Here: $f(\boldsymbol{y}) = f(\boldsymbol{y}|\mathbb{1}_M(\boldsymbol{y}))f(\mathbb{1}_M(\boldsymbol{y}))$.
- R-Package `selectiveInference`: valid inference after model selection with Lasso, LARS or Forward Stepwise Regression.

3.4 Misspecification, quasi-likelihood and estimating equations

3.4 Misspecification, quasi-likelihood and estimating equations

Previously: full(*genuine*) likelihood-inference given a parametric statistical model, that is, a family of distributions or densities with parameter $\boldsymbol{\theta} \in \Theta$.
Previous basic assumption: There is a "true" $\boldsymbol{\theta}_0 \in \Theta$ such that $\mathbb{P}_{\theta_0}$ is the distribution of the data generating process $\mathbb{P}_0$, i.e. $\mathbb{P}_{\theta_0} = \mathbb{P}_0$ applies.



3.4 Misspecification, quasi-likelihood and estimating equations

Questions:

3.4.1 What happens with likelihood inference in $\mathcal{P}$, if the data generating process $\mathbb{P}_0 \notin \mathcal{P}$ *(incorrect specification)*?

3.4.2 What happens if the distribution type is incorrectly specified, however, the expected value is correctly specified *(quasi-likelihood)*?

3.4.2 Can the likelihood be dropped and can one instead start directly from the quasi-maximum likelihood estimation equations?

$$\mathbf{qs}(\widehat{\boldsymbol{\theta}}|\boldsymbol{X}) \overset{!}{=} \mathbf{0}$$

I.e., is there something like a generalized score equation $\mathbf{qs}(\boldsymbol{\theta}|\boldsymbol{X})$?

3.4 Misspecification, quasi-likelihood and estimating equations

Example 3.8 (Linear Model) We look again at the standard assumption

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d.}{\sim} N(0,\sigma^2)$$

or

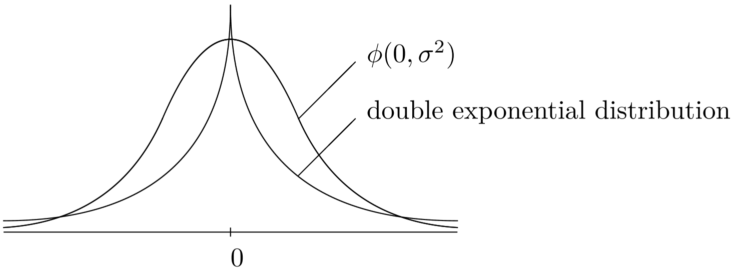
$$\boldsymbol{Y}|\boldsymbol{X} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n), \quad \boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2),$$

i.e $\mathcal{P} = \{\phi(\boldsymbol{Y}|\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n) : (\boldsymbol{\beta}, \sigma^2) \in \mathbb{R}^p \times (0,\infty)\}$.
Possible incorrect specifications:

3.4 Misspecification, quasi-likelihood and estimating equations

(a) $N(0,\sigma^2)$-assumption for ε_i is wrong, e.g. the *true* distribution could be the double exponential distribution (Laplace distribution) with heavy-tails which is therefore less sensitive to outliers:

$$f(\varepsilon_i) \propto \exp\bigl(-|\varepsilon_i/\sigma|\bigr).$$



3.4 Misspecification, quasi-likelihood and estimating equations

(b) The covariance structure is wrong, i.e. $\text{Cov}(\boldsymbol{Y}) \neq \sigma^2\boldsymbol{I}_n$. Assume that the true covariance structure is e.g. $\text{Cov}(\boldsymbol{Y}) = \sigma^2\boldsymbol{W}$.

  - $\boldsymbol{W} = \text{diag}(W_1, \dots, W_n)$ (heteroscedastic errors) or
  - $\boldsymbol{W}$ with off-diagonal elements (correlated errors).

(c) The expected value structure is wrong: $\mathbb{E}[\boldsymbol{Y}] \neq \boldsymbol{X}\boldsymbol{\beta}$, because

  - nonlinear effects are missing but necessary, e.g. $x\beta_1 + x^2\beta_2$ or $\beta \log x$,
  - lack of important regressors.

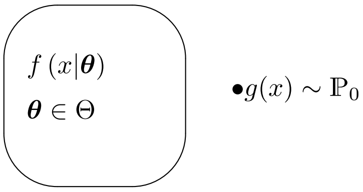
3.4.1 ML estimation in the case of misspecification

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.1 ML estimation in the case of misspecification

We restrict ourselves to the iid case: Let $X_1, \dots, X_n$ i.i.d. $X \sim g(x)$ and $g(x)$ the true density. As a statistical model let us consider the family of densities

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

If a $\boldsymbol{\theta}_0 \in \Theta$ exists with $g(x) \equiv f(x|\boldsymbol{\theta}_0)$, then the model is correctly specified. If there is no $\boldsymbol{\theta}_0 \in \Theta$ with $g(x) \equiv f(x|\boldsymbol{\theta}_0)$, then the model is incorrectly specified. Idea: search a $\boldsymbol{\theta}_0 \in \Theta$, such that $f_{\boldsymbol{\theta}_0}$ is "next to" g.


$f\,(x|\boldsymbol{\theta})$
$\boldsymbol{\theta} \in \Theta$
$\bullet g(x) \sim \mathbb{P}_0$

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.1 ML estimation in the case of misspecification

Definition 3.9 (Kullback-Leibler-Distance)
The Kullback-Leibler-Distance *of g and $f_{\boldsymbol{\theta}}$ is defined by*

$$D(g, f_{\boldsymbol{\theta}}) = \mathbb{E}_g \left[\log \frac{g(X)}{f(X|\boldsymbol{\theta})} \right],$$

i.e. (note: $D(g, f_{\boldsymbol{\theta}}) \neq D(f_{\boldsymbol{\theta}}, g)$)

$$D(g, f_{\boldsymbol{\theta}}) = \int \log \frac{g(x)}{f(x|\boldsymbol{\theta})}\, g(x)\, dx$$

for X continuous. The expected value $\mathbb{E}_g$ is calculated with respect to the "true" density or probability function $g(x)$. Note, that X is here a scalar random variable.

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.1 ML estimation in the case of misspecification

The following applies:

$$D(g, f_{\boldsymbol{\theta}}) \geq 0$$

with

$$D(g, f_{\boldsymbol{\theta}_0}) = 0 \quad \Leftrightarrow \quad g \equiv f_{\boldsymbol{\theta}_0}.$$

Therefore:

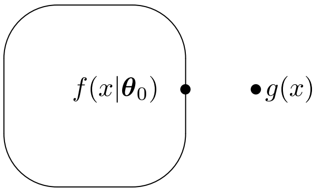
$$D(g, f_{\boldsymbol{\theta}_0}) = 0 \text{ for some } \boldsymbol{\theta}_0 \quad \Leftrightarrow \quad \text{Model specified correctly}$$

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.1 ML estimation in the case of misspecification

Proof with Jensen's inequality.
Let $\boldsymbol{\theta}_0$ the minimizer of the Kullback-Leibler distance:

$$\begin{aligned} \boldsymbol{\theta}_0 &= \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \Big(\mathbb{E}_g[\log g(X)] - \mathbb{E}_g[\log f(X|\boldsymbol{\theta})] \Big) \\ &= \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\, \mathbb{E}_g \Big[\log f(X|\boldsymbol{\theta}) \Big]. \end{aligned}$$

The density $f_{\boldsymbol{\theta}_0}$ is then next to g in the sense of the Kullback-Leibler distance.



3.4 Misspecification, quasi-likelihood and estimating equations
3.4.1 ML estimation in the case of misspecification

The law of large numbers provides the conjecture

$$\begin{tikzcd}
\frac{1}{n}\sum_{i=1}^{n}\log\, f(X_i|\boldsymbol{\theta}) \arrow[r, "\text{argmax}"] \arrow[d, "\mathbb{P}"'] & \hat{\boldsymbol{\theta}}_n \arrow[d, "\mathbb{P}?"] \\
\mathbb{E}_g[\,\log\, f(X|\boldsymbol{\theta})\,] \arrow[r, "\text{argmax}"] & \boldsymbol{\theta}_0
\end{tikzcd}$$

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.1 ML estimation in the case of misspecification

Theorem 3.10 (Asymptotic properties of the ML estimator in the case of misspecification)

Let $X_1, ..., X_n$ iid. from X. With regularity assumptions similar as in Theorem 3.7, it holds:

1. *Consistency: Let $\boldsymbol{\theta}_0$ be a (local) maximizer of*

$$\mathbb{E}_g[\,\log\, f(X|\boldsymbol{\theta})\,]$$

(or a minimizer of $D(g, f_{\boldsymbol{\theta}})$). Then there exists a sequence $\hat{\boldsymbol{\theta}}_n$ of („Quasi-") ML estimators, that are local maximizers of

$$\frac{1}{n}\sum_{i=1}^{n}\log\, f(X_i|\boldsymbol{\theta})$$

with

$$\hat{\boldsymbol{\theta}}_n \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_0.$$

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.1 ML estimation in the case of misspecification

2. Asymptotic normality:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N\left(\,\mathbf{0}, \boldsymbol{qi}^{-1}(\boldsymbol{\theta}_0)\,\mathrm{Cov}(\boldsymbol{s}_1(\boldsymbol{\theta}_0|X))\,\boldsymbol{qi}^{-1}(\boldsymbol{\theta}_0)\,\right)$$

with (X is one observation)

$$\mathrm{Cov}(\boldsymbol{s}_1(\boldsymbol{\theta}_0|X)) = \mathbb{E}_g\Big[\underbrace{\left(\frac{\partial\,\log\, f(X|\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right)}_{\boldsymbol{s}_1(\boldsymbol{\theta}_0|X)}\underbrace{\left(\frac{\partial\,\log\, f(X|\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right)^{\top}}_{\boldsymbol{s}_1(\boldsymbol{\theta}_0|X)^{\top}}\Big]$$

and the (Quasi-) Fisher-Information

$$\boldsymbol{qi}(\boldsymbol{\theta}) = \mathbb{E}_g\left[-\frac{\partial^2\log\, f(X|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^{\top}}\right].$$

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.1 ML estimation in the case of misspecification

Remarks

- Proof and regularity conditions: Pawitan p. 372 ff

- If the model specification is correct, i.e. $g(x) \equiv f(x|\boldsymbol{\theta}_0)$,

$$\mathrm{Cov}(\boldsymbol{s}_1(\boldsymbol{\theta}_0|X)) = \boldsymbol{qi}(\boldsymbol{\theta}_0) = \boldsymbol{i}(\boldsymbol{\theta}_0)$$

  and one obtains the usual asymptotic normal distribution and asymptotic unbiasedness of the ML estimator.

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.1 ML estimation in the case of misspecification

- Informally, one gets

$$\hat{\boldsymbol{\theta}}_n \overset{a}{\sim} N\Big(\boldsymbol{\theta}_0, \underbrace{\frac{1}{n}\boldsymbol{qi}^{-1}(\boldsymbol{\theta}_0)\,\mathrm{Cov}(\boldsymbol{s}_1(\boldsymbol{\theta}_0|X))\,\boldsymbol{qi}^{-1}(\boldsymbol{\theta}_0)}_{\boldsymbol{V}(\boldsymbol{\theta}_0)}\Big),$$

and $\boldsymbol{V}(\boldsymbol{\theta}_0)$ is estimated by

$$\widehat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}}_n) = \boldsymbol{J}^{-1}(\hat{\boldsymbol{\theta}}_n|\boldsymbol{x})\,\boldsymbol{C}(\hat{\boldsymbol{\theta}}_n|\boldsymbol{x})\,\boldsymbol{J}^{-1}(\hat{\boldsymbol{\theta}}_n|\boldsymbol{x}) \qquad \text{(„Sandwich"-Matrix)}$$

with

$$\begin{aligned} \boldsymbol{C}(\hat{\boldsymbol{\theta}}_n|\boldsymbol{x}) &= \sum_{i=1}^{n} \boldsymbol{s}_1(\hat{\boldsymbol{\theta}}_n|x_i)\,\boldsymbol{s}_1^{\top}(\hat{\boldsymbol{\theta}}_n|x_i) && \begin{array}{l}n\text{-times empirical covariance}\\ \text{matrix of } \boldsymbol{s}_1(\boldsymbol{\theta}_0|X)\end{array} \\ \boldsymbol{J}(\hat{\boldsymbol{\theta}}_n|\boldsymbol{x}) &= -\sum_{i=1}^{n} \underbrace{\frac{\partial^2 \log\, f(x_i|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^{\top}}}_{\frac{\partial^2\,\ell(\boldsymbol{\theta}|x_i)}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^{\top}}}\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} && \begin{array}{l}\text{empirical observed information}\\ \text{matrix of the sample}\end{array} \end{aligned}$$

Remarks

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.1 ML estimation in the case of misspecification

1. There are extensions to the i.i.d case with $f_i(x_i|\boldsymbol{\theta})$ in place of $f(x_i|\boldsymbol{\theta})$.
2. With two parameter blocks $\widetilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}, \boldsymbol{\alpha})^{\top}$, the score function is

$$\boldsymbol{s}(\boldsymbol{\theta}, \boldsymbol{\alpha}|\boldsymbol{x}) = \begin{pmatrix} s_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\alpha}|\boldsymbol{x}) \\ s_{\boldsymbol{\alpha}}(\boldsymbol{\theta}, \boldsymbol{\alpha}|\boldsymbol{x}) \end{pmatrix} = \begin{pmatrix} s_{\boldsymbol{\theta}}(\widetilde{\boldsymbol{\theta}}|\boldsymbol{x}) \\ s_{\boldsymbol{\alpha}}(\widetilde{\boldsymbol{\theta}}|\boldsymbol{x}) \end{pmatrix}.$$

If only the parameter $\boldsymbol{\theta}$ is of interest, and if $\mathbb{E}_g[\boldsymbol{s}_{\boldsymbol{\theta}}(\widetilde{\boldsymbol{\theta}}_0|\boldsymbol{X})] = \boldsymbol{0}$ is fulfilled (although the likelihood is misspecified), then

$$\hat{\boldsymbol{\theta}}_n \overset{a}{\sim} N\Big(\boldsymbol{\theta}_0, \hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}}_n)\Big) \quad \Rightarrow \quad \text{Quasi-Likelihood.}$$

3.4.2 Quasi-Likelihood und Estimating Equations

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.2 Quasi-Likelihood und Estimating Equations

Question: Can we estimate parameters of interest – such as the mean μ in the iid case or the covariate vector $\boldsymbol{\beta}$ in the regression case – with consistent and asymptotically normally distributed estimators, if the statistical model is only partially incorrectly specified or incompletely specified?
Example 3.9 Let $Y_1, \ldots, Y_n$ i.i.d. as $Y \sim f(y|\mu, \sigma^2)$, f symmetric about μ, but not normal, e.g.

$$\mathbb{P}_0 = \left\{ f(y|\mu_0) = \frac{1}{2\sigma}\, e^{-|y-\mu_0|/\sigma} \right\} \text{(Laplace- or double exponential).}$$

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.2 Quasi-Likelihood und Estimating Equations

Idea: maximize the log-likelihood

$$\mathrm{ql}(\mu|\boldsymbol{y}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 + \text{const}$$

i.e. use the normal log-likelihood as a quasi-log-likelihood.

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.2 Quasi-Likelihood und Estimating Equations

This leads to the quasi-score function

$$\mathrm{qs}(\mu|\boldsymbol{y}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu).$$

It is

$$\frac{\partial}{\partial\mu}\mathbb{E}_0[\log f(\boldsymbol{Y}|\mu_0)] = \mathbb{E}_0[\,\mathrm{qs}(\mu_0|\boldsymbol{Y})\,] = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\underbrace{\mathbb{E}_0[Y_i]}_{=\mu_0} - \mu_0) = 0,$$

so the minimizer of the Kullback-Leibler distance is equal to that true μ_0. Setting the quasi-score function to zero, the QML estimator is $\hat{\mu}_{\mathrm{QML}} = \bar{y}$. Since $\mathbb{E}_0[\bar{Y}] = \mu_0$, the estimator is unbiased, and because of Theorem 3.10, consistent and asymptotic normally distributed. However, $\bar{Y}$ is no longer an (asymptotically) efficient estimator (the Rao-Cramer limit is not reached).

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.2 Quasi-Likelihood und Estimating Equations

Example 3.10 (Linear model)
Standard–assumption:

$$y_i|\boldsymbol{x}_i \sim N(\boldsymbol{x}_i^\top\boldsymbol{\beta}, \sigma^2)$$

bzw.

$$\boldsymbol{Y}|\boldsymbol{X} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}_n)\ .$$

Possible incorrect specifications:

(a) assumption of normal distribution is wrong,

(b) assumed covariance structure $\mathrm{Cov}(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}_n$ is false,

(c) assumed expected value, $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{X\beta}$, is incorrect.

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.2 Quasi-Likelihood und Estimating Equations

(a): $\boldsymbol{Y}$ is not normally distributed, but the covariance structure and the expected value are correctly specified. There is a $\boldsymbol{\beta}_0$, such that $\mathbb{E}_0[\boldsymbol{Y}] = \boldsymbol{X\beta}_0$ is the correct expectation.

$$\mathbf{qs}(\boldsymbol{\beta}) = \frac{1}{\sigma^2}\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X\beta})$$

$$\mathbb{E}_0[\mathbf{qs}(\boldsymbol{\beta}_0)] = \mathbf{0}.$$

$\mathbb{E}_0[\mathbf{qs}(\boldsymbol{\beta}_0)]$ is the expected value with respect to the true distribution $\mathbb{P}_0$ of $\boldsymbol{Y}$.

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.2 Quasi-Likelihood und Estimating Equations

Therefore

$$\hat{\boldsymbol{\beta}}_{\mathrm{QML}} = \hat{\boldsymbol{\beta}}_{\mathrm{KQ}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$$

with

$$\begin{aligned}\mathbb{E}_0(\hat{\boldsymbol{\beta}}_{\mathrm{QML}}) &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbb{E}_0[\boldsymbol{Y}] = \boldsymbol{\beta}_0 \quad \text{(unbiased)},\\ \mathrm{Cov}_0(\hat{\boldsymbol{\beta}}_{\mathrm{QML}}) &= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1},\end{aligned}$$

thus

$$\hat{\boldsymbol{\beta}}_{\mathrm{QML}} \overset{a}{\sim} N(\boldsymbol{\beta}_0, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1})$$

as under the normal assumption.

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.2 Quasi-Likelihood und Estimating Equations

(b): The true covariance matrix is $\sigma^2\boldsymbol{W}$ instead of $\sigma^2\boldsymbol{I}_n$:

$\mathbb{P}_0 : \boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}_0, \sigma^2\boldsymbol{W})$, but $\mathbf{qs}(\boldsymbol{\beta})$ and $\hat{\boldsymbol{\beta}}_{\mathrm{QML}}$ as before.

$$\begin{array}{rcl}
\mathbb{E}_0[\mathbf{qs}(\boldsymbol{\beta}_0)] & = & \mathbf{0} \\
\mathbb{E}_0[\hat{\boldsymbol{\beta}}_{\mathrm{QML}}] & = & (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0 \\
\mathrm{Cov}_0(\hat{\boldsymbol{\beta}}_{\mathrm{QML}}) & = & (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathrm{Cov}_0(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\
& = & \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\
(& \neq & \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\ )
\end{array}$$

$\hat{\boldsymbol{\beta}}_{\mathrm{QML}}$ is consistent, but not efficient. An efficient estimator would be the weighted LS or Aitken estimator $\hat{\boldsymbol{\beta}}_{\mathrm{AITKEN}} = (\boldsymbol{X}^\top\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{W}^{-1}\boldsymbol{Y}$. $\mathrm{Cov}_0(\hat{\boldsymbol{\beta}}_{\mathrm{AITKEN}}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{W}^{-1}\boldsymbol{X})^{-1} = \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\beta}_0)$ under $\mathbb{P}_0$.

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.2 Quasi-Likelihood und Estimating Equations

(c): The true expected value is not equal to $\boldsymbol{X}\boldsymbol{\beta}_{\mathbf{0}}$:

$$\begin{array}{rl}
\text{true expected value:} & \mathbb{E}_0[\boldsymbol{Y}] = \boldsymbol{\mu}_0 = \boldsymbol{X}_0\boldsymbol{\beta}_0 \\
\Rightarrow \text{true model:} & \boldsymbol{Y} \sim N(\boldsymbol{X}_0\boldsymbol{\beta}_0, \sigma^2\boldsymbol{I}_n)
\end{array}$$

(if normal assumption and $\mathrm{Cov}_0(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}_n$ is correct). Then

$$\hat{\boldsymbol{\beta}}_{QML} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$$

$$\mathbb{E}_0[\hat{\boldsymbol{\beta}}_{QML}] = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}_0\boldsymbol{\beta}_0 \neq \boldsymbol{\beta}_0\ .$$

Thus $\hat{\boldsymbol{\beta}}_{QML}$ is biased, but provides the best approximating linear model with design matrix $\boldsymbol{X}$. The covariance matrix is then given by:

$$\begin{array}{rcl}
\mathrm{Cov}_0(\hat{\boldsymbol{\beta}}_{QML}) & = & (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\underbrace{\mathrm{Cov}_0(\boldsymbol{Y})}_{\sigma^2\boldsymbol{I}_n}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\
& = & \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}.
\end{array}$$

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.2 Quasi-Likelihood und Estimating Equations

Conclusion from the examples:

- If the likelihood or the variance structure is incorrectly specified, but the expected value

  $$\mathbb{E}_0[Y_i] = \mu_i = \boldsymbol{x}_i^\top\boldsymbol{\beta}$$

  is correctly specified, one obtains consistent estimators for μ or $\boldsymbol{\beta}$.

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.2 Quasi-Likelihood und Estimating Equations

- It is sufficient to find the root of the quasi-score function

  $$\mathrm{qs}(\hat{\mu}|\boldsymbol{y}) \overset{!}{=} 0 \quad \text{resp.} \quad \boldsymbol{qs}(\hat{\boldsymbol{\beta}}) \overset{!}{=} \mathbf{0}$$

  If for the "true" μ_0 or $\boldsymbol{\beta}_0$

  $$\mathbb{E}_0[\mathrm{qs}(\mu_0|\boldsymbol{Y})] = 0 \quad \text{resp.} \quad \mathbb{E}_0[\boldsymbol{qs}(\boldsymbol{\beta}_0)] = \mathbf{0}$$

  holds, then the roots $\hat{\mu}$, resp. $\hat{\boldsymbol{\beta}}$, are consistent and asymptotically normally distributed estimators for μ_0 or $\boldsymbol{\beta}_0$.

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.2 Quasi-Likelihood und Estimating Equations

$\Rightarrow$ Idea of estimating equations:
Define an *estimating function* or *quasi-score function*

$$\mathbf{qs}(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{n} \psi_i(y_i, \boldsymbol{\theta})$$

such that for the "true" parameter $\boldsymbol{\theta}_0$

$$\mathbb{E}_0[\mathbf{qs}(\boldsymbol{\theta}_0|\boldsymbol{y})] = \sum_{i=1}^{n} \mathbb{E}_0[\psi_i(y_i, \boldsymbol{\theta}_0)] = \mathbf{0}$$

holds. Then the *Quasi-ML-estimator* or *"M–estimator"*, defined as root

$$\mathbf{qs}(\hat{\boldsymbol{\theta}}_{QML}|\boldsymbol{y}) \overset{!}{=} \mathbf{0} \quad (\textit{estimation equation})$$

of the estimating function $\mathbf{qs}(\boldsymbol{\theta}|\boldsymbol{y})$, is consistent and asymptotic normally distributed.

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.2 Quasi-Likelihood und Estimating Equations

Example 3.11 (generalized regression) Let

$$\begin{aligned} \mathbb{E}_0[Y_i] = \mu_i(\boldsymbol{\beta}) & \qquad \text{corretly specified,} \\ \mathrm{Var}_0(Y_i) = \phi\, v_i(\boldsymbol{\beta}) & \qquad \text{(possibly) incorrectly specified.} \end{aligned}$$

One makes an assumption regarding the estimating equation, but not for the distribution of Y_i.

3.4 Misspecification, quasi-likelihood and estimating equations
3.4.2 Quasi-Likelihood und Estimating Equations

$$\begin{aligned} \mathbf{qs}(\boldsymbol{\beta}) &= \frac{1}{\phi} \sum_{i=1}^{n} \left(\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right) v_i(\boldsymbol{\beta})^{-1} \underbrace{(y_i - \mu_i(\boldsymbol{\beta}))}_{\mathbb{E}_0(y_i) - \mu_i(\boldsymbol{\beta}) = 0} \\ &\propto \sum_{i=1}^{n} \left(\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right) v_i(\boldsymbol{\beta})^{-1} (y_i - \mu_i(\boldsymbol{\beta})). \end{aligned}$$

Therefore $\mathbb{E}_0[\mathbf{qs}(\boldsymbol{\beta})] = \mathbf{0}$ and

$$\mathbf{qs}(\hat{\boldsymbol{\beta}}) \overset{!}{=} \mathbf{0}\ .$$

$\Rightarrow$ $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normally distributed.
An extension leads to *"generalized estimating equations" (GEE)* (as in a GLM: $\mu_i(\boldsymbol{\beta}) = h(\boldsymbol{x}_i^\top \boldsymbol{\beta})$).

Chapter 4

Bayesian Inference

Bayes inference

Objectives of Chapter 4: Introduction to Bayes inference

4.1 Overview

4.1 Overview

- *"Definition" Bayesian Inference:* Fit a probabilistic model to data.
- *Result:* Probability distribution for the *parameters of the model* and other unobserved quantities, for example predictions for new observations, missing data.
- The chapter is mainly based on the book of Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, Donald B. Rubin, *Bayesian Data Analysis*, Third Edition. If a direct reference or citation is made, the abbreviation **(GCSDVR)** is used. See also `http://www.stat.columbia.edu/~gelman/book/`

4.1 Overview

Idealized Process of Bayesian Data Analysis (GCSDVR, page 3):

1. Establish a *full* probability model or a joint probability distribution for all observable and unobservable quantities. For this, knowledge of the underlying scientific problem and the data generating process is helpful.
2. Calculation of the posterior distribution of the unobservable quantities (parameters, missing data, ...): conditional probability distribution of the unobservable quantities given the observed data.
3. Model diagnosis: fit, sensitivity (with regard to the assumptions in 1.).

Result: "coherent system", which allows the (coherent) calculation of any interesting probability concerning the unobservable quantities.
Examples for interesting probabilities.

4.1 Overview

- $P(\theta_j \in [0.2, 0.5]$: probability statement for a parameter. Parameters are random variables in Bayesian inference.
- $P(x^* > 10)$: probability statement for a prediction.
- $P(\theta_j > \theta_k)$: probability statement on two parameters.
- $P(\log(\theta_j + \theta_k) > 0)$: probability statement on transformations of parameters.

4.2 Exchangeability

4.2 Exchangeability

Exchangeability („Interchangeability ") is an important concept for statistical modeling. It goes back to de Finetti.

Definition 4.1 (Finite Exchangeability)
The random variables $X_1, \ldots, X_n$ *are* exchangeable *with respect to the Probability measure* $\mathbb{P}$*, if*

$$\mathbb{P}(x_1, \ldots, x_n) = \mathbb{P}(x_{\pi(1)}, \ldots, x_{\pi(n)})$$

for all permutations

$$\pi : \{1, \ldots, n\} \ \rightarrow \ \{1, \ldots, n\}$$

applies. If there is a density f *for* $\mathbb{P}$*, then the following applies accordingly:*

$$f(x_1, \ldots, x_n) = f(x_{\pi(1)}, \ldots, x_{\pi(n)}).$$

4.2 Exchangability

Definition 4.2 (Infinite Exchangability)
The infinite sequence $X_1, X_2, \ldots$ *is* exchangable*, if each finite subsequence is exchangeable.*

Comment. Analogous to the above definitions *conditional exchangeability* can be defined, for example in the regression case for $Y_1|\boldsymbol{x}_1, \ldots, Y_n|\boldsymbol{x}_n$.

4.2 Exchangeability

Example 4.1

1. *If* $X_1, X_2, \ldots, X_n$ *are independent and identically distributed Bernoulli random variables, then they are exchangeable conditional on* θ*, where* $\theta = P(X_i = 1)$*. In that case*

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

and so the X_i*'s are exchangeable.*

2. *Let* $p(\beta_1, \beta_2, \ldots, \beta_k|\phi)$ *a prior on the regression coefficients of a (G)LM. Often, there is no reason* not *to assume an exchangeable prior, e.g. (using an obvious but sloppy notation)*

$$p(\beta_1, \beta_2, \ldots, \beta_k|\phi = (\mu_j = 0, \sigma_j^2 = 1000, j = 1, \ldots, k)) = \prod_{j=1}^{k} N(0, 1000) \ ,$$

i.e. the product of k *independent normal densities with mean 0 and variance 1000 ("uninformative" prior).* ϕ *(and its distribution!) is in general unknown, but here it is prespecified, i.e. set to fixed values.*

4.2 Exchangeability

Example 4.2 (example 4.1 cont.)

1. *Now, let the prior of regression coefficients in a (G)LM*

$$p(\beta_1, \beta_2, \ldots, \beta_k|\boldsymbol{\mu} = 0, c = 1000, \boldsymbol{\Sigma}) = MVN(\mathbf{0}, 1000\boldsymbol{\Sigma}) \ ,$$

where

$$\boldsymbol{\Sigma}_{st} = \begin{cases} 1 & \text{if } s = t \\ 0.5 & \text{if } s \neq t \ . \end{cases}$$

Also, $p(\beta_1, \beta_2, \ldots, \beta_k|\cdot)$ *is exchangeable. Note, that* $\boldsymbol{\Sigma}$ *is in fact a (uniform) correlation matrix.*

4.2 Exchangeability

Example 4.3 ((GCSDVR, page 5))
Hierarchical modeling *In Chapter 5 and subsequent chapters, we focus on hierarchical models (also called multilevel models), which are used when information is available on several different levels of observational units. In a hierarchical model, it is possible to speak of **exchangeability** at each level of units. For example, suppose two medical treatments are applied, in separate randomized experiments, to patients in several different cities. Then, if no other information were available, it would be reasonable to treat the patients within each city as exchangeable and also treat the results from different cities as themselves exchangeable. In practice it would make sense to include, as explanatory variables at the city level, whatever relevant information we have on each city, as well as the explanatory variables mentioned before at the individual level, and then the conditional distributions given these explanatory variables would be exchangeable.*

4.2 Exchangeability

Example 4.4 (Mixture of iid distributions)
Let

$$p(\theta|\phi) = \prod_{j=1}^{k} p(\theta_j|\phi)$$

and (ϕ unknown)

$$p(\theta) = \int \prod_{j=1}^{k} p(\theta_j|\phi)p(\phi)d\phi$$

*the mixture. A related theoretical result, so-called de Finetti's theorem, states that in the limit as $k \to \infty$, **any** suitably well-behaved exchangeable distribution on $(\theta_1, \dots, \theta_k)$ can be expressed as a mixture of independent and identical distributions (note: the theorem does not hold in general when k is finite). This is called* representation theorem*.*

Note **(GCSDVR), page 105**: As a simple counterexample to the above mixture model, consider the probabilities of a given die landing on each of its six faces. The probabilities $\theta_1, \dots, \theta_6$ are exchangeable, but the six parameters θ_j are constrained to sum to 1 and so cannot be modeled with a mixture of independent identical distributions; nonetheless, they can be modeled exchangeably.
Summary:

4.2 Exchangeability

- Exchangeability can be a property for the distribution of the *data*
- Exchangeability can be a property for the distribution of the *parameters*
- Exchangeability is often implicitly assumed without saying a word about it.

