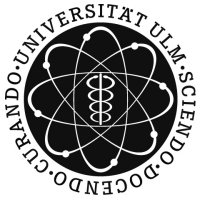


ulm university universität
uulm

Mathematical Statistics

Lecture Notes

Prof. Dr. Evgeny Spodarev

Ulm

2024

Preface

The present lecture notes aim to give an introduction to different aspects of modern statistics. They are, in their current state, a result of holding lectures on statistics at Ulm University in the years 2010-2023 for students of mathematical bachelor's and master's programs.
The goal of the lectures is to provide an overview of typical problem settings and approaches to statistical inference. Additionally it aims to present a middle ground between practically orientated applied statistical monographs (which are usually mathematically sparse) and arid books on mathematical statistics. Whether I actually succeeded in finding said middle ground, shall be decided by the reader.
I would like to thank my colleagues at the Institute of Stochastics for their support and exhilarating discussions during the making of these notes. A special thanks goes to Linus Lach for the English translation of the German version and the creation of figures which accompany the text. I am also indebted to Tobias Brosch for the initial creation of the German LaTeX–version and to Viet Hoang for the many corrections.

Ulm, July 11, 2025
Evgeny Spodarev

Contents

Chapter 1

Point Estimation

Let $(x_1, \ldots, x_n)$ be a given sample. Assume that it is a realization of a random sample $(X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ are independent identically distributed (i.i.d.) random variables with unknown distribution F. Further assume that F is an element of a parametric family of distributions given by $\{F_\theta : \theta \in \Theta\}$. Here $\theta = (\theta_1, \ldots, \theta_m) \in \Theta$ denotes the *m-dimensional parameter vector* of the distribution F_θ, and $\Theta \subset \mathbb{R}^m$ is the so called *parameter space* (a Borel subset of $\mathbb{R}^m$, which is composed of all valid parameter values). The parametrization $\theta \mapsto F_\theta$ is set to be *identifiable*, under the assumption that $F_{\theta_1} \neq F_{\theta_2}$ for $\theta_1 \neq \theta_2$.
An important task in statistics, discussed in this chapter, is the estimation of the parameter vector θ (or a part of θ) on the basis of a given sample $(x_1, \ldots, x_n)$. In this context, the described procedure is called *point estimation* with respect to a *point estimator* $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^m$, which is a valid sample function. Usually one assumes that

$$P\left(\hat{\theta}(X_1, \ldots, X_n) \in \Theta\right) = 1\,,$$

even though exceptions exist. The probability space $(\Omega, \mathcal{F}, P)$ on which the random sample is defined has yet to be specified thoroughly. Here, the so called *canonical probability space* comes into play, which is defined by

$$\Omega = \mathbb{R}^\infty\,, \qquad \mathcal{F} = \mathcal{B}_\mathbb{R}^\infty = \mathcal{B}_\mathbb{R} \otimes \mathcal{B}_\mathbb{R} \otimes \ldots$$

with probability measure P given by

$$P\left(\{\omega = (\omega_1, \ldots, \omega_n, \ldots) \in \mathbb{R}^\infty : \omega_{i_1} \leq x_{i_1}, \ldots, \omega_{i_k} \leq x_{i_k}\}\right) = \prod_{j=1}^{k} F_\theta(x_{i_j})$$

for all $k \in \mathbb{N}$ and $1 \leq i_1 < \cdots < i_k$. In order to emphasize that P depends on θ, the notation P_θ, E_θ and Var_θ is introduced for the measure P as well as the expectation E and variance Var with respect to P.

On the canonical probability space $(\Omega, \mathcal{F}, P_\theta)$ it holds that $X_i(\omega) = \omega_i$ (projection on the i'th coordinate), $i = 1, \ldots, n$, with

$$P_\theta(X_i \leq x_i) = P_\theta(\{\omega \in \Omega : \omega_i \leq x_i\}) = F_\theta(x_i)\,, \quad i = 1, \ldots, n, \quad x_i \in \mathbb{R}\,.$$

1.1 Parametric families of reference distributions

In the lecture "Elementary Probability Theory" some parametric families have already been introduced. In this section, more parametric families of distributions that play a special role (e.g. as reference distributions in estimation theory, statistical tests and confidence intervals) will be presented.

1.1.1 Gamma distribution

First, consider the following special functions:

1. The *Gamma function:*

$$\Gamma(p) = \int_0^\infty x^{p-1} e^{-x} dx, \qquad \text{for } p > 0\,.$$

   The following properties hold:

   - $\Gamma(1) = 1$
   - $\Gamma(1/2) = \sqrt{\pi}$,
   - $\Gamma(p+1) = p\Gamma(p)$ for all $p > 0$,
   - $\Gamma(n+1) = n!$ for all $n \in \mathbb{N}$.

2. The *Beta function:*

$$B(p, q) = \int_0^1 t^{p-1}(1-t)^{q-1}\, dt\,, \quad p, q > 0\,.$$

   The following properties hold:

   - $B(p, q) = B(q, p)$,
   - $B(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$ for all $p, q > 0$,

Definition 1.1.1. The *Gamma distribution* with parameters $\lambda > 0$ and $p > 0$ is an absolutely continuous distribution with probability density function

$$f_X(x) = \begin{cases} \frac{\lambda^p x^{p-1}}{\Gamma(p)} e^{-\lambda x}\,, & x \geq 0\,, \\ 0\,, & x < 0\,. \end{cases} \tag{1.1}$$

Denote by $X \sim \Gamma(\lambda, p)$ a random variable X which is Gamma distributed with parameters λ and p. Obviously $X \geq 0$ almost surely (a.s.).

Exercise 1.1.2. Show that (1.1) is indeed a probability density function.

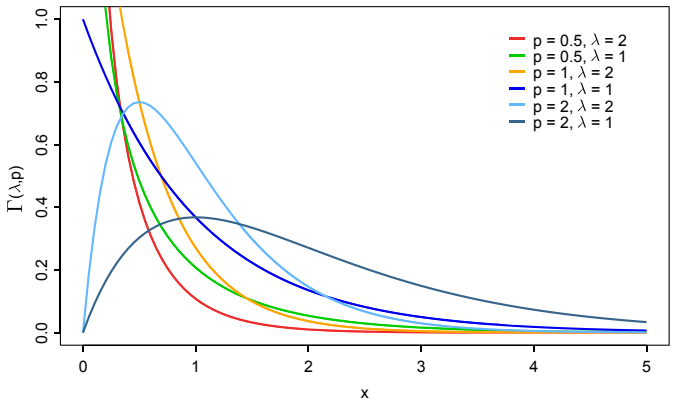


Figure 1.1: Probability density function of the Gamma distribution with various choices for the parameters $\lambda > 0$ and $p > 0$.

Example 1.1.3.

1. The Gamma distribution is often used for modeling small and medium sized insurance claims.

2. If $p = 1$, then $\Gamma(\lambda, 1) = Exp(\lambda)$, i.e. the *Exponential distribution* with parameter $\lambda > 0$.

Theorem 1.1.4. Let $X \sim \Gamma(\lambda, p)$.

1. The moment generating function $\Psi_X(s)$ of X is given by

$$\Psi_X(s) = \mathrm{E}e^{sX} = \frac{1}{(1 - s/\lambda)^p}, \quad s < \lambda.$$

   The characteristic function $\varphi_X(s)$ of X is given by

$$\varphi_X(s) = \mathrm{E}e^{isX} = \frac{1}{(1 - is/\lambda)^p}, \quad s \in \mathbb{R}.$$

2. The k-th moments of X are given by

$$\mathrm{E}X^k = \frac{p(p+1) \cdot \cdots \cdot (p+k-1)}{\lambda^k}, \quad k \in \mathbb{N}.$$

Proof

1. Consider

$$\begin{aligned}\Psi_X(s) = \int_0^\infty e^{sx} f_X(x)\,dx &= \frac{\lambda^p}{\Gamma(p)} \int_0^\infty x^{p-1} e^{\overbrace{(s-\lambda)}^{<0}x}\,dx \\ &\underset{-(s-\lambda)x=y}{=} \frac{\lambda^p}{\Gamma(p)} \int_0^\infty \frac{y^{p-1}}{(-(s-\lambda))^p} e^{-y}\,dy = \frac{\lambda^p \Gamma(p)}{\Gamma(p)(\lambda-s)^p} \\ &= \left(\frac{\lambda}{\lambda-s}\right)^p = \frac{1}{(1-s/\lambda)^p}\,, \quad \lambda > s\,.\end{aligned}$$

If $s \in \mathbb{C}$ and $\mathrm{Re}\,(s) < \lambda$, then $\Psi_X(s)$ is holomorphic on D, where $D \coloneqq \{s = x + iy \in \mathbb{C} : x < \lambda\}$. It holds that

$$\Psi_X(s) = \varphi_X(-is)\,,$$

for $s = it$, $0 < \lambda$, which implies that

$$\varphi_X(t) = \Psi_X(it)\,, \qquad t \in \mathbb{R}.$$

Ultimately, this yields

$$\varphi_X(t) = \frac{1}{(1-it/\lambda)^p}\,, \quad t \in \mathbb{R}\,.$$

2.

$$\mathrm{E}X^k = \Psi^{(k)}(0) \Longrightarrow \mathrm{E}X^k = \frac{p \cdot (p+1) \cdot \ldots \cdot (p+k-1)}{\lambda^k}\,, \quad k \in \mathbb{N}\,.$$

□

Corollary 1.1.5 (Stability of the Gamma distribution)**.** If $X \sim \Gamma(\lambda, p_1)$, $Y \sim \Gamma(\lambda, p_2)$ and X and Y are independent, then $X + Y \sim \Gamma(\lambda, p_1 + p_2)$.

Proof It holds that

$$\begin{aligned}\varphi_{X+Y}(s) &= \varphi_X(s) \cdot \varphi_Y(s) \\ &= \frac{1}{(1-is/\lambda)^{p_1}} \cdot \frac{1}{(1-is/\lambda)^{p_2}} \\ &= \left(\frac{1}{1-is/\lambda}\right)^{p_1+p_2} \\ &= \varphi_{\Gamma(\lambda, p_1+p_2)}(s)\,.\end{aligned}$$

Since the characteristic function uniquely determines the distribution of a random variable, $X + Y \sim \Gamma(\lambda, p_1 + p_2)$ holds. □

Example 1.1.6. Let $X_1, \ldots, X_n \sim Exp(\lambda)$ be independent. By Corollary 1.1.5 it holds that $X = X_1 + \ldots + X_n \sim \Gamma(\lambda, \underbrace{1 + \ldots + 1}_{n}) = \Gamma(\lambda, n)$, since $Exp(\lambda) = \Gamma(\lambda, 1)$. This special case of the Gamma distribution is also called *Erlang distribution* with parameters $\lambda > 0$ and $n \in \mathbb{N}$. Notation: $X \sim Erl(\lambda, n)$.

$$\text{In summary:} \quad Erl(\lambda, n) = \Gamma(\lambda, n), \quad \lambda > 0,\ n \in \mathbb{N}.$$

Interpretation: In risk theory the random variables X_i represent interarrival times for the individual damages. Here $X = \sum_{i=1}^{n} X_i$ represents the occurrence time of the n-th loss with $X \sim Erl(\lambda, n)$.