4.3 Bayesian Inference At a Glance

4.3 Bayesian Inference At a Glance

Notation:

- X: observed data
- $\widetilde{X}$: unobserved data
- θ: parameter

Aim:

- Probability statements conditional on observed data
- prediction and predictive inference

Basic components in Bayesian Inference:

- $p(\theta)$ prior distribution
- $f(x|\theta)$ data distribution, likelihood (often also: $p(x|\theta)$)
- $f(\theta|x)$ posterior distribution (often also: $p(\theta|x)$)
- $f(\widetilde{x}|x)$ predictive distribution (often also: $p(\widetilde{x}|x)$)

4.3 Bayesian Inference At a Glance

According to Bayes' theorem, the joint distribution of (θ, x) is equal to

$$f(\theta, x) = f(x|\theta) \cdot p(\theta),$$

therefore

$$f(\theta|x) = \frac{f(\theta, x)}{f(x)} = \frac{f(x|\theta)p(\theta)}{f(x)},$$

in which

$$\begin{aligned} f(x) &= \sum_{\theta \in \Theta} f(x|\theta)p(\theta), \quad \text{if } \theta \text{ is discrete,} \\ f(x) &= \int_{\Theta} f(x|\theta)p(\theta)\, d\theta, \quad \text{if } \theta \text{ is continuous.} \end{aligned}$$

4.3 Bayesian Inference At a Glance

Unnormalized posterior:

$$f(\theta|x) \propto f(x|\theta)p(\theta)$$

Prior predictive distribution before observing the data x:

$$f(x) = \int_{\Theta} f(\theta, x)\, d\theta = \int_{\Theta} f(x|\theta)\; p(\theta)\; d\theta$$

A posteriori predictive distribution after observing data x for a new observation $\widetilde{x}$:

$$\begin{aligned} f(\widetilde{x}|x) &= \int_{\Theta} f(\widetilde{x}, \theta|x)\, d\theta = \int_{\Theta} f(\widetilde{x}|\theta, x)\; f(\theta|x)\; d\theta \\ &= \int_{\Theta} f(\widetilde{x}|\theta)\; f(\theta|x)\; d\theta, \end{aligned}$$

since $\widetilde{x}$ is (assumed to be) conditionally independent of x given θ .

4.3 Bayesian Inference At a Glance

Likelihood and odds ratios:

The ratio of the posterior density evaluated at points θ_1 and θ_2 is called the posterior odds of θ_1 compared to θ_2:

$$\frac{f_{\theta|x}(\theta_1|x)}{f_{\theta|x}(\theta_2|x)} = \frac{\frac{f(x|\theta_1)f(\theta_1)}{f(x)}}{\frac{f(x|\theta_2)f(\theta_2)}{f(x)}} = \frac{f(x|\theta_1)}{f(x|\theta_2)} \cdot \frac{f(\theta_1)}{f(\theta_2)},$$

Therefore:

$$\text{Posterior-Odds} = \text{Prior-Odds} \times \text{likelihood ratio.}$$

4.3 Bayesian Inference At a Glance

- The posterior distribution is often only known as being proportional to the product of likelihood and prior, but the normalization constant cannot be computed. It is often not a well-known distribution such as the normal distribution.
- Although the posterior distribution can often not be analyzed analytically, samples, i.e. random numbers, can be drawn from it. The random samples can then be used to estimate interesting quantities through the empirical analogues, e.g. expectations can be estimated by arithmetic means. Or one can use e.g. kernel density estimates.
- Methods for sampling from the posterior are the key for Bayesian inference in models with many parameters.

4.4 Recap: Models with One Parameter

4.4 Recap: Models with One Parameter

- In the following, θ is a scalar.
- The prior distribution can have more than one parameter.
- In that case, the following concepts work well:
  - Conjugate prior distributions
  - Reference priors or reference analysis

4.4 Recap: Models with One Parameter

- Conjugate prior distributions (prior distributions which lead to a posterior in the same distribution family) are available for example for (one-parameter) exponential families.

  Advantage: analytical calculations possible, no simulation required.

  Disadvantage: usually not available for complex models, therefore rather used as a building block in more complicated models.

4.4 Recap: Models with One Parameter

Reference priors and mutual information:

- *Idea:* choose prior such that the data (likelihood) dominates the posterior distribution even in the case of low sample sizes (”let the data speak for themselves”).
- First, the data distribution is parametrized by a family with sufficient statistics $T(X)$. Then, a prior $p(\theta)$ is searched that maximizes its Kullback-Leibler divergence from the posterior $f(\theta|t)$. For that, the K-L divergence is averaged over the distribution of T (i.e., the expectation is taken):

$$\begin{aligned}
D(f(\theta|t), p(\theta)) &= \int f(\theta|t) \log\left\{\frac{f(\theta|t)}{p(\theta)}\right\} d\theta \\
\mathbb{E}_T[D(f(\theta|t), p(\theta))] &= \int f(t) \int f(\theta|t) \log\left\{\frac{f(\theta|t)}{p(\theta)}\right\} d\theta dt \\
&= \int\int f(\theta, t) \log\left\{\frac{f(\theta, t)}{p(\theta) f(t)}\right\} d\theta dt
\end{aligned}$$

  The last expression is the *mutual information* between T and θ. Maximizing this information means that ”one gets most out of the data” for estimating θ. If θ and T are independent → mutual information is 0.

4.4 Recap: Models with One Parameter

- To calculate reference priors is mathematically often not tractable. Therefore one switches to asymptotic considerations using n independent copies of T, T^n, and letting $n \to \infty$.
- One can show that if there is a BAN estimator for θ, then the so-called Jeffreys prior

$$p(\theta) \propto \sqrt{I(\theta)}\ ,$$

  where $I(\theta)$ is the usual Fisher information based on the log-likelihood (with $n = 1$), is a reference prior.
- The Jeffreys prior is invariant to bijective transformations of θ.
- Generally, the mutual information is transformation invariant.
- The Jeffreys prior may or may not be a conjugated prior.
- Transforming the ideas from the scalar case to multiparameter models is beyond the scope of this lecture.

4.4 Recap: Models with One Parameter

Example 4.1 (Binomial- and Negative Binomial distribution)

1. *Binomial distribution:* the likelihood is

$$f(x|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}.$$

As reference prior, the Jeffreys prior can be used: Beta $\left(\frac{1}{2}, \frac{1}{2}\right)$,

$$p(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}.$$

Let $y = \sum_{i=1}^{n} x_i$. Then the reference posterior is:

$$\begin{aligned} f(\theta|x) &\propto f(x|\theta)p(\theta) \\ &\propto \theta^{y}(1-\theta)^{n-y}\theta^{-1/2}(1-\theta)^{-1/2} \\ &= \theta^{y-1/2}(1-\theta)^{n-y-1/2}. \end{aligned}$$

Remark 1: the Beta(α, β) *density is*

$$p(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},\ \theta \in [0,1]\ .$$

Remark 2: the Fisher information for a Bernoulli(θ) *is* $\frac{1}{\theta(1-\theta)}$.

4.4 Recap: Models with One Parameter

This corresponds to the kernel of the density of a Beta $\left(\frac{1}{2}+y, \frac{1}{2}+n-y\right)$–distribution.

Therefore, $f(\theta|x)$ is also for the extreme cases $y = 0$ or $y = n$ still a "proper" density, i.e. it is integrable and integrates to 1.

If one uses the Haldane prior instead,

$$p(\theta) \propto \theta^{-1}(1-\theta)^{-1},$$

which represents an improper prior, 'Beta$(0,0)$', the posterior Beta$(y, n-y)$ is not proper for the extreme cases $y = 0$ or $y = n$.

4.4 Recap: Models with One Parameter

2. *Negative binomial distribution:* Let X be the number of attempts until $y \geq 1$ successes occur. Then the likelihood is

$$f(x|\theta) \propto \binom{x-1}{y-1}\theta^y(1-\theta)^{x-y} \quad \text{für } x \geq y.$$

The reference prior is Jeffreys prior for the geometric distribution (this corresponds to $y = 1$):

$$p(\theta) \propto \theta^{-1}(1-\theta)^{-1/2},$$

from which we get the reference posterior

$$f(\theta|x) \propto \theta^{y-1}(1-\theta)^{x-y-1/2}\ .$$

Thus, a Beta$(y, x - y + 1/2)$–distribution, results. Since $y \geq 1$ and $x \geq y$, this is also always a proper posterior.

4.5 Multi-Parameter Models

4.5.1 Normal distribution

4.5 Multi-Parameter Models 4.5.1 Normal Distribution

Let $X_1, \ldots, X_n|\mu, \sigma^2 \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ with μ, σ^2 unknown and $x = (x_1, \ldots, x_n)$ the observed data.

(i) *Joint posterior distribution of $\mu, \sigma^2|x$:*

$$f(\mu, \sigma^2|x) \propto f(x|\mu, \sigma^2) \cdot p(\mu, \sigma^2)$$

(ii) *Conditional posterior distributions $\mu|\sigma^2, x$ and $\sigma^2|\mu, x$:*

$$f(\mu|\sigma^2, x) \propto f(\mu, \sigma^2|x) \quad \text{and} \quad f(\sigma^2|\mu, x) \propto f(\mu, \sigma^2|x)$$

(iii) *Marginal posterior distribution of $\mu|x$:*

$$f(\mu|x) = \int f(\mu, \sigma^2|x)\ d\sigma^2 = \int f(\mu|\sigma^2, x) f(\sigma^2|x)\ d\sigma^2$$

(iv) *Marginal posterior distribution of $\sigma^2|x$:*

$$f(\sigma^2|x) = \int f(\mu, \sigma^2|x)\ d\mu = \int f(\sigma^2|\mu, x) f(\mu|x)\ d\mu$$

Remark: The proportionalities in (ii) follow e.g. from $f(\mu|\sigma^2, x) = f(\mu, \sigma^2|x)/f(\sigma^2|x)$ and the fact that $f(\sigma^2|x)$ does not depend on μ, the argument on the left hand side $f(\mu|\sigma^2, x)$.

4.5 Multi-Parameter Models 4.5.1 Normal Distribution

I. Non-informative Prior Distribution

If only one of the two parameters is unknown, one often chooses the following prior distributions (Jeffreys prior distributions):

$$\begin{aligned} \sigma^2 \text{ known:} \quad & p(\mu) \propto \text{ const}, \\ \mu \text{ known:} \quad & p(\sigma^2) \propto (\sigma^2)^{-1}. \end{aligned}$$

One possibility to turn this into a multidimensional prior:

$$p(\mu, \sigma^2) = p(\mu) \cdot p(\sigma^2) \propto (\sigma^2)^{-1},$$

i.e. we assume independent priors for μ and σ^2.
The joint posterior distribution $f(\mu, \sigma^2|x)$ is then:

4.5 Multi-Parameter Models
4.5.1 Normal Distribution

$$\begin{aligned}
f(\mu, \sigma^2|x) &\propto \text{likelihood } \times \text{ prior} \\
&= \left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \sigma^{-1} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right)\right] \cdot (\sigma^2)^{-1} \\
&\propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right) \\
&= \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2} \left((n-1)s^2 + n(\overline{x} - \mu)^2\right)\right)
\end{aligned}$$

with $s^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2/(n-1)$.

Remark: Generally, we write the posterior ignoring the normalization constant, hence the $\propto$ notation.
The conditional posterior of μ, $f(\mu|\sigma^2, x)$, can be derived from the case with known variance σ^2: $f(\mu|\sigma^2, x)$ is the density of a $N(\overline{x}, \sigma^2/n)$. For the derivation we use that every conditional density is proportional to the joint density:

4.5 Multi-Parameter Models
4.5.1 Normal Distribution

$$\begin{aligned}
f(\mu|\sigma^2, x) &\propto f(\mu, \sigma^2|x) \\
&\propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2} \left((n-1)s^2 + n(\overline{x} - \mu)^2\right)\right) \\
& \text{ ignore all terms } \textbf{not} \text{ depending on } \mu \\
&\propto \exp\left(-\frac{1}{2\sigma^2} \left(n(\overline{x} - \mu)^2\right)\right) \\
&= \exp\left(-\frac{1}{2\frac{\sigma^2}{n}} \left((\overline{x} - \mu)^2\right)\right)
\end{aligned}$$

The last expression is the **kernel** of the normal density $N(\overline{x}, \sigma^2/n)$.
For the marginal posterior $f(\sigma^2|x)$ one has

4.5 Multi-Parameter Models
4.5.1 Normal Distribution

$$\begin{aligned}
f(\sigma^2|x) &= \int f(\mu, \sigma^2|x)\, d\mu \\
&\propto \int \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2} \left((n-1)s^2 + n(\overline{x} - \mu)^2\right)\right) d\mu \\
&\propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right) \int \exp\left(-\frac{1}{2\sigma^2} n(\overline{x} - \mu)^2\right) d\mu.
\end{aligned}$$

4.5 Multi-Parameter Models
4.5.1 Normal Distribution

It applies

$$\int \exp\left(-\frac{1}{2\sigma^2} n(\overline{x} - \mu)^2\right) d\mu = \sqrt{2\pi\, \frac{\sigma^2}{n}}$$

and thus

$$\begin{aligned} f(\sigma^2|x) &\propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right)\sqrt{2\pi\,\frac{\sigma^2}{n}} \\ &\propto (\sigma^2)^{\frac{-n-2}{2}}(\sigma^2)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right) \\ &= (\sigma^2)^{-\frac{n+1}{2}} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right). \end{aligned}$$

4.5 Multi-Parameter Models
4.5.1 Normal Distribution

The kernel of the density is the kernel of an inverse gamma distribution with the parameters $(n-1)/2$ and $(n-1)s^2/2$.
Inverse gamma:

$$p(y|a,b) = \frac{b^a}{\Gamma(a)} y^{-a-1} \exp\big(-\frac{b}{y}\big)$$

σ^2 takes the role of y, $a = (n-1)/2$, $b = (n-1)s^2/2$.
Because of

$$f(\mu, \sigma^2|x_1, \dots, x_n) = f(\mu|\sigma^2, x_1, \dots, x_n) \cdot f(\sigma^2|x_1, \dots, x_n)$$

the joint posterior distribution of $\mu, \sigma^2|x_1, \dots, x_n$ can now be simulated as follows:

4.5 Multi-Parameter Models
4.5.1 Normal Distribution

Algorithm 2: Direct simulation of the joint posterior distribution with non-informative prior

```
Repeat for $s = 1, \dots, S$:
    1. Draw $(\sigma^2)^{(s)}$ from IG $\big(\frac{n-1}{2}, \frac{n-1}{2}s^2\big)$.
    2. Draw $(\mu)^{(s)}$ from $N(\overline{x}, \frac{1}{n}(\sigma^2)^{(s)})$.
One gets pairs $\big[(\mu^{(1)}, (\sigma^2)^{(1)}), \dots, (\mu^{(S)}, (\sigma^2)^{(S)})\big]$.
```

4.5 Multi-Parameter Models
4.5.1 Normal Distribution

The previous algorithm can be extended to make draws from the *posterior predictive distribution* of a new observation $\tilde{x}$:

$$\begin{aligned} f(\widetilde{x}|x) &= \int\int f(\widetilde{x}|\mu, \sigma^2, x) f(\mu, \sigma^2|x)\, d\mu\, d\sigma^2 \\ &= \int\int f(\widetilde{x}|\mu, \sigma^2) f(\mu, \sigma^2|x)\, d\mu\, d\sigma^2. \end{aligned}$$

The first of the two densities in the integral is the data density, here a normal. The second is the posterior density. Given actual draws $((\mu)^{(s)}, (\sigma^2)^{(s)})$ for (μ, σ^2) one simply has to draw

$$\tilde{x}^{(s)} \sim N((\mu)^{(s)}, (\sigma^2)^{(s)})\ .$$

Alternatively, calculations show that $f(\widetilde{x}|x)$ is a t-distribution with location parameter $\bar{x}$, scale parameter $(1 + 1/n)^{1/2}s$ and $(n-1)$ degrees of freedom, so direct draws are also possible.
But: this general idea can be used in any case, even if the direct calculation of the density is complicated. Only draws from the data model (and of the posterior) must be possible.

4.5 Multi-Parameter Models
4.5.1 Normal Distribution

$\boldsymbol{\sigma^2}$ as a nuisance parameter
If one is only interested in μ, there are two options for simulation:

1. Simulate the joint posterior $f(\mu, \sigma^2|x)$ according to the above algorithm and consider only the draws $\mu^{(1)}, \dots, \mu^{(S)}$.

2. Calculate directly the marginal posterior $f(\mu|x)$:

$$f(\mu|x) = \int_0^\infty f(\mu, \sigma^2|x)\ d\sigma^2.$$

4.5 Multi-Parameter Models 4.5.1 Normal Distribution

Regarding 2.:
If one substitutes $z = A/(2\sigma^2)$ with $A = (n-1)s^2 + n(\mu - \overline{x})^2$ we get with

$$\sigma^2 = \frac{1}{2}Az^{-1} \quad \text{and} \quad d\sigma^2 = -2A^{-1}\sigma^4 dz = -\frac{1}{2}Az^{-2}dz$$

for $f(\mu|x)$:

$$\begin{aligned}
\int_0^\infty f(\mu, \sigma^2|x)\ d\sigma^2 &\propto \int_0^\infty A^{-\frac{n+2}{2}} z^{\frac{n+2}{2}}\ \exp(-z)\ A\ z^{-2}dz \\
&= \int_0^\infty A^{-\frac{n}{2}} z^{\frac{n-2}{2}} \exp(-z)\ dz \\
&= A^{-\frac{n}{2}} \int_0^\infty z^{\frac{n-2}{2}} \exp(-z)\ dz.
\end{aligned}$$

4.5 Multi-Parameter Models 4.5.1 Normal Distribution

In general, for $a > 0$ and $m > -1$:

$$\int_0^\infty x^m \exp(-ax)\ dx = \frac{\Gamma(m+1)}{a^{m+1}}.$$

It follows that the integral is constant with respect to μ and therefore

$$\begin{aligned}
f(\mu|x) &\propto A^{-\frac{n}{2}} \\
&= \left[(n-1)s^2 + n(\mu - \overline{x})^2\right]^{-\frac{n}{2}} \\
&= \left[1 + \frac{(\mu - \overline{x})^2}{(n-1)s^2/n}\right]^{-\frac{n}{2}} \\
&= \left[1 + \frac{1}{n-1}\left(\frac{\mu - \overline{x}}{\frac{s}{\sqrt{n}}}\right)^2\right]^{-\frac{n}{2}}
\end{aligned}$$

4.5 Multi-Parameter Models 4.5.1 Normal Distribution

which is the kernel of the density of a scaled non-central t-distribution with scale parameter $m = s/\sqrt{n}$, location parameters $l = \overline{x}$ and $\nu = n - 1$ degrees of freedom. Generally, the kernel of the density of such a general t-distribution is

$$\text{kernel}(f(\theta)) = \left[1 + \frac{1}{\nu}\left(\frac{\theta - l}{m}\right)^2\right]^{-\frac{\nu+1}{2}}.$$

4.5 Multi-Parameter Models 4.5.1 Normal Distribution

Remarks:

1. Instead of σ^2, the so-called *precision* $\kappa = (\sigma^2)^{-1}$ can also be used. Using $p(\mu, \kappa) \propto (\kappa)^{-1}$, it follows that $\kappa|x \sim \text{Gamma}\left(\frac{n-1}{2}, \frac{n-1}{2}s^2\right)$.

2. Instead of an inverse gamma distribution, the special case of an scaled inverse χ^2-distribution inv-χ^2 is used (see below).

4.5 Multi-Parameter Models 4.5.1 Normal Distribution

II. Conjugate Prior Distribution
According to remark 2, use the scaled inverse $\chi^2(\nu_0, \sigma_0^2)$-distribution as prior.
Advantage: Better interpretation (we will see this later in the case of informative priors).
Disadvantage: This approach is "non-standard".
Random numbers from a scaled inverse χ^2 -distribution can be obtained/simulated as follows:

4.5 Multi-Parameter Models 4.5.1 Normal Distribution

Algorithm 3: simulation of $\boldsymbol{\theta \sim}$ inv-$\boldsymbol{\chi^2(\nu_0, \sigma_0^2)}$

```
1. Draw $X^* \sim \chi^2(\nu_0)$.
2. Set $\theta = \frac{\nu_0 \sigma_0^2}{X^*}$.
```

4.5 Multi-Parameter Models 4.5.1 Normal Distribution

The following applies:

$$\text{inv-}\chi^2(\nu_0, \sigma_0^2) \ = \ \text{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2\right).$$

This can be verified with the theorem for density transformations: define $\alpha = \nu_0/2$ and $\beta = 1/2$, so that $X^* \sim \text{Gamma}(\alpha, \beta)$. The inverse of the transformation in step 2 is

$$X^* = g^{-1}(\theta) = \frac{\nu_0 \sigma_0^2}{\theta}$$

and the corresponding derivative with respect to θ is

$$(g^{-1})'(\theta) = -\frac{\nu_0 \sigma_0^2}{\theta^2}.$$

4.5 Multi-Parameter Models 4.5.1 Normal Distribution

One thus obtains:

$$f(\theta) = f_{X^*}(g^{-1}(\theta)) \cdot |(g^{-1})'(\theta)| = \frac{(\beta \nu_0 \sigma_0^2)^\alpha}{\Gamma(\alpha)}\, \theta^{-\alpha-1} \exp\left(-\beta \nu_0 \sigma_0^2/\theta\right).$$

This is the density of an inverse gamma distribution with parameters $(\alpha, \beta\nu_0\sigma_0^2)$, which equals the density of the desired inverse χ^2-distribution. One possible prior is now

$$p(\mu, \sigma^2) = p(\mu|\sigma^2) \cdot p(\sigma^2)$$

with

$$\mu|\sigma^2 \ \sim \ N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right) \quad \text{and} \quad \sigma^2 \ \sim \ \text{inv-}\chi^2(\nu_0, \sigma_0^2)$$

4.5 Multi-Parameter Models 4.5.1 Normal Distribution

For this one briefly writes: N-inv-$\chi^2\left(\mu_0, \frac{\sigma_0^2}{\kappa_0}; \nu_0, \sigma_0^2\right)$. The priors are now interdependent. So now let $(\mu, \sigma^2) \sim$ N-inv-$\chi^2\left(\mu_0, \frac{\sigma_0^2}{\kappa_0}; \nu_0, \sigma_0^2\right)$. The **prior density** is then

$$\begin{aligned} p(\mu, \sigma^2) \quad &= \quad \frac{1}{\sqrt{2\pi\sigma^2\kappa_0^{-1}}} \exp\left(-\frac{1}{2\frac{\sigma^2}{\kappa_0}}(\mu - \mu_0)^2\right) \\ &\times \quad \frac{(\frac{1}{2}\nu_0\sigma_0^2)^{\nu_0/2}}{\Gamma(\nu_0/2)} (\sigma^2)^{-(\frac{\nu_0}{2}+1)} \exp\left(-\frac{1}{2}\frac{\nu_0\sigma_0^2}{\sigma^2}\right) \\ &\propto \quad (\sigma^2)^{-\frac{1}{2}} (\sigma^2)^{-(\frac{\nu_0}{2}+1)} \exp\left(-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2 + \kappa_0(\mu - \mu_0)^2\right]\right). \end{aligned}$$

The (hyper-)parameters κ_0, ν_0, σ_0^2 and μ_0 have to be fixed at certain values.

Remark: The marginal prior for μ is a t-distribution.

4.5 Multi-Parameter Models
4.5.1 Normal Distribution

The **joint posterior**, given data $x = (x_1, \dots, x_n)$ from $N(\mu, \sigma^2)$, is

$$\begin{aligned} f(\mu, \sigma^2|x) &\propto (\sigma^2)^{-\frac{1}{2}}(\sigma^2)^{-(\frac{\nu_0}{2}+1)} \exp\left(-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2 + \kappa_0(\mu-\mu_0)^2\right]\right) \\ &\quad \times(\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\left((n-1)s^2 + n(\overline{x}-\mu)^2\right)\right). \end{aligned}$$

4.5 Multi-Parameter Models
4.5.1 Normal Distribution

One can show (see exercise) that the posterior is again N-inv-χ^2–distributed with updated parameters

$$\begin{aligned} \mu_n &= \left(\frac{\kappa_0}{\kappa_0+n}\right)\mu_0 + \left(\frac{n}{\kappa_0+n}\right)\overline{x}, \\ \kappa_n &= \kappa_0 + n, \\ \nu_n &= \nu_0 + n, \\ \nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\overline{x}-\mu_0)^2. \end{aligned}$$

4.5 Multi-Parameter Models
4.5.1 Normal Distribution

The interpretation of the parameters is as follows:

- μ_n is the weighted mean of the sample mean and prior expected value. In the limiting cases $\kappa_0 \to \infty$ one gets $\mu_n = \mu_0$ and for $n \to \infty$, we have $\mu_n = \overline{x}$.
- ν_n are the posterior degrees of freedom as the sum of prior degrees of freedom and sample size.
- The posterior-sum of squares $\nu_n\sigma_n^2$ can be partitioned into the prior-sum of squares $\nu_0\sigma_0^2$, the sum of squares $(n-1)s^2$ of the sample and a term quantifying the uncertainty that occurs by the difference between the sample mean and the prior expected value.

4.5 Multi-Parameter Models
4.5.1 Normal Distribution

The **conditional posterior** of $\mu|\sigma^2, x$ is

$$\begin{aligned} \mu|\sigma^2, x &\sim N\left(\mu_n, \frac{\sigma^2}{\kappa_n}\right) \\ &\hat{=} N\left(\frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\overline{x}, \frac{\sigma^2}{\kappa_0+n}\right) \\ &\hat{=} N\left(\frac{\frac{\kappa_0}{\sigma^2}\mu_0 + \frac{n}{\sigma^2}\overline{x}}{\frac{\kappa_0}{\sigma^2}+\frac{n}{\sigma^2}}, \frac{1}{\frac{\kappa_0}{\sigma^2}+\frac{n}{\sigma^2}}\right). \end{aligned}$$

The weights κ_0/σ^2 and n/σ^2 correspond to the prior and data precision. The **marginal posterior** of $\sigma^2|x$ is

$$\sigma^2|x \sim \text{inv-}\chi^2(\nu_n, \sigma_n^2).$$

4.5 Multi-Parameter Models
4.5.1 Normal Distribution

This enables the simulation of the joint posterior distribution:

Algorithm 4: Direct simulation of the joint posterior distribution with a conjugated prior

1. Draw $(\sigma^2)^*$ from inv-$\chi^2(\nu_n, \sigma_n^2)$.
2. Draw μ^* from $N\left(\mu_n, \frac{(\sigma^2)^*}{\kappa_n}\right)$.

The algorithm can be extended to make draws from the *posterior predictive distribution* of a new observation $\tilde{x}$ as discussed previously.

4.5 Multi-Parameter Models 4.5.1 Normal Distribution

The **marginal posterior** of $\mu|x$ is

$$f(\mu|x) \propto \left[1 + \frac{\kappa_n(\mu - \mu_n)^2}{\nu_n \sigma_n^2}\right]^{-(\nu_n+1)/2}.$$

This corresponds to a t-distribution with ν_n degrees of freedom, location parameter μ_n and scale parameter σ_n^2/κ_n.

4.5.2 Dirichlet-Multinomial Model

4.5 Multi-Parameter Models 4.5.2 Dirichlet-Multinomial Model

The *Dirichlet-Multinomial-Model* is used for data $y_1, \dots, y_n$ from a *multinomial distribution.* The multinomial distribution is a generalization of the Binomial distribution for more than two possible events and fixed sample size n. For example, one could ask people in a survey for their party preferences.

4.5 Multi-Parameter Models 4.5.2 Dirichlet-Multinomial Model

A multinomial random variable Y can have k possible outcomes (for example CDU/CSU, SPD, FDP, Greens, Leftists, Others). The random variable X_j, $1 \leq j \leq k$, denotes the number of observations with category j in the sample, such that $\sum_{j=1}^k X_j = n$. The parameter $\theta_j = \mathbb{P}(Y = j) \in [0,1]$ for $Y \in \{1, \dots, k\}$ with $\sum_{j=1}^k \theta_j = 1$ denotes the probability for the category j.

4.5 Multi-Parameter Models 4.5.2 Dirichlet-Multinomial Model

The likelihood of $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^\top$ given the observed data $\boldsymbol{x} = (x_1, \dots, x_k)^\top$ is

$$L(\boldsymbol{\theta}) = f(\boldsymbol{x}|\boldsymbol{\theta}) \propto \prod_{j=1}^k \theta_j^{x_j}.$$

Because of the restriction $\sum_{j=1}^k \theta_j = 1$, there are in fact only $k-1$ Parameters, because the k-th parameter can be determined by $\theta_k = 1 - \theta_1 - \dots \theta_{k-1}$. Hence, the likelihood can also be written in the form

$$L(\boldsymbol{\theta}) \propto \left(\prod_{j=1}^{k-1} \theta_j^{x_j}\right) (1 - \theta_1 - \dots - \theta_{k-1})^{x_k}$$

4.5 Multi-Parameter Models 4.5.2 Dirichlet-Multinomial Model

The distribution is conjugate to the multinomial distribution is the so-called *Dirichlet distribution*, a generalization of the Beta distribution, given by

$$\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^\top \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_k) = D(\boldsymbol{\alpha}),$$

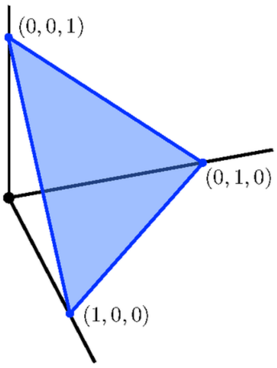
with density function

$$p(\boldsymbol{\theta}) = \frac{\Gamma(\alpha_1 + \dots + \alpha_k)}{\Gamma(\alpha_1) \cdot \dots \cdot \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \cdot \dots \cdot \theta_k^{\alpha_k - 1},$$

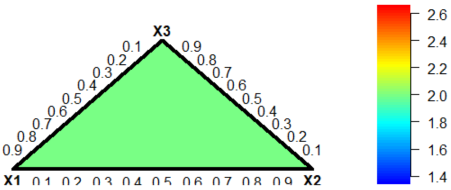
where $\alpha_j > 0$ for all $j = 1, \dots, k$ and again $\theta_j \in [0,1]$ with $\sum_{j=1}^k \theta_j = 1$. The Dirichlet distribution thus specifies a distribution on a $(k-1)$-dimensional simplex.

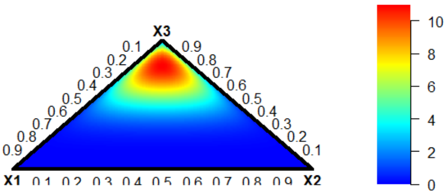
Remark: The assumption $\theta_j \in [0,1]$ can be changed to $\theta_j \in (0,1)$ to exclude the edge cases. The $(k-1)$-dimensional simplex is then open.

4.5 Multi-Parameter Models



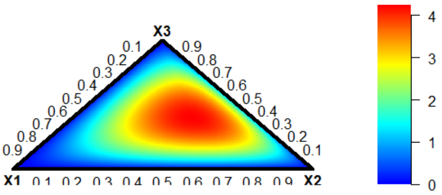
Contour plots:

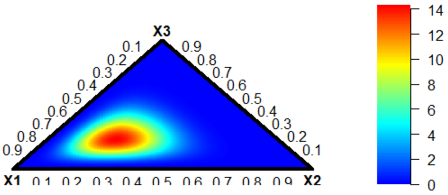




Dirichlet(1,1,1)

Dirichlet(1.5, 1.5, 5)





Dirichlet (1.5, 2, 2)

Dirichlet (8,4,4)

```
> ddirichlet(c(.1,.2,.7), c(1.5,1.5,5)) # density of Dir(1.5,1.5,5) at (.1,.2,.7)
[1] 9.07897
> ddirichlet(c(.1,.1,.8), c(1.5,1.5,5))
[1] 10.9519
> ddirichlet(c(.1,.2,.8), c(1.5,1.5,5))
[1] 0
```

Properties of the Dirichlet distribution:
Define $\alpha_0 = \sum_{j=1}^{k} \alpha_j$.

4.5 Multi-Parameter Models
4.5.2 Dirichlet-Multinomial Model

- Moments:

$$\begin{aligned}
\mathbb{E}(\theta_j) &= \frac{\alpha_j}{\alpha_0}, \\
\mathrm{Var}(\theta_j) &= \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)}, \\
\mathrm{Cov}(\theta_i, \theta_j) &= -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)} \ ,
\end{aligned}$$

  where the restriction $\sum_{j=1}^{k} \theta_j = 1$ implies the negative correlation.

4.5 Multi-Parameter Models
4.5.2 Dirichlet-Multinomial Model

- Mode exists under the assumption $\alpha_j > 1 \forall j$:

$$\mathrm{Mode}(\boldsymbol{\theta})_j = \frac{\alpha_j - 1}{\alpha_0 - k}$$

  is the j-th component of the k-dimensional mode.

- Every marginal distribution is again a Dirichlet distribution, for example

$$(\theta_i, \theta_j, 1 - \theta_i - \theta_j) \ \sim \ \mathrm{Dirichlet}(\alpha_i, \alpha_j, \alpha_0 - \alpha_i - \alpha_j).$$

  In particular

$$\theta_j \ \sim \ \mathrm{Beta}(\alpha_j, \alpha_0 - \alpha_j).$$

4.5 Multi-Parameter Models
4.5.2 Dirichlet-Multinomial Model

- The conditional distributions are also Dirichlet distributions. If $h < k$ and $(\theta_1, \ldots, \theta_k)^\top \sim \mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_k)$, then

$$\frac{1}{\sum_{i=1}^{h} \theta_i} (\theta_1, \ldots, \theta_h)^\top | (\theta_{h+1}, \ldots, \theta_k)^\top \sim \ \mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_h) \ .$$

  Alternatively, define

$$\tilde{\theta}_i = \frac{\theta_i}{1 - \sum_{r=1}^{m-1} \theta_r} \quad , \ m \leq \ i \leq k,$$

  then given the realizations $\theta_1, \ldots, \theta_{m-1}$,

$$(\tilde{\theta}_m, \ldots, \tilde{\theta}_k)^\top \ \sim \ \mathrm{Dirichlet}(\alpha_m, \ldots, \alpha_k).$$

4.5 Multi-Parameter Models
4.5.2 Dirichlet-Multinomial Model

Algorithm 5: Simulation from the Dirichlet distribution

```
• Simulation 1:
    1. Draw $Z_1, Z_2, \dots, Z_k$ from (independent) gamma distributions with
       parameters $(\alpha_1, 1), \dots, (\alpha_k, 1)$.
    2. Set
           $\theta_j = \frac{Z_j}{\sum_{i=1}^k Z_i}$ .
• Simulation 2 (”Stick Breaking Prior”):
    1. Draw $\theta_1 \sim \mathrm{Beta}(\alpha_1, \alpha_0 - \alpha_1)$.
    2. For $j = 2, \dots, k-1$:
        (i) Draw $Z_j \sim \mathrm{Beta}(\alpha_j, \sum_{i=j+1}^k \alpha_i)$.
        (ii) Set $\theta_j = \left(1 - \sum_{i=1}^{j-1} \theta_i\right) Z_j$.
    3. Set $\theta_k = 1 - \sum_{i=1}^{k-1} \theta_i$.
```

Applications of the Dirichlet distribution:

4.5 Multi-Parameter Models
4.5.2 Dirichlet-Multinomial Model

- Compositional data, rates, proportions, percentages, ppm
- LDA (latent dirichlet allocation model), see https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
- As prior for a multinomial distribution

4.5 Multi-Parameter Models
4.5.2 Dirichlet-Multinomial Model

```
> # Stick breaking. Dir(1.5, 1.5, 5)
> # "true" probabilities=expectations
> alpha <- c(1.5, 1.5, 5)
> alpha.0 <- sum(alpha)
> print("True expectations:")
[1] "True expectations:"
> print(alpha/alpha.0)
[1] 0.1875 0.1875 0.6250
>
> n <- 1000
> dir.draws <- matrix(nrow=n, ncol=3, data=0)
>
> for ( s in 1:n) {
+
+   theta.1 <- rbeta(n=1, alpha[1], alpha.0-alpha[1] )
+   Z <- rbeta(1, alpha[2], alpha[3])
+   theta.2 <- (1-theta.1) * Z
+   theta.3 <- 1-theta.1-theta.2
+   dir.draws[s,] <- c(theta.1, theta.2, theta.3)
+ }
>
```

```
> print("sampled expectations:")
[1] "sampled expectations:"
> colMeans(dir.draws)
[1] 0.1880414 0.1872584 0.6247002
```

4.5 Multi-Parameter Models
4.5.2 Dirichlet-Multinomial Model

R example, used libraries: Compositional, MCMCpack, DirichletReg

```
# visualize
dr <- DR_data(dir.draws)
plot(dr)
# Regression
h <- DirichReg(formula=dr~1, data=as.data.frame(dir.draws), model="common" )
[1] "Coefficients:"
> exp(as.numeric(coef(h)))
[1] 1.464163 1.443149 4.839917
```

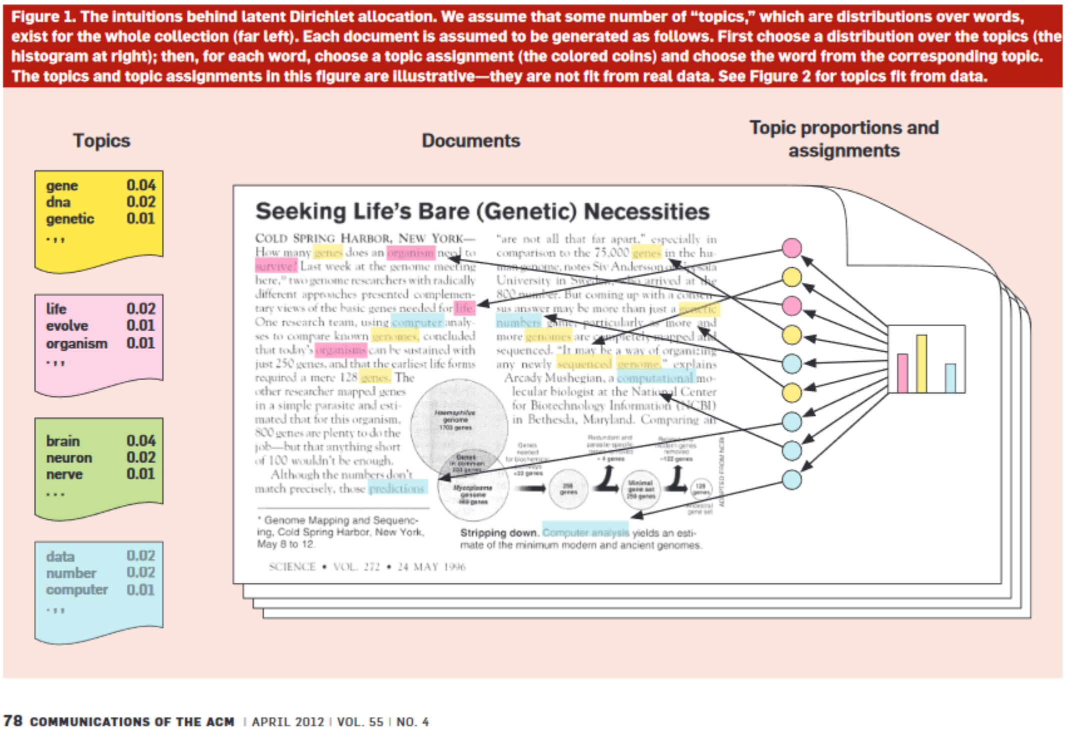


Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.

Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
Redundant and parasite-specific genes removed −4 genes
Related and modern genes removed −122 genes
256 genes
Minimal gene set 250 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

78 COMMUNICATIONS OF THE ACM | APRIL 2012 | VOL. 55 | NO. 4

LDA (Latent Dirichlet Allocation): Probabilistic topic modeling for text data

4.5 Multi-Parameter Models
4.5.2 Dirichlet-Multinomial Model

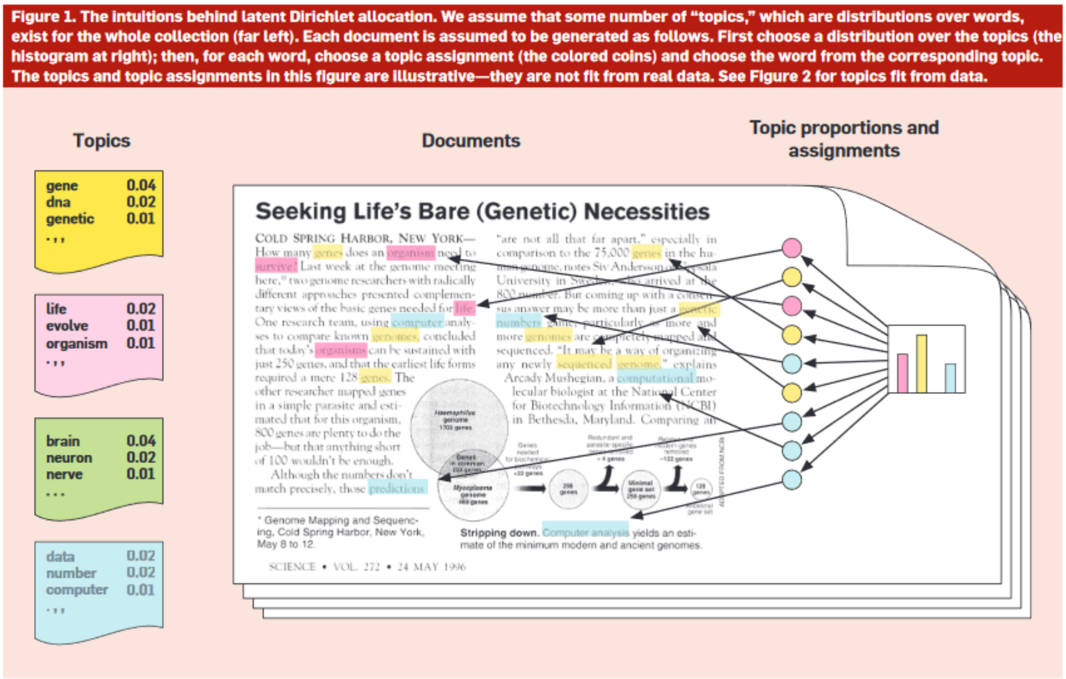


Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.

Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Seeking Life's Bare (Genetic) Necessities
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

78 COMMUNICATIONS OF THE ACM | APRIL 2012 | VOL. 55 | NO. 4

4.5 Multi-Parameter Models
4.5.2 Dirichlet-Multinomial Model

Back to our D-M-model...
For $\boldsymbol{x}|\boldsymbol{\theta} \sim \text{Multinomial}(n; \theta_1, \ldots, \theta_k)$ and $\boldsymbol{\theta}|\boldsymbol{\alpha} \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$, the **posterior distribution** of $\boldsymbol{\theta}$ is

$$\begin{aligned} f(\boldsymbol{\theta}|\boldsymbol{x}) &\propto L(\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \\ &\propto \prod_{j=1}^{k} \theta_j^{x_j} \cdot \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} \\ &= \prod_{j=1}^{k} \theta_j^{x_j + \alpha_j - 1}, \end{aligned}$$

i.e. the posterior is a $\text{Dirichlet}(x_1 + \alpha_1, \ldots, x_k + \alpha_k)$.

4.5 Multi-Parameter Models
4.5.2 Dirichlet-Multinomial Model

Interpretation of the posterior The posterior expectation

$$\mathbb{E}[\theta_j|\boldsymbol{x}] = \frac{x_j + \alpha_j}{\sum_{i=1}^{k} x_i + \sum_{i=1}^{k} \alpha_i} = \frac{x_j + \alpha_j}{n + \alpha_0}$$

can be paraphrased to

$$\mathbb{E}[\theta_j|\boldsymbol{x}] = \frac{\alpha_0}{\alpha_0 + n} \cdot \underbrace{\frac{\alpha_j}{\alpha_0}}_{\substack{\text{Prior} \\ \text{expected value}}} + \frac{n}{\alpha_0 + n} \cdot \underbrace{\frac{x_j}{n}}_{\text{MLE}} .$$

The parameter α_0 can be written as "prior number of observations" and α_j can be interpret as "prior successes" for category j .

4.5 Multi-Parameter Models
4.5.2 Dirichlet-Multinomial Model

Comment.

1. The choice $\alpha_1 = \ldots = \alpha_k = 0$ corresponds to a uniform distribution for $\{\log\theta_j\}_{j=1}^k$. In this case the posterior is only proper if $x_j \geq 1,\ j = 1, \ldots, p$.
2. The choice of $\alpha_1 = \ldots = \alpha_k = 1/2$ corresponds to Jeffreys' Prior.
3. The choice $\alpha_1 = \ldots = \alpha_k = 1$ corresponds to a uniform prior on the simplex.

4.5 Multi-Parameter Models
4.5.2 Dirichlet-Multinomial Model

Comment. The Dirichlet distribution is also suitable as a prior in the analysis of contingency tables with a multinomial survey scheme:

x_1	x_2
x_3	x_4

($n = x_1 + x_2 + x_3 + x_4$). Extensions to restricted multinomial distributions are possible (loglinear models).

4.5 Multi-Parameter Models
4.5.2 Dirichlet-Multinomial Model

Ad hoc procedure: add 1/2 for each entry of the contingency table and then calculate the maximum likelihood estimator; this corresponds to $\alpha_1 = \ldots = \alpha_k = 3/2$ and the posterior mode estimator

$$\begin{aligned}\text{Mode}(\boldsymbol{\theta}|\boldsymbol{x})_j &= \frac{x_j + \alpha_j - 1}{\sum_{i=1}^k x_i + \sum_{i=1}^k \alpha_i - k} \\ &= \frac{x_j + \frac{1}{2}}{\sum_{i=1}^k x_i + \frac{1}{2}k} \,.\end{aligned}$$

4.5.3 Multivariate Normal Distribution

4.5 Multi-Parameter Models
4.5.3 Multivariate Normal Distribution

Notation:

- $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ is p-dimensional random vector.
- $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$ is p-dimensional mean vector.
- The symmetric and positive definite (notation: $\boldsymbol{\Sigma} > 0$) matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_{pp} \end{pmatrix}$$

is the $p \times p$-dimensional covariance matrix.

4.5 Multi-Parameter Models
4.5.3 Multivariate Normal Distribution

- An observation $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$ is MVN($\boldsymbol{\mu}, \boldsymbol{\Sigma}$) (multivariate normally distributed), if

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right).$$

- The likelihood for n i.i.d. realizations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &\propto |\boldsymbol{\Sigma}|^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\right) \\ &= |\boldsymbol{\Sigma}|^{-n/2} \exp\left(-\frac{1}{2}\,\text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_0)\right) \end{aligned}$$

with $\boldsymbol{S}_0 = \sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top$.

4.5 Multi-Parameter Models 4.5.3 Multivariate Normal Distribution

- The second identity results from the transformations (tr is the trace-operator)

$$\begin{aligned}
\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}) &= \operatorname{tr}\left(\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right)\\
&= \operatorname{tr}\left(\sum_{i=1}^{n}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\right)\\
&= \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\right)\\
&= \operatorname{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_0).
\end{aligned}$$

4.5 Multi-Parameter Models 4.5.3 Multivariate Normal Distribution

I. Conjugated prior distribution for unknown $\boldsymbol{\mu}$ and known $\boldsymbol{\Sigma}$
The conjugate **prior distribution** for $\boldsymbol{\mu}$ if $\boldsymbol{\Sigma}$ is known is

$$\boldsymbol{\mu} \;\sim\; \mathrm{MVN}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$$

with $\boldsymbol{\Lambda}_0 > 0$. The **posterior** for $\boldsymbol{\mu}$ is then

$$f(\boldsymbol{\mu}|\boldsymbol{x},\boldsymbol{\Sigma}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Lambda}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right).$$

4.5 Multi-Parameter Models 4.5.3 Multivariate Normal Distribution

The term in the exponential is a quadratic form in $\boldsymbol{\mu}$. Completing the quadratic form and pulling out constant factors leads to

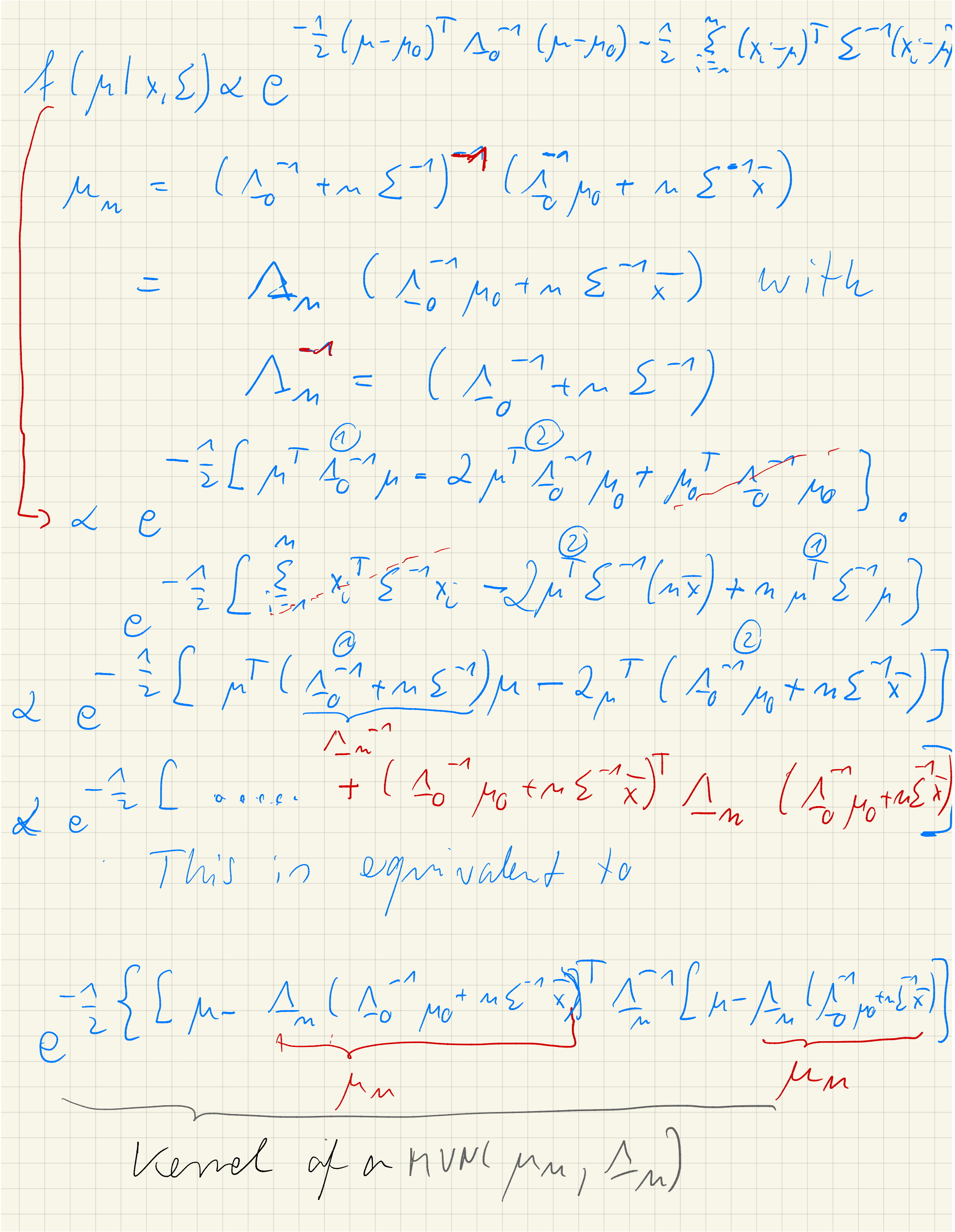
$$f(\boldsymbol{\mu}|\boldsymbol{x},\boldsymbol{\Sigma}) \;\sim\; \mathrm{MVN}(\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n)$$

with

$$\begin{aligned}
\boldsymbol{\mu}_n &= \left(\boldsymbol{\Lambda}_0^{-1}+n\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\mu}_0+n\boldsymbol{\Sigma}^{-1}\overline{\boldsymbol{x}}\right),\\
\boldsymbol{\Lambda}_n^{-1} &= \boldsymbol{\Lambda}_0^{-1}+n\boldsymbol{\Sigma}^{-1}
\end{aligned}$$

and $\overline{\boldsymbol{x}} = (\sum_{i=1}^{n}\boldsymbol{x}_i)/n$ in analogy to the univariate case.

4.5 Multi-Parameter Models 4.5.3 Multivariate Normal Distribution



$$f(\mu|x,\Sigma) \propto e^{-\frac{1}{2}(\mu-\mu_0)^T \Lambda_0^{-1}(\mu-\mu_0) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu}$$

$$\mu_n = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x})$$

$$= \Lambda_n (\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x}) \quad \text{with}$$

$$\Lambda_n^{-1} = (\Lambda_0^{-1} + n\Sigma^{-1})$$

$$\propto e^{-\frac{1}{2}[\overset{①}{\mu^T\Lambda_0^{-1}\mu} = 2\overset{②}{\mu^T\Lambda_0^{-1}\mu_0} + \cancel{\mu_0^T\Lambda_0^{-1}\mu_0}]} \cdot e^{-\frac{1}{2}[\cancel{\sum_{i=1}^{n} x_i^T\Sigma^{-1}x_i} - 2\overset{②}{\mu^T\Sigma^{-1}(n\bar{x})} + n\overset{①}{\mu^T\Sigma^{-1}\mu}]}$$

$$\propto e^{-\frac{1}{2}[\overset{①}{\mu^T\underbrace{(\Lambda_0^{-1}+n\Sigma^{-1})}_{\Lambda_n^{-1}}\mu} - 2\overset{②}{\mu^T(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x})}]}$$

$$\propto e^{-\frac{1}{2}[\,\ldots\ldots + (\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x})^T \Lambda_n (\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x})]}$$

This is equivalent to

$$\underbrace{e^{-\frac{1}{2}\{[\mu - \underbrace{\Lambda_n(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x})}_{\mu_n}]^T \Lambda_n^{-1}[\mu - \underbrace{\Lambda_n(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x})}_{\mu_n}]}}_{\text{Kernel of a MVN}(\mu_n,\ \Lambda_n)}$$

The **conditional** and **marginal posterior** for *subsets* of $\boldsymbol{\mu}$ follow from the properties of the multivariate normal distribution: look at the partitions

4.5 Multi-Parameter Models
4.5.3 Multivariate Normal Distribution

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{pmatrix}, \quad \boldsymbol{\mu}_n = \begin{pmatrix} \boldsymbol{\mu}_n^{(1)} \\ \boldsymbol{\mu}_n^{(2)} \end{pmatrix}, \quad \boldsymbol{\Lambda}_n = \begin{pmatrix} \boldsymbol{\Lambda}_n^{(11)} & \boldsymbol{\Lambda}_n^{(12)} \\ \boldsymbol{\Lambda}_n^{(21)} & \boldsymbol{\Lambda}_n^{(22)} \end{pmatrix}.$$

Then we get the conditional distributions

4.5 Multi-Parameter Models
4.5.3 Multivariate Normal Distribution

$$\boldsymbol{\mu}^{(1)}|\boldsymbol{\mu}^{(2)}, \boldsymbol{x} \sim \text{MVN}\left(\boldsymbol{\mu}_n^{(1)} + \boldsymbol{\beta}^{1|2}\left(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}_n^{(2)}\right), \boldsymbol{\Lambda}_n^{1|2}\right)$$

with

$$\begin{aligned}
\boldsymbol{\beta}^{1|2} &= \boldsymbol{\Lambda}_n^{(12)} \left(\boldsymbol{\Lambda}_n^{(22)}\right)^{-1}, \\
\boldsymbol{\Lambda}^{1|2} &= \boldsymbol{\Lambda}_n^{(11)} - \boldsymbol{\Lambda}_n^{(12)} \left(\boldsymbol{\Lambda}_n^{(22)}\right)^{-1} \boldsymbol{\Lambda}_n^{(21)}
\end{aligned}$$

and (e.g.) the marginal distribution

$$\boldsymbol{\mu}^{(1)} \sim \text{MVN}\left(\boldsymbol{\mu}_n^{(1)}, \boldsymbol{\Lambda}_n^{(11)}\right).$$

$\boldsymbol{\beta}^{1|2}$ is the regression coefficient (matrix) of regressing subset $\boldsymbol{\mu}^{(1)}$ on subset $\boldsymbol{\mu}^{(2)}$.
The **joint posterior distribution** of a new observation $\widetilde{\boldsymbol{x}}$ and $\boldsymbol{\mu}$ (still we assume that the covariance is known) is

4.5 Multi-Parameter Models
4.5.3 Multivariate Normal Distribution

$$f(\widetilde{\boldsymbol{x}}, \boldsymbol{\mu}|\boldsymbol{x}) = \text{MVN}(\widetilde{\boldsymbol{x}}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \cdot \text{ MVN}(\boldsymbol{\mu}|\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n).$$

In the exponential one gets a quadratic form in $(\widetilde{\boldsymbol{x}}, \boldsymbol{\mu})$. $(\widetilde{\boldsymbol{x}}, \boldsymbol{\mu})$ are jointly multivariate normal distributed, and therefore it follows that $\widetilde{\boldsymbol{x}}$ has as marginal distribution also a multivariate normal distribution. The required parameters can be calculated using the iterated expectation and variance:

$$\mathbb{E}[\widetilde{\boldsymbol{x}}|\boldsymbol{x}] = \mathbb{E}[\mathbb{E}[\widetilde{\boldsymbol{x}}|\boldsymbol{\mu}, \boldsymbol{x}]|\boldsymbol{x}] = \mathbb{E}[\boldsymbol{\mu}|\boldsymbol{x}] = \boldsymbol{\mu}_n$$

and

$$\begin{aligned}
\text{Var}(\widetilde{\boldsymbol{x}}|\boldsymbol{x}) &= \mathbb{E}[\text{Var}(\widetilde{\boldsymbol{x}}|\boldsymbol{\mu}, \boldsymbol{x})|\boldsymbol{x}] + \text{Var}[\mathbb{E}(\widetilde{\boldsymbol{x}}|\boldsymbol{\mu}, \boldsymbol{x})|\boldsymbol{x}] \\
&= \mathbb{E}[\boldsymbol{\Sigma}|\boldsymbol{x}] + \text{Var}[\boldsymbol{\mu}|\boldsymbol{x}] \\
&= \boldsymbol{\Sigma} + \boldsymbol{\Lambda}_n.
\end{aligned}$$

4.5 Multi-Parameter Models
4.5.3 Multivariate Normal Distribution

II. Conjugated prior distribution for unknown $\boldsymbol{\mu}$ and unknown $\boldsymbol{\Sigma}$
In section 4.5.1-II (slide 428) we used as conjugate prior distribution for the parameters μ and σ^2

$$\mu|\sigma^2 \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right) \quad \text{and} \quad \sigma^2 \sim \text{inv-}\chi^2(\nu_0, \sigma_0^2),$$

in short

$$\mu, \sigma^2 \sim \text{N-inv-}\chi^2\left(\mu_0, \frac{\sigma_0^2}{\kappa_0}; \nu_0, \sigma_0^2\right).$$

4.5 Multi-Parameter Models
4.5.3 Multivariate Normal Distribution

Here we use the multivariate analogue

$$\boldsymbol{\mu}|\boldsymbol{\Sigma} \sim \text{MVN}\left(\boldsymbol{\mu}_0, \frac{1}{\kappa_0}\boldsymbol{\Sigma}\right) \quad \text{and} \quad \boldsymbol{\Sigma} \sim \text{inv-Wishart}_{\nu_0}(\boldsymbol{\Lambda}_0^{-1}),$$

in short

$$\boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \text{MVN-inv-Wishart}\left(\boldsymbol{\mu}_0, \frac{1}{\kappa_0}\boldsymbol{\Lambda}_0; \nu_0, \boldsymbol{\Lambda}_0\right).$$

The **joint prior density** is then

$$\begin{aligned}
p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto\ & |\boldsymbol{\Sigma}|^{-\left(\frac{\nu_0+p}{2}+1\right)} \\
& \cdot \exp\left(-\frac{1}{2}\,\text{tr}(\boldsymbol{\Lambda}_0\boldsymbol{\Sigma}^{-1}) - \frac{\kappa_0}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right).
\end{aligned}$$

Here ν_0 denotes the prior number of degrees of freedom, κ_0 the prior number of measurements on the $\mathbf{\Sigma}$-scale. The **joint posterior distribution** of $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ is then

4.5 Multi-Parameter Models
4.5.3 Multivariate Normal Distribution

$$\boldsymbol{\mu}, \mathbf{\Sigma} | \boldsymbol{x} \sim \text{MVN-inv-Wishart}\left(\boldsymbol{\mu}_n, \frac{1}{\kappa_n}\mathbf{\Lambda}_n; \nu_n, \mathbf{\Lambda}_n\right).$$

with

$$\begin{aligned}
\boldsymbol{\mu}_n &= \frac{\kappa_0}{\kappa_0 + n}\boldsymbol{\mu}_\mathbf{0} + \frac{n}{\kappa_0 + n}\overline{\boldsymbol{x}}, \\
\kappa_n &= \kappa_0 + n, \\
\nu_n &= \nu_0 + n, \\
\mathbf{\Lambda}_n &= \mathbf{\Lambda}_0 + \boldsymbol{S} + \frac{\kappa_0 n}{\kappa_0 + n}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)^\top,
\end{aligned}$$

where $\boldsymbol{S} = \sum_{i=1}^n (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top$ by analogy to the univariate case.