Definition 1.1.7 (χ^2 distribution)**.** X is a χ^2 distributed random variable with k degrees of freedom (Notation: $X \sim \chi^2_k$), if $X \stackrel{d}{=} X_1^2 + \ldots + X_k^2$, where $X_1, \ldots, X_k \sim N(0, 1)$ are i.i.d. random variables.

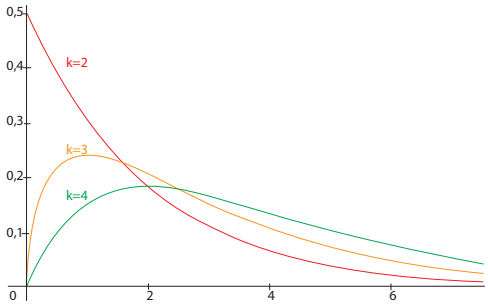


Figure 1.2: Probability density function of the χ^2_k distribution with $k = 2, 3, 4$ degrees of freedom.

Theorem 1.1.8 (χ^2 distribution: Special case of the Gamma distribution with $\lambda = 1/2$, $p = k/2$)**.** If $X \sim \chi^2_k$, then

1. $X \sim \Gamma(1/2, k/2)$, i.e.

$$f_X(x) = \begin{cases} \dfrac{x^{k/2-1} e^{-x/2}}{2^{k/2}\Gamma(k/2)}, & x \geq 0 \\ 0, & x < 0 \end{cases}. \tag{1.2}$$

2. In particular $\mathrm{E}X = k$, $\mathrm{Var}\, X = 2k$.

Proof

1. Let $X = X_1^2 + \ldots + X_k^2$ with $X_i \sim N(0,1)$ i.i.d. random variables. Calculating the *distribution function of* X_i^2 by [33, Satz 3.6.4] and using
$$f_{X_1^2}(x) = \frac{1}{2\sqrt{x}} \left(f_{X_1}(\sqrt{x}) + f_{X_1}(-\sqrt{x}) \right)$$
yields
$$\begin{aligned} P(X_1^2 \leq x) &\underset{y^2=t}{=} \int_0^x \left(\frac{1}{\sqrt{2\pi}} e^{\frac{-t}{2}} \frac{1}{2\sqrt{t}}\, dt + \frac{1}{\sqrt{2\pi}} e^{-t/2} \frac{1}{2\sqrt{t}} \right) dt \\ &= \int_0^x \frac{(1/2)^{-1/2} t^{1/2-1}}{\Gamma(1/2)} e^{-t/2}\, dt\,, \qquad x \geq 0\,. \end{aligned}$$
Thus $X_1^2 \sim \Gamma(1/2, 1/2) \Longrightarrow X \sim \Gamma(1/2, \underbrace{1/2 + \ldots + 1/2}_{k \text{ times}}) = \Gamma(1/2, k/2)$ and therefore (1.2) with respect to the density holds.

2. Because of the additivity of the expected value and the independence of the X_i, it holds that
$$\mathrm{E}X = k \cdot \mathrm{E}X_1^2\,, \quad \mathrm{Var}\, X = k \mathrm{Var}\, X_1^2\,, \quad \mathrm{E}(X_1^2) = \mathrm{E}(\Gamma(1/2, 1/2)) = 1 \text{ by Theorem 1.1.4, 2}\,.$$
Indeed,
$$\begin{aligned} \mathrm{E}(X_1^2) = \frac{1/2}{1/2} = 1\,, \quad \mathrm{E}(X_1^4) = \frac{1/2(1/2+1)}{(1/2)^2} = \frac{3/4}{1/4} = 3, \\ \mathrm{Var}\, X_1^2 = \mathrm{E}(X_1^4) - \left(\mathrm{E}(X_1^2)\right)^2 = 3 - 1 = 2. \end{aligned}$$

□

1.1.2 Student's[1] t distribution

Definition 1.1.9. Let X and Y be independent random variables, where $X \sim N(0,1)$ and $Y \sim \chi_r^2$. The random variable
$$U \stackrel{d}{=} \frac{X}{\sqrt{Y/r}}$$
is called *Student* or t *distributed with* r *degrees of freedom.* Notation: $U \sim t_r$.

Theorem 1.1.10 (Probability density function of the t distribution)**.** If $X \sim t_r$, then

1.
$$f_X(x) = \frac{1}{\sqrt{r} B\left(\frac{r}{2}, \frac{1}{2}\right)} \cdot \frac{1}{\left(1 + \frac{x^2}{r}\right)^{\frac{r+1}{2}}}\,, \qquad x \in \mathbb{R}\,.$$

[1] Named after the mathematician William Sealy Gosset, who signed his work under the pseudonym "Student".

2. $\mathrm{E}X = 0\,, \quad \mathrm{Var}\, X = \frac{r}{r-2}\,, \quad r \geq 3.$

Remark 1.1.11.

1. Student's t distribution is symmetric. In particular,

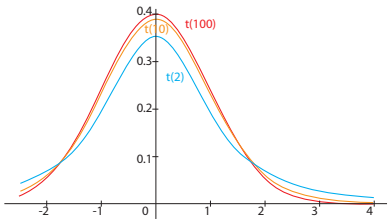


Figure 1.3: Probability density function f of the t distribution for $r = 2, 10, 100$

$$t_{r,\alpha} = -t_{r,1-\alpha}, \quad \alpha \in (0,1),$$

where $t_{r,\alpha}$ is the α quantile of the Student's distribution with r degrees of freedom.

2. For $r \to \infty$ it holds that $f_r(x) \to \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,, \quad x \in \mathbb{R}$. (Proof: Exercise)

3. For $r = 1$ the t distribution coincides with the standard Cauchy distribution, i.e. it holds that $t_1 = Cauchy(0,1)$ with probability density function $f(x) = \frac{1}{\pi(1+x^2)}$. The expected value of t_1 doesn't exist.

Proof of Theorem 1.1.10:

1. It holds that $X := \varphi(Y,Z)$, where $\varphi(x,y) = \frac{x}{\sqrt{y/r}}$ and $V = (Y,Z)$ is a two dimensional random vector with $Y \sim N(0,1)$ and $Z \sim \chi^2_r$ independent of Y. The *density transformation theorem* [33, Theorem 3.6.6] states that

$$f_{\varphi(V)}(x) = f_V(\varphi^{-1}(x))|J|,$$

where $|J| = |\det J|$, where $J = \left(\frac{\partial \varphi_i^{-1}(x)}{\partial x_j}\right)_{i,j=1}^{n}$ denotes the Jacobi matrix of the function $\varphi = (\varphi_1, \ldots, \varphi_n) : \mathbb{R}^n \to \mathbb{R}^n$. Computing φ^{-1}, where $\varphi : (x,y) \mapsto (v,w)$ as above, with $v = \frac{x}{\sqrt{y/r}}$, $w = y$ yields

$$v = \frac{x}{\sqrt{\frac{y}{r}}} \Rightarrow x = v\sqrt{\frac{y}{r}} = v\sqrt{\frac{w}{r}}\,.$$

Thus,

$$\varphi^{-1} : (v,w) \mapsto \left(v\sqrt{\frac{w}{r}}, w\right),$$

and the Jacobi matrix is given by

$$J = \begin{pmatrix} \frac{\partial \varphi_1^{-1}}{\partial v} & \frac{\partial \varphi_1^{-1}}{\partial w} \\ \frac{\partial \varphi_2^{-1}}{\partial v} & \frac{\partial \varphi_2^{-1}}{\partial w} \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{w}{r}} & \frac{v}{2\sqrt{wr}} \\ 0 & 1 \end{pmatrix}.$$

For $V = (Y, Z)$ with Y and Z independent it follows that

$$f_V(x,y) = f_Y(x) \cdot f_Z(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \cdot \frac{y^{r/2-1} e^{-y/2}}{\Gamma(r/2) 2^{r/2}} = \frac{y^{r/2-1} e^{-\frac{y+x^2}{2}}}{2^{\frac{r+1}{2}} \Gamma(1/2) \Gamma(r/2)}$$

for all $x \in \mathbb{R}$ and $y > 0$. The density transformation theorem ultimately yields

$$\begin{aligned} f_X(v) &= \int_0^\infty f_{\varphi(V)}(u,w) dw = \int_0^\infty f_V(\varphi^{-1}(v,w)) |J| \, dw \\ &= \int_0^\infty \frac{e^{-(v^2 \frac{w}{r} + w)/2} w^{r/2-1}}{2^{\frac{r+1}{2}} \Gamma(1/2) \Gamma(r/2)} \sqrt{w/r} \, dw \\ &= \frac{1}{\sqrt{r} 2^{\frac{r+1}{2}} \Gamma(1/2) \Gamma(r/2)} \cdot \int_0^\infty w^{\frac{r-1}{2}} e^{-\overbrace{\frac{\frac{v^2}{r}+1}{2} \cdot w}^{=t}} dw \\ &\underset{w = \frac{2t}{v^2/r+1}}{=} \frac{1}{\sqrt{r} 2^{\frac{r+1}{2}} \Gamma(1/2) \Gamma(r/2)} \cdot \int_0^\infty \frac{2^{\frac{r-1}{2}+1} t^{\frac{r-1}{2}}}{(v^2/r+1)^{\frac{r-1}{2}+1}} e^{-t} dt \\ &= \frac{2^{\frac{r+1}{2}} \Gamma(\frac{r+1}{2})}{(\frac{v^2}{r}+1)^{\frac{r+1}{2}} \sqrt{r} 2^{\frac{r+1}{2}} \Gamma(1/2) \Gamma(r/2)} \\ &= \frac{1}{\sqrt{r} B(r/2, 1/2) (1 + v^2/r)^{\frac{r+1}{2}}} \end{aligned}$$

2. Exercise.

□

1.1.3 Fisher-Snedecor distribution (F distribution)

Definition 1.1.12. Let $X \overset{d}{=} \frac{U_r/r}{U_s/s}$, where $U_r \sim \chi_r^2$, $U_s \sim \chi_s^2$, $r, s \in \mathbb{N}$, U_r, U_s are independent. Then, X is *Fisher or F distributed* with $r, s \in \mathbb{N}$ degrees of freedom. Notation: $X \sim F_{r,s}$.