4.5 Multi-Parameter Models
4.5.3 Multivariate Normal Distribution

The interpretation of the parameters of slide 435 can be directly transferred: the posterior expectation is a weighted mean of the sample mean vector and the prior expectation. The total dispersion matrix $\mathbf{\Lambda}_n$ can be partitioned into a prior dispersion matrix, an empirical dispersion matrix and a dispersion between prior expectation and sample mean (vector).

4.5 Multi-Parameter Models
4.5.3 Multivariate Normal Distribution

The **marginal posterior** of $\boldsymbol{\mu}$ is a multivariate t-distribution with parameters $\boldsymbol{\mu}_n$ and $\mathbf{\Lambda}_n/(\kappa_n \cdot (\nu_n - p + 1))$, the marginal posterior for $\mathbf{\Sigma}$ an inverse Wishart distribution with parameters ν_n and $\mathbf{\Lambda}_n^{-1}$. To simulate from the joint posterior or from the predictive distribution, the following algorithm can be used:

4.5 Multi-Parameter Models
4.5.3 Multivariate Normal Distribution

Algorithm 6: Simulation from the joint posterior and the predictive distribution with conjugated prior

```
For $s = 1, \ldots, S$:
    1. Draw $\mathbf{\Sigma}^{(s)} | \boldsymbol{x} \sim \text{inv-Wishart}_{\nu_n}\left(\mathbf{\Lambda}_n^{-1}\right)$.
    2. Draw $\boldsymbol{\mu}^{(s)} | \mathbf{\Sigma}^{(s)}, \boldsymbol{x} \sim \text{MVN}\left(\boldsymbol{\mu}_n, \frac{1}{\kappa_n}\mathbf{\Sigma}^{(s)}\right)$.
    3. Draw $\widetilde{\boldsymbol{x}}^{(s)} | \boldsymbol{\mu}^{(s)}, \mathbf{\Sigma}^{(s)}, \boldsymbol{x} \sim \text{MVN}\left(\boldsymbol{\mu}^{(s)}, \mathbf{\Sigma}^{(s)}\right)$.
Then $(\boldsymbol{\mu}^{(s)}, \mathbf{\Sigma}^{(s)})$ are draws from the joint posterior density, $\tilde{\boldsymbol{x}}^{(s)}$ a draw from the posterior predictive distribution.
```

Remark: density function and random generation from the Inverse Wishart distribution are e.g. implemented in the R package MCMCpack: riwish, diwish

III. Non-informative prior for unknown $\boldsymbol{\mu}$ and unknown $\mathbf{\Sigma}$

4.5 Multi-Parameter Models
4.5.3 Multivariate Normal Distribution

As a non-informative prior for unknown $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$, multivariate Jeffreys' prior is often used:

$$p(\boldsymbol{\mu}, \mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-(p+1)/2}.$$

This corresponds to the limit cases $\kappa_0 \to 0, \nu_0 \to 1, |\mathbf{\Lambda}_0| \to 0$ of the conjugate prior. For the posterior, the **conditional distribution** of $\boldsymbol{\mu}$

$$\boldsymbol{\mu} | \mathbf{\Sigma}, \boldsymbol{x} \sim \text{MVN}\left(\overline{\boldsymbol{x}}, \frac{1}{n}\mathbf{\Sigma}\right),$$

4.5 Multi-Parameter Models
4.5.3 Multivariate Normal Distribution

and for the **marginal distributions**

$$\boldsymbol{\Sigma}|\boldsymbol{x} \sim \text{inv-Wishart}_{n-1}(\boldsymbol{S}),$$
$$\boldsymbol{\mu}|\boldsymbol{x} \sim \text{mv-t}_{n-p}\left(\overline{\boldsymbol{x}}, \frac{1}{n(n-p)}\boldsymbol{S}\right).$$

In the following example, we will consider the bivariate normal distribution. With this example we illustrate an important simulation strategy for Bayesian inference, i.e. sampling from posterior distributions, by the so-called Gibbs sampling. The algorithm itself is first introduced generically.

4.5 Multi-Parameter Models
4.5.3 Multivariate Normal Distribution

Algorithm 7: Gibbs-Sampler for draws from a posterior distribution

Given: a multidimensional continuous random vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$ with *target* (posterior) distribution $f(\boldsymbol{\theta}|\boldsymbol{x})$ from which samples shall be generated.

Problem: Samples from the (joint) posterior cannot be drawn or are difficult to draw (target distribution is not a well-known distribution and/or many parameters).

Remark: $\theta_1, \dots, \theta_p$ can be scalar components, but also subvectors, i.e. $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k)$, $k < p$.

The goal is to create a Markov chain $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \dots$ with starting value $\boldsymbol{\theta}^{(0)}$ and stationary distribution $f(\boldsymbol{\theta}|\boldsymbol{x})$. Let $\boldsymbol{\theta}^{(t)}$ the current state of the Markov chain. Let further $\boldsymbol{\theta}^{(t)}$ be divided into k subvectors $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_2^{(t)}, \dots, \boldsymbol{\theta}_k^{(t)})$. Define

$$\boldsymbol{\theta}_{-s}^{(t)} = \left(\boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_2^{(t)}, \dots, \boldsymbol{\theta}_{s-1}^{(t)}, \boldsymbol{\theta}_{s+1}^{(t-1)}, \dots, \boldsymbol{\theta}_k^{(t-1)}\right)$$

for $s = 1, \dots, k$.

4.5 Multi-Parameter Models
4.5.3 Multivariate Normal Distribution

Algorithm 7: Gibbs-Sampler continued

Furthermore, the *full conditional distributions ("full conditionals")* are given by

$$f(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{-s}, \boldsymbol{x}) = \frac{f\left(\boldsymbol{\theta}_s, \boldsymbol{\theta}_{-s}, \boldsymbol{x}\right)}{\int f\left(\boldsymbol{\theta}_s, \boldsymbol{\theta}_{-s}, \boldsymbol{x}\right) d\boldsymbol{\theta}_s}$$

and draws from them are feasible.

Remark: remember that all conditional distributions are proportional to the joint distribution.

4.5 Multi-Parameter Models
4.5.3 Multivariate Normal Distribution

Algorithm 7: Gibbs-Sampler continued

```
Then the next state $\boldsymbol{\theta}^{(t+1)}$ is reached by updating component-wise:
Step 1: draw $\boldsymbol{\theta}_1^{(t+1)} \sim f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_{-1}^{(t)}, \boldsymbol{x})$
Step 2: draw $\boldsymbol{\theta}_2^{(t+1)} \sim f(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_3^{(t)}, \dots, \boldsymbol{\theta}_k^{(t)}, \boldsymbol{x})$
Step 3: draw $\boldsymbol{\theta}_3^{(t+1)} \sim f(\boldsymbol{\theta}_3|\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \boldsymbol{\theta}_4^{(t)}, \dots, \boldsymbol{\theta}_k^{(t)}, \boldsymbol{x})$
  ⋮
Step k: draw $\boldsymbol{\theta}_k^{(t+1)} \sim f(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{-k}^{(t+1)}, \boldsymbol{x})$.
Repeat these steps "enough" times.
```

4.5 Multi-Parameter Models
4.5.3 Multivariate Normal Distribution

After a certain number of repetitions, $\boldsymbol{\theta}^{(t)}$ can be assumed to be a draw from the target distribution $f(\boldsymbol{\theta}|\boldsymbol{x})$. In contrast to the above "direct" simulation algorithms, however, the draws are now dependent.

4.5 Multi-Parameter Models
4.5.3 Multivariate Normal Distribution

Example 4.2. (Bivariate Normal Distribution) Let $\boldsymbol{x}$ have a bivariate normal distribution with mean vector $(\mu_1, \mu_2)^\top$ and covariance matrix

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \text{ und } \rho \text{ known.}$$

With a non-informative prior $p(\mu_1, \mu_2) \propto$ const for μ_1, μ_2 , the posterior is reduced to the likelihood for given data $\boldsymbol{x} = ((x_{11}, x_{12})^\top, \dots, (x_{n1}, x_{n2})^\top)$,

$$L(\mu_1, \mu_2) = \left(\frac{1}{2\pi}\right)^n \left(1-\rho^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2(1-\rho^2)}\,A\right),$$

in which

$$A = \sum_{i=1}^n \left[(x_{i1}-\mu_1)^2 - 2\rho(x_{i1}-\mu_1)(x_{i2}-\mu_2) + (x_{i2}-\mu_2)^2\right]$$

4.5 Multi-Parameter Models
4.5.3 Multivariate Normal Distribution

We calculate the full conditional distributions $\mu_1|\mu_2, \boldsymbol{x}$ and $\mu_2|\mu_1, \boldsymbol{x}$. Of course, for symmetry reasons, it is sufficient to show this for $\mu_1|\mu_2, \boldsymbol{x}$. Because of

$$f(\mu_1|\mu_2, x) = \frac{f(\mu_1, \mu_2|\boldsymbol{x})}{f(\mu_2|\boldsymbol{x})} = \frac{f(\mu_1, \mu_2, \boldsymbol{x})}{f(\mu_2, \boldsymbol{x})} \propto f(\mu_1, \mu_2|x)\ ,$$

it is sufficient to look only on terms in the joint posterior that depend on the variable in the conditional distribution (but e.g. those who depend on the condition can be ignored). One gets

$$f(\mu_1|\mu_2, \boldsymbol{x}) \propto \exp\left(-\frac{1}{2(1-\rho^2)}\, n\left[\mu_1^2 - 2\mu_1(\overline{x}_1 + \rho(\mu_2 - \overline{x}_2))\right]\right)$$

with $\bar{x}_j = (\sum_{i=1}^n x_{ij})/n$ for $j = 1, 2$.

4.5 Multi-Parameter Models
4.5.3 Multivariate Normal Distribution

A quadratic completion of the term in square brackets by $\overline{x}_1 + \rho(\mu_2 - \overline{x}_2)$ provides the result

$$p(\mu_1|\mu_2, \boldsymbol{x}) \propto \exp\left(-\frac{1}{2\frac{1-\rho^2}{n}}\left(\mu_1 - [\overline{x}_1 + \rho(\mu_2 - \overline{x}_2)]\right)^2\right).$$

This corresponds to the kernel of a $N(\overline{x}_1 + \rho(\mu_2 - \overline{x}_2), (1-\rho^2)/n)$-distribution. The associated Gibbs sampler has the form:

1. Choose a starting value $\mu_2^{(0)}$.
2. For $s = 1, \ldots, S$:
   (a) Draw $\mu_1^{(s)}|\mu_2^{(s-1)} \sim N\left(\overline{x}_1 + \rho\left(\mu_2^{(s-1)} - \overline{x}_2\right), \frac{1-\rho^2}{n}\right)$.
   (b) Draw $\mu_2^{(s)}|\mu_1^{(s)} \sim N\left(\overline{x}_2 + \rho\left(\mu_1^{(s)} - \overline{x}_1\right), \frac{1-\rho^2}{n}\right)$.

4.6 Bayesian Linear Model

4.6 Bayesian Linear Model

Model:

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

where $\boldsymbol{y} \in \mathbb{R}^n, \boldsymbol{X} \in \mathbb{R}^{n\times p}, \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\varepsilon} \in \mathbb{R}^n$ *Assumptions and notation:*

$$\begin{aligned} p &= \operatorname{rank}(\boldsymbol{X}) \\ \boldsymbol{\varepsilon} &= (\varepsilon_1, \ldots, \varepsilon_n)^\top, \quad \varepsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2) \end{aligned}$$

4.6 Bayesian Linear Model

Bayesian:

$$\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{X} \ \sim \ \text{MVN}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$$

Likelihood:

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) &\propto |\sigma^2\boldsymbol{I}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right) \\ &= (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right) \end{aligned}$$

4.6.1 Non Informative Prior Distribution

4.6 Bayesian Linear Model 4.6.1 Non-Informative Prior Distribution

The **non-informative prior**

$$p(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-1}$$

is particularly useful in the case of $p \ll n$. For the **joint posterior** it follows:

$$f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X}) \propto (\sigma^2)^{-(\frac{n}{2}+1)} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right).$$

Let

$$\begin{aligned} &\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}, \\ &\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \underbrace{\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top}_{\boldsymbol{H}}\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}, \\ &\widehat{\boldsymbol{\varepsilon}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{y} - \widehat{\boldsymbol{y}}. \end{aligned}$$

4.6 Bayesian Linear Model 4.6.1 Non-Informative Prior Distribution

It is known from the theory of linear models, that

$$\boldsymbol{X}^\top\widehat{\boldsymbol{\varepsilon}} = 0, \quad \widehat{\boldsymbol{y}}^\top\widehat{\boldsymbol{\varepsilon}} = 0.$$

This results in the following transformations:

$$\begin{aligned}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) &= [(\boldsymbol{y}-\widehat{\boldsymbol{y}})+(\widehat{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})]^\top[(\boldsymbol{y}-\widehat{\boldsymbol{y}})+(\widehat{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})]\\ &= \widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}+(\widehat{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})^\top(\widehat{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})+2(\widehat{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})^\top\widehat{\boldsymbol{\varepsilon}}\\ &= \widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}+(\widehat{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})^\top(\widehat{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})\\ &\overset{\widehat{\boldsymbol{y}}=\boldsymbol{X}\widehat{\boldsymbol{\beta}}}{=} \widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}+(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}),\end{aligned}$$

so that the posterior can be written as

$$f(\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\boldsymbol{X}) \propto (\sigma^2)^{-(\frac{n}{2}+1)}\exp\left(-\frac{1}{2\sigma^2}\left(\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}+(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right)\right).$$

4.6 Bayesian Linear Model 4.6.1 Non-Informative Prior Distribution

The **conditional posterior** of $\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}$ is

$$f(\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}) \propto \exp\left(-\frac{1}{2\sigma^2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right),$$

as $\widehat{\boldsymbol{\varepsilon}}=\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ does not depend on $\boldsymbol{\beta}$. One identifies the above expression as the kernel of a multivariate normal distribution, precisely as

$$f(\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}) \;\sim\; \text{MVN}(\widehat{\boldsymbol{\beta}},\sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}).$$

Remark: In linear model theory, we have the result

$$f(\widehat{\boldsymbol{\beta}}|\sigma^2,\boldsymbol{y},\boldsymbol{X}) \;\sim\; \text{MVN}(\boldsymbol{\beta},\sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}).$$

4.6 Bayesian Linear Model 4.6.1 Non-Informative Prior Distribution

The **marginal posterior** of σ^2 is obtained by integrating out $\boldsymbol{\beta}$ from the joint posterior or, more simply, using Bayes' theorem

$$f(\sigma^2|\boldsymbol{y},\boldsymbol{X}) = \frac{f(\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\boldsymbol{X})}{f(\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X})}.$$

The normalization constant for the conditional posterior of $\boldsymbol{\beta}$ is $(\sigma^2)^{-p/2}$, so that

$$\begin{aligned}f(\sigma^2|\boldsymbol{y},\boldsymbol{X}) &\propto (\sigma^2)^{-(\frac{n}{2}+1)}(\sigma^2)^{\frac{p}{2}}\exp\left(-\frac{1}{2\sigma^2}\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}\right)\\ &= (\sigma^2)^{-(\frac{n-p}{2}+1)}\exp\left(-\frac{1}{2\sigma^2}\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}\right).\end{aligned}$$

This is the kernel of an inv-$\chi^2\left(n-p,\frac{\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}}{n-p}\right)$ or IG $\left(\frac{n-p}{2},\frac{\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}}{2}\right)$-distribution. The following applies:

$$\mathbb{E}[\sigma^2|\boldsymbol{y},\boldsymbol{X}] = \frac{n-p}{n-p-2}\cdot\frac{\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}}{n-p} = \frac{\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}}{n-p-2}.$$

4.6 Bayesian Linear Model 4.6.1 Non-Informative Prior Distribution

Algorithm 8: Direct simulation of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma^2}$ in the Bayesian linear model

For $t=1,\dots,T$:

1. Draw $\left(\sigma^2\right)^{(t)}$ from $f\left(\sigma^2|\boldsymbol{y},\boldsymbol{X}\right)$, i.e. from inv-$\chi^2\left(n-p,\frac{\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}}{n-p}\right)$.
2. Draw $\boldsymbol{\beta}^{(t)}$ from $f\left(\boldsymbol{\beta}|\left(\sigma^2\right)^{(t)},\boldsymbol{y},\boldsymbol{X}\right)$, i.e. from MVN $\left(\widehat{\boldsymbol{\beta}},\left(\sigma^2\right)^{(t)}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right)$, in which $\widehat{\boldsymbol{\beta}}=(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$.

4.6 Bayesian Linear Model 4.6.1 Non-Informative Prior Distribution

An alternative to direct simulation is the use of the Gibbs sampler. In addition to the full conditional density of $\boldsymbol{\beta}$, the full conditional of σ^2,

$$\begin{aligned} f(\sigma^2|\boldsymbol{\beta}, \boldsymbol{y},\boldsymbol{X}) &\propto p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X}) \\ &\propto (\sigma^2)^{-(\frac{n}{2}+1)} \exp\left(-\frac{1}{2\sigma^2}\left(\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}} + (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right)\right), \end{aligned}$$

is used for simulation. For fixed $\boldsymbol{\beta}$ this is the kernel of a scaled inv-$\chi^2\left(n, \frac{\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}+(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{n}\right)$-distribution.

4.6 Bayesian Linear Model 4.6.1 Non-Informative Prior Distribution

This also allows the **marginal posterior** of $\boldsymbol{\beta}$ to be derived:

$$\begin{aligned} f(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) = \frac{f(\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\boldsymbol{X})}{f(\sigma^2|\boldsymbol{y},\boldsymbol{X},\boldsymbol{\beta})} &= \frac{f(\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X})\cdot f(\sigma^2|\boldsymbol{y},\boldsymbol{X})}{f(\sigma^2|\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{X})} \\ \propto \frac{\exp\left(-\frac{1}{2\sigma^2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right)}{\left[\frac{\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}+(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{n}\right]^{n/2}\exp\left(-\frac{1}{2\sigma^2}[\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}+(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})]\right)}&. \end{aligned}$$

so that:

$$f(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) \propto \left[\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}+(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right]^{-n/2}.$$

Remark: Note the marginal density $f(\sigma^2|\boldsymbol{y},\boldsymbol{X})$ can be ignored, since no terms depend on the argument $\boldsymbol{\beta}$ on the LHS.

4.6 Bayesian Linear Model 4.6.1 Non-Informative Prior Distribution

If one sets

$$\widehat{\sigma}^2_\varepsilon = \frac{\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}}{n-p} \;\Leftrightarrow\; \widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}} = (n-p)\widehat{\sigma}^2_\varepsilon,$$

one gets

$$\begin{aligned} f(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) &\propto \left[(n-p)\widehat{\sigma}^2_\varepsilon + (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right]^{-n/2} \\ &= \left((n-p)\widehat{\sigma}^2_\varepsilon\cdot\left[1+\frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{(n-p)\widehat{\sigma}^2_\varepsilon}\right]\right)^{-\frac{n}{2}} \\ &\propto \left[1+\frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{(n-p)\widehat{\sigma}^2_\varepsilon}\right]^{-\frac{(n-p)+p}{2}}. \end{aligned}$$

4.6 Bayesian Linear Model 4.6.1 Non-Informative Prior Distribution

This corresponds to the kernel of a multivariate t-distribution with $n-p$ degrees of freedom, location parameter $\widehat{\boldsymbol{\beta}}$ and scale parameter $\sigma^2_\varepsilon(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$, such that

$$\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X} \;\sim\; \text{mv-}t_{n-p}\left(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2_\varepsilon(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right).$$

Finally we consider the **predictive distribution** for $\widetilde{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}$. Let

- m the number of new observations,
- $\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{x}_1},\ldots,\widetilde{\boldsymbol{x}_m})^\top$ new observations of regressors of dimension $m\times p$, e.g. $\widetilde{\boldsymbol{x}_1}$ is $p\times 1$, $\widetilde{\boldsymbol{x}_1}^\top$ is then $1\times p$.

- $\widetilde{\boldsymbol{y}} = (\widetilde{y_1}, \dots, \widetilde{y_m})^\top$ is the vector of the forecasts of dimension $m \times 1$.

For the simulation we can extend algorithm 8 as follows:

4.6 Bayesian Linear Model 4.6.1 Non-Informative Prior Distribution

```
Algorithm 9: Direct simulation of the predictive distribution in the Bayesian linear model
For $t = 1, \dots, T$ :
    1. Draw $\left(\sigma^2\right)^{(t)}$ from $f\left(\sigma^2|\boldsymbol{y}, \boldsymbol{X}\right)$, i.e. from inv-$\chi^2\left(n-p, \frac{\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}}{n-p}\right)$.
    2. Draw $\boldsymbol{\beta}^{(t)}$ from $f\left(\boldsymbol{\beta}|\left(\sigma^2\right)^{(t)}, \boldsymbol{y}, \boldsymbol{X}\right)$, i.e. from MVN $\left(\widehat{\boldsymbol{\beta}}, \left(\sigma^2\right)^{(t)}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right)$, where $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$.
    3. For $i = 1, \dots, m$:
        Draw $\widetilde{y}_i \sim MVN\left(\widetilde{\boldsymbol{x}_i}^\top\boldsymbol{\beta}^{(t)}, (\sigma^2)^{(t)}\right)$, where $\widetilde{\boldsymbol{x}_i}^\top$ is the $i$-th row of $\widetilde{\boldsymbol{X}}$ .
```

4.6 Bayesian Linear Model 4.6.1 Non-Informative Prior Distribution

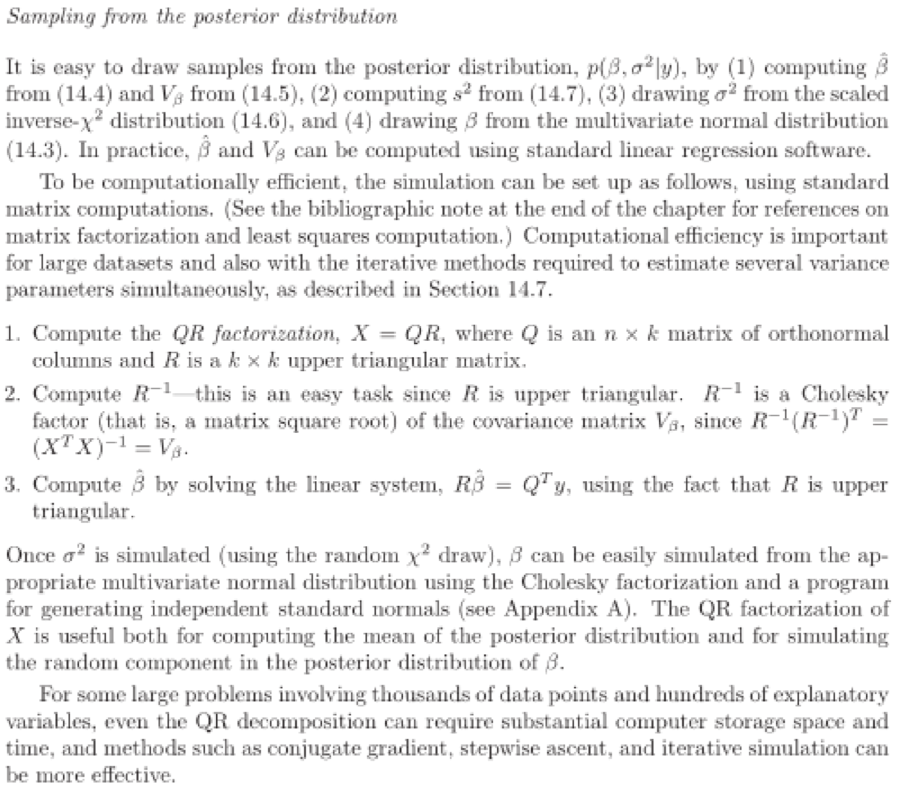
An analytical calculation is even possible::

$$f(\widetilde{\boldsymbol{y}}|\boldsymbol{y}, \boldsymbol{X}, \widetilde{\boldsymbol{X}}) \sim \text{mv-t}_{n-p}\left[\widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}}, \hat{\sigma}^2_\varepsilon\left(\widetilde{\boldsymbol{X}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\widetilde{\boldsymbol{X}}^\top + \boldsymbol{I}\right)\right]$$

in analogy to the calculation of forecasts and forecast intervals for linear models from a frequentist point of view.

4.6 Bayesian Linear Model 4.6.1 Non-Informative Prior Distribution

Computational remark (from the BDA book, 3rd edition, p. 356):



Sampling from the posterior distribution

It is easy to draw samples from the posterior distribution, $p(\beta, \sigma^2|y)$, by (1) computing $\hat{\beta}$ from (14.4) and V_β from (14.5), (2) computing s^2 from (14.7), (3) drawing σ^2 from the scaled inverse-χ^2 distribution (14.6), and (4) drawing β from the multivariate normal distribution (14.3). In practice, $\hat{\beta}$ and V_β can be computed using standard linear regression software.

To be computationally efficient, the simulation can be set up as follows, using standard matrix computations. (See the bibliographic note at the end of the chapter for references on matrix factorization and least squares computation.) Computational efficiency is important for large datasets and also with the iterative methods required to estimate several variance parameters simultaneously, as described in Section 14.7.

1. Compute the *QR factorization*, $X = QR$, where Q is an $n \times k$ matrix of orthonormal columns and R is a $k \times k$ upper triangular matrix.
2. Compute R^{-1}—this is an easy task since R is upper triangular. R^{-1} is a Cholesky factor (that is, a matrix square root) of the covariance matrix V_β, since $R^{-1}(R^{-1})^T = (X^TX)^{-1} = V_\beta$.
3. Compute $\hat{\beta}$ by solving the linear system, $R\hat{\beta} = Q^Ty$, using the fact that R is upper triangular.

Once σ^2 is simulated (using the random χ^2 draw), β can be easily simulated from the appropriate multivariate normal distribution using the Cholesky factorization and a program for generating independent standard normals (see Appendix A). The QR factorization of X is useful both for computing the mean of the posterior distribution and for simulating the random component in the posterior distribution of β.

For some large problems involving thousands of data points and hundreds of explanatory variables, even the QR decomposition can require substantial computer storage space and time, and methods such as conjugate gradient, stepwise ascent, and iterative simulation can be more effective.

4.6.2 Conjugate Prior Distribution

4.6 Bayesian Linear Model 4.6.2 Conjugate Prior Distribution

Use of the **conjugate prior**

$$\sigma^2 \sim \text{inv-}\chi^2(\kappa_0, \sigma_0^2) \ , \quad \beta|\sigma^2 \sim \text{MVN}(\boldsymbol{\beta}_0, \sigma^2\boldsymbol{\Sigma}_0)$$

or

$$\boldsymbol{\beta}, \sigma^2 \sim \text{MVN-inv-}\chi^2(\boldsymbol{\beta}_0, \sigma_0^2\boldsymbol{\Sigma}_0; \kappa_0, \sigma_0^2)$$

results in the **joint posterior**

$$\sigma^2|\boldsymbol{y}, \boldsymbol{X} \sim \text{inv-}\chi^2(\kappa_n, \sigma_n^2) \ , \quad \boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X} \sim \text{MVN}(\boldsymbol{\beta}_n, \sigma^2\boldsymbol{\Sigma}_n)$$

or

$$\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X} \sim \text{MVN-inv-}\chi^2(\boldsymbol{\beta}_n, \sigma_n^2\boldsymbol{\Sigma}_n; \kappa_n, \sigma_n^2) \ ,$$

4.6 Bayesian Linear Model 4.6.2 Conjugate Prior Distribution

where

$$\begin{aligned}
\boldsymbol{\beta}_n &= (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{X}^\top\boldsymbol{y}) \ , \\
\boldsymbol{\Sigma}_n &= (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{X}^\top\boldsymbol{X})^{-1} \ , \\
\kappa_n &= \kappa_0 + n \ , \\
\sigma_n^2 &= (\boldsymbol{\beta}_0^\top\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0 - \boldsymbol{\beta}_n^\top\boldsymbol{\Sigma}_n^{-1}\boldsymbol{\beta}_n + \boldsymbol{y}^\top\boldsymbol{y} + \kappa_0\sigma_0^2)/(\kappa_0 + n) \ .
\end{aligned}$$

The **conditional posterior** of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} \ |\sigma^2, \boldsymbol{y}, \boldsymbol{X} \quad \sim \quad \text{MVN}(\boldsymbol{\beta}_n, \sigma^2\boldsymbol{\Sigma}_n),$$

and the **marginal posterior** of σ^2 is

$$\sigma^2|\boldsymbol{y}, \boldsymbol{X} \quad \sim \quad \text{inv-}\chi^2(\kappa_n, \sigma_n^2).$$

4.6 Bayesian Linear Model 4.6.2 Conjugate Prior Distribution

The full conditionals are

$$\begin{aligned}
\sigma^2|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{X} &\sim \text{inv-}\chi^2(\kappa_0 + n + p, \sigma_0^2/\kappa_0 + \widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}/n + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)) \\
\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X} &\sim \text{MVN}\left(\left(\frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{\Sigma}_0^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{y} + \frac{1}{\sigma^2}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0\right), \right. \\
&\qquad \left. \left(\frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{\Sigma}_0^{-1}\right)^{-1}\right)
\end{aligned}$$

4.6.3 Special Cases and Extensions

4.6 Bayesian Linear Model 4.6.3 Special Cases and Extensions

1. **Ridge-Regression**

   Note: It generally makes sense to standardize covariates (without intercept) to eliminate differences in the scale. Furthermore, one centers the response so that the intercept vanishes. Let the final design matrix and response then $\boldsymbol{X}^*, \boldsymbol{y}^*$. Now

   $$\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}, \ \boldsymbol{\varepsilon} \sim N(0, \sigma^2\boldsymbol{I}).$$

   The ridge estimator is given by

   $$\hat{\boldsymbol{\beta}}_{\text{R}} = [(\boldsymbol{X}^*)^\top\boldsymbol{X}^* + \lambda\boldsymbol{I}]^{-1}(\boldsymbol{X}^*)^\top\boldsymbol{y}^*$$

   with $\lambda > 0$.

4.6 Bayesian Linear Model 4.6.3 Special Cases and Extensions

This can be interpreted in a Bayesian way as follows: Let

$$p(\boldsymbol{\beta}^*) \sim N(0, \tau^2 \boldsymbol{I}),$$

i.e. the components of $\boldsymbol{\beta}^*$ are uncorrelated in the prior (and thus also independent because of the normal assumption). Then the conditional posterior $f\left(\boldsymbol{\beta}^* | \boldsymbol{y}^*, \boldsymbol{X}^*, \sigma^2, \tau^2\right)$ is

$$\text{MVN}\left(\left[(\boldsymbol{X}^*)^\top \boldsymbol{X}^* + \frac{\sigma^2}{\tau^2}\boldsymbol{I}\right]^{-1} (\boldsymbol{X}^*)^\top \boldsymbol{y}^*, \sigma^2 \left((\boldsymbol{X}^*)^\top \boldsymbol{X}^* + \frac{\sigma^2}{\tau^2}\boldsymbol{I}\right)^{-1}\right).$$

The parameter λ corresponds to the quotient σ^2/τ^2.

4.6 Bayesian Linear Model 4.6.3 Special Cases and Extensions

2. **Unequal variances of the error variables / dependent error variables**
   Generally:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$$
$$\boldsymbol{y} \sim \text{MVN}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{y}}),\ \boldsymbol{\Sigma}_{\boldsymbol{y}} = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$$

   Problem: Specification of the prior distribution of $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$.

4.6 Bayesian Linear Model 4.6.3 Special Cases and Extensions

Possible ways out:

(a) Parametrize

$$\boldsymbol{\Sigma}_{\boldsymbol{y}} = \sigma^2 \boldsymbol{Q}_{\boldsymbol{y}}$$

with $\boldsymbol{Q}_{\boldsymbol{y}}$ known and $p(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-1}$. This case can be reduced to the model from section 4.6.1 by considering the model

$$\underbrace{\boldsymbol{Q}^{-1/2}\boldsymbol{y}}_{\boldsymbol{y}^*} = \underbrace{\boldsymbol{Q}^{-1/2}\boldsymbol{X}}_{\boldsymbol{X}^*}\beta + \underbrace{\boldsymbol{Q}^{-1/2}\varepsilon}_{\boldsymbol{\varepsilon}^*}$$

Then we get again a homoscedastic regression model in $\boldsymbol{y}^*, \boldsymbol{X}^*, \boldsymbol{\varepsilon}^*$.

4.6 Bayesian Linear Model 4.6.3 Special Cases and Extensions

(b) Weighted regression:

$$\boldsymbol{\Sigma}_{\boldsymbol{y}} = \text{diag}(\sigma^2 w_i^{-1})_{1 \le i \le n}$$

can be seen as a special case of (a) .

4.6 Bayesian Linear Model 4.6.3 Special Cases and Extensions

(c) Correlations: write

$$\boldsymbol{\Sigma}_{\boldsymbol{y}} = \boldsymbol{SRS} \text{ with } \boldsymbol{S} = \text{diag}(\sigma_1, \ldots, \sigma_p)$$

with, for example

$$p(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2) = \prod_{j=1}^{p} p(\sigma_j^2) \quad \text{and} \quad p(\sigma_j^2) \sim \text{inv-}\chi^2(\nu_j, \sigma_{0j}^2).$$

The specification of the correlation matrix remains to be done. Prior specifications in particular must guarantee positive definiteness. A simple variant is the use of (positive) ”equi-correlation”, which may be a reasonable assumption for cluster data:

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix},$$

4.6 Bayesian Linear Model 4.6.3 Special Cases and Extensions

where $\rho \sim U(0,1)$ forces a positive correlation. For repeated measurements, an autoregressive covariance structure may be used. E.g., for 1st order autocorrelated error terms

$$\varepsilon_t = \rho\varepsilon_{t-1} + Z_t,\ Z_t \sim N(0,\sigma^2),$$

we get

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{p-1} \\ \rho & 1 & \rho & \cdots & \rho^{p-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & \cdots & 1 \end{bmatrix}.$$

Other decompositions are based on the Cholesky- or spectral decomposition and may be relatively complex.

4.6 Bayesian Linear Model 4.6.3 Special Cases and Extensions

(d) Switching to models with random effects: models with random effects (linear mixed models, generalized linear mixed models) lead to structured covariance matrices with a relatively low number of parameters.

But: the model equation changes (contains now also random effects) and $\boldsymbol{\Sigma_y} \neq \boldsymbol{\Sigma_\varepsilon}$, i.e. we get into a different class of models.

4.7 Bayesian Generalized Linear Model

4.7 Bayesian Generalized Linear Model

Structure of GLMs: The response follows a distribution from a simple exponential family ($i = 1,\dots,n$)

$$f(y_i|\theta_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi_i}\right)\cdot c(y_i,\phi_i) \tag{4.1}$$

or

$$f(y_i|\theta_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)}\right)\cdot c(y_i,\phi_i),$$

where in many cases $\phi_i \equiv \phi$ (Bernoulli-, Poisson distribution). It is

$$\mu_i = \mathbb{E}[y_i|\theta_i] = b'(\theta_i)\ ,\quad \mathrm{Var}(y_i|\theta_i) = b''(\theta_i)\phi_i$$

and θ_i the canonical parameter. With a link function g and response function $h = g^{-1}$ let

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top\boldsymbol{\beta}. \tag{4.2}$$

4.7 Bayesian Generalized Linear Model

Example 4.3 (Logit-Model) With $\mu_i = \mathbb{P}(y_i = 1)$

$$\begin{aligned} f(y_i|\mu_i) &= \mu_i^{y_i}(1-\mu_i)^{1-y_i} \\ &= \exp\big(y_i\log(\mu_i) + (1-y_i)\log(1-\mu_i)\big) \\ &= \exp\left(y_i \underbrace{\log\left(\frac{\mu_i}{1-\mu_i}\right)}_{\theta_i} + \log(1-\mu_i)\right) \end{aligned}$$

with

$$\theta_i = \log\left(\frac{\mu_i}{1-\mu_i}\right) \Leftrightarrow \mu_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}\ .$$

This corresponds to (4.1) with $\phi_i = 1,\ c(y_i, \phi_i) = 1$,

4.7 Bayesian Generalized Linear Model

$$
\begin{aligned}
b(\theta_i) &= -\log\left(1 - \frac{\exp(\theta_i)}{1+\exp(\theta_i)}\right)\\
&= -\log\left(\frac{1}{1+\exp(\theta_i)}\right)\\
&= \log\left(1+\exp(\theta_i)\right)
\end{aligned}
$$

and

$$
b'(\theta_i) = \frac{1}{1+\exp(\theta_i)} \cdot \exp(\theta_i) = \mu_i.
$$

4.7 Bayesian Generalized Linear Model

- A suitable prior distribution for $\boldsymbol{\beta}$ in (4.2) is
$$
\boldsymbol{\beta} \ \sim \ \mathrm{MVN}(\boldsymbol{\beta}_0, \boldsymbol{B}_0)
$$
with $\boldsymbol{B}_0 > 0$ (see section 4.5.3 on the multivariate Normal distribution with known covariance matrix).
- $\boldsymbol{\beta}$ influences $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ through the predictor $\boldsymbol{\mu(\beta)} = h(\boldsymbol{X\beta})$, where the application of h is component-by-component.
- For $\boldsymbol{B}_0^{-1} \to \boldsymbol{0}$ one obtains a non-informative prior.

4.7 Bayesian Generalized Linear Model

Using the representation (4.1) as an exponential family with canonical parameters, the posterior distribution is

$$
\begin{aligned}
f(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) &\propto \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top \boldsymbol{B}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right) \cdot \prod_{i=1}^n \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi_i}\right)\\
&= \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top \boldsymbol{B}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right) \exp\left(\sum_{i=1}^n \frac{y_i\theta_i - b(\theta_i)}{\phi_i}\right)\\
&= \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top \boldsymbol{B}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + \sum_{i=1}^n \frac{y_i\theta_i - b(\theta_i)}{\phi_i}\right).
\end{aligned}
$$

4.7 Bayesian Generalized Linear Model

Example 4.4 (Logit-Model) In the case of the logit model, the posterior is therefore

$$
\begin{aligned}
f(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) &\propto \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top \boldsymbol{B}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right)\\
&\quad \times \prod_{i=1}^n \mu_i(\boldsymbol{\beta})^{y_i}(1-\mu_i(\boldsymbol{\beta}))^{1-y_i}\\
&= \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top \boldsymbol{B}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right)\\
&\quad \times \prod_{i=1}^n h(\boldsymbol{x}_i^\top\boldsymbol{\beta})^{y_i}(1-h(\boldsymbol{x}_i^\top\boldsymbol{\beta}))^{1-y_i}\\
&= \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top \boldsymbol{B}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right)\\
&\quad \times \prod_{i=1}^n \left(\frac{\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}\right)^{y_i}\left(\frac{1}{1+\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}\right)^{1-y_i}.
\end{aligned}
$$

4.7 Bayesian Generalized Linear Model

Problem: The kernel of the posterior does not correspond to the kernel of a known distribution. The posterior distribution is therefore not analytically accessible. Possible ways out are

1. Approximation or
2. Exploration of the posterior by generating samples from the posterior.

We consider Solution 2 below. There are several possibilities. A proposal from Gamerman is very well established(1997), which is a variant of the *Metropolis-Hastings-Algorithm.* [1]

4.7.1 A MCMC algorithm: Metropolis-Hastings

4.7 Bayesian Generalized Linear Model 4.7.1 A MCMC algorithm: Metropolis-Hastings

First, a description of the basic problem follows, without going further into the details of the underlying mathematical theory. It is known that for $X_i \overset{i.i.d}{\sim} \pi$, $i = 1, \dots, n$, where π denotes a distribution, characteristics of interest of this distribution (moments, density etc., provided they exist) can be estimated by simulating random numbers according to π, i.e. a Monte Carlo estimate can be obtained. Examples are

$$\begin{aligned}
\widehat{\mathbb{E}[X]} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
\widehat{\mathbb{E}[X^2]} &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 \\
\mathbb{E}[\widehat{\exp(X)}] &= \frac{1}{n}\sum_{i=1}^{n} \exp(x_i) \ .
\end{aligned}$$

4.7 Bayesian Generalized Linear Model 4.7.1 A MCMC algorithm: Metropolis-Hastings

- This is little "exciting", because, if π is known, the expected value is also usually known. However, suppose one considers a (nonlinear) function of X, e.g. $g(X) = X^2$. Then the density of the transformation $g(X)$ can still be determined analytically, but it may be complex to calculate the expected value.
- In the case that X is multidimensional, the analytical determination of such parameters is often impossible and at higher dimensions can be too unstable by using e.g. numerical integration methods.

4.7 Bayesian Generalized Linear Model 4.7.1 A MCMC algorithm: Metropolis-Hastings

Under suitable conditions, the above Monte-Carlo estimate can be extended to

$$\mathbb{E}[\widehat{g(X)}] = \frac{1}{n}\sum_{i=1}^{n} g(x_i).$$

(this is a general principle, so not necessarily Bayesian, as long as π is not, for example, a posterior distribution.)

4.7 Bayesian Generalized Linear Model 4.7.1 A MCMC algorithm: Metropolis-Hastings

It should be noted, however, that this procedure is, compared to the exact solution, afflicted with a Monte Carlo error. Another essential prerequisite is that random numbers from π can be drawn. Methods for generating iid random numbers are for example

- the inversion method,

[1] Gamerman (1997): *Sampling from the posteriori distribution in generalized linear models.* Statistics and Computing **7**, pp. 57-68.

- rejection sampling or
- importance Sampling.

Especially with higher dimensions, however, these are often not applicable or only very difficult to use.

4.7 Bayesian Generalized Linear Model 4.7.1 A MCMC algorithm: Metropolis-Hastings

Markov Chain Monte Carlo (MCMC) is an alternative method. The aim is to generate a Markov chain $(X_0, \ldots, X_n)$ of (dependent!) random numbers whose distribution converges to the distribution of interest, i.e. π is the stationary or invariant distribution of the Markov chain. The *ergodic theorem* then allows estimates of the form

$$\widehat{\mathbb{E}[X]} = \frac{1}{n - \text{burnin}} \sum_{i=\text{burnin}+1}^{n} x_i \quad \text{or} \quad \widehat{\mathbb{E}[g(X)]} = \frac{1}{n - \text{burnin}} \sum_{i=\text{burnin}+1}^{n} g(x_i),$$

where $x_0, \ldots, x_{\text{burnin}}$ denote values at the beginning of the sequence, before the chain is in the stationary distribution, and therefore are "thrown away" .

4.7 Bayesian Generalized Linear Model 4.7.1 A MCMC algorithm: Metropolis-Hastings

Practical implementation: Start with a starting value x_0 and then draw for $i = 1, \ldots, n$ values $X_i \sim P(\cdot|X_{i-1})$, where P is the Markov transition kernel which only depends on the current state of the chain.

The kernel, respectively the Markov chain, must fulfill the following requirements:

1. The Markov chain is homogeneous.
2. The Markov chain is irreducible.
3. The Markov chain is aperiodic.
4. The Markov chain is positively recurrent.

4.7 Bayesian Generalized Linear Model 4.7.1 A MCMC algorithm: Metropolis-Hastings

- We consider Markov chains in discrete time with discrete or continuous state space, usually a subset of $\mathbb{R}^p$.
- For general state spaces, "more technique" is required, but no new ideas.
- For the case considered here, the target distribution π is always given, e.g. as posterior distribution (up to a normalization constant).

4.7 Bayesian Generalized Linear Model 4.7.1 A MCMC algorithm: Metropolis-Hastings

Univariate Metropolis-Hastings

We now describe the *Metropolis-Hastings algorithm* (in short: *MH*) to generate a Markov chain as described above for the univariate case; this algorithm contains the Gibbs sampler as a special case. Let π be the density of the target distribution from which we want to simulate, and let q a suitable so-called proposal density from which the new chain states are generated, i.e.

$$X_i \sim q(\cdot|x_{i-1}),$$

for example $q_{x_i|x_{i-1}} = N(x_{i-1}, 1)$ or $q_{x_i|x_{i-1}} = U(x_{i-1} - c, x_{i-1} + c)$.

4.7 Bayesian Generalized Linear Model 4.7.1 A MCMC algorithm: Metropolis-Hastings

The proposals are not always accepted but only with a certain *acceptance probability* $\alpha(x_{i-1}, x_i)$. For the MH algorithm, this probability is

$$\alpha(x_{i-1}, x_i) = \min\left(1, \frac{\pi(x_i) \cdot q(x_{i-1}|x_i)}{\pi(x_{i-1}) \cdot q(x_i|x_{i-1})}\right).$$

If x_i is not accepted, then one sets $x_i \leftarrow x_{i-1}$, i.e. the old state is maintained.

- A major benefit of this procedure is that the (mostly unknown) normalization constant of π cancels out, i.e. the MH algorithm can be especially applied for these cases.
- The construction of α ensures that conditions 1 to 4 are met.

4.7 Bayesian Generalized Linear Model 4.7.1 A MCMC algorithm: Metropolis-Hastings

For $q(x_{i-1}|x_i) = q(x_i|x_{i-1})$ the MH algorithm is reduced to the *Metropolis-Algorithm* with

$$\alpha(x_{i-1}, x_i) = \min\left(1, \frac{\pi(x_i)}{\pi(x_{i-1})}\right),$$

i.e. if the target density evaluated at the point x_i is greater than at x_{i-1}, the new proposal is always accepted, otherwise only with probability $\pi(x_i)/\pi(x_{i-1})$. If one sets the acceptance probability to one, one obtains the Gibbs sampler.

4.7 Bayesian Generalized Linear Model 4.7.1 A MCMC algorithm: Metropolis-Hastings

- The MH algorithm tends to accept new values in high density areas (relevant areas).
- The probability of acceptance should not be too low to receive regular state changes in the chain.
- However, it should also not be too high, i.e. the variance of the proposal distribution shouldn't be too low to sufficiently explore the density π over the whole support.

4.7 Bayesian Generalized Linear Model 4.7.1 A MCMC algorithm: Metropolis-Hastings

Algorithm 10: Univariate Metropolis-Hastings algorithm

```
Set starting value $X_0$.
For $i = 1, \dots, n$:
    1. Draw $X_i$ from $q(\cdot|x_{i-1})$.
    2. Draw $U \sim U(0,1)$; accept if
           $U \leq \alpha(x_{i-1}, x_i)$,
       otherwise set $x_i \leftarrow x_{i-1}$.
```

4.7 Bayesian Generalized Linear Model 4.7.1 A MCMC algorithm: Metropolis-Hastings

Multivariate Metropolis-Hastings

- The generalization to the multivariate case is (in principle) easy, for example using

$$q(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}) = \text{MVN}(\boldsymbol{x}_{i-1}, \boldsymbol{\Sigma}).$$

- One problem here is the choice of the ”tuning matrix” $\boldsymbol{\Sigma}$ that controls the acceptance probability. Often, one chooses $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$.
- Several runs are started and each time the acceptance rate is calculated. The variances of the proposal density are then varied until ”reasonable” acceptance rates are achieved.

In the case of **Bayesian GLMs**, there is a variant that automatically generates useful proposal densities, which is described in the following section.

4.7.2 Metropolis-Hastings with IWLS Proposal Density

4.7 Bayesian Generalized Linear Model 4.7.2 Metropolis-Hastings with IWLS Proposal Density

From the lecture Generalized Regression, the *Fisher-Scoring* algorithm is known as an algorithm to compute maximum likelihood estimates:

Example 4.5 (Fisher-Scoring in the Logit-Model) The score function in the logit model (with canonical link function) is

$$\boldsymbol{s}(\boldsymbol{\beta}) = \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})).$$

The Fisher information is

4.7 Bayesian Generalized Linear Model 4.7.2 Metropolis-Hastings with IWLS Proposal Density

$$\boldsymbol{F}(\boldsymbol{\beta}) = \boldsymbol{X}^\top \mathrm{diag}\Big(\mu_i(\boldsymbol{\beta})(1-\mu_i(\boldsymbol{\beta}))\Big)\boldsymbol{X}.$$

Let $\widehat{\boldsymbol{\beta}}$ denote the ML-estimator, then

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = \Big[\boldsymbol{X}^\top \mathrm{diag}\Big(\mu_i(\widehat{\boldsymbol{\beta}})(1-\mu_i(\widehat{\boldsymbol{\beta}}))\Big)\boldsymbol{X}\Big]^{-1}.$$

The Fisher scoring algorithm then has the form

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{(k+1)} &= \widehat{\boldsymbol{\beta}}^{(k)} + \boldsymbol{F}^{-1}(\widehat{\boldsymbol{\beta}}^{(k)})\boldsymbol{s}(\widehat{\boldsymbol{\beta}}^{(k)}) \\ &= \widehat{\boldsymbol{\beta}}^{(k)} + \left[\boldsymbol{X}^\top \mathrm{diag}\Big(\mu_i(\widehat{\boldsymbol{\beta}}^{(k)})(1-\mu_i(\widehat{\boldsymbol{\beta}}^{(k)}))\Big)\boldsymbol{X}\right]^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}^{(k)})).\end{aligned}$$

4.7 Bayesian Generalized Linear Model 4.7.2 Metropolis-Hastings with IWLS Proposal Density

In general, Fisher scoring can be described as follows: define pseudo-observations $\widetilde{\boldsymbol{y}} = (\widetilde{y}_1(\boldsymbol{\beta}), \ldots, \widetilde{y}_n(\boldsymbol{\beta}))^\top$, where

$$\widetilde{y}_i(\boldsymbol{\beta}) = \boldsymbol{x}_i^\top\boldsymbol{\beta} + D_i^{-1}(y_i - \mu_i)$$

with

$$D_i(\boldsymbol{\beta}) = \frac{\partial h(\eta_i)}{\partial \eta_i} = \frac{\partial h(\boldsymbol{x}_i^\top\boldsymbol{\beta})}{\partial \boldsymbol{x}_i^\top\boldsymbol{\beta}} \quad \text{und } \eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}.$$

In the special case of the logit model

$$D_i(\boldsymbol{\beta}) = \mu_i(1-\mu_i) = \mu_i(\boldsymbol{\beta})(1-\mu_i(\boldsymbol{\beta})).$$

4.7 Bayesian Generalized Linear Model 4.7.2 Metropolis-Hastings with IWLS Proposal Density

Combine these entries to $\boldsymbol{D} = \mathrm{diag}(D_1, \ldots, D_n)$. Define further

$$w_i(\boldsymbol{\beta}) = D_i^2(\boldsymbol{\beta})[\mathrm{Var}(y_i)]^{-1} \text{ and } \boldsymbol{W} = \mathrm{diag}(w_1(\boldsymbol{\beta}), \ldots, w_n(\boldsymbol{\beta})).$$

In the Logit-Model:

$$w_i(\boldsymbol{\beta}) = \frac{[\mu_i(1-\mu_i)]^2}{\mu_i(1-\mu_i)} = \mu_i(1-\mu_i).$$

Then the Fisher scoring can be written as iteratively (re)-weighted least squares *(IWLS)* estimator

$$\hat{\boldsymbol{\beta}}^{(k+1)} = (\boldsymbol{X}^\top\boldsymbol{W}^{(k)}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{W}^{(k)}\widetilde{\boldsymbol{y}}^{(k)}.$$

4.7 Bayesian Generalized Linear Model 4.7.2 Metropolis-Hastings with IWLS Proposal Density

From the analogy of least squares and maximum likelihood estimation in the normal distribution case, this can be interpreted as

$$\widetilde{\boldsymbol{y}}^{(k)} \sim \mathrm{MVN}\left(\boldsymbol{X}\boldsymbol{\beta}, (\boldsymbol{W}^{-1})^{(k)}\right).$$

Bayesian version: combine that with the prior distribution $\boldsymbol{\beta} \sim \mathrm{MVN}(\boldsymbol{\beta}_0, \boldsymbol{B}_0)$.
For that, iterate:

4.7 Bayesian Generalized Linear Model 4.7.2 Metropolis-Hastings with IWLS Proposal Density

1. Let the current state be $\boldsymbol{\beta}^{(t)}$; calculate

$$\widetilde{\boldsymbol{y}}^{(t)} = \boldsymbol{X}^\top \boldsymbol{\beta}^{(t)} + \boldsymbol{D}^{-1}(\beta^{(t)})(\boldsymbol{y} - \boldsymbol{\mu}(\beta^{(t)})).$$

2. Draw $\boldsymbol{\beta}^* \sim \mathrm{MVN}(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{C}^{(t+1)})$ with

$$\begin{aligned}\boldsymbol{\beta}^{(t+1)} &= (\boldsymbol{B}_0^{-1} + \boldsymbol{X}^\top \boldsymbol{W}(\boldsymbol{\beta}^{(t)})\boldsymbol{X})^{-1} \\ &\qquad \cdot [\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{X}^\top \boldsymbol{W}(\beta^{(t)})\widetilde{\boldsymbol{y}}(\beta^{(t)})], \\ \boldsymbol{C}^{(t+1)} &= (\boldsymbol{B}_0^{-1} + \boldsymbol{X}^\top \boldsymbol{W}(\beta^{(t)})\boldsymbol{X})^{-1}.\end{aligned}$$

4.7 Bayesian Generalized Linear Model 4.7.2 Metropolis-Hastings with IWLS Proposal Density

3. Accept $\boldsymbol{\beta}^*$ with probability

$$\alpha(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^*) = \min\left(1, \frac{f(\boldsymbol{\beta}^*|\boldsymbol{X})}{f(\boldsymbol{\beta}^{(t)}|\boldsymbol{X})} \times \frac{q(\boldsymbol{\beta}^{(t)}|\boldsymbol{\beta}^*)}{q(\boldsymbol{\beta}^*|\boldsymbol{\beta}^{(t)})}\right),$$

where $q(\boldsymbol{\beta}^{(t)}|\boldsymbol{\beta}^*)$ is the value of the density of

$$\begin{aligned}\mathrm{MVN}\Bigg(&\Big(\boldsymbol{B}_0^{-1} + \boldsymbol{X}^\top \boldsymbol{W}(\boldsymbol{\beta}^*)\boldsymbol{X}\Big)^{-1}\Big(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{X}^\top \boldsymbol{W}(\boldsymbol{\beta}^*)\widetilde{\boldsymbol{y}}(\boldsymbol{\beta}^*)\Big), \\ &\Big(\boldsymbol{B}_0^{-1} + \boldsymbol{X}^\top \boldsymbol{W}(\boldsymbol{\beta}^*)\boldsymbol{X}\Big)^{-1}\Bigg)\end{aligned}$$

evaluated at the point $\boldsymbol{\beta}^{(t)}$ ...

4.7 Bayesian Generalized Linear Model 4.7.2 Metropolis-Hastings with IWLS Proposal Density

and analogously $q(\boldsymbol{\beta}^*|\boldsymbol{\beta}^{(t)})$ is the value of the density of

$$\begin{aligned}\mathrm{MVN}\Bigg(&\Big(\boldsymbol{B}_0^{-1} + \boldsymbol{X}^\top \boldsymbol{W}(\boldsymbol{\beta}^{(t)})\boldsymbol{X}\Big)^{-1}\Big(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{X}^\top \boldsymbol{W}(\boldsymbol{\beta}^{(t)})\widetilde{\boldsymbol{y}}(\boldsymbol{\beta}^{(t)})\Big), \\ &\Big(\boldsymbol{B}_0^{-1} + \boldsymbol{X}^\top \boldsymbol{W}(\boldsymbol{\beta}^{(t)})\boldsymbol{X}\Big)^{-1}\Bigg)\end{aligned}$$

evaluated at the point $\boldsymbol{\beta}^*$.

4.8 Bayesian Generalized Linear Mixture Models

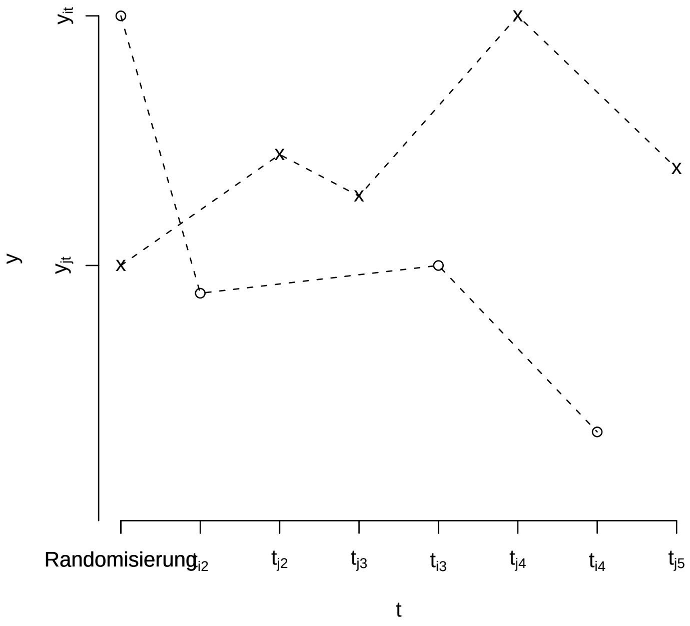
4.8 Bayesian Generalized Linear Mixture Models

The predictor of the GLM from the previous section is expanded here. In the following, we will focus on cluster and longitudinal data. In the case of the latter, the data structure for an individual (statistical unit) i is as follows:

Response	Covariates			Time of observation
y_{i1}	x_{i11}	$\dots$	x_{ip1}	t_{i1}
$\vdots$				$\vdots$
y_{iT_i}	x_{i1T_i}	$\dots$	x_{ipT_i}	t_{iT_i}

Here $y_{i1}, \dots, y_{iT_i}$ are correlated, T_i can vary and the observation times can vary from individual to individual. The observation times should, however, not be informative for the response. The following figure shows schematically such a data situation, which could for example show up in a controlled study.

4.8 Bayesian Generalized Linear Mixture Models



4.8 Bayesian Generalized Linear Mixture Models

- This type of data pose a challenge not just because of the dependency of the observations, but also because of possible missing data and drop-outs.

- Often, longitudinal data also occurs in combination with survival data, for example $y_{it_i}, \dots, y_{iT_i}$ can be the course of a biomarker (possibly measured with measurement errors). The question is then whether the biomarker course is prognostic for the survival time. This leads to the so-called *joint modeling approach.*

- GLMMs can handle this type of data well if one accepts the so-called ”conditional independence assumption”, which is introduced in the following.

4.8 Bayesian Generalized Linear Mixture Models

1. Extend the predictor to

$$\eta_{it} = \boldsymbol{x}_{it}^\top \underbrace{\boldsymbol{\beta}}_{\text{fixed effects}} + \boldsymbol{z}_{it}^\top \underbrace{\boldsymbol{\alpha}_i}_{\text{random effects}}$$

assuming that

$$\boldsymbol{\alpha}_i \;\sim\; \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}).$$

Here $\boldsymbol{x}_{it}^\top = (x_{i1}, \dots, x_{ip_t})$ is the covariate vector, and $\boldsymbol{z}_{it}^\top \in \mathbb{R}^{1\times q}$ can contain covariates from $\boldsymbol{x}_{it}$ and, for example, the time t itself.