Lemma 1.1.13. Let $X \sim F_{r,s}$, $r, s \in \mathbb{N}$. Then, X is absolutely continuously distributed with probability density function

$$f_X(x) = \frac{x^{r/2-1}}{B(r/2, s/2)(r/s)^{-r/2}(1 + (r/s) \cdot x)^{\frac{r+s}{2}}} \cdot I(x > 0),$$

where $I_B(x)$ denotes the indicator function of the set B.

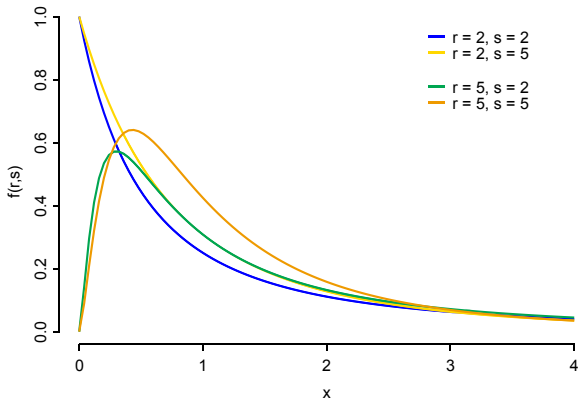


Figure 1.4: Probability density function of the $F_{r,s}$ distribution for various choices of r and s.

Proof Note that for $U_r \sim \chi^2_r$ the probability density function is given by

$$f_{U_r}(x) = \frac{x^{r/2-1}e^{-x/2}}{\Gamma(r/2)2^{r/2}}\,, \quad x > 0\,, \quad r \in \mathbb{N}.$$

Consequently,

$$P\left({}^{U_r}/_r \leq x\right) = P(U_r \leq rx) = F_{U_r}(rx),$$

and therefore

$$\begin{aligned} f_{U_r/r}(x) = (F_{U_r}(rx))' = r \cdot f_{U_r}(rx) &= \frac{r(rx)^{r/2-1}e^{\frac{-rx}{2}}}{\Gamma(r/2)2^{r/2}} \cdot I(x > 0) \\ &= \frac{r^{r/2}x^{r/2-1}e^{-r/2\cdot x}}{\Gamma(r/2)2^{r/2}} \cdot I(x > 0)\,. \end{aligned}$$

By the density transformation theorem for the ratio of two random variables [33, Theorem 3.6.9., 2] it holds that

$$f_{\frac{U_r/r}{U_s/s}}(x) = \int_0^\infty t f_{U_r/r}(xt) \cdot f_{U_s/s}(t)\,dt \cdot I(x > 0)\,.$$

Hence,

$$\begin{aligned}
f_X(x) &= \int_0^\infty t \frac{r^{r/2}(tx)^{r/2-1}e^{-\frac{rtx}{2}}}{\Gamma(r/2)2^{r/2}} \cdot \frac{s^{s/2}t^{s/2-1}e^{-st/2}}{\Gamma(s/2)2^{s/2}}\,dt \\
&= \frac{r^{r/2}s^{s/2}x^{r/2-1}}{\Gamma(r/2)\Gamma(s/2)2^{\frac{r+s}{2}}} \cdot \int_0^\infty t^{r/2+s/2-1} e^{-\overbrace{\frac{rx+s}{2}}^{=y} t}\,dt \\
&= \frac{r^{r/2}s^{s/2}x^{r/2-1}}{\Gamma(r/2)\Gamma(s/2)} \cdot \int_0^\infty \frac{y^{\frac{r+s}{2}-1}}{(rx+s)^{\frac{r+s}{2}}} \cdot e^{-y}\,dy \\
&\underset{t=\frac{y}{\frac{rx+s}{2}}}{=} \frac{r^{r/2}s^{s/2}x^{r/2-1}}{\Gamma(r/2)\Gamma(s/2)s^{\frac{r+s}{2}}(1+\frac{r}{s}\cdot x)^{\frac{r+s}{2}}} \cdot \Gamma\left(\frac{r+s}{2}\right) \\
&= \frac{(r/s)^{r/2}x^{r/2-1}}{B(r/2,s/2)(1+\frac{r}{s}x)^{\frac{r+s}{2}}} \cdot I(x>0)\,.
\end{aligned}$$

□

Remark 1.1.14. Let $X \sim F_{r,s},\ r,s \in \mathbb{N}$, with probability density function f_X.

1. Some graphs of the F distribution are shown in Figure 1.4.
2. Some properties of the F distribution:

   Lemma 1.1.15. Let $X \sim F_{r,s}$, $r,s \in \mathbb{N}$. Then,

   (a)
   $$\mathrm{E}X = \frac{s}{s-2}, \qquad s \geq 3\,.$$

   (b)
   $$\operatorname{Var} X = \frac{2s^2(r+s-2)}{r(s-4)(s-2)^2}\,, \qquad s \geq 5\,.$$

   (c) Denote by $F_{r,s,\alpha}$ the α-quantile of the $F_{r,s}$ distribution. Then,
   $$F_{r,s,\alpha} = \frac{1}{F_{s,r,1-\alpha}}\,, \quad \alpha \in (0,1).$$

   Exercise 1.1.16. Prove Lemma 1.1.15.

3. The following approximation formula holds for quantiles $F_{r,s,\alpha}$ (cf.

Abramowitz, Stegun (1972)): $F_{r,s,\alpha} \approx e^{\omega}$, where

$$\omega = 2\left(\frac{\alpha(h+a)^{1/2}}{h} - \left(\frac{1}{r-1} - \frac{1}{s-1}\right)\cdot\left(a + \frac{5}{6} - \frac{2}{3h}\right)\right),$$
$$h = 2\left(\frac{1}{r-1} + \frac{1}{s-1}\right)^{-1},$$
$$a = \frac{z_\alpha^2 - 3}{6}$$

and z_α is the α quantile of the $N(0,1)$ distribution.

1.2 Methods for obtaining point estimators

The following introductory examples were given in the lecture "Elementary Probability Theory and Statistics".

Definition 1.2.1.

1. The function $\hat{F}_n(x) = \#\{x_i : x_i \le x\,,\, i = 1,\dots,n\}/n$ for all $x \in \mathbb{R}$ is called *empirical distribution function of a realized sample* $(x_1,\dots,x_n)$. Here $\hat{F}_n : \mathbb{R}^{n+1} \to [0,1]$ holds, since $\hat{F}_n(x) = \varphi(x_1,\dots,x_n,x)$.

2. The random variable $\hat{F}_n : \Omega \times \mathbb{R} \to [0,1]$ which is indexed by $x \in \mathbb{R}$ is called *empirical distribution function of the random sample* given by $(X_1,\dots,X_n)$, if

$$\hat{F}_n(x,\omega) = \hat{F}_n(x) = \frac{1}{n}\#\{X_i,\, i = 1,\dots,n : X_i(\omega) \le x\}\,, \quad x \in \mathbb{R}\,.$$

Equivalently to Definition 1.2.1 it can be shown that

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \le x)\,, \quad x \in \mathbb{R},$$

where

$$I(x \in A) = \begin{cases} 1, & x \in A \\ 0, & \text{otherwise.} \end{cases}$$

It holds that

$$\hat{F}_n(x) = \begin{cases} 1, & x \ge x_{(n)}\,, \\ \frac{i}{n}, & x_{(i)} \le x < x_{(i+1)}\,, \quad i = 1,\dots,n-1\,, \\ 0, & x < x_{(1)}\,. \end{cases}$$

for $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$.
The height of the jump at $x_{(i)}$ is equal to the relative frequency f_i of $x_{(i)}$. If $x_{(i)} = x_{(i+1)}$ for a $i \in \{1,\dots,n\}$, the value i/n does not occur (cf. [33, Section 6.3.2]).

1.2.1 Plug-In estimator

Based on the empirical distribution function $\hat{F}_n$, the *Plug-in method* yields the class of *Plug-in estimators*. Let $M := \{F : F \text{ is a distribution function}\}$.

Definition 1.2.2. Let the parameter θ of the distribution F be given as a functional $T : M \to \mathbb{R}$ of F, i.e $\theta = T(F)$. Then, $\hat{\theta} = T(\hat{F}_n)$ is called the *Plug-in estimator* for θ.

Definition 1.2.3. Let F be an arbitrary distribution function. The functional $T : M \to \mathbb{R}$ is called *linear*, if

$$T(aF_1+bF_2) = aT(F_1)+bT(F_2) \qquad \text{for all } a, b \in \mathbb{R}_+ , a+b = 1 , \quad F_1, F_2 \in M .$$

Consider a special class of linear functionals given by

$$T(F) = \int_{\mathbb{R}} r(x)\, dF(x) ,$$

where $r(x)$ is an arbitrary continuous function with $\mathbb{E}\,(r(X)) < \infty$. An example for such T is given by

$$\mathrm{E}X^k = \int_{\mathbb{R}} x^k\, dF(x) , \qquad k \in \mathbb{N} .$$

Lemma 1.2.4. The Plug-in estimator for $\theta = \int_{\mathbb{R}} r(x)\, dF(x)$ is given by

$$\hat{\theta} = \int_{\mathbb{R}} r(x)\, d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} r(x_i).$$

Exercise 1.2.5. Prove Lemma 1.2.4!

Example 1.2.6 (Plug-in estimator)**.**

1. $\bar{X}_n$ is a Plug-in estimator for the expected value μ.
2. *Plug-in estimator* for $\sigma^2 = \operatorname{Var} X$: It holds that $\operatorname{Var} X = \mathrm{E}X^2 - (\mathrm{E}X)^2$ and therefore
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{n-1}{n} S_n^2 .$$
3. *Estimator for skewness and kurtosis* $\hat{\gamma}_1$ and $\hat{\gamma}_2$ (cf. [33, Section 6.4.4]) are Plug-in estimators, since the coefficient of skewness is defined as
$$\gamma_1 = \mathrm{E}\left(\frac{X-\mu}{\sigma}\right)^3$$
where $\mu = \mathrm{E}X$, $\sigma^2 = \operatorname{Var} X$, implies
$$\hat{\gamma}_1 \overset{\substack{\mu \mapsto \bar{X}_n \\ \sigma^2 \mapsto \hat{\sigma}^2}}{=} \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^3}{(\hat{\sigma}_n^2)^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right)^{3/2}} .$$
The construction of $\hat{\gamma}_2$ can be done in the same spirit.