- *Remark: in the linear mixed model, additional error terms ε_{it} are included in the predictor. These error terms may be iid or have an additional correlation structure, e.g. they may be autocorrelated.*

4.8 Bayesian Generalized Linear Mixture Models

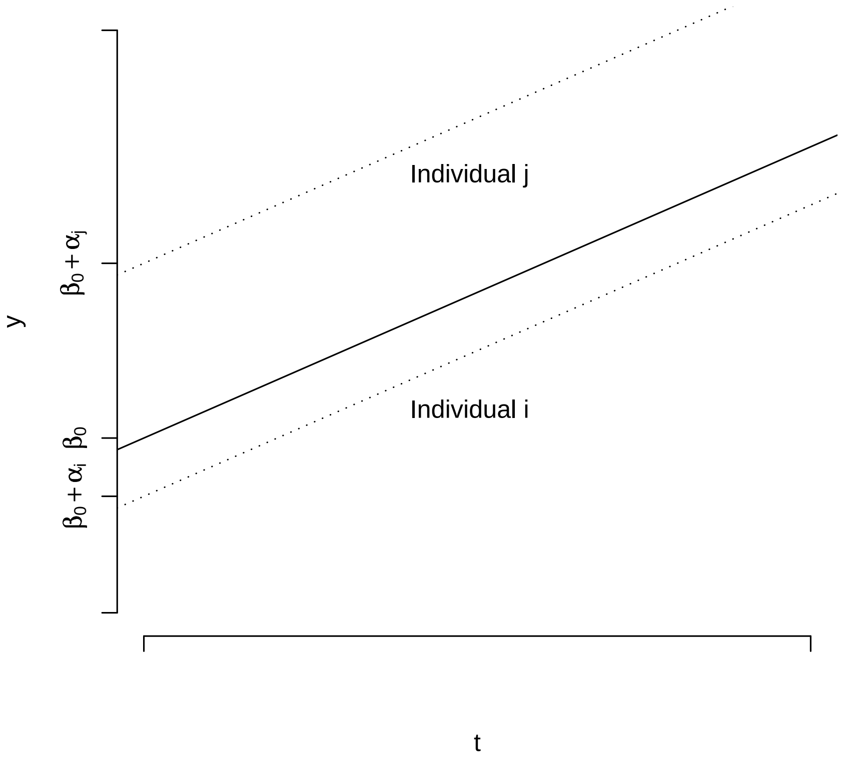
Example 4.6 (Random Intercept Model) Let

$$\eta_{it_i} = \beta_0 + \beta_1 t_i + \alpha_i,\ \alpha_i \ \sim\ N(0, \sigma_\alpha^2),$$

then for an individual i we have:

$$\begin{pmatrix} \eta_{it_{i_1}} \\ \vdots \\ \eta_{it_{T_i}} \end{pmatrix} = \begin{pmatrix} 1 & t_{i_1} \\ \vdots & \vdots \\ 1 & t_{i_{T_i}} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \alpha_i.$$

4.8 Bayesian Generalized Linear Mixture Models



4.8 Bayesian Generalized Linear Mixture Models

2. We make the assumption of conditional independence

$$y_{it} \perp\!\!\!\perp y_{it'} \,\big|\, \boldsymbol{\alpha}_i, \boldsymbol{\beta}$$

for all $t \neq t'$. This allows the representation of the joint distribution of $(y_{i1}, \dots, y_{iT_i})$ *conditional on the random effects* $\boldsymbol{\alpha}_i$ *(and* $\boldsymbol{\beta}$*)* as the product of the conditional distributions

$$f(y_{i1}, \dots, y_{iT_i} |\, \boldsymbol{\alpha}_i, \boldsymbol{\beta}) = \prod_{t=1}^{T_i} f(y_{it}|\boldsymbol{\alpha}_i, \boldsymbol{\beta}).$$

- *Remark: Of course, after integrating out the random effects, the observations* $(y_{i1}, \dots, y_{iT_i})$ *are correlated.*

4.8 Bayesian Generalized Linear Mixture Models

Remark: Without this conditional independence assumption, GLMMs would lose clearly their attractiveness.
The full setup for individual i in matrix notation is then:

$$\boldsymbol{y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iT_i} \end{pmatrix}, \quad \boldsymbol{X}_i = \begin{pmatrix} x_{i11} & \dots & x_{ip1} \\ \vdots & & \\ x_{i1T_i} & \dots & x_{ipT_i} \end{pmatrix},$$

$$\boldsymbol{Z}_i = \begin{pmatrix} z_{i11} & \dots & z_{iq1} \\ \vdots & & \\ z_{i1T_i} & \dots & z_{iqT_i} \end{pmatrix},$$

and

$$\boldsymbol{\eta}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\alpha}_i.$$

4.8 Bayesian Generalized Linear Mixture Models

The full setup for all n individuals in matrix notation is then:

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_n \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{Z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{Z}_n \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_n \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{Z}_1 & \cdots & \mathbf{0} \\ \boldsymbol{X}_2 & \mathbf{0} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \mathbf{0} \\ \boldsymbol{X}_n & \mathbf{0} & \cdots & \boldsymbol{Z}_n \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_n \end{pmatrix}.$$

Additionally, we can abbreviate $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_n)^\top$.

4.8 Bayesian Generalized Linear Mixture Models

The Bayesian approach for GLMMs is basically the same as for Bayesian GLMs with

$$\boldsymbol{\beta} \sim \mathrm{MVN}(\boldsymbol{\beta}_0, \boldsymbol{B}_0) \quad \text{and} \quad \boldsymbol{\alpha} \sim \mathrm{MVN}(\mathbf{0}, \underbrace{\mathrm{diag}(\boldsymbol{\Sigma}, \dots, \boldsymbol{\Sigma})}_{(n\cdot q)\times(n\cdot q)})$$

or

$$\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{pmatrix} \sim \mathrm{MVN}\left(\begin{pmatrix} \boldsymbol{\beta}_0 \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}, \underbrace{\mathrm{diag}(\boldsymbol{B}_0, \boldsymbol{\Sigma}, \dots, \boldsymbol{\Sigma})}_{(p+nq)\times(p+nq)} \right).$$

Here $\mathrm{diag}(\boldsymbol{\Sigma}, \dots, \boldsymbol{\Sigma})$ or $\mathrm{diag}(\boldsymbol{B}_0, \boldsymbol{\Sigma}, \dots, \boldsymbol{\Sigma})$ denote block diagonal matrices.
Comment.

4.8 Bayesian Generalized Linear Mixture Models

(i) In more complex situations, e.g. if individuals are not independent, $\operatorname{diag}(\boldsymbol{B}_0, \boldsymbol{\Sigma}, \dots, \boldsymbol{\Sigma})$ can be replaced by a non-block diagonal matrix (example: space-time correlated data).

(ii) GLMMs are high dimensional when n is large. Specific Algorithms for finding estimates are necessary.

4.8 Bayesian Generalized Linear Mixture Models

- In addition, a (hyper-) prior distribution is constructed for $\boldsymbol{\Sigma}$ because the unobserved $\boldsymbol{\alpha}_i$ are latent variables, i.e. the $\boldsymbol{\alpha}_i$ are to be treated like the ε_i in the linear model and for these we have introduced a prior for the variance.
- For example, the prior could be $\boldsymbol{\Sigma} \sim \text{inv-Wishart}_{\nu_0}(\boldsymbol{\Lambda}_0^{-1})$, i.e.

$$p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu_0+q+1)/2} \exp\left(-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_0)\right),$$

4.8 Bayesian Generalized Linear Mixture Models

with resulting posterior distribution

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}|\boldsymbol{y}, \boldsymbol{X}) \propto & \left[\prod_{i=1}^{n}\left\{\prod_{j=1}^{T_i} f(y_{it_j}|\boldsymbol{\beta}, \boldsymbol{\alpha}_i)\right\}\right] \times \\ & \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top \boldsymbol{B}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right) \times \\ & |\boldsymbol{\Sigma}|^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \boldsymbol{\alpha}_i^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}_i\right) \times \\ & |\boldsymbol{\Sigma}|^{-(\nu_0+q+1)/2} \exp\left(-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_0)\right). \end{aligned}$$

4.8 Bayesian Generalized Linear Mixture Models

A possible algorithm for simulating the posterior is then a blockwise Gibbs sampler.

1. *Full conditional for the $\boldsymbol{\beta}$-Block:*

$$f(\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{\Sigma}, \boldsymbol{y}, \boldsymbol{X}) \propto \left[\prod_{i=1}^{n}\prod_{j=1}^{T_i} f(y_{it_j}|\cdot)\right] \cdot \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top \boldsymbol{B}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right)$$

   can be treated as in a Bayesian GLM if one additionally uses an offset $\boldsymbol{z}_i^T\boldsymbol{\alpha}_i$:

$$\widetilde{y}_i(\boldsymbol{\beta}^{(t-1)}|\boldsymbol{\alpha}_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}^{(t-1)} + \boldsymbol{z}_i^\top \boldsymbol{\alpha}_i + D_i^{-1}[y_i - \mu_i(\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\alpha}_i)]$$

   or

$$\widetilde{y}_i(\boldsymbol{\beta}^{(t-1)}|\boldsymbol{\alpha}_i) - \boldsymbol{z}_i^\top \boldsymbol{\alpha}_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}^{(t-1)} + D_i^{-1}[y_i - \mu_i(\boldsymbol{\beta}^{(t-1)}, \boldsymbol{\alpha}_i)]$$

   Define $\widetilde{\widetilde{y}}_i(\boldsymbol{\beta}^{(t-1)}|\boldsymbol{\alpha}_i) = \widetilde{y}_i(\boldsymbol{\beta}^{(t-1)}|\boldsymbol{\alpha}_i) - \boldsymbol{z}_i^\top\boldsymbol{\alpha}_i$, then we can apply the IWLS Metropolis Hastings algorithm from 4.7.2 for drawing from this full conditional.

4.8 Bayesian Generalized Linear Mixture Models

2. *Full conditionals for the $\boldsymbol{\alpha}_i$-blocks:* For $i = 1, \dots, n$ one obtains

$$f(\boldsymbol{\alpha}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{y}, \boldsymbol{X}) \propto \left[\prod_{i=1}^{n}\prod_{j=1}^{T_i} f(y_{it_j}|\cdot)\right] \exp\left(-\boldsymbol{\alpha}_i^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}_i\right).$$

This can again be interpreted as a Bayesian GLM with offset $\boldsymbol{x}_i^\top \boldsymbol{\beta}$ and the IWLS-MH algorithm with proposal distribution

$$\begin{aligned}\mathrm{MVN}\Big(&[\boldsymbol{\Sigma}^{-1} + \boldsymbol{z}_i \boldsymbol{W}_i(\boldsymbol{\alpha}_{\boldsymbol{i}}^{(t-1)})\boldsymbol{z}_i^\top]^{-1} \boldsymbol{z}_i \boldsymbol{W}_i(\boldsymbol{\alpha}_{\boldsymbol{i}}^{(t-1)})[\widetilde{y}_i(\boldsymbol{\alpha}_i^{(t-1)}) - \boldsymbol{x}_i^\top \boldsymbol{\beta}],\\ &[\boldsymbol{\Sigma}^{-1} + \boldsymbol{z}_i \boldsymbol{W}_i(\boldsymbol{\alpha}_{\boldsymbol{i}}^{(t-1)})\boldsymbol{z}_i^\top]^{-1}\Big)\end{aligned}$$

can be applied.

4.8 Bayesian Generalized Linear Mixture Models

3. *Full-Conditional for* $\boldsymbol{\Sigma}$*:*

$$\begin{aligned} f(\boldsymbol{\Sigma}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{y}, \boldsymbol{X}) \propto\ & |\boldsymbol{\Sigma}|^{-(n+\nu_0+q+1)/2} \\ & \cdot \exp\left(-\frac{1}{2}\,\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_0 + \sum_{i=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_i^\top\right)\right)\end{aligned}$$

corresponds to the kernel of an inv-Wishart$_{\nu_0+n}(\boldsymbol{\Lambda}_0 + \sum_{i=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_i^\top)$-distribution (implicit dimension $q \times q$); random numbers form this distribution can be simulated directly with a suitable random number generator.

4.8 Bayesian Generalized Linear Mixture Models

- All in all, we have a block-wise Gibbs sampler with

$$\begin{array}{ccccccc} 1 & + & n & + & 1 & = & n+2 \\ \boldsymbol{\beta} & & \{\boldsymbol{\alpha}_i\}_{i=1}^{n} & & \boldsymbol{\Sigma} & & \end{array}$$

  blocks.

- Since the updates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_i$ involve an acceptance mechanism, the algorithm is sometimes called *Metropolis-Hastings-within-Gibbs*-algorithm, as the individual blocks (1. and 2.) are each using a Metropolis-Hastings update. At the end of a cycle, however, the $n+2$ blocks are formally accepted with probability 1 and therefore its a Gibbs sampler.

Addendum: Bayesian statistics in R with Stan

We model a cat's heart weight using the `cats` data from the `Mass` package. We evaluate the following model equations (notation might differ from previous material):

$$\begin{aligned} Hwt_i &= \beta_0 + \beta_1 \cdot Bwt_i + \beta_2 \cdot Sex_i + \epsilon_i \\ Hwt_i &= \beta_0 + \beta_1 \cdot \sqrt{Bwt_i} + \beta_2 \cdot Sex_i + \epsilon_i \\ Hwt_i &= \beta_0 + \beta_1 \cdot Bwt_i + \beta_2 \cdot Sex_i + \beta_3 \cdot Bwt_i \cdot Sex_i + \epsilon_i \end{aligned}$$

```
# Load data
library(MASS)
data <- cats

# Configure models in Stan
library(rstan)
# Model matrices
```

```
X1 <- model.matrix(~Bwt + Sex, data)
X2 <- model.matrix(~sqrt(Bwt) + Sex, data)
X3 <- model.matrix(~Bwt + Sex + Bwt*Sex, data)
N = nrow(data)
stan.data1 <- list(N = N, K = 3 , X = X1, Hwt = data$Hwt)
stan.data2 <- list(N = N, K = 3 , X = X2, Hwt = data$Hwt)
stan.data3 <- list(N = N, K = 4 , X = X3, Hwt = data$Hwt)

rstan_options(auto_write = TRUE)
options(mc.cores = parallel::detectCores())
```

Body weight likely has positive effect on heart weight. The effect probably varies depending on the individual due to genetic factors. A prior distribution with a high variance is a good choice. Male sex likely results in a higher heart weight, although this effect might indirectly stem from a higher body weight. We therefore cannot make a prior assumption about the isolated effect which we aim to estimate here. A distribution with a high variance is a suitable choice here.

```
# Beta priors:

# Uninformative normal prior (high variance)
stanmodel.normal.highvar.prior = "

data {
int N; // the number of observations
int K; // the number of columns in the model matrix
real Hwt[N]; // the response
matrix[N,K] X; // the model matrix
}

parameters {
vector[K] beta; // the regression parameters
real<lower=0.01> sigma; // the standard deviation
}

model {
beta[1] ~normal(0, 50);
for (i in 2:K)
beta[i] ~normal(0, 50);
sigma ~uniform(0,100);
Hwt ~normal(X * beta, sigma);
}

generated quantities {
real deviance; // deviance
for (n in 1:N)
deviance = -2 * normal_lpdf(Hwt | X * beta, sigma);
```

```
// normal_lpf = pointwise likelihood per observation
} "
# Low information normal prior (lower variance)
stanmodel.normal.lowvar.prior = "

data {
int N; // the number of observations
int K; //t he number of columns in the model matrix
real Hwt[N]; // the response
matrix[N,K] X; // the model matrix
}

parameters {
vector[K] beta; //the regression parameters
real<lower=0.01> sigma; //the standard deviation
}

model {
beta[1] ~normal(0, 10);
for (i in 2:K)
beta[i] ~normal(0, 2);
sigma ~uniform(0,100);
Hwt ~normal(X * beta, sigma);
}

generated quantities {
real deviance; // deviance
for (n in 1:N) deviance = -2* normal_lpdf(Hwt | X* beta, sigma);
// normal_lpf = pointwise likelihood per observation
} "
# Uninformative Cauchy prior (location = 0, scale = 10 for intercept, 2.5
for remaining parameter
# (recommendation by Gelman et. al)
stanmodel.cauchy.prior = "

data {
int N; // the number of observations
int K; // the number of columns in the model matrix
real Hwt[N]; // the response
matrix[N,K] X; // the model matrix
}

parameters {
vector[K] beta; // the regression parameters
real<lower=0.01> sigma; // the standard deviation
}
```

```
model {
beta[1] ~cauchy(0, 10);
for (i in 2:K)
beta[i] ~cauchy(0, 2.5);
sigma ~uniform(0,100);
Hwt ~normal(X * beta, sigma); }

generated quantities {
real deviance; // deviance
for (n in 1:N)
deviance = -2* normal_lpdf(Hwt | X * beta, sigma);
// normal_lpf = pointwise likelihood per observation
} "
stan.norm.highvar.1 <- stan(model_code = stanmodel.normal.highvar.prior,
data = stan.data1, iter = 1000, warmup = 100, control = list(adapt_delta
= 0.99))
stan.norm.highvar.2 <- stan(model_code = stanmodel.normal.highvar.prior,
data = stan.data2, iter = 1000, warmup = 100, control = list(adapt_delta
= 0.99))
stan.norm.highvar.3 <- stan(model_code = stanmodel.normal.highvar.prior,
data = stan.data3, iter = 1000, warmup = 100, control = list(adapt_delta
= 0.99))

stan.norm.lowvar.1 <- stan(model_code = stanmodel.normal.lowvar.prior,
data = stan.data1, iter = 1000, warmup = 100, control = list(adapt_delta
= 0.99))
stan.norm.lowvar.2 <- stan(model_code = stanmodel.normal.lowvar.prior,
data = stan.data2, iter = 1000, warmup = 100, control = list(adapt_delta
= 0.99))
stan.norm.lowvar.3 <- stan(model_code = stanmodel.normal.lowvar.prior,
data = stan.data3, iter = 1000, warmup = 100, control = list(adapt_delta
= 0.99))

stan.cauchy.1 <- stan(model_code = stanmodel.cauchy.prior, data =
stan.data1, iter = 1000, warmup = 100, control = list(adapt_delta =
0.99))
stan.cauchy.2 <- stan(model_code = stanmodel.cauchy.prior, data =
stan.data2, iter = 1000, warmup = 100, control = list(adapt_delta =
0.99))
stan.cauchy.3 <- stan(model_code = stanmodel.cauchy.prior, data =
stan.data3, iter = 1000, warmup = 100, control = list(adapt_delta =
0.99))
```

Model evaluation and selection:

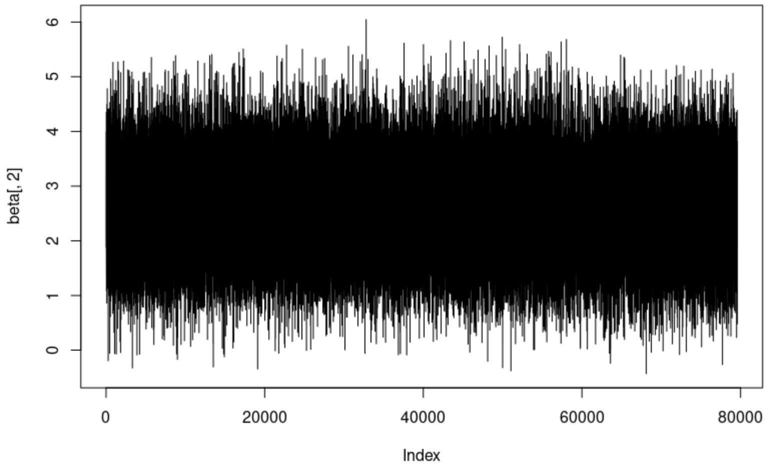
```
# Results decisively depend on prior assumptions, as the amount of data
is low (n = 144)
# Due to missing background knowledge, an unformative prior is preferable
# ---------------------------------------
# Deviance information criterion

# Extraction from Stan data

# 1. model
dev.norm.lowvar.1 = median(extract(stan.norm.lowvar.1)$deviance) #
deviance = -2 * log likelihood
dev.thetahat.norm.lowvar.1 = -2* sum(dnorm(data$Hwt,
mean = X1 %*% apply(extract(stan.norm.lowvar.1)$beta, MARGIN = 2, median)

# mean = X*Beta
, sd = sqrt(median(extract(stan.norm.lowvar.1)$sigma)), log = TRUE))
# sd = sqrt(sigma)
DIC.norm.lowvar.1 = 2 * dev.norm.lowvar.1 - dev.thetahat.norm.lowvar.1

##
dev.norm.highvar.1 = median(extract(stan.norm.highvar.1)$deviance) #
deviance = -2 * log likelihood
dev.thetahat.norm.highvar.1 = -2* sum(dnorm(data$Hwt, mean = X1 %*%
apply(extract(stan.norm.highvar.1)$beta, MARGIN = 2, median) , sd =
sqrt(median(extract(stan.norm.highvar.1)$sigma)), log = TRUE))
DIC.norm.highvar.1 = 2 * dev.norm.highvar.1 - dev.thetahat.norm.highvar.1


##
dev.cauchy.1 = median(extract(stan.cauchy.1)$deviance) # deviance = -2 *
log likelihood
dev.theta.hat.cauchy.1 = -2* sum(dnorm(data$Hwt, mean = X1 %*%
apply(extract(stan.cauchy.1)$beta, MARGIN = 2, median) , sd =
sqrt(median(extract(stan.cauchy.1)$sigma)), log = TRUE))
DIC.cauchy.1 = 2 * dev.cauchy.1 - dev.theta.hat.cauchy.1

# ---------------------------------------
# 2. model

dev.norm.lowvar.2 = median(extract(stan.norm.lowvar.2)$deviance) #
deviance = -2 * log likelihood
dev.thetahat.norm.lowvar.2 = -2* sum(dnorm(data$Hwt, mean = X2 %*%
apply(extract(stan.norm.lowvar.2)$beta, MARGIN = 2, median) , sd =
sqrt(median(extract(stan.norm.lowvar.2)$sigma)), log = TRUE))
DIC.norm.lowvar.2 = 2 * dev.norm.lowvar.2 - dev.thetahat.norm.lowvar.2

##
dev.norm.highvar.2 = median(extract(stan.norm.highvar.2)$deviance) #
deviance = -2 * log likelihood
dev.thetahat.norm.highvar.2 = -2* sum(dnorm(data$Hwt, mean = X2 %*%
apply(extract(stan.norm.highvar.2)$beta, MARGIN = 2, median) , sd =
sqrt(median(extract(stan.norm.highvar.2)$sigma)), log = TRUE))
DIC.norm.highvar.2 = 2 * dev.norm.highvar.2 - dev.thetahat.norm.highvar.2


##
dev.cauchy.2 = median(extract(stan.cauchy.2)$deviance) # deviance = -2 *
log likelihood
dev.theta.hat.cauchy.2 = -2* sum(dnorm(data$Hwt, mean = X2 %*%
apply(extract(stan.cauchy.2)$beta, MARGIN = 2, median) , sd =
sqrt(median(extract(stan.cauchy.2)$sigma)), log = TRUE))
DIC.cauchy.2 = 2 * dev.cauchy.2 - dev.theta.hat.cauchy.2

# 3. model
dev.norm.lowvar.3 = median(extract(stan.norm.lowvar.3)$deviance) #
deviance = -2 * log likelihood
dev.thetahat.norm.lowvar.3 = -2* sum(dnorm(data$Hwt, mean = X3 %*%
apply(extract(stan.norm.lowvar.3)$beta, MARGIN = 2, median) , sd =
sqrt(median(extract(stan.norm.lowvar.3)$sigma)), log = TRUE))
DIC.norm.lowvar.3 = 2 * dev.norm.lowvar.3 - dev.thetahat.norm.lowvar.3

##
dev.norm.highvar.3 = median(extract(stan.norm.highvar.3)$deviance) #
deviance = -2 * log likelihood
dev.thetahat.norm.highvar.3 = -2* sum(dnorm(data$Hwt, mean = X3 %*%
apply(extract(stan.norm.highvar.3)$beta, MARGIN = 2, median) , sd
= sqrt(median(extract(stan.norm.highvar.3)$sigma)), log = TRUE))
DIC.norm.highvar.3 = 2 * dev.norm.highvar.3 - dev.thetahat.norm.highvar.3


##
dev.cauchy.3 = median(extract(stan.cauchy.3)$deviance) # deviance = -2 *
log likelihood
dev.theta.hat.cauchy.3 = -2* sum(dnorm(data$Hwt, mean = X3 %*%
apply(extract(stan.cauchy.3)$beta, MARGIN = 2, median) , sd =
sqrt(median(extract(stan.cauchy.3)$sigma)), log = TRUE))
DIC.cauchy.3 = 2 * dev.cauchy.3 - dev.theta.hat.cauchy.3

# Model selection

named.list <- function(...) {
l <- setNames( list(...) , as.character( match.call()[-1]) )
return(l)
}

DIC = rbind(DIC.norm.highvar.1, DIC.norm.highvar.2)
DIC = named.list(
DIC.norm.highvar.1,
DIC.norm.highvar.2,
DIC.norm.highvar.3,
DIC.norm.lowvar.1,
DIC.norm.lowvar.2,
DIC.norm.lowvar.3,
DIC.cauchy.1,
DIC.cauchy.2,
DIC.cauchy.3)

DIC
# DIC does not vary substantially
# Best model according to deviance information criterion:
DIC[which(DIC == min(unlist(DIC)))]

# ---------------------------------------
# We build the chosen model from the ground up
stan.final = stan(model_code = stanmodel.normal.highvar.prior, data
= stan.data3, iter = 20000, warmup= 100, control = list(adapt_delta =
0.99))

stan.final
ml.fit.3
```

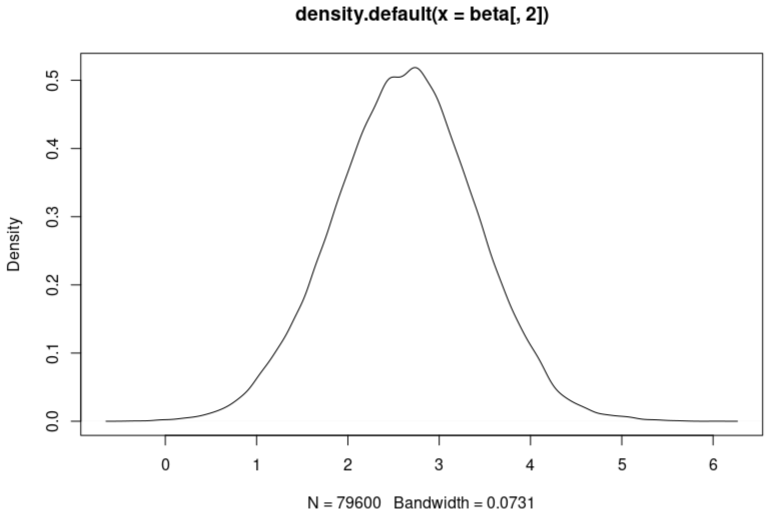
Output:

```
> stan.final
Inference for Stan model: d82885041c12e28b3bcdc374b9cc4391.
4 chains, each with iter=20000; warmup=100; thin=1;
post-warmup draws per chain=19900, total post-warmup draws=79600.

mean se_mean sd 2.5% 25% 50% 75% 97.5% n_eff Rhat
beta[1] 3.00 0.02 1.85 -0.58 1.77 2.99 4.24 6.64 9394 1
beta[2] 2.63 0.01 0.78 1.09 2.11 2.63 3.15 4.13 9470 1
beta[3] -4.18 0.02 2.07 -8.24 -5.58 -4.19 -2.78 -0.12 9066 1
beta[4] 1.68 0.01 0.84 0.04 1.12 1.68 2.25 3.33 8991 1
sigma 1.45 0.00 0.09 1.30 1.39 1.45 1.51 1.64 72494 1
deviance 515.08 0.02 3.21 510.83 512.72 514.43 516.73 523.07 18129 1
lp__ -124.86 0.01 1.59 -128.81 -125.68 -124.53 -123.68 -122.75 18048 1

Samples were drawn using NUTS(diag_e) at Tue Jul 5 16:01:44 2022.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).

> ml.fit.3
Call:
lm(formula = Hwt ~Bwt + Sex + Bwt * Sex, data = data)
```

```
Coefficients:
(Intercept) Bwt SexM Bwt:SexM
2.981 2.636 -4.165 1.676

# Bayesian and Frequentist model are similar

# Parameters
beta = extract(stan.final)$beta
plot(beta[,1], type = "l") # intercept
plot(density(beta[,1]))
plot(beta[,2], type = "l") # body weight
plot(density(beta[,2]))
plot(beta[,3], type = "l") # sex (binary)
plot(density(beta[,3]))
plot(beta[,4], type = "l") # body weight * sex
plot(density(beta[,3]))

# Errors
sigma = extract(stan.norm.highvar.1)$sigma
plot(sigma, type = "l")
plot(density(sigma))

# Marginal posteriors exhibit good mixing and convergence
# Posterior densities are relatively smooth
```

Exemplary diagnostic plots for the body weight parameter:




density.default(x = beta[, 2])
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
5
6
N = 79600 Bandwidth = 0.0731

Chapter 5

Bootstrap

Bootstrap

Literature on the subject:

- Efron B., Tibshirani R.J.: An Introduction to the Bootstrap (1993)
- Hall P.: The Bootstrap and Edgeworth Expansion (1992)
- Davison A.C.: Recent Developments in Bootstrap Methodology, Statistical Science (2003), Vol. 18, No. 2, pp. 141-157

5.1 Introduction

5.1 Bootstrap: introduction

- *Bootstrap*: a piece of leather or other strong material at the back of a boot that you use to help you pull the boot on.
- „Pull oneself out of the swamp by own hair” → Baron Munchausen (with horse). Also, "pull oneself up by one’s bootstraps" — improve one’s position by one’s own efforts.
- Computer-aided method
- Based on repeated draws (*resampling*) from the observed data.
- *Aim:* estimate variance, bias, distribution of a statistic $T = T(X_1, \dots, X_n)$, confidence intervals, tests.

5.1 Bootstrap: introduction

- When? In situations, where
  - (a) asymptotic statements are questionable (small sample sizes),
  - (b) analytical derivations are very complicated or impossible, for example, if no parametric distribution assumptions can be made → bootstrap for nonparametric estimations.
- Does „bootstrap” always work? No, not always (bootstrap can be *inconsistent*), but often.

5.1 Bootstrap: introduction

5.1.1 Basic idea

5.1 Bootstrap: introduction
5.1.1 Basic idea

One-sample problem: $X = (X_1, \ldots, X_n)$, $X_i \overset{i.i.d.}{\sim} F$, F is unknown
Statistic of interest: $T(X)$
Observed data: $x = (x_1, x_2, \ldots, x_n) \to T(x)$
Bootstrap samples: draw *randomly* n times *with* replacement from $(x_1, \ldots, x_n)$. We obtain

$$x^* = (x_1^*, x_2^*, \ldots, x_n^*) \to T(x^*).$$

Example: $x = (1, 2, 5)$, $n = 3$. $x^* = (1, 1, 5)$ is a possible bootstrap sample.
Thus:

5.1 Bootstrap: introduction
5.1.1 Basic idea

(1) Values from the original sample x can be included in the bootstrap sample

  (i) once,

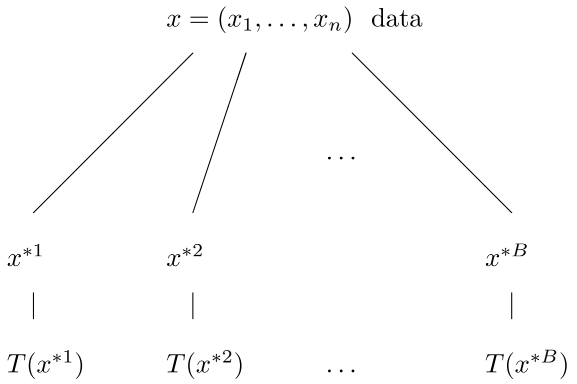
  (ii) multiple times,

  (iii) not at all.

(2) The bootstrap sample also has a sample size of n.

5.1 Bootstrap: introduction
5.1.1 Basic idea

Sketch:



B: number of bootstrap samples
With the calculated statistics $T(x^{*1}), \dots, T(x^{*B})$, statements can be made about the distribution of T, for example

5.1 Bootstrap: introduction
5.1.1 Basic idea

$$\mathrm{Var}_F(T) \;\approx\; \widehat{\mathrm{Var}}_{\mathrm{Boot}}(T) = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left[T(x^{*b}) - \bar{T}_{\mathrm{Boot}} \right]^2 \right\}$$

where

$$\bar{T}_{\mathrm{Boot}} = \frac{1}{B} \sum_{b=1}^{B} T(x^{*b}).$$

5.1.2 Empirical distribution function and the plug-in principle

5.1 Bootstrap: introduction
5.1.2 Empirical distribution function and the plug-in principle

$X = (X_1, \dots, X_n)$, $X_i \overset{i.i.d.}{\sim} F$, F is unknown
$x = (x_1, x_2, \dots, x_n)$ data
Empirical distribution function:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x),$$

where I is the indicator function.
Plug-in-principle: Replace F with $\hat{F}_n$.

5.1 Bootstrap: introduction
5.1.2 Empirical distribution function and the plug-in principle

Example 5.1

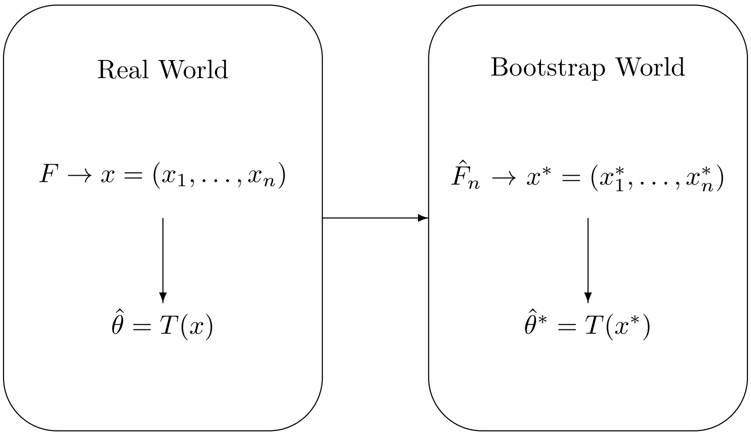
$$
\begin{aligned}
T(F) &= \mu = \int x \mathrm{d}F(x) \\
T(\hat{F}_n) &= \int x \mathrm{d}\hat{F}_n(x) \\
&= \sum_{i=1}^{n} x_i \hat{P}_n(X = x_i) \qquad \textit{(assuming all } x_i \textit{ are different)} \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}
\end{aligned}
$$

Plug-in-principle makes sense if *no further information about F* is available except the sample.
→ „nonparametric setup"

5.1.3 Real world and bootstrap world

5.1 Bootstrap: introduction
5.1.3 Real world and bootstrap world

Once again one sample case:


Real World
$F \to x = (x_1, \dots, x_n)$
$\hat{\theta} = T(x)$
Bootstrap World
$\hat{F}_n \to x^* = (x_1^*, \dots, x_n^*)$
$\hat{\theta}^* = T(x^*)$

5.1 Bootstrap: introduction
5.1.3 Real world and bootstrap world

- The unknown distribution F yields x as a random sample.
- The empirical distribution $\hat{F}_n$ yields x^* as a random bootstrap sample.
- The statistic of interest $\hat{\theta} = T(x)$ is a function of the random sample.
- The bootstrap counterpart $\hat{\theta}^* = T(x^*)$ is a function of the bootstrap sample.

⇒ In general, F or $\hat{F}_n$ in the above figure can be replaced with an *estimated* probability model P or $\hat{P}_n$.

5.1.4 The ideal bootstrap distribution

5.1 Bootstrap: introduction
5.1.4 The ideal bootstrap distribution

Data $x = (x_1, x_2, \dots, x_n)$.
Question: How many different bootstrap samples are there?

Example 5.2
Let $x = (1, 2, 5)$. The ordering does not matter here. Because $n = 3$ there are 10 different bootstap samples (if all x_i are different):

$$\begin{gathered}(1,1,1), (2,2,2), (5,5,5), (1,1,2), (1,1,5),\\ (2,2,5), (1,2,2), (1,5,5), (2,5,5), (1,2,5).\end{gathered}$$

5.1 Bootstrap: introduction 5.1.4 The ideal bootstrap distribution

- The ideal bootstrap estimate is the one that results from covering all possible bootstrap samples.
- For example, the ideal bootstrap estimate for the variance of $\hat{\theta} = \text{median}(X)$ in example 5.2 would be the variance over 10 bootstrap samples.
- However, it must be taken into account that the samples are drawn with different probabilities:

5.1 Bootstrap: introduction 5.1.4 The ideal bootstrap distribution

5.1 Bootstrap: introduction 5.1.4 The ideal bootstrap distribution

Consider $\hat{\theta} = \text{median}(X)$. Then $\hat{\theta}(x) = 2$ is the estimate from the sample and

$$\begin{aligned}\text{Var}_{\hat{F}_n}(\hat{\theta}^*) &= \left(\frac{1}{3}\right)^3 \Big\{(1-c)^2 + (2-c)^2 + (5-c)^2\\ &\quad +3 \cdot \big[(1-c)^2 + (1-c)^2 + (2-c)^2 + (2-c)^2 + (5-c)^2\\ &\quad + (5-c)^2\big] + 6 \cdot (2-c)^2\Big\}\\ &= 2.32,\end{aligned}$$

where

$$\begin{aligned}c = \bar{\hat{\theta}}^* &= \left(\frac{1}{3}\right)^3 [1 + 2 + 5 + 3 \cdot (1 + 1 + 2 + 2 + 5 + 5) + 6 \cdot 2]\\ &= \left(\frac{1}{3}\right)^3 [8 + 3 \cdot 16 + 12] = \left(\frac{1}{3}\right)^3 \cdot 68 = \frac{68}{27} \approx 2.5\end{aligned}$$

is the mean of all estimated medians.

5.1 Bootstrap: introduction 5.1.4 The ideal bootstrap distribution

Generally speaking, there are $\binom{2n-1}{n}$ possible bootstrap samples, provided that all n data points $x_1, \ldots, x_n$ are *different*:

$$\begin{aligned}n = 3: &\quad \binom{5}{3} = 10\\ n = 15: &\quad \binom{29}{15} = 77\ 558\ 760\\ n = 20: &\quad \binom{39}{20} = 68\ 923\ 264\ 410\end{aligned}$$

This means that, when n is not very small, it is pracitcally not possible to use the ideal bootstrap distribution. Instead, one is satisfied with $B \ll \binom{2n-1}{n}$ samples of all possible bootstrap samples.

5.2 Bootstrap estimation of standard errors

5.2 Bootstrap estimation of standard errors

- One sample case: $X = (X_1, \ldots, X_n)$, $X_i \overset{\text{i.i.d.}}{\sim} F$, F is unknown

- Data: $x = (x_1, \dots, x_n)$
- The aim of this section is to estimate the standard error of an estimator $\hat{\theta} = \hat{\theta}(X)$ for $\theta = T(F)$. Here, $\hat{\theta}(X)$ *can* be the plug-in-estimator $T(\hat{F}_n)$, but not necessarily.
- *Question:* How good is the estimate $\hat{\theta}$?

5.2.1 Bootstrap algorithm for estimating the standard error

5.2 Bootstrap estimation of standard errors 5.2.1 Bootstrap algorithm for estimating the standard error

Algorithm 11: Bootstrap algorithm for estimating the standard error

1. Generate B bootstrap samples $x^{*1}, \dots, x^{*B}$.
2. Compute $\hat{\theta}^*(b)$, $b = 1, \dots, B$.
3. Estimate the standard error $\mathrm{se}_F(\hat{\theta}) = \sqrt{\mathrm{Var}_F(\hat{\theta})}$ with

$$\widehat{\mathrm{se}}_B = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left[\hat{\theta}^*(b) - \hat{\theta}^*(\cdot) \right]^2 \right\}^{\frac{1}{2}}$$

$$\text{where} \quad \hat{\theta}^*(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^*(b).$$

5.2 Bootstrap estimation of standard errors 5.2.1 Bootstrap algorithm for estimating the standard error

The bootstrap estimate of the standard error $\mathrm{se}_F(\hat{\theta})$ of an estimate $\hat{\theta}$ (where data is from F) is thus the standard error for random samples of *size* n drawn from $\hat{F}_n$ *with* replacement.
The following holds:

$$\lim_{B \to \infty} \widehat{\mathrm{se}}_B = \mathrm{se}_{\hat{F}_n}(\hat{\theta}^*).$$

The ideal bootstrap estimate $\mathrm{se}_{\hat{F}_n}(\hat{\theta}^*)$ and the approximation $\widehat{\mathrm{se}}_B$ are often called *nonparametric bootstrap estimate* because they are only based on $\hat{F}_n$ and $\hat{F}_n$ is the nonparametric estimate for F .
→ Section 5.2.3: parametric bootstrap (F is no longer estimated by $\hat{F}_n$).

5.2 Bootstrap estimation of standard errors 5.2.1 Bootstrap algorithm for estimating the standard error

Example 5.3
Two (quasi-) continuous features Y and Z are recorded over n individuals, i.e.

$$X = ((Y_1, Z_1), (Y_1, Z_1), \dots, (Y_n, Z_n)), \quad (Y_i, Z_i) \overset{i.i.d.}{\sim} F_{Y,Z}.$$

5.2 Bootstrap estimation of standard errors 5.2.1 Bootstrap algorithm for estimating the standard error

Goal: estimate the standard error of the correlation coefficient of Y and Z.

5.2.2 Number of replications

The number of replications B is determined by the following considerations:

5.2 Bootstrap estimation of standard errors 5.2.2 Number of replications

(i) Practical considerations: if $\hat{\theta}(x^*)$ is a complex function of x^*, then B will have to be smaller, as opposed to the case when $\hat{\theta}(x^*)$ is a simple function of x^*.

(ii) Precision considerations: it is true that

$$\mathrm{Var}(\widehat{\mathrm{se}}_B) > \mathrm{Var}\big(\underbrace{\mathrm{se}_{\hat{F}_n}(\hat{\theta}^*)}_{\text{ideal bootstrap estimate}}\big).$$

The question is by how much the variance of $\widehat{\mathrm{se}}_B$ is greater than the variance of $\mathrm{se}_{\hat{F}_n}$.

5.2 Bootstrap estimation of standard errors 5.2.2 Number of replications

From theoretical considerations, it follows that $B = 200$ is usually sufficient for estimating a standard error in a one-sample problem. For confidence intervals, significantly more replications are needed ($B \approx 2000$).

5.2.3 Parametric Bootstrap

5.2 Bootstrap estimation of standard errors 5.2.3 Parametric bootstrap

Definition 5.1
The bootstrap parametric estimate of the standard error is defined by

$$se_{\hat{F}_{n,par}}(\hat{\theta}^*)\ ,$$

where $\hat{F}_{n,par}$ is an estimate of F, derived from a parametric model.

5.2 Bootstrap estimation of standard errors 5.2.3 Parametric bootstrap

Example 5.4
Let $X = ((Y_1, Z_1)', ..., (Y_n, Z_n)')$ with

$$\begin{pmatrix} Y_i \\ Z_i \end{pmatrix} \overset{i.i.d.}{\sim} F_{Y,Z}\ .$$

Assumption: $F_{Y,Z}$ is a bivariate normal distribution and

$$\begin{aligned} \hat{\mu} &= \begin{pmatrix} \bar{y} \\ \bar{z} \end{pmatrix}, \\ \hat{\Sigma} &= \frac{1}{n}\begin{pmatrix} \sum_{i=1}^n (y_i-\bar{y})^2 & \sum_{i=1}^n (y_i-\bar{y})(z_i-\bar{z}) \\ \sum_{i=1}^n (y_i-\bar{y})(z_i-\bar{z}) & \sum_{i=1}^n (z_i-\bar{z})^2 \end{pmatrix}. \end{aligned}$$

5.2 Bootstrap estimation of standard errors 5.2.3 Parametric bootstrap

This means that we now use $\hat{F}_{n,par} = N_2(\hat{\mu}, \hat{\Sigma})$ as an estimate of F, and instead of taking bootstrap samples from the data, let's take bootstrap samples from this bivariate Normal distribution:

$$\left.\begin{aligned} x^{*1} &= ((Y_1^{*1}, Z_1^{*1})', \ldots, (Y_n^{*1}, Z_n^{*1})') \\ &\vdots \\ x^{*B} &= ((Y_1^{*B}, Z_1^{*B})', \ldots, (Y_n^{*B}, Z_n^{*B})') \end{aligned}\right\} \sim N_2(\hat{\mu}, \hat{\Sigma}).$$

After that, it's business as usual!

5.2 Bootstrap estimation of standard errors 5.2.3 Parametric bootstrap

Example 5.5 (Standard error for the estimate of the correlation coefficient θ)

(i) Comparison with the formula for the bivariate normal distribution:

$$\widehat{\mathrm{se}}_{N_2(\mu,\Sigma)}(\hat{\theta}) = \frac{1-\hat{\theta}^2}{\sqrt{n-3}}\ .$$

(ii) Comparison after Fisher transformation:

$$\hat{\xi} = \frac{1}{2}\log\left(\frac{1+\hat{\theta}}{1-\hat{\theta}}\right) \overset{approx.}{\sim} N\left[\frac{1}{2}\log\left(\frac{1+\theta}{1-\theta}\right), \left(\frac{1}{\sqrt{n-3}}\right)^2\right].$$

To exploit this result, inference could be performed for $\hat{\xi}$ and then converted to the true correlation coefficient θ by transforming backwards.

5.2.4 An example in which the nonparametric bootstrap does not work

5.2 Bootstrap estimation of standard errors
5.2.4 An example where the nonparametric bootstrap does not work

- Consider $X = (X_1, \dots, X_n)$ with $X_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, \theta)$.
- Let the maximum $\hat{\theta}_{\text{ML}} = X_{(n)}$ be known.
- The probability that $X_{(n)}$ is *not* in the bootstrap sample is $\left(1 - \frac{1}{n}\right)^n$.
- Then, the probability that $X_{(n)}$ is in the bootstrap sample is

$$1 - \left(1 - \frac{1}{n}\right)^n \to 1 - e^{-1} \approx 0.632 \quad \text{for } n \to \infty \; .$$

5.2 Bootstrap estimation of standard errors
5.2.4 An example in which the nonparametric bootstrap does not work

- That is $P(\hat{\theta}^* = \hat{\theta}_{\text{ML}}) \approx 0.632$ for $n \to \infty$, so the distribution of $\hat{\theta}^*$ puts a probability mass of 0.632 on the ML estimator. This is thus reproduced and there is no information gain from these samples!
- *Problem:* $\hat{F}_n$ is not a good estimate for F in the extreme ranges of F.
- In contrast, for parametric bootstrap the following applies

$$X^* = (X_1^*, \dots, X_n^*) \text{ with } X_i^* \sim \text{Unif}(0, \hat{\theta}_{ML})$$

and therefore

$$\mathbb{P}(\hat{\theta}^* = \hat{\theta}_{ML}) = 0 \; .$$

- *So:* nonparametric bootstrap can go wrong!

5.2.5 Two-sample problem for independent samples

5.2 Bootstrap estimation of standard errors
5.2.5 Two-sample problem for independent samples

Let

$$\left.\begin{array}{l} Y_1, \dots, Y_n \overset{\text{i.i.d.}}{\sim} F \\ Z_1, \dots, Z_m \overset{\text{i.i.d.}}{\sim} G \end{array}\right\} \text{independent, for example } \left\{\begin{array}{l} F: \text{ treatment} \\ G: \text{ control} \end{array}\right.$$

and $X = (Y_1, \dots, Y_n, Z_1, \dots, Z_m)$ or $x = (y_1, \dots, y_n, z_1, \dots, z_m)$.
Aim: estimate the standard error of the estimate for the difference $\theta = \underbrace{\mu_Y}_{\mathbb{E}(Y_i)} - \underbrace{\mu_Z}_{\mathbb{E}(Z_i)}$.

5.2 Bootstrap estimation of standard errors
5.2.5 Two-sample problem for independent samples

Consider

$$\hat{\theta} = \bar{y} - \bar{z} \; .$$

Procedure for the b-th bootstrap sample:

$$\begin{array}{rcll} y^{*b} & = & (y_1^{*b}, \dots, y_n^{*b}) & \text{randomly with replacement from } \hat{F}_n \\ z^{*b} & = & (z_1^{*b}, \dots, z_m^{*b}) & \text{randomly with replacement from } \hat{G}_m \end{array}$$

5.2 Bootstrap estimation of standard errors
5.2.5 Two-sample problem for independent samples

Estimation:

$$\underbrace{\widehat{\text{se}}_{F,G}(\hat{\theta})}_{\text{Real world}} = \underbrace{\text{se}_{\hat{F}_n, \hat{G}_m}(\hat{\theta}^*)}_{\substack{\text{Ideal} \\ \text{estimate in} \\ \text{the bootstrap} \\ \text{world}}} \approx \underbrace{\widehat{\text{se}}_B}_{\substack{\text{Approx. of} \\ \text{the ideal} \\ \text{bootstrap} \\ \text{estimate}}} = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left[\hat{\theta}^*(b) - \hat{\theta}^*(\cdot) \right]^2 \right\}^{\frac{1}{2}}$$

with

$$\hat{\theta}^*(b) = \bar{y}^{*b} - \bar{z}^{*b} = \frac{1}{n}\sum_{i=1}^{n} y_i^{*b} - \frac{1}{m}\sum_{i=1}^{m} z_i^{*b}$$

and

$$\hat{\theta}^*(\cdot) = \frac{1}{B}\sum_{b=1}^{B}(\bar{y}^{*b} - \bar{z}^{*b}) = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*(b) \ .$$

5.2.6 Bootstrap for time series

Consider the time series $y_1, y_2, \dots, y_T$ and the centered time series $z_1, z_2, \dots, z_T$ with $z_t = y_t - \bar{y}$ for $t = 1, \dots, T$.
Assumption: z_t is an AR(1)-process

5.2 Bootstrap estimation of standard errors 5.2.6 Bootstrap for time series

$$z_t = \beta z_{t-1} + \varepsilon_t \quad (t = 2, \dots, T)$$

with initial condition z_1, $|\beta| < 1$ and $\varepsilon_t \overset{\text{i.i.d.}}{\sim} F$ for $t = 2, \dots, T$. F is unknown and $\mathbb{E}(\varepsilon_t) = 0$. The least squares estimate for β is:

$$\sum_{t=2}^{T}(z_t - \beta z_{t-1})^2 \rightarrow \min_{\beta} \rightarrow \hat{\beta}.$$

(Since no distribution assumption was made here, ML estimation is not possible.)
Goal: estimate for $\text{se}_{F,\beta}(\hat{\beta})$.
Idea: calculate residuals

5.2 Bootstrap estimation of standard errors 5.2.6 Bootstrap for time series

$$\left.\begin{array}{rcl} \hat{\varepsilon}_2 & = & z_2 - \hat{\beta} z_1, \\ & \vdots & \\ \hat{\varepsilon}_T & = & z_T - \hat{\beta} z_{T-1}. \end{array}\right\} T-1 \text{ residuals}$$

Denote with $\hat{F}_{T-1}$ the empirical distribution function of the $\hat{\varepsilon}_2, \dots, \hat{\varepsilon}_T$. The *b-th bootstrap sample* is obtained as follows:

5.2 Bootstrap estimation of standard errors 5.2.6 Bootstrap for time series

(i) Draw $\varepsilon_2^{*b}, \dots, \varepsilon_T^{*b}$ randomly with replacement from $\hat{F}_{T-1}$.

(ii) Compute recursively

$$\begin{array}{rcl} z_1 & = & y_1 - \bar{y} \\ z_2^{*b} & = & \hat{\beta} z_1 + \varepsilon_2^{*b} \\ z_3^{*b} & = & \hat{\beta} z_2^{*b} + \varepsilon_3^{*b} \\ & \vdots & \\ z_T^{*b} & = & \hat{\beta} z_{T-1}^{*b} + \varepsilon_T^{*b}. \end{array}$$

(iii) Compute $\hat{\beta}^{*b}$ by least-squares from $z_2^{*b}, \dots, z_T^{*b}$.

5.2 Bootstrap estimation of standard errors
5.2.6 Bootstrap for time series

Thus:

$$\widehat{\text{se}}_{F,\beta}(\hat{\beta}) = \text{se}_{\hat{F}_{T-1},\hat{\beta}}(\hat{\beta}^*) \approx \widehat{\text{se}}_B(\hat{\beta}^*) = \left\{\frac{1}{B-1}\sum_{b=1}^{B}\left[\hat{\beta}^{*b} - \hat{\beta}^*(\cdot)\right]^2\right\}^{\frac{1}{2}}$$

with

$$\hat{\beta}^*(\cdot) = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}^{*b}\ .$$

Alternative idea: „Moving Block Bootstrap" (Efron und Tibshirani, 1993).

5.3 Bootstrap in regression models

5.3.1 Bootstrap in the linear model

5.3 Bootstrap in regression models
5.3.1 Bootstrap in the linear model

Data: $(y_i, \boldsymbol{x}_i^\top)$, $i = 1, \dots, n$, for response y_i and covariates $\boldsymbol{x}_i^\top \in \mathbb{R}^{1\times p}$.
We present three bootstrap variants based on the linear model

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

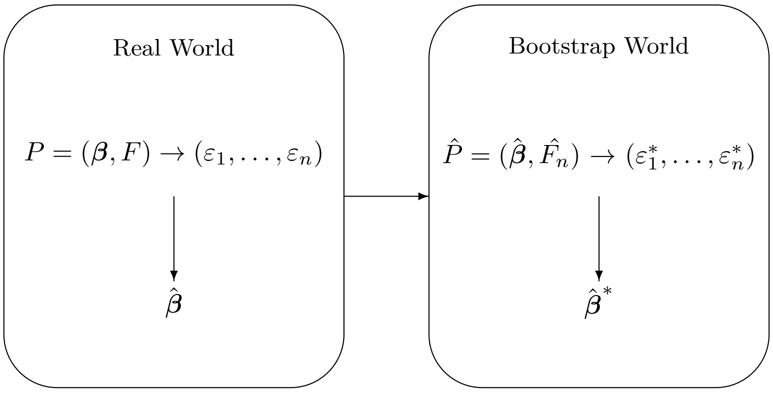
with

$$\varepsilon_i \overset{\text{i.i.d.}}{\sim} F \quad \text{und} \quad \mathbb{E}(\varepsilon_i) = 0$$

for $i = 1, \dots, n$. Extensions to GLMs are possible.

5.3 Bootstrap in regression models
5.3.1 Bootstrap in the linear model

Variant 1: *(non-parametric)* bootstrap of the residuals. We consider a probability model ("Real World") $P = (\boldsymbol{\beta}, F)$, where $\boldsymbol{\beta}$ is the regression parameter and F is the distribution of the residuals.


Real World
$P = (\boldsymbol{\beta}, F) \to (\varepsilon_1, \dots, \varepsilon_n)$
$\hat{\boldsymbol{\beta}}$
Bootstrap World
$\hat{P} = (\hat{\boldsymbol{\beta}}, \hat{F}_n) \to (\varepsilon_1^*, \dots, \varepsilon_n^*)$
$\hat{\boldsymbol{\beta}}^*$

5.3 Bootstrap in regression models
5.3.1 Bootstrap in the linear model

1. Step: Compute $\hat{\boldsymbol{\beta}}$ with the least squares method: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$.

2. Step: Calculate the residuals $\hat{\boldsymbol{\varepsilon}} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

3. Step: For the empirical distribution of residuals $\hat{F}_n$, assign probability mass of $\frac{1}{n}$ on $\hat{\varepsilon}_i$, $i = 1, \dots, n$ (without any further restriction, all residuals are different).

4. Step:
- Draw a sample $\boldsymbol{\varepsilon}^* = (\varepsilon_1^*, \dots, \varepsilon_n^*)$ from $\hat{F}_n$.

- Calculate „new“ bootstrap target variables
$$y_i^* = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + \varepsilon_i^*$$
for $i = 1, \ldots, n$, i.e.
$$\boldsymbol{y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}^*.$$
- Compute the bootstrap least square estimator
$$\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}^*.$$

Result: in this special case, bootstrap leads to the same result as the classical estimation method.
Reason:

5.3 Bootstrap in regression models
5.3.1 Bootstrap in the linear model

$$\begin{aligned}
\mathrm{Var}_{\hat{F}_n}(\hat{\boldsymbol{\beta}}^*) &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \mathrm{Var}_{\hat{F}_n}(\boldsymbol{y}^*)\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\
&= \hat{\sigma}_F^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1},
\end{aligned}$$

since $\mathrm{Var}_{\hat{F}_n}(\boldsymbol{y}^*) = \mathrm{Var}_{\hat{F}_n}(\boldsymbol{\varepsilon}^*) = \hat{\sigma}_F^2 \boldsymbol{I}$ with $\hat{\sigma}_F^2 = \frac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}}{n}$ (model with a constant).
Therefore:

$$\widehat{\mathrm{se}}_F(\hat{\beta}_j) = \mathrm{se}_{\hat{F}}(\hat{\beta}_j^*) = \widehat{\mathrm{se}}_\infty(\hat{\beta}_j^*) = \hat{\sigma}_F \sqrt{[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}]_{jj}}\ .$$

Note: it was assumed that $\boldsymbol{X}$ is a matrix of non-random values (for example a design matrix in the controlled experiment). If $\boldsymbol{y}$ and $\boldsymbol{X}$ are random, it is better to use the following vector sampling.

5.3 Bootstrap in regression models
5.3.1 Bootstrap in the linear model

Variant 2: *vector sampling („Bootstrapping Pairs“)*
Bootstrap samples are drawn randomly from the pairs $(y_1, \boldsymbol{x}_1^\top), \ldots, (y_n, \boldsymbol{x}_n^\top)$ with replacement. Then, each $\hat{\boldsymbol{\beta}}^*$ is calculated again and the rest continues as in the variant 1.
Rule of thumb: bootstrapping of pairs is less susceptible to the violations of assumptions for the (linear) model than bootstrapping of the residuals.

5.3 Bootstrap in regression models
5.3.1 Bootstrap in the linear model

Variant 3: *parametric bootstrap*
A distribution assumption is made for the errors

$$\varepsilon_i \sim F_{par}, \qquad \text{for example } \ \varepsilon_i \sim N(0, \sigma^2),$$

1. Step: Calculate $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_F^2$.

2. Step: Set $y_i^* = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + \varepsilon_i^*$, where $\varepsilon_i^* \sim N(0, \hat{\sigma}_F^2)$.

3. Step: Continue as in variant 1.

Again, no bootstrap is necessary if the normal assumption is made for the errors and if $\hat{\boldsymbol{\beta}}$ is the usual least squares estimate.
Conclusion: Bootstrap is only useful for variant 2.

5.3 Bootstrap in regression models
5.3.1 Bootstrap in the linear model

5.3.2 Bootstrap in the generalized linear model

5.3 Bootstrap in regression models
5.3.2 Bootstrap in the generalized linear model

- Extensions to generalized linear models are possible in principle. However, the question arises which residuals should be used in the case of bootstrapping of residuals, if
- y is a count variable (Poisson distribution) or
  y is binary (Binomial distribution).
- Conceptually, vector sampling is substantially easier here.

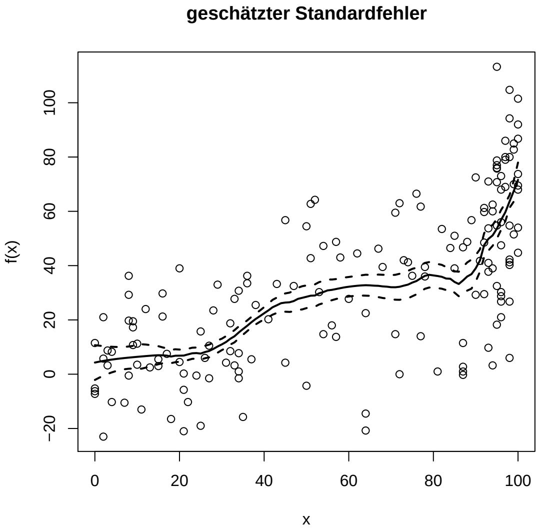
5.3.3 Other applications

5.3 Bootstrap in regression models
5.3.3 Other applications

Example 5.6
Nonparametric regression, for example LOESS

$$y = f(x) + \varepsilon.$$

Here the standard error is estimated point by point.



5.4 Bias estimation using bootstrap

5.4 Bias estimation using bootstrap

Let $X_1, \dots, X_n$ i.i.d. with distribution function F, where F is unknown, $\theta = T(F)$ and $\hat{\theta} = T(\hat{F_n})$ or $\hat{\theta} = \hat{\theta}(x)$.The bias of $\hat{\theta}$ is

$$\text{Bias}_F(\hat{\theta}, \theta) = \mathbb{E}_F(\hat{\theta}) - \theta = \mathbb{E}_F(\hat{\theta}) - T(F)\ .$$

We get the bootstrap bias estimate again with the usual principle:

- Replace F with $\hat{F_n}$,
- Replace $\hat{\theta}$ with $\hat{\theta}^*$,
- Replace θ with $\hat{\theta} = T(\hat{F_n})$.

So:

5.4 Bias estimation using bootstrap

$$\widehat{\text{Bias}}_F(\hat{\theta}, \theta) = \text{Bias}_{\hat{F}_n}(\hat{\theta}^*, \hat{\theta}) = \mathbb{E}_{\hat{F}_n}[\hat{\theta}^*] - T(\hat{F}_n)\ .$$

Remark
$\hat{\theta}$ may or may not be the plug-in estimate.