4. The *empirical coefficient of correlation* ϱ_{XY} is a Plug-in estimator, since

$$\hat{\varrho}_{XY} = \frac{S^2_{XY}}{\sqrt{S^2_{XX}}\sqrt{S^2_{YY}}} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2 \sum_{i=1}^n (Y_i - \bar{Y}_n)^2}}\,.$$

Indeed

$$\varrho_{XY} = \frac{\mathrm{E}(X - \mathrm{E}X)(Y - \mathrm{E}Y)}{\sqrt{\mathrm{Var}\, X \cdot \mathrm{Var}\, Y}} = \frac{\mathrm{E}(XY) - \mathrm{E}X \cdot \mathrm{E}Y}{\sqrt{(\mathrm{E}X^2 - (\mathrm{E}X)^2)(\mathrm{E}Y^2 - (\mathrm{E}Y)^2)}}$$

and therefore, considering the linear functionals

$$T_1(F) = \int x\, dF(x),\ T_2(F) = \int x^2\, dF(x),\ T_{12}(G) = \int xy\, dG(x,y)$$

$$\varrho_{XY} = \frac{T_{12}(F_{XY}) - T_1(F_X) \cdot T_1(F_Y)}{\sqrt{(T_2(F_X) - (T_1(F_X))^2)\,(T_2(F_Y) - (T_1(F_Y))^2)}}\,.$$

$\hat{\varrho}_{XY}$ is obtained by replacing F_X, F_Y and F_{XY} in T_1, T_2 and T_{12} with $\hat{F}_{n,X}$, $\hat{F}_{n,Y}$ and $\hat{F}_{n,XY}$:

$$\hat{\varrho}_{XY} = \frac{T_{12}(\hat{F}_{n,XY}) - T_1(\hat{F}_{n,X}) \cdot T_1(\hat{F}_{n,Y})}{\sqrt{\left(T_2(\hat{F}_{n,X}) - \left(T_1(\hat{F}_{n,X})\right)^2\right)\left(T_2(\hat{F}_{n,Y}) - \left(T_1(\hat{F}_{n,Y})\right)^2\right)}}\,.$$

1.2.2 Method of moments estimator

In the following let $(X_1, \ldots, X_n)$ be a sample of i.i.d. random variables X_i with distribution function $F \in \{F_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^m$ (parametric model). Assume that the parametrisation $\theta \mapsto F_\theta$ is distinguishable, i.e. $F_\theta \neq F_{\theta'} \iff \theta \neq \theta'$.

Goal: Construction of an estimator $\hat{\theta}(X_1, \ldots, X_n)$ for $\theta = (\theta_1, \ldots, \theta_m)$. [33, Theorem 4.5.6[2]] implies that under certain conditions on F (e.g. uniform distribution on a compact interval) the underlying distribution can be determined, if the moments $\mathrm{E}X^k\,, \quad k \in \mathbb{N}$ are known. The *method of moments estimation* is based on the idea of estimating F by using the moments and was introduced by Karl Pearson in the end of the 19th century.

Assumptions: There exists $r \geq m$ such that $\mathrm{E}_\theta |X_i|^r < \infty$. Assume that the moments $\mathrm{E}_\theta X_i^k = g_k(\theta)$, $k = 1, \ldots, r$ are given as functions of the parameter vector $\theta = (\theta_1, \ldots, \theta_m) \in \Theta$.

[2] **Theorem 4.5.6.** Let X be a random variable with values in $C \subset \mathbb{R}$, i.e.

$$P(X \in C) = 1.$$

If $C \subset [a, b]$, $a < b$, then $\{\mu_k\}_{k \in \mathbb{N}}$ defines P_X uniquely.

Moment equation system: $\hat{\mu}_k = g_k(\theta)$, $k = 1, \dots, r$, where $\hat{\mu}_k$ are the k-th empirical moments defined by $\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$.

Definition 1.2.7. If the system above is uniquely solvable for θ, then its solution $\hat{\theta}(X_1, \dots, X_n)$ is called *moment estimator (M-estimator)* of θ.

Lemma 1.2.8. Let $g = (g_1, \dots, g_r) : \Theta \to C \subset \mathbb{R}^r$ be a bijective function, and let its inverse function $g^{-1} : C \to \Theta$ be continuous. Then the moment estimator $\hat{\theta}(X_1, \dots, X_n)$ of θ is strongly consistent.

Proof It holds that $\hat{\theta}(X_1, \dots, X_n) = g^{-1}(\hat{\mu}_1, \dots, \hat{\mu}_r) \xrightarrow[n\to\infty]{\text{a.s.}} \theta$, since $\hat{\mu}_k \xrightarrow[n\to\infty]{\text{a.s.}} g_k(\theta)$, $k = 1, \dots, r$ (strong consistency of the empirical moments) and g^{-1} is continuous. □

Remark 1.2.9.

1. Under certain conditions with respect to the regularity of F_θ the moment estimator $\hat{\theta}(X_1, \dots, X_n)$ for θ is asymptotically normally distributed:
$$\sqrt{n}\left(\hat{\theta}(X_1, \dots, X_n) - \theta\right) \xrightarrow[n\to\infty]{d} N(0, \Sigma)\,,$$
where $N(0, \Sigma)$ is the multivariate normal distribution with covariance matrix
$$\Sigma = G^T \mathrm{E}(YY^T)G$$
with
$$Y = (X, X^2, \dots, X^r)^T\,, \quad X \overset{d}{=} X_i\,,$$
and
$$G = \left(\frac{\partial g_i^{-1}}{\partial \theta_j}\right)_{\substack{i=1\dots r\,,\\ j=1\dots m}}\,.$$

2. Other properties for the moment estimator do not hold in general (e.g. not all moment estimators are unbiased (cf. Example 1.2.10, 1)).

3. Sometimes $r > m$ equations in the moment equation system are necessary in order to obtain a moment estimator. It can occur for example, if some $g_i = const$, i.e they do not provide additional information about θ (cf. Example 1.2.10, 2)).

Example 1.2.10.

1. *Normal distribution:* $X_i \overset{d}{=} X\,, \quad i = 1, \dots, n\,, \quad X \sim N(\mu, \sigma^2)$; The goal is to obtain a moment estimator for μ and σ^2, so $\theta = (\mu, \sigma^2)$. It holds that
$$\begin{aligned} g_1(\mu, \sigma^2) &= \mathrm{E}_\theta X = \mu\,, \\ g_2(\mu, \sigma^2) &= \mathrm{E}_\theta X^2 = \mathrm{Var}_\theta\, X + (\mathrm{E}_\theta X)^2 = \sigma^2 + \mu^2\,. \end{aligned}$$

Consider the system of equations

$$\begin{cases} \frac{1}{n}\sum_{i=1}^n X_i = \mu\,, \\ \frac{1}{n}\sum_{i=1}^n X_i^2 = \mu^2 + \sigma^2\,. \end{cases}$$

It follows that

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n\,, \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^n X_i^2 - \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i^2 - \bar{X}_n^2\right) \\ &= \frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2 = \frac{n-1}{n} S_n^2\,, \end{aligned}$$

hence, the moment estimators are given by $\hat{\mu} = \bar{X}_n\,, \quad \hat{\sigma}^2 = \frac{n-1}{n} S_n^2$. Note that $\hat{\sigma}^2$ is not unbiased, since

$$\mathrm{E}_\theta \hat{\sigma}^2 = \frac{n-1}{n} \cdot \mathrm{E}_\theta S_n^2 = \frac{n-1}{n}\sigma^2\,.$$

2. *Uniform distribution:* $X_i \overset{d}{=} X$, $i = 1, \ldots, n$, $X \sim U[-\theta, \theta]$, $\theta > 0$. The goal is to obtain a moment estimator for θ. It holds that

$$\begin{aligned} g_1(\theta) &= \mathrm{E}_\theta X = 0\,, \\ g_2(\theta) &= \mathrm{E}_\theta X^2 = \mathrm{Var}_\theta\, X = \frac{(\theta - (-\theta))^2}{12} = \frac{(2\theta)^2}{12} = \frac{\theta^2}{3}\,. \end{aligned}$$

Thus, the following system of equations can be set up:

$$\begin{cases} \frac{1}{n}\sum_{i=1}^n X_i = 0 \qquad \text{(useless)}\,, \\ \frac{1}{n}\sum_{i=1}^n X_i^2 = \frac{\theta^2}{3}\,. \end{cases}$$

Solving the above system of equations for θ yields the moment estimator $\hat{\theta} = \sqrt{\frac{3}{n}\sum_{i=1}^n X_i^2}$. Here, two equations for the estimation of one parameter θ were necessary, i.e. $r = 2 > m = 1$.

1.2.3 Maximum-likelihood estimator

Maximum-likelihood estimators were discovered by Carl Friedrich Gauss (beginnig of the 19th century) and Sir Ronald Fisher (1922). Assume that all distributions in the parametric family $\{F_\theta : \theta \in \Theta\}$ are either discrete or continuous.

Definition 1.2.11. Consider the random sample $X = (X_1, \ldots, X_n)$.

1. Let X_i, $i = 1, \ldots, n$, be absolutely continuous random variables with probability density function $f_\theta(x)$. Then,

$$L(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} f_\theta(x_i)\,, \qquad (x_1, \ldots, x_n) \in \mathbb{R}^n\,, \quad \theta \in \Theta$$

   is called *likelihood funktion* of the sample $(x_1, \ldots, x_n)$.

2. Let X_i, $i = 1, \ldots, n$, be discrete random variables with probability mass function $p_\theta(x) = P_\theta(X_i = x)\,,\ x \in C$, where C is the range of X. Then,

$$L(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} p_\theta(x_i)\,, \qquad (x_1, \ldots, x_n) \in C^n\,, \quad \theta \in \Theta$$

   is called *Likelihood function* of the sample $(x_1, \ldots, x_n)$.

By this definition

- *the discrete case yields* $L(x_1, \ldots, x_n, \theta) = P_\theta(X_1 = x_1, \ldots, X_n = x_n)$
- *the continuous case yields*

$$\begin{aligned} L(x_1, \ldots, x_n, \theta) \prod_{i=1}^{n} \Delta x_i & \\ &= f_{(X_1,\ldots,X_n),\theta}(x_1, \ldots, x_n)\Delta x_1 \cdot \ldots \cdot \Delta x_n \\ &\approx P_\theta(X_1 \in [x_1\,,\, x_1 + \Delta x_1], \ldots, X_n \in [x_n\,,\, x_n + \Delta x_n]) \end{aligned}$$

   for $\Delta x_i \to 0$, $i = 1, \ldots, n$.

The goal is to construct an estimator θ such that the probability

$$P_\theta(X_1 = x_1, \ldots, X_n = x_n) \ \text{ resp. } \ P_\theta(X_i \in [x_i, x_i + \Delta x_i],\ i = 1, \ldots, n)$$

is maximized. This procedure is called *Maximum-likelihood method*.

Definition 1.2.12. Assume that the maximization problem given by $L(x_1, \ldots, x_n, \theta) \mapsto \max_{\theta \in \Theta}$ is uniquely solvable. Then,

$$\hat{\theta}(x_1, \ldots, x_n) = \operatorname*{argmax}_{\theta \in \Theta} L(x_1, \ldots, x_n, \theta)$$

is called *Maximum-Likelihood estimator* of θ (*ML estimator*).

Remark 1.2.13.

1. There are only very few cases in which the ML estimator $\hat{\theta}$ for θ is explicitly expressible. In most cases, the constant factor of the likelihood function is omitted. By taking the logarithm of the remaining function
$$\log L(x_1, \ldots, x_n, \theta)$$
the so called *log-likelihood function* is obtained.
Consequently
$$\prod_{i=1}^{n} f_\theta(x_i) \quad \text{resp.} \quad \prod_{i=1}^{n} p_\theta(x_i)$$
turn into sums
$$\sum_{i=1}^{n} \log f_\theta(x_i) \quad \text{resp.} \quad \sum_{i=1}^{n} \log p_\theta(x_i)\,,$$
which are easier to differentiate with respect to θ. To compute the maximum of the log-likelihood function, one considers the first order conditions
$$\frac{\partial \log L(x_1, \ldots, x_n, \theta)}{\partial \theta_j} = 0\,, \qquad j = 1, \ldots, m\,,$$
which are a necessary condition for an extremum of $\log L$ (and thus of L, since the logarithm is monotonically increasing). If this system is uniquely solvable and the obtained solution $\hat{\theta}(X_1, \ldots, X_n)$ is a maximum, it is also the ML estimator.

2. In most applied cases, the ML estimators need to be calculated by numerical methods.

Example 1.2.14.

1. *Bernoulli distribution:* Let $X_i \sim Bernoulli(p)$ i.i.d., $i = 1, \ldots, n$, with $p \in [0, 1]$. Since
$$X_i = \begin{cases} 1\,, & \text{with probability } p, \\ 0\,, & \text{else,} \end{cases}$$
where the respective probability mass function is given by
$$p_\theta(x) = p^x (1-p)^{1-x}\,, \qquad x \in \{0, 1\}\,.$$
The *likelihood function of the random sample* $(X_1, \ldots, X_n)$ is given by
$$\begin{aligned} L(x_1, \ldots, x_n, \theta) &= \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} \\ &= p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i} \overset{\text{def.}}{=} h(p). \end{aligned}$$

(a) If $\sum_{i=1}^n x_i = 0$ ($\iff x_1 = x_2 = \ldots = x_n = 0$), then $h(p) = (1-p)^n$ is maximized at $p = 0$. The ML estimator is then given by $\hat{p}(0, \ldots, 0) = 0$.

(b) If $\sum_{i=1}^n x_i = n$ ($\iff x_1 = x_2 = \ldots = x_n = 1$), then $h(p) = p^n$ is maximized at $p = 1$. The ML estimator is then given by $\hat{p}(1, 1, \ldots, 1) = 1$.

(c) If $0 < \sum_{i=1}^n x_i < n$, then

$$\log L(x_1, \ldots, x_n, p) = n\bar{x}_n \log p + n(1 - \bar{x}_n) \log(1 - p) = n \cdot g(p)\,.$$

Since $g(p) \underset{p \to 0,1}{\longrightarrow} -\infty$ and

$$\frac{\partial g(p)}{\partial p} = \frac{\bar{x}_n}{p} + \frac{1 - \bar{x}_n}{1 - p} \cdot (-1) = \frac{\bar{x}_n}{p} + \frac{\bar{x}_n - 1}{1 - p} = 0$$

$\iff (1-p)\bar{x}_n + (\bar{x}_n - 1)p = 0 \iff p = \bar{x}_n$, the continuity of g implies that g attains exactly one $\operatorname{argmax}_p g(p) = \bar{x}_n$.

Thus, the ML estimator is given by $\hat{p}(X_1, \ldots, X_n) = \bar{X}_n$.

2. *Uniform distribution:* Let $X_i \sim U[0, \theta]$, $i = 1, \ldots, n$, i.i.d. with $\theta > 0$. The goal is to obtain a ML estimator for θ. It holds that

$$f_{X_i}(x) = {}^1\!/\!_\theta \cdot I(x \in [0, \theta])\,, \qquad i = 1, \ldots, n\,.$$

Thus, the likelihood function is given by

$$\begin{aligned} L(x_1, \ldots, x_n, \theta) &= \begin{cases} ({}^1\!/\!_\theta)^n\,, & 0 \le x_1, \ldots, x_n \le \theta \\ 0\,, & \text{else} \end{cases} \\ &= \begin{cases} ({}^1\!/\!_\theta)^n\,, & \text{if } \min\{x_1, \ldots, x_n\} \ge 0 \\ & \text{and } \max\{x_1, \ldots, x_n\} \le \theta \\ 0\,, & \text{else} \end{cases} \\ &= g(\theta)\,, \qquad \theta > 0. \end{aligned}$$

Therefore, $\hat{\theta} = \operatorname{argmax}_{\theta > 0} g(\theta) = \max\{x_1, \ldots, x_n\} = x_{(n)}$. So the ML estimator is given by $\hat{\theta}(X_1, \ldots, X_n) = X_{(n)}$.

It can be shown that under certain conditions, the ML estimator is weakly consistent and asymptotically normal distributed.

Definition 1.2.15. Let

$$L(x, \theta) = \begin{cases} f_\theta(x)\,, & \text{if continuous,} \\ p_\theta(x)\,, & \text{if discrete} \end{cases}$$

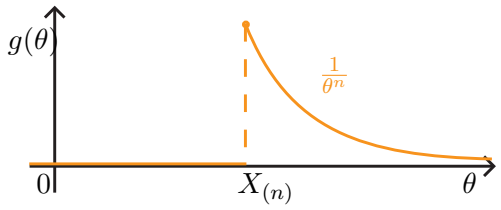


Figure 1.5: Illustration of the function g.

be the likelihood function of X. For $\theta, \theta' \in \Theta$ and $X \stackrel{d}{=} X_i$, $P_\theta(L(X, \theta') = 0) = 0$ define the *Kullback-Leibler information* (distance) $H(P_\theta, P_{\theta'})$ in the continuous case as

$$H(P_\theta, P_{\theta'}) = \mathrm{E}_\theta \log L(X, \theta) - \mathrm{E}_\theta \log L(X, \theta') = \int_{\mathbb{R}} \log \frac{L(x, \theta)}{L(x, \theta')} \cdot L(x, \theta)\, dx.$$

If $P_\theta(L(X, \theta') = 0) > 0$, then define $H(P_\theta, P_{\theta'}) = \infty$. In the discrete case take the sum over all non-trivial $p_\theta(x)$ instead of the integral.

The following lemma will show that $H(\cdot\,, \cdot)$ has the properties $H(P_\theta, P_{\theta'}) = 0 \iff \theta = \theta'$ and $H(P_\theta, P_{\theta'}) \geq 0 \quad \forall\, \theta, \theta' \in \Theta$. It is, on the other hand, easy to prove that $H(P_\theta, P_{\theta'})$ is not symmetric with respect to θ and θ'. Thus, $H(\cdot\,, \cdot)$ is not a metric.

Lemma 1.2.16. It holds that

1. $H(P_\theta, P_{\theta'})$ is well-defined and ≥ 0.
2. If $H(P_\theta, P_{\theta'}) = 0$, then $\theta = \theta'$.

Proof Consider the continuous case $P_\theta\,, \quad \theta \in \Theta$ (the discrete case can be shown in the same spirit).

1. Define
$$f(x) = \begin{cases} \frac{L(x,\theta)}{L(x,\theta')}\,, & \text{if } L(x, \theta') > 0\,, \\ 1\,, & \text{else.} \end{cases}$$
If $P_\theta(L(X, \theta') = 0) = 0$, then $P_\theta(L(X, \theta') > 0) = 1$. On the other hand, if $H(P_\theta, P_{\theta'}) = \infty > 0$, then H is well-defined and positive. With probability 1 it holds that $L(x, \theta) = f(x) \cdot L(x, \theta')$.
Let $g(x) = 1 - x + x \log x\,, \quad x > 0$. It can be shown that g is convex with $g(x) \geq 0$. Indeed, it holds that
$$g'(x) = -1 + \log x + 1 = \log x\,, g''(x) = 1/x > 0\,.$$

Thus, g admits exactly one zero at $x = 1$, which is a minimum. Consider $g(f(X))\,, \quad X \sim L(x, \theta')$. Then,

$$\begin{aligned} 0 \leq \mathrm{E}_{\theta'} g(f(X)) &= 1 - \mathrm{E}_{\theta'} f(X) + \mathrm{E}_{\theta'} \left(f(X) \log f(X)\right) \\ &= 1 - \int \frac{L(x,\theta)}{L(x,\theta')} \cdot L(x,\theta')\, dx + \int \frac{L(x,\theta)}{L(x,\theta')} \cdot \log \frac{L(x,\theta)}{L(x,\theta')} \cdot L(x,\theta') dx \\ &= H(P_\theta, P_{\theta'})\,. \end{aligned}$$

Therefore, $H(P_\theta, P_{\theta'}) \geq 0$, which was to be shown.

2. If $H(P_\theta, P_{\theta'}) = 0 \quad \Longrightarrow \quad \mathrm{E}_{\theta'} g(f(X)) = 0\,, \quad g(f(X)) \geq 0$. Thus, $L(x,\theta')$-almost surely $g(f(X)) = 0 \quad \Longrightarrow f(X) \overset{\theta'\text{-a.s.}}{=} 1$, which implies either $L(X,\theta') = 0$ or $L(x,\theta) = L(x,\theta')$ for $L(x,\theta')$-almost all x and therefore $P_\theta = P_{\theta'}$.

□

Example 1.2.17.

1. Let $\Theta = \mathbb{R}_+$ and $\{P_\lambda, \lambda > 0\}$ be the family of exponential distributions with parameter $\lambda > 0$ and probability density functions $L(x,\lambda) = \lambda e^{-\lambda x} I(x \geq 0)$. Computing the Kullback-Leibler information $H(P_\lambda, P'_\lambda)$ for any $\lambda, \lambda' > 0$ yields

$$\begin{aligned} H(P_\lambda, P'_\lambda) &= \int_0^\infty \log\left(\frac{\lambda e^{-\lambda x}}{\lambda' e^{-\lambda' x}}\right) \lambda e^{-\lambda x} dx \\ &= \log\left(\frac{\lambda}{\lambda'}\right) \cdot \underbrace{\int_0^\infty \lambda e^{-\lambda x} dx}_{=1} - (\lambda - \lambda') \underbrace{\int_0^\infty x \lambda e^{-\lambda x} dx}_{=\frac{1}{\lambda}} \\ &= \log\left(\frac{\lambda}{\lambda'}\right) - \frac{\lambda - \lambda'}{\lambda} \\ &= \frac{\lambda'}{\lambda} - 1 - \log\left(\frac{\lambda'}{\lambda}\right). \end{aligned}$$

For $\lambda = \lambda'$ we get $H(P_\lambda, P_\lambda) = 1 - 1 - \log(1) = 0$.

2. It may also happen that $H(P_\theta, P_{\theta'}) = +\infty$ for absolutely continuous distributions P_θ. As an example, consider the family $\{U[0,\theta], \theta > 0\}$ of uniform distributions on $[0,\theta]$ with the likelihood $L(x,\theta) = \frac{I(x \in [0,\theta])}{\theta}$.

Then,

$$\begin{aligned} H(P_\theta, P_{\theta'}) &= \frac{1}{\theta}\int_0^\theta \log\left(\frac{\frac{1}{\theta}}{\frac{1}{\theta'}I(x\in[0,\theta'])}\right)dx \\ &= \frac{1}{\theta}\int_0^\theta \log\left(\frac{\theta'}{\theta I(x\in[0,\theta'])}\right)dx \\ &= \begin{cases} \log\left(\frac{\theta'}{\theta}\right), & \text{if } \theta' \geq \theta, \\ +\infty, & \text{if } \theta' < \theta. \end{cases} \end{aligned}$$

Theorem 1.2.18 (Weak consistency of ML estimators)**.** Let $m = 1$ and Θ be an open interval in $\mathbb{R}$. Furthermore, let $L(x_1, \ldots, x_n, \theta)$ be *unimodal*, i.e. for the ML estimator $\hat{\theta}$ of θ it holds that

$$\begin{cases} \forall\, \theta < \hat{\theta}(x_1, \ldots, x_n) & \Longrightarrow L(x_1, \ldots, x_n, \theta) \text{ is increasing} \\ \forall\, \theta > \hat{\theta}(x_1, \ldots, x_n) & \Longrightarrow L(x_1, \ldots, x_n, \theta) \text{ is decreasing} \end{cases}$$

(i.e. $\max_{\theta\in\Theta} L(x_1, \ldots, x_n, \theta)$ exists and is unique). Then,

$$\hat{\theta}(X_1, \ldots, X_n) \xrightarrow[n\to\infty]{P} \theta.$$

Proof For the weak consistency (the convergence in probability) of $\hat{\theta}$ to hold, the following needs to be shown:

$$P_\theta\left(\left|\hat{\theta}(X_1, \ldots, X_n) - \theta\right| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0\,, \qquad \varepsilon > 0\,. \tag{1.3}$$

Let $\varepsilon > 0$: $\theta \pm \varepsilon \in \Theta$ be arbitrary. Then, the Kullback-Leibler information satisfies $H(P_\theta, P_{\theta\pm\varepsilon}) > \sigma > 0$, because of the distinguishability of the parametrization of P_θ and Lemma 1.2.16. Consider $\{|\hat{\theta} - \theta| \leq \varepsilon\}$. In order to show (1.3), it is sufficient to find a lower bound for $P_\theta(|\hat{\theta} - \theta| \leq \varepsilon)$, which converges to 1 for $n \to \infty$. By unimodality it holds that

$$\begin{aligned} &\left\{|\hat{\theta} - \theta| < \varepsilon\right\} \\ &\quad \overset{\text{unimod}}{\supseteq} \left\{L(X_1, \ldots, X_n, \theta) \in \Big(L(X_1, \ldots, X_n, \theta - \varepsilon), L(X_1, \ldots, X_n, \theta + \varepsilon)\Big)^{3}\right\} \\ &\quad = \cup \left\{\frac{L(X_1, \ldots, X_n, \theta)}{L(X_1, \ldots, X_n, \theta \pm \varepsilon)} > 1\right\} \overset{\sigma>0\Rightarrow e^{n\delta}>1}{\supseteq} \cup \left\{\frac{L(X_1, \ldots, X_n, \theta)}{L(X_1, \ldots, X_n, \theta \pm \varepsilon)} > e^{n\delta}\right\} \\ &\quad = \cup \left\{\frac{1}{n}\log\frac{L(X_1, \ldots, X_n, \theta)}{L(X_1, \ldots, X_n, \theta \pm \varepsilon)} > \sigma\right\} = A_+ \cup A_-\,, \end{aligned}$$

where

$$A_\pm = \left\{\frac{1}{n}\log\frac{L(X_1, \ldots, X_n, \theta)}{L(X_1, \ldots, X_n, \theta \pm \varepsilon)} > \sigma\right\}\,.$$

[3] This means an interval with these endpoints, even though we don't immediately know which one is larger.

Hence,

$$P_\theta\left(|\hat{\theta}-\theta|<\varepsilon\right) \geq P_\theta(A_+ \cup A_-)\,.$$

Showing that

$$\lim_{n\to\infty} P_\theta(A_\pm) = 1 \tag{1.4}$$

then implies

$$1 \geq \lim_{n\to\infty} P_\theta(A_+ \cup A_-) \geq \lim_{n\to\infty} P_\theta(A_\pm) = 1,$$

in particular this yields

$$\lim_{n\to\infty} P_\theta(A_+ \cup A_-) = 1$$

and

$$1 \geq \lim_{n\to\infty} P_\theta\left(|\hat{\theta}-\theta|<\varepsilon\right) \geq 1.$$

which implies that

$$\lim_{n\to\infty} P_\theta\left(|\hat{\theta}-\theta|>\varepsilon\right) \leq 1 - \underbrace{\lim_{n\to\infty} P_\theta\left(|\hat{\theta}-\theta|<\varepsilon\right)}_{=1} = 0,$$

i.e. $\hat{\theta} \xrightarrow[n\to\infty]{P} \theta$.

In the following it will now be shown that $P_\theta(A_+) \underset{n\to\infty}{\longrightarrow} 1$ (similar for $P_\theta(A_-) \underset{n\to\infty}{\longrightarrow} 1$):

1. Let $H(P_\theta, P_{\theta+\varepsilon}) < \infty$ and

$$f(x) = \begin{cases} \dfrac{L(x,\theta)}{L(x,\theta+\varepsilon)}\,, & \text{if } L(x,\theta+\varepsilon) > 0\,, \\ 1\,, & \text{else.} \end{cases}$$

By Definition 1.2.15, it holds that $P_\theta(L(X_1,\theta+\varepsilon)>0) = 1$. Furthermore, the strong law of large numbers implies

$$\begin{aligned} &\frac{1}{n}\log\frac{L(X_1,\ldots,X_n,\theta)}{L(X_1,\ldots,X_n,\theta+\varepsilon)} = \frac{1}{n}\sum_{i=1}^{n}\log\frac{L(X_i,\theta)}{L(X_i,\theta+\varepsilon)} = \frac{1}{n}\sum_{i=1}^{n}\log f(X_i) \\ &\xrightarrow[n\to\infty]{\text{f.s.}} \mathrm{E}_\theta \log f(X_1) = \int L(x,\theta)\cdot\log\frac{L(x,\theta)}{L(x,\theta+\varepsilon)}\,dx = H(P_\theta, P_{\theta+\varepsilon}) > \sigma > 0, \end{aligned}$$

since $\log f(X_1) \in L^1(\Omega,\mathcal{F},P)$ and

$$\mathrm{E}_\theta \log f(X_1) = H(P_\theta, P_{\theta+\varepsilon}) < \infty \quad\Longrightarrow\quad P(A_+) \underset{n\to\infty}{\longrightarrow} 1\,.$$

2. Let $H(P_\theta, P_{\theta+\varepsilon}) = \infty$ and $P_\theta(L(X_1, \theta+\varepsilon) = 0) = 0$, then

$$f(x) \overset{\text{a.s.}}{=} \frac{L(x,\theta)}{L(x,\theta+\varepsilon)}$$

with respect to the distribution P_{X_1}. Now, $\log\min\{f(X_1), c\} \in L^1(\Omega, \mathcal{F}, P)$ for all $c > 0$. Thus, similarly to 1 it holds that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \log\min\{f(X_i), c\} &\xrightarrow[n\to\infty]{\text{a.s.}} \mathrm{E}_\theta \log\min\{f(X_1), c\} \in (0,\infty) \\ &\xrightarrow[c\to\infty]{} H(P_\theta, P_{\theta+\varepsilon}) = \infty\end{aligned}$$

and therefore

$$\begin{aligned}A_+ &\supset \left\{\frac{1}{n}\sum_{i=1}^{n} \log\min\{f(X_i), c\} > \sigma\right\} \\ \Longrightarrow P(A_+) &\geq P\left(\frac{1}{n}\sum_{i=1}^{n} \log\min\{f(X_i), c\} > \sigma\right) \xrightarrow[n\to\infty]{} 1\,.\end{aligned}$$

3. Let $H(P_\theta, P_{\theta+\varepsilon}) = \infty$ and $P_\theta(L(X_1, \theta+\varepsilon) = 0) = a > 0$. Then,

$$\begin{aligned}P_\theta\left(\frac{1}{n}\log\frac{L(X_1,\ldots,X_n,\theta)}{L(X_1,\ldots,X_n,\theta+\varepsilon)} = \infty\right) &= 1 - P\left(\frac{1}{n}\log\frac{L(X_1,\ldots,X_n,\theta)}{L(X_1,\ldots,X_n,\theta+\varepsilon)} < \infty\right) \\ &= 1 - P\left(\bigcap_{i=1}^{n}\{L(X_i,\theta+\varepsilon) > 0\}\right) \\ &\overset{X_i \text{ i.i.d.}}{=} 1 - (1-a)^n \xrightarrow[n\to\infty]{} 1.\end{aligned}$$

In summary, $P(A_+) \xrightarrow[n\to\infty]{} 1$.

□

Definition 1.2.19. Let $X = (X_1, \ldots, X_n)$ be a random sample of i.i.d. random variables $X_i \sim F_\theta\,, \quad \theta \in \Theta$. Let $L(x,\theta)$ be the likelihood function of X_i. Then,

$$I(\theta) = \mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log L(X_1,\theta)\right)^2, \quad \theta \in \Theta \tag{1.5}$$

is called the *Fisher information* of the sample $(X_1, \ldots, X_n)$.

From now on it will be assumed that $0 < I(\theta) < \infty$. In the following some necessary conditions with respect to the asymptotically normal distribution of the ML estimator will be presented.

1. $\Theta \subset \mathbb{R}$ is an open interval ($m = 1$).

2. It holds that $P_\theta \neq P_{\theta'}$ if and only if $\theta \neq \theta'$.

3. The family $\{P_\theta,\, \theta \in \Theta\}$, $\theta \in \Theta$ consists only of discrete or continuous distributions and no mixtures.

4. $B = \text{supp}\ \ L(x,\theta) = \{x \in \mathbb{R} : L(x,\theta) > 0\}$ does not depend on $\theta \in \Theta$. Here supp f denotes the support of f, which is defined as

$$\text{supp}\ f = \{x \in \mathbb{R} :\ f(x) \neq 0\},$$

and the likelihood function $L(x,\theta)$ is given by

$$L(x,\theta) = \begin{cases} p(x,\theta)\,, & \text{in the discrete case,} \\ f(x,\theta)\,, & \text{in the continuous case,} \end{cases} \tag{1.6}$$

where $p(x,\theta)$ resp. $f(x,\theta)$ denotes the probability mass or density function of P_θ.

5. The mapping $L(x,\theta)$ is three times continuously differentiable and

$$0 = \frac{d^k}{d\theta^k} \int_B L(x,\theta)\, dx = \int_B \frac{\partial^k}{\partial\theta^k} L(x,\theta)\, dx\,, \quad k = 1, 2,\ \theta \in \Theta\,.$$

Since the integral of $L(x,\theta)$ is equal to 1, the above derivative is equal to 0. In the discrete case, the integral is replaced by a sum over all values $x \in \mathbb{R}$ with positive probability mass $p(x,\theta) > 0$.

6. For all $\theta_0 \in \Theta$ there exists a constant $\sigma_{\theta_0} > 0$ and a measurable function $g_{\theta_0} : B \to [0,\infty)$, such that

$$\left| \frac{\partial^3 \log L(x,\theta)}{\partial\theta^3} \right| \leq g_{\theta_0}(x)\,, \qquad \forall x \in B\,, \quad |\theta - \theta_0| < \sigma_{\theta_0}\,,$$

where $\mathrm{E}_{\theta_0}\, g_{\theta_0}(X_1) < \infty$.

Remark 1.2.20. It holds that

$$n \cdot I(\theta) = \mathrm{Var}_\theta \left(\frac{\partial}{\partial\theta} \log L(X_1, \ldots, X_n, \theta) \right)\,,$$

where

$$L(X_1, \ldots, X_n, \theta) = \prod_{i=1}^{n} L(X_i, \theta) \tag{1.7}$$

is the likelihood function of the sample $(X_1, \ldots, X_n)$ with $L(X_i,\theta)$ given in (1.6).

Proof Note that

$$\begin{aligned}\frac{\partial}{\partial\theta}\log L(X_1,\dots,X_n,\theta) &= \frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log L(X_i,\theta) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log L(X_i,\theta)\\ &= \sum_{i=1}^{n}\frac{L'(X_i,\theta)}{L(X_i,\theta)}\,.\end{aligned}$$

Furthermore

$$\begin{aligned}\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log L(X_1,\dots,X_n,\theta)\right) &= \sum_{i=1}^{n}\mathrm{E}_\theta\frac{L'(X_i,\theta)}{L(X_i,\theta)}\\ &= \sum_{i=1}^{n}\int_B\frac{L'(X,\theta)}{L(X,\theta)}\cdot L(X,\theta)\,dx\\ &\overset{5)}{=} 0\,.\end{aligned}$$

In summary

$$\begin{aligned}\mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log L(X_1,\dots,X_n,\theta)\right) &= \mathrm{Var}_\theta\left(\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log L(X_i,\theta)\right)\\ &\overset{X_i\ \text{i.i.d}}{=} \sum_{i=1}^{n}\mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log L(X_i,\theta)\right)\\ &\overset{X_i\ \text{i.i.d}}{=} n\cdot\mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log L(X_1,\theta)\right)\\ &= n\cdot\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log L(X_1,\theta)\right)^2 = n\cdot I(\theta).\end{aligned}$$

□

Example 1.2.21. Let $X_i\sim N(\mu,\sigma^2),\ i=1,\dots,n$. For $\theta=\mu$ the Fisher information is given by $I(\mu)=\frac{1}{\sigma^2}$ assuming that σ^2 is known.
Indeed

$$\begin{aligned}L(X_1,\mu) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(X_1-\mu)^2}{2\sigma^2}\right\},\\ \log L(X_1,\mu) &= -\log(\sqrt{2\pi}\sigma) - \frac{(X_1-\mu)^2}{(2\sigma^2)},\\ \frac{\partial\log L(X_1,\mu)}{\partial\mu} &= -\frac{2(X_1-\mu)}{2\sigma^2}\cdot(-1) = \frac{X_1-\mu}{\sigma^2},\end{aligned}$$

Hence,

$$I(\mu)=\mathbb{E}_\mu\left(\frac{\partial\log L(X_1,\mu)}{\partial\mu}\right)^2 = \frac{1}{\sigma^4}\mathbb{E}_\mu(X_1-\mu)^2 = \frac{1}{\sigma^4}\cdot\ \sigma^2 = \frac{1}{\sigma^2}.$$

By Remark 1.2.20 and [33, Theorem 7.3.2., 4)] with $\hat{\mu} = \overline{X}_n$, it holds that $\mathrm{Var}_\mu\left(\frac{\partial}{\partial_\mu} \log L(X_1, \ldots, X_n, \mu)\right) = \frac{n}{\sigma^2} = \frac{1}{\mathrm{Var}_\mu(\hat{\mu})}$. This means that little information about μ (small values of $I(\mu)$) leads to an increasing variance when estimating μ and vice versa.

Theorem 1.2.22. Let $(X_1, \ldots, X_n)$ be a random sample of i.i.d. random variables fulfilling conditions 1) to 6) and $0 < I(\theta) < \infty\,, \quad \theta \in \Theta$. Let $\hat{\theta}(X_1, \ldots, X_n)$ be a weakly consistent ML estimator for θ. Then, the ML estimator $\hat{\theta}(X_1, \ldots, X_n)$ is also asymptotically normally distributed, in particular

$$\sqrt{n \cdot I(\theta)}\left(\hat{\theta}(X_1, \ldots, X_n) - \theta\right) \xrightarrow[n\to\infty]{d} Y \sim N(0,1)\,.$$

Proof Denote by $l_n(\theta) = \log L(X_1, \ldots, X_n, \theta)$ the log-likelihood function, $\theta \in \Theta$. Let $l_n^{(k)}$ denote the k-th derivative of l_n with respect to θ, i.e.

$$l_n^{(k)}(\theta) = \frac{d^k}{d\theta^k} l_n(\theta)\,, \qquad k = 1, 2, 3\,.$$

Since $\hat{\theta}$ is a ML estimator $l_n^{(1)}(\hat{\theta}) = 0$ must hold. Considering the Taylor expansion of $l_n^{(1)}(\hat{\theta})$ in a neighborhood of θ yields

$$0 = l_n^{(1)}(\hat{\theta}) = l_n^{(1)}(\theta) + (\hat{\theta} - \theta) \cdot l_n^{(2)}(\theta) + (\hat{\theta} - \theta)^2 \cdot \frac{l_n^{(3)}(\theta^*)}{2}\,,$$

where θ^* is between θ and $\hat{\theta}$. Note that

$$-(\hat{\theta} - \theta)\left(l_n^{(2)}(\theta) + (\hat{\theta} - \theta)\frac{l_n^{(3)}(\theta^*)}{2}\right) = l_n^{(1)}(\theta),$$

and consequently

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{\frac{l_n^{(1)}(\theta)}{\sqrt{n}}}{-\frac{l_n^{(2)}(\theta)}{n} - (\hat{\theta} - \theta)\frac{l_n^{(3)}(\theta^*)}{2n}}.$$

By showing

1. $$\frac{l_n^{(1)}(\theta)}{\sqrt{n}} \xrightarrow[n\to\infty]{d} N(0, I(\theta))\,,$$

2. $$-\frac{l_n^{(2)}(\theta)}{n} \xrightarrow[n\to\infty]{\text{a.s.}} I(\theta)\,,$$

3.
$$(\hat{\theta}-\theta) \xrightarrow[n\to\infty]{P} 0 \quad \text{and} \quad \frac{l_n^{(3)}(\theta^*)}{2n}$$
is bounded, i.e.
$$\text{there exists a } c>0: \qquad \lim_{n\to\infty} P_\theta\left(\left|\frac{l_n^{(3)}(\theta^*)}{2n}\right| < c\right) = 1\,,$$

it can be followed that
$$(\hat{\theta}-\theta)\cdot\frac{l_n^{(3)}(\theta^*)}{2n} \xrightarrow[n\to\infty]{P} 0\,, \text{ since } \left|\frac{l_n^{(3)}(\theta^*)}{n}\right| \le c \text{ with high probability}$$

and hence
$$\sqrt{n}(\hat{\theta}-\theta) = \frac{\frac{l_n^{(1)}(\theta)}{\sqrt{n}}}{-\frac{l_n^{(2)}(\theta)}{n} - (\hat{\theta}-\theta)\frac{l_n^{(3)}(\theta^*)}{2n}} \xrightarrow[n\to\infty]{d} Z_1 \sim N\left(0, \frac{1}{I(\theta)}\right)$$

by Slutskys Theorem. Ultimately this yields
$\sqrt{n}\sqrt{I(\theta)}(\hat{\theta}-\theta) \xrightarrow[n\to\infty]{d} Y \sim N(0,1)$.

1. The central limit Theorem implies
$$\frac{l_n^{(1)}(\theta)}{\sqrt{n}} = \frac{\sum_{i=1}^n \frac{\partial}{\partial\theta}\log L(X_i,\theta)}{\sqrt{n}} \xrightarrow[n\to\infty]{d} Y_1 \sim N\Bigg(0, \underbrace{\mathrm{Var}_\theta\Big(\frac{\partial}{\partial\theta}L(X_i,\theta)\Big)}_{=I(\theta)}\Bigg)$$
since $\frac{\partial}{\partial\theta}\log L(X_i,\theta)$ are i.i.d. random variables with expectation 0 (cf. Remark 1.2.20).

2.
$$\begin{aligned}
-\frac{1}{n}l_n^{(2)}(\theta) &= -\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log L(X_i,\theta)\\
&= \frac{1}{n}\sum_{i=1}^n \frac{\left(L^{(1)}(X_i,\theta)\right)^2 - L(X_i,\theta)\cdot L^{(2)}(X_i,\theta)}{\left(L(X_i,\theta)\right)^2}\\
&= \frac{1}{n}\sum_{i=1}^n \left(\frac{L^{(1)}(X_i,\theta)}{L(X_i,\theta)}\right)^2 - \frac{1}{n}\sum_{i=1}^n \frac{L^{(2)}(X_i,\theta)}{L(X_i,\theta)}\\
&\xrightarrow[n\to\infty]{\text{a.s.}} \mathrm{E}_\theta\left(\frac{L^{(1)}(X_1,\theta)}{L(X_1,\theta)}\right)^2 - \mathrm{E}_\theta\left(\frac{L^{(2)}(X_1,\theta)}{L(X_1,\theta)}\right) = I(\theta)
\end{aligned}$$
by the law of large numbers, where
$$L^{(k)}(X_i,\theta) = \frac{\partial^k}{\partial\theta^k}L(X_i,\theta)$$

is the k-th derivative of the likelihood function with respect to θ, and

$$\mathrm{E}_\theta\left(\frac{L^{(2)}(X_1,\theta)}{L(X_1,\theta)}\right) = \int_B \frac{\partial^2}{\partial\theta^2} L(x,\theta)\,dx \overset{5)}{=} \frac{d^2}{d\theta^2}\int_B L(x,\theta)\,dx = 0\,.$$

3. By the weak consistency of $\hat{\theta}$ we have $\hat{\theta} \xrightarrow[n\to\infty]{P} \theta$. Following this it can be shown that

$$\frac{l_n^{(3)}(\theta^*)}{n}(\hat{\theta}-\theta) \xrightarrow[n\to\infty]{P} 0\,.$$

Note that $\hat{\theta} \xrightarrow[n\to\infty]{P} \theta$ implies that for all $\varepsilon > 0$

$$P\left(|\hat{\theta}-\theta| \le \varepsilon\right) \xrightarrow[n\to\infty]{} 1\,,$$

which means that $|\hat{\theta}-\theta| \le \sigma_\theta$, with high probability $\sigma_\theta > 0$, as required in Condition 6. Thus, for all θ with $|\hat{\theta}-\theta| < \sigma_\theta$

$$\left|\frac{l_n^{(3)}(\theta)}{n}\right| \le \frac{1}{n}\sum_{i=1}^{n} \underbrace{\left|\frac{\partial^3}{\partial\theta^3}\log L(X_i,\theta)\right|}_{\le g_\theta(X_i)} \le \frac{1}{n}\sum_{i=1}^{n} g_\theta(X_i) \xrightarrow[n\to\infty]{\text{a.s.}} \mathrm{E}_\theta\, g_\theta(X_1) < \infty.$$

Consequently, there exists a constant $c > 0$ such that

$$P_\theta\left(\left|\frac{l_n^{(3)}(\theta^*)}{n}\right| < c\right) \xrightarrow[n\to\infty]{} 1,$$

and hence

$$\frac{l_n^{(3)}(\theta^*)}{n}(\hat{\theta}-\theta) \xrightarrow[n\to\infty]{P} 0\,.$$

□

1.2.4 Bayesian estimation

Let $(X_1,\ldots,X_n)$ be a random sample, where X_i are i.i.d. random variables with distribution function F_θ, $\theta \in \Theta$. The distribution F_θ can be either discrete or continuous. Additionally, let θ be a realization of a random variable $\tilde{\theta}$ with distribution $Q(\cdot)$ on the measurable space $(\Theta, \mathcal{B}_\Theta)$, which is either discrete with probability mass function $q(\cdot)$ on absolutely continuous with probability density function $q(\cdot)$. As usual, both cases will be handled simultaneously with integration being replaced by summation in the discrete case.

Definition 1.2.23. The distribution $Q(\cdot)$ is called *prior or apriori distribution* of the parameter θ (of $\tilde{\theta}$) (prior means "prior to the experiment $(X_1, \dots, X_n)$").

Definition 1.2.24. The *posterior distribution* of the parameter θ (of $\tilde{\theta}$) is given by the probability mass/density function $q_{X_1,\dots,X_n}(\theta, X_1, \dots, X_n)$, which is defined by

$$\begin{cases} P(\tilde{\theta} = \theta \,|\, X_1 = x_1, \dots, X_n = x_n)\,, & \text{if } Q \text{ is discrete,} \\ f_{\tilde{\theta}|X_1,\dots,X_n}(\theta, x_1, \dots, x_n)\,, & \text{if } Q \text{ is continuous.} \end{cases}$$

Here,

$$\begin{aligned} P(\tilde{\theta} = \theta \,|\, X_1 = x_1, \dots, X_n = x_n) &= \frac{P(\tilde{\theta} = \theta, X_1 = x_1, \dots, X_n = x_n)}{P(X_1 = x_1, \dots, X_n = x_n)} \\ &= \frac{P_\theta(X_i = x_i\,,\ i = 1, \dots, n) \cdot q(\theta)}{\sum_{\theta_1 \in \Theta} P_{\theta_1}(X_i = x_i,\ i = 1, \dots, n) \cdot q(\theta_1)} \end{aligned}$$

by the *Bayes formula*, resp.

$$\begin{aligned} f_{\tilde{\theta}|X_1,\dots,X_n}(\theta, x_1, \dots, x_n) &= \frac{f_{(\tilde{\theta},X_1,\dots,X_n)}(\theta, x_1, \dots, x_n)}{f_{X_1,\dots,X_n}(x_1, \dots, x_n)} \\ &= \frac{L(x_1, \dots, x_n, \theta) \cdot q(\theta)}{\int_\Theta L(x_1, \dots, x_n, \theta_1) \cdot q(\theta_1)\, d\theta_1}, \end{aligned}$$

where $L(x_1, \dots, x_n, \theta)$ is the likelihood funtion defined in (1.7).

Definition 1.2.25. A *loss function* $V : \Theta^2 \to \mathbb{R}_+$ is a Θ^2 measurable function.

Loss functions are used as follows: Denote by $\mathrm{E}_* V(\tilde{\theta}, a)$ the *expected loss* (mean risk), which occurs from estimating θ with a, where E_* is the expectation with respect to the *posterior distribution* of $\tilde{\theta}$. Note that $\mathrm{E}_* V(\tilde{\theta}, a)$ is a function of a and $x_1, \dots, x_n$, since the sample $(x_1, \dots, x_n)$ is an explicit part of the posteriori distribution. In particular, it holds that

$$\mathrm{E}_* V(\tilde{\theta}, a) = \varphi(x_1, \dots, x_n, a)\,.$$

Definition 1.2.26. An estimator $\hat{\theta}$ is called a *Bayes estimator* of θ, if

$$\hat{\theta}(x_1, \dots, x_n) = \operatorname*{argmin}_a \mathrm{E}_* V(\tilde{\theta}, a) \tag{1.8}$$

exists and is unique.

Remark 1.2.27.

1. Sometimes $\hat{\theta} \notin \Theta$, which is attributable to $\varphi(x_1, \dots, x_n, a)$ attaining its minimum outside of Θ.

2. The name "Bayes approach" honors the English mathematician Thomas Bayes (1702–1761), who only introduced the idea behind the Bayes formula given by

$$P(B_i|A) = \frac{P(A|B_i) \cdot P(B_i)}{\sum_j P(A|B_j) \cdot P(B_j)}. \tag{1.9}$$

The actual discovery of (1.9) was by Pierre-Simon Laplace (1749–1827) (end of the 18th century). The formula was explicitly used in the derivation of the *posterior distribution* of $\tilde{\theta}$.

3. The approach in Definition 1.2.26 is usually only realizable by numeric minimization. There are only very few cases where an analytic solution of the minimization problem stated in (1.8) can be computed.

Example 1.2.28 (Quadratic loss function)**.** If $V(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2$, then

$$\begin{aligned}\operatorname*{argmin}_a \left(\varphi(x_1, \ldots, x_n, a)\right) &= \operatorname*{argmin}_a \left(\mathrm{E}_*(\tilde{\theta} - a)^2\right) \\ &= \operatorname*{argmin}_a \left(\mathrm{E}_*\tilde{\theta}^2 - 2a\mathrm{E}_*\tilde{\theta} + a^2\right) \\ &= \mathrm{E}_*\tilde{\theta}.\end{aligned}$$

The *Bayes estimator* of $\hat{\theta}(x_1, \ldots, x_n)$ for θ is thus given by $\mathrm{E}_*\tilde{\theta}$.

Example 1.2.29 (Bernoulli distribution)**.** Let $(X_1, \ldots, X_n)$ be an i.i.d. random sample of random variables $X_i \sim Bernoulli(p)$, $p \in (0, 1)$. Furthermore let $\tilde{p} \sim Beta(\alpha, \beta)$, $\alpha, \beta > 0$ be the prior distribution, with probability mass function

$$q(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)} \cdot I_{[0,1]}(p)\,.$$

The posterior distribution of $\tilde{p}$ is then given by

$$q^*(p) = f_{\tilde{p}|X_1=x_1,\ldots,X_n=x_n}(p) = \frac{P_p(X_1 = x_1, \ldots, X_n = x_n) \cdot q(p)}{\int_0^1 P_{p_1}(X_1 = x_1, \ldots, X_n = x_n) \cdot q(p_1)\, dp_1}\,.$$

It is always possible to calculate the posterior distribution with respect to a function $g(X_1, \ldots, X_n)$ instead of the vector $(X_1, \ldots, X_n)$.
Here, $Y = g(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$ denotes the number of successful trials within n experiments, where

$$X_i = \begin{cases} 1\,, & \text{with probability } p\,, \\ 0\,, & \text{else.} \end{cases}$$

Therefore,

$$q^*(p) = f_{\tilde{p}|Y=k}(p) = \frac{P_p(Y=k)\cdot q(p)}{\int_0^1 P_{p_1}(Y=k)q(p_1)\,dp_1}$$
$$\underset{\text{falls } \tilde{p}=p}{\overset{Y\sim Bin(n,p),}{=}} \frac{\binom{n}{k}p^k(1-p)^{n-k}\cdot(B(\alpha,\beta))^{-1}\cdot p^{\alpha-1}(1-p)^{\beta-1}}{\frac{\binom{n}{k}}{B(\alpha,\beta)}\cdot\int_0^1 p_1^{k+\alpha-1}(1-p_1)^{n-k+\beta-1}\,dp_1}$$
$$= \frac{p^{k+\alpha-1}(1-p)^{n-k+\beta-1}}{B(k+\alpha,n-k+\beta)}\,, \qquad p\in[0,1]\,.$$

holds for the posterior distribution with respect to Y. Hence, the posterior distribution of $\tilde{p}$ under the condition $Y=k$ is given by

$$Beta(k+\alpha, n-k+\beta).$$

For the *Bayes estimator* it holds that

$$\hat{p}(x_1,\ldots,x_n) = \mathrm{E}_*\tilde{p} = \int_0^1 p\cdot q^*(p)\,dp = \frac{\int_0^1 p^{k+\alpha}(1-p)^{n-k+\beta-1}\,dp}{B(k+\alpha,n-k+\beta)}$$
$$= \frac{B(k+\alpha+1,n-k+\beta)}{B(k+\alpha,n-k+\beta)}$$
$$= \ldots = \frac{k+\alpha}{\alpha+\beta+n}$$
$$= \frac{\sum_{i=1}^n x_i+\alpha}{\alpha+\beta+n}$$
$$= \frac{\alpha+n\bar{x}_n}{\alpha+\beta+n}\,.$$

Interpretation:

$$\hat{p}(X_1,\ldots,X_n) = \underbrace