5.4 Bias estimation using bootstrap

In general, the ideal bootstrap bias estimate $\text{Bias}_{\hat{F}_n}$ must be again approximated by a Monte Carlo simulation:
If we have B independent bootstrap samples, then we can estimate the bias as

$$\widehat{\text{Bias}}_B = \hat{\theta}^*(\cdot) - \underbrace{T(\hat{F}_n)}_{\hat{\theta}}$$

with

$$\hat{\theta}^*(\cdot) = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*(b)$$

Thus, $\widehat{\text{se}}_B$ and $\widehat{\text{Bias}}_B$ can be calculated at the same time from the same bootstrap samples.

5.4.1 Bias correction

5.4 Bias estimation using bootstrap

A bias-corrected estimator is obtained from

5.4.1 Bias correction

$$\begin{aligned}\overline{\theta} &= \hat{\theta} - \widehat{\text{Bias}}_B \\ &= \hat{\theta} - [\hat{\theta}^*(\cdot) - \hat{\theta}] \\ &= 2\hat{\theta} - \hat{\theta}^*(\cdot)\ .\end{aligned}$$

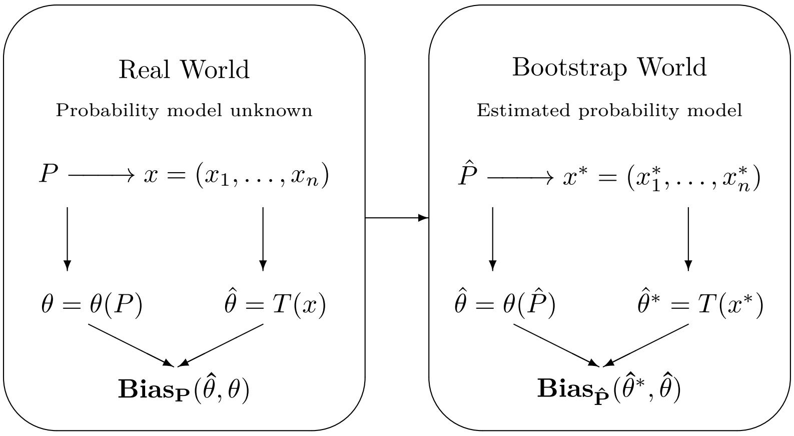
5.4 Bias estimation using bootstrap
5.4.1 Bias correction

Remarks:

1. $\hat{\theta}^*(\cdot)$ itself is *not* a bias-corrected estimate.
2. $\overline{\theta}$ can have a much greater variance than $\hat{\theta}$. Therefore, a bias correction in practice can be „dangerous“.
3. Bias estimation is more difficult than variance or standard error estimation.
4. Jackknife variance and bias estimates fail when T is not „smooth“. For example with the median. Bootstrap works here unless the distribution is „odd“. For example, when the variance is not finite.

5.4 Bias estimation using bootstrap
5.4.1 Bias correction

General scheme for any probability model P:


Real World
Probability model unknown
$P \longrightarrow x = (x_1, \ldots, x_n)$
$\theta = \theta(P)$
$\hat{\theta} = T(x)$
$\mathbf{Bias_P}(\hat{\theta}, \theta)$
Bootstrap World
Estimated probability model
$\hat{P} \longrightarrow x^* = (x_1^*, \ldots, x_n^*)$
$\hat{\theta} = \theta(\hat{P})$
$\hat{\theta}^* = T(x^*)$
$\mathbf{Bias_{\hat{P}}}(\hat{\theta}^*, \hat{\theta})$

5.5 Bootstrap confidence intervals

5.5.1 Introduction

5.5 Bootstrap confidence intervals
5.5.1 Introduction

Normal 90%-confidence interval:

$$\hat{\theta} \pm 1.645 \cdot \widehat{\text{se}}.$$

Normal 95%-confidence interval:

$$\hat{\theta} \pm 1.96 \cdot \widehat{\text{se}}.$$

5.5 Bootstrap confidence intervals
5.5.1 Introduction

Here, $\widehat{\text{se}}$ can also be a bootstrap estimate.
The reason for this is mostly:

$$Z = \frac{\hat{\theta} - \theta}{\widehat{\text{se}}} \overset{\text{approx.}}{\sim} N(0, 1) \quad \text{(asymptotic statement)} \ .$$

The asymptotic distribution is (approximately) independent of θ; Z is called the approximate pivot.
If n is small, the quantiles of the normal distribution can be replaced with the quantiles of the t-distribution

$$\hat{\theta} \pm t_{n-1}^{(1-\alpha/2)} \cdot \widehat{\text{se}} \ .$$

5.5.2 Bootstrap t-interval

5.5 Bootstrap confidence intervals
5.5.2 Bootstrap t-interval

Idea: avoid assumption of the normal distribution, estimate distribution of Z from the data. This is described in the following sections.

Consider

$$Z = \frac{\hat{\theta} - \theta}{\widehat{\text{se}}} \ , \tag{3}$$

where $\widehat{\mathrm{se}}$ initially represents any „reasonable” estimate of the standard error of $\hat{\theta}$.
Idea: estimate the distribution of Z as follows:

5.5 Bootstrap confidence intervals
5.5.2 Bootstrap t-interval

1. Generate B bootstrap samples $x^{*1}, \ldots, x^{*B}$.

2. Compute
$$Z^*(b) = \frac{\hat{\theta}^*(b) - \hat{\theta}}{\widehat{\mathrm{se}}^*(b)} \ ,$$
where $\widehat{\mathrm{se}}^*(b)$ is an estimate of the standard error of $\hat{\theta}^*(b)$. Arrange $Z^*(b)$ in an ascending order.

5.5 Bootstrap confidence intervals
5.5.2 Bootstrap t-interval

3. Estimate the quantiles $\hat{t}^{(\alpha)}$ and $\hat{t}^{(1-\alpha)}$ (for a $(1-2\alpha)$-confidence interval) as
$$\frac{\#\left\{Z^*(b) \leq \hat{t}^{(\alpha)}\right\}}{B} = \alpha \ .$$
where $\#A$ is the cardinality of the set A.

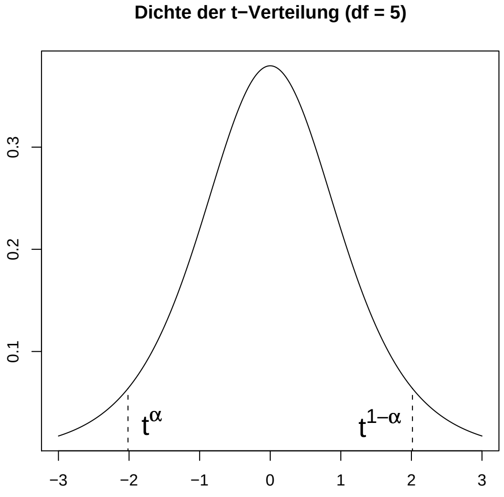
   Example: For $B = 1000$, $\hat{t}^{(0.05)}$ is the 50-th value of the ordered $Z^*(b)$ values, $\hat{t}^{(0.95)}$ is the 950-th value of the ordered $Z^*(b)$-values.

5.5 Bootstrap confidence intervals
5.5.2 Bootstrap t-interval

4. The bootstrap t-interval for the confidence level $1 - 2\alpha$ is then
$$\left[\hat{\theta} - \hat{t}^{(1-\alpha)} \cdot \widehat{\mathrm{se}},\ \hat{\theta} - \hat{t}^{(\alpha)} \cdot \widehat{\mathrm{se}}\right]$$
where $\widehat{\mathrm{se}}$ is from formula (3).

5.5 Bootstrap confidence intervals
5.5.2 Bootstrap t-interval

Analogy with the t-distribution:
$$\left[\hat{\theta} - t^{1-\alpha} \cdot \widehat{\mathrm{se}},\ \hat{\theta} + t^{1-\alpha} \cdot \widehat{\mathrm{se}}\right] \qquad (t^{1-\alpha} = -t^{\alpha}) \ .$$



Note: If $B\alpha$ is not an integer and $\alpha \leq \frac{1}{2}$, then choose $k = \lfloor (B+1)\alpha \rfloor$, that is the largest integer $\leq (B+1)\alpha$. The empirical quantiles are then the k-th and the $(B+1-k)$-th value of the ordered $Z^*(b)$ values.

5.5 Bootstrap confidence intervals 5.5.2 Bootstrap t-interval

Problems:

5.5 Bootstrap confidence intervals 5.5.2 Bootstrap t-interval

1. The bootstrap t-interval can be greatly influenced by outliers.

2. Consider again

$$Z^*(b) = \frac{\hat{\theta}^*(b) - \hat{\theta}}{\widehat{\text{se}}^*(b)} \,.$$

How can one estimate $\widehat{\text{se}}^*(b)$?

5.5 Bootstrap confidence intervals 5.5.2 Bootstrap t-interval

(i) If $\hat{\theta}$ is the mean, then:

$$\widehat{\text{se}}^*(b) = \frac{1}{n} \left\{ \sum_{i=1}^{n} (x_i^{*b} - \bar{x}^{*b})^2 \right\}^{\frac{1}{2}} \quad \text{(Plug-in-estimate).}$$

5.5 Bootstrap confidence intervals 5.5.2 Bootstrap t-interval

(ii) If $\hat{\theta}$ is complicated or no standard formula is available:
→ *Nested bootstrap:*
A bootstrap estimate of the standard error for *each* bootstrap sample. For example, $B = 1000$ und $B^* = 50$

$$BB^* = 1000 \cdot 50 = 50\,000$$

samples are necessary. So, we sample on two nested levels: Real World → Bootstrap World → Nested Bootstrap World.
Advantage: this process can be parallelized.

5.5 Bootstrap confidence intervals 5.5.2 Bootstrap t-interval

3. The bootstrap t-interval is affected by the scale of the parameter, it is not invariant to transformations. With small samples in a nonparametric setup, irregular behavior occurs; here, however, a transformation of the parameters delivers more reliable results.

5.5 Bootstrap confidence intervals 5.5.2 Bootstrap t-interval

Example 5.7 (Transformation of the correlation coefficient)
Let θ be the correlation coefficient. We can get confidence interval for θ in the following 2 ways:

(i) Bootstrap t-interval for θ directly.

(ii) Bootstrap t-interval for

$$\phi = \frac{1}{2} \log \left(\frac{1+\theta}{1-\theta} \right) \quad \textit{(Fisher Z-transformation)}$$

and then transform back the endpoints using the inverse

$$\theta = \frac{e^{2\phi} - 1}{e^{2\phi} + 1},$$

which provides a shorter (= better) confidence interval than in (i).

Result:

5.5 Bootstrap confidence intervals
5.5.2 Bootstrap t-interval

1. Use bootstrap t-interval only for simple problems when θ is a localisation parameter, for example median, trimmed mean or quantile.

2. In complex cases, variance stabilization is necessary.

5.5.3 Bootstrap percentile interval

5.5 Bootstrap confidence intervals
5.5.3 Bootstrap percentile interval

Idea: directly use the empirical distribution of the estimator $\hat{\theta}^*$ from B bootstrap samples.
So:

5.5 Bootstrap confidence intervals
5.5.3 Bootstrap percentile interval

1. Draw $\begin{array}{cccl} x^{*1}, & \ldots, & x^{*B} & B \text{ bootstrap replications} \\ \downarrow & & \downarrow & \\ \hat{\theta}^*(1), & \ldots, & \hat{\theta}^*(B) & \text{with } \hat{\theta}^*(b) = T(x^{*b}). \end{array}$

2. Sort the $\hat{\theta}^*(b)$ in ascending order: $\hat{\theta}^*_{(1)}, \ldots, \hat{\theta}^*_{(B)}$.

3. Compute $B\alpha$ and $B(1-\alpha)$ (or for non-integer number, a modification as in section 5.5.2) and denote with $\hat{\theta}_B^{*(\alpha)}$ or. $\hat{\theta}_B^{*(1-\alpha)}$ the values at the respective positions in the sorted sequence of the bootstrap estimates. Then,

$$\left[\hat{\theta}_{\text{lower}}, \hat{\theta}_{\text{upper}}\right] = \left[\hat{\theta}_B^{*(\alpha)}, \hat{\theta}_B^{*(1-\alpha)}\right]$$

   is the approximate $(1-2\alpha)$-confidence interval.

5.5 Bootstrap confidence intervals
5.5.3 Bootstrap percentile interval

Example: for $B = 2000$ and $\alpha = 0.05$ choose the 100- and 1900-th value from the ordered list.
Alternatively: denote with $\hat{G}_B$ the empirical distribution of $\hat{\theta}^*$. Then

$$\left[\hat{\theta}_{\text{lower}}, \hat{\theta}_{\text{upper}}\right] = \left[\hat{G}^{-1}(\alpha), \hat{G}^{-1}(1-\alpha)\right] .$$

5.5 Bootstrap confidence intervals
5.5.3 Bootstrap percentile interval

Advantages of the percentile method:

(i) It is invariant to (strictly monotonic) transformations.

(ii) It is *range-preserving*, i.e. the percentile interval is in permissible range of the parameter. *Example:* for the correlation coefficient, the interval from the percentile method lies in the range $[-1, 1]$.

Problem: usually insufficient coverage, i.e. the intervals are frequently too optimistic.

5.6 Cross-validation and prediction errors

5.6 Cross-validation and prediction errors

- Prediction error in the regression model (using a quadratic loss function) refers to the expected difference between future and predicted response,

$$\mathbb{E}(Y - \hat{Y})^2.$$

- In an (unordered) classification problem, the prediction error is defined as the probability of a misclassification,
$$P(\hat{Y} \neq Y)\ .$$
- This section deals with estimating the prediction error in both settings.

5.6 Cross-validation and prediction errors

- A possible estimate of the *in*-sample prediction error in the regression model is
$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
or
$$\frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$
where p denotes the number of predictor variables.
- This estimate is however too optimistic, i.e. the true prediction error is underestimated, since the same data is used for the fitting and assessment of the model.
- The test data here is therefore the same as the training data.

5.6 Cross-validation and prediction errors

Ideal situation:

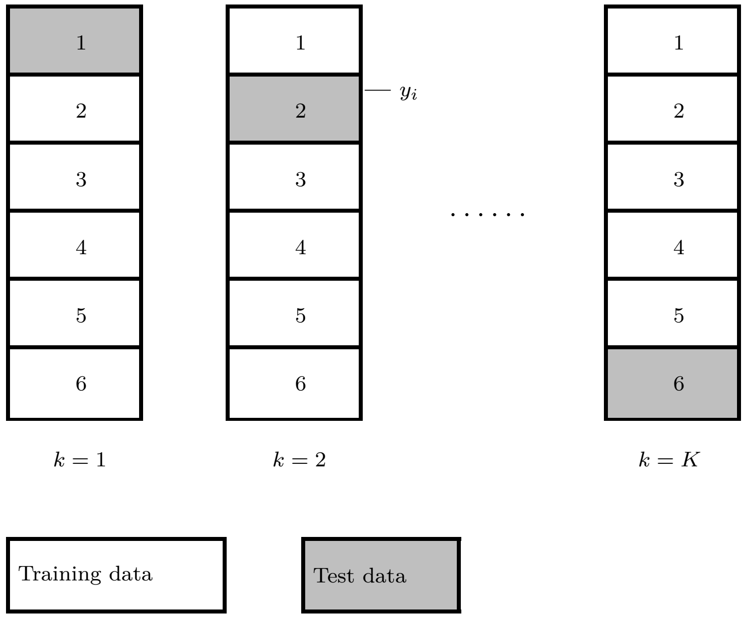
- Use *new* data $y_1^0, \dots, y_m^0$ as test data.
- Use the model estimated from the training data $y_1, \dots, y_n$ for prediction of $y_1^0, \dots, y_m^0$.
- Estimate the prediction error via
$$\frac{1}{m}\sum_{i=1}^{m}(y_i^0 - \hat{y}_i^0)^2.$$

5.6 Cross-validation and prediction errors

- Mostly, however, no additional data is available. If it is, heterogeneity problems can arise. This is where cross-validation comes in.
- For larger data sets, the data set is divided into two (train/test) or three parts (train/validation/test). The latter is usually required in settings where also hyperparameters have to be estimated (e.g. in machine learning).
- For smaller data sets, a popular method is the *k-fold cross validation*. Here the data is divided into K approximately equal parts and for each $k = 1, \dots, K$ the predictions of the k-th part of the data are based on a model relying on the other $K-1$ parts (if hyperparameter estimation is needed, nested cross-validation may be necessary, but we will not treat that case further).

5.6 Cross-validation and prediction errors

Illustration for $K = 6$:



5.6 Cross-validation and prediction errors

- Let $k(i)$ be the part that contains the i-th observation y_i. For example, $k(i) = 2$ in the above graph. Then $\hat{y}_i^{-k(i)}$ denotes the prediction for y_i, calculated without the part $k(i)$, i.e. without the part that contains y_i.
- The estimation of the prediction error by cross-validation is given by

$$\text{CV} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i^{-k(i)} \right)^2 .$$

- Sometimes one uses $K = n$, which is called *"leave-one-out cross validation" (LOOCV)*. However, this is very time-consuming for large n and complex regression procedures.

5.6.1 Bootstrap estimation of the prediction error

5.6 Cross-validation and prediction errors
5.6.1 Bootstrap estimation of the prediction error

Using the example of regression, we will explain how bootstrap can be used in estimating the prediction error. Consider

$$\left.\begin{matrix} Y_1 & \boldsymbol{Z}_1^\top \\ Y_2 & \boldsymbol{Z}_2^\top \\ \vdots & \\ Y_n & \boldsymbol{Z}_n^\top \end{matrix}\right\} \text{ Data } x$$

Goal: prediction of a new observation $(Y_0 | \boldsymbol{Z}_0^\top, x)$ from the population distribution F.

5.6 Cross-validation and prediction errors
5.6.1 Bootstrap estimation of the prediction error

- *So:*

  $x \to \text{Model}|x \to \text{Predictor } \eta_x(\boldsymbol{Z}_0) \text{ for } Y_0 \to \text{Prediction error.}$

- The prediction $\eta_x(\boldsymbol{Z}_0)$ is thus based on the model based on x.
- The *prediction error* for $\eta_x(\boldsymbol{Z}_0)$ is defined by

  $$\operatorname{err}(x, F) \equiv \mathbb{E}_{0F}(Q(Y_0, \eta_x(\boldsymbol{Z}_0))$$

  where Q is the loss function, for example $Q[y, \eta] = (y - \eta)^2$.
- $\mathbb{E}_{0F}$ stands for the expectation on a new observation $(Y_0, \boldsymbol{Z}_0^\top)$ from F.
- The *(apparent error in sample)* is

  $$\operatorname{err}(x, \hat{F}_n) = \mathbb{E}_{0\hat{F}_n}(Q(Y_0, \eta_x(\boldsymbol{Z}_0)) = \frac{1}{n}\sum_{i=1}^{n} Q[y_i, \eta_x(\boldsymbol{z}_i)].$$

- However, this error is too optimistic.

5.6 Cross-validation and prediction errors
5.6.1 Bootstrap estimation of the prediction error

- With the plug-in principle, one gets an improved estimate as follows: Let $x^{*1}, \ldots, x^{*B}$ B be bootstrap samples with

  $$\begin{aligned} x^{*1} &= \{(y_1^{*1}, \boldsymbol{z}_1^{*1^\top}), \ldots, (y_n^{*1}, \boldsymbol{z}_n^{*1^\top})\} \\ &\vdots \\ x^{*B} &= \{(y_1^{*B}, \boldsymbol{z}_1^{*B^\top}), \ldots, (y_n^{*B}, \boldsymbol{z}_n^{*B^\top})\}\ . \end{aligned}$$

- Then (for any b)

  $$\operatorname{err}(x^{*b}, \hat{F}_n) = \frac{1}{n}\sum_{i=1}^{n} Q[y_i, \eta_{x^{*b}}(\boldsymbol{z}_i)]$$

  is a plug-in estimate for $\operatorname{err}(x, F)$, where y_i and $\boldsymbol{z}_i$ are from the original sample. Thus, the model calculated on the basis of x^{*b} is used to estimate the prediction error in the original sample.

5.6 Cross-validation and prediction errors
5.6.1 Bootstrap estimation of the prediction error

But we would like an estimate for the (*average prediction error*) $\mathbb{E}_F[\operatorname{err}(x, F)]$:

$$\begin{tikzcd} \mathbb{E}_F[\operatorname{err}(x, F)] \arrow[d, "\text{ideal bootstrap estimate}"] \\ \mathbb{E}_{\hat{F}_n}[\operatorname{err}(x^*, \hat{F}_n)] = \mathbb{E}_{\hat{F}_n}\left\{\frac{1}{n}\sum_{i=1}^{n} Q[y_i, \eta_{x^*}(\boldsymbol{z}_i)]\right\} \arrow[d, "\text{approx. bootstrap estimate}"] \\ \hat{\mathbb{E}}_{\hat{F}_n}[\operatorname{err}(x^*, \hat{F}_n)] = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{n}\sum_{i=1}^{n} Q[y_i, \eta_{x^{*b}}(\boldsymbol{z}_i)]\ . \end{tikzcd} \qquad (4)$$

5.6 Cross-validation and prediction errors
5.6.1 Bootstrap estimation of the prediction error

- Compare
$$\hat{\mathbb{E}}_{\hat{F}_n}[\mathrm{err}(x^*, \hat{F}_n)]$$
with the so-called *bootstrap sample error*
$$\hat{\mathbb{E}}_{\hat{F}_n}[\mathrm{err}(x^*, \hat{F}_n^*)] = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{n}\sum_{i=1}^{n} Q[y_i^{*b}, \eta_{x^{*b}}(\boldsymbol{z}_i^{*b})].$$
- Here $\hat{F}_n^*$ is the empirical distribution function that results from x^*.
- Again, this error is generally too optimistic.

For a bias correction, one considers the *average optimism*:

5.6 Cross-validation and prediction errors
5.6.1 Bootstrap estimation of the prediction error

$$w(F) = \mathbb{E}_F(\mathrm{err}(x, F)) - \underbrace{\mathbb{E}_F(\mathrm{err}(x, \hat{F}_n))}_{\text{average apparent error}}$$

$$\Big\downarrow \text{ plug-in principle, ideal bootstrap estimate}$$

$$w(\hat{F}_n) = \mathbb{E}_{\hat{F}_n}(\mathrm{err}(x^*, \hat{F}_n)) - \mathbb{E}_{\hat{F}_n}(\mathrm{err}(x^*, \hat{F}_n^*))$$

$$\Big\downarrow \text{ approximate bootstrap estimate}$$

$$\hat{w}(\hat{F}_n) = \frac{1}{Bn}\left\{\sum_{b=1}^{B}\sum_{i=1}^{n} Q[y_i, \eta_{x^{*b}}(\boldsymbol{z}_i)] - \sum_{b=1}^{B}\sum_{i=1}^{n} Q[y_i^{*b}, \eta_{x^{*b}}(\boldsymbol{z}_i^{*b})]\right\} .$$

5.6 Cross-validation and prediction errors
5.6.1 Bootstrap estimation of the prediction error

- The final estimate of the average prediction error
$$\mathbb{E}_F(\mathrm{err}(x, F)) = \mathbb{E}_F(\mathrm{err}(x, \hat{F}_n)) + w(F)$$
is carried out by
$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} Q[y_i, \eta_x(\boldsymbol{z_i})]}_{\text{in sample, original data}} + \hat{w}(\hat{F}_n)\ . \tag{5}$$
- *Conclusion:* (5) is better than (4).

5.6.2 The 0.632 bootstrap estimator

5.6 Cross-validation and prediction errors
5.6.2 The 0.632 bootstrap estimator

Idea: use only the cases that are not included in the bootstrap sample to estimate the prediction error. The probability that a case is in the bootstrap sample is given by

$$1 - \left(1 - \frac{1}{n}\right)^n \approx 0.632.$$

5.6 Cross-validation and prediction errors

5.6.2 The 0.632 bootstrap estimator

The estimated error is then

$$\begin{aligned}\widehat{\text{err}}^{0.632} &= \underbrace{\text{err}(x,\hat{F}_n)}_{\text{apparent error}} + 0.632\,(\hat{\varepsilon}_0 - \text{err}(x,\hat{F}_n)) \\ &= 0.368\,\text{err}(x,\hat{F}_n) + 0.632\,\hat{\varepsilon}_0\end{aligned}$$

with

$$\hat{\varepsilon}_0 = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{B_i}\sum_{b\in C_i} Q[y_i, \eta_{x^{*b}}(\boldsymbol{z}_i)]\right\}.$$

Here
C_i: set of all indices of the bootstrap samples, that do not contain observation i.
B_i: number of bootstrap samples, that do not contain observation i.

5.7 Other methods

5.7.1 Wild bootstrap

5.7 Other methods
5.7.1 Wild bootstrap

- Suggested by Wu (1986): Jackknife, Bootstrap and other resampling methods in regression analysis (with discussions), Annals of Statistics, Volume 14, 1261-1350.
- Idea: bootstrap the residuals $\hat{\varepsilon}_i^* = V_i^*\hat{\varepsilon}_i$, with V_i^* *i.i.d.* from the two-point distribution

$$\begin{aligned}P(V_i^* &= \frac{\sqrt{5}+1}{2}) = \frac{\sqrt{5}-1}{2\sqrt{5}} \\ P(V_i^* &= -\frac{\sqrt{5}-1}{2}) = \frac{\sqrt{5}+1}{2\sqrt{5}}\ .\end{aligned}$$

- The following holds:

$$\begin{aligned}E(V_i^*) &= \frac{\sqrt{5}+1}{2}\frac{\sqrt{5}-1}{2\sqrt{5}} - \frac{\sqrt{5}-1}{2}\frac{\sqrt{5}+1}{2\sqrt{5}} = 0 \\ \text{Var}(V_i^*) &= \frac{5+\sqrt{5}+1}{4}\frac{\sqrt{5}-1}{2\sqrt{5}} + \frac{5-\sqrt{5}+1}{2\sqrt{5}}\frac{\sqrt{5}+1}{2\sqrt{5}} = 1.\end{aligned}$$

5.7.2 Bayesian Bootstrap

5.7 Other methods
5.7.2 Bayesian bootstrap

- Weighting of the original sample and calculation of weighted statistics.
- Weights are drawn randomly: draw $(n-1)$ random numbers $u_1,\ldots,u_{n-1}$ from uniform distribution on $(0,1)$, order the numbers $u_{(0)}, u_{(1)}, \ldots, u_{(n-1)}, u_{(n)}$, with $u_{(0)} = 0$ and $u_{(n)} = 1$. Calculate the n gaps/weights $g_i = u_{(i)} - u_{(i-1)}$, $i = 1,\ldots,n-1$, leading to $g = (g_1,\ldots,g_n)$ as probability vector, and use these for the weighing of the original sample
- Note that for example $\text{E}(G_i) = 1/n$
- Rubin (1981): weights only on observed values not on unobserved ones. Simulates the posterior distribution of the statistics under an improper Dirichlet prior distribution. Can be understood as a critique of bootstrap and Bayesian bootstrap.

5.7.3 Bootstrap tests

5.7 Other methods
5.7.3 Bootstrap tests

- The basic idea is similar to that of permutation tests.
- Permutation tests : draw without replacement. Bootstrap: draw with replacement.
- Simple case : two-sample hypothesis. Assumptions:

$$\begin{array}{rcl} Y_i & \sim & F(.) \quad i.i.d., i = 1, ..., n \\ Z_j & \sim & G(.) \quad i.i.d., i = 1, ..., m \end{array}$$

- Null hypothesis:

$$H_0 : F = G \ .$$

- Data: $\boldsymbol{y} = (y_1, \ldots, y_n)$ and $\boldsymbol{z} = (z_1, \ldots, z_m)$. Combined sample: $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{z})$. Under H_0 both samples and the combined sample come from the same distribution.
- Requires test statistic t, that is sensitive to differences in F and G (keyword: power of a test)

5.7 Other methods
5.7.3 Bootstrap tests

- After t has been set, we simulate the distribution of t under the null hypothesis
- Let (e.g.) $t(\boldsymbol{x}) = \bar{y} - \bar{z}$ with $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{z} = \frac{1}{m}\sum_{j=1}^{m} z_j$.
- Under $H_0 : F = G = F_0$ we can draw bootstrap samples of size $n + m$ from the combined sample $\boldsymbol{x}$:
  1. Draw B bootstrap samples $\boldsymbol{x}^{*b}$ of size $n + m$ with replacement from $\boldsymbol{x}$
  2. Consider $\boldsymbol{x}^{*b}$ as $(\boldsymbol{y}^{*b}, \boldsymbol{z}^{*b})$ and use the first n observations for the computation of $\bar{y}^{*b}$ and the remaining m observations for the computation of $\bar{z}^{*b}$ and calculate $t(\boldsymbol{x}^{*b}) = \bar{y}^{*b} - \bar{z}^{*b}$ for $b = 1, \ldots, B$.
  3. The bootstrap p-value (also often: ASL, achieved significance level) is then

$$\text{p-value}_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B} 1\{|t(\boldsymbol{x}^{*b})| \geq |t(\boldsymbol{x})|\}$$

5.7 Other methods
5.7.3 Bootstrap tests

- Studentized statistics are usually preferred, for example

$$t(\boldsymbol{x}) = \frac{\bar{y} - \bar{z}}{s_y^2/n + s_z^2/m}$$

  with $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ and $s_z^2 = \frac{1}{m-1}\sum_{j=1}^{m}(z_j - \bar{z})^2$
- Permutation test: the pooled sample is also the basis. However, the sample is drawn without replacement, i.e. a permutation of the pooled sample is considered.
- Power of the test depends on the alternative H_1 .
- Special cases need special test statistics.
- $H_0 : \mu_y = \mu_z$ (equality of expected values): bootstrap test is possible, permutation test only under the additional assumption $F = G$.

5.7.4 Bagging

- Bagging: bootstrap aggregation, see Hastie et al., Elements of Statistical Learning, Chapters 8 and 15

- Example: calculation of the average of forecasts / predictions based on bootstrap samples.

- Only useful if the forecast is non-linear or adaptive function of the covariates, e.g. regression/classification trees.

- Hope: variance reduction of forecasts, especially for unstable processes, such as trees

- Average of B iid random variables $X_1, \ldots, X_B$ with variance σ^2 has variance σ^2/B, for correlated RVs with the same pairwise correlation $\rho > 0$ the variance is equal to

$$\begin{aligned} \mathrm{Var}(\bar{X}) &= 1/B^2(B\sigma^2 + 2B(B-1)/2\rho\sigma^2) \\ &= \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \ . \end{aligned}$$

  Only the second term approaches 0 for $B \to \infty$. Random forest: further decorrelation of trees by random and limited selection of split variables at each step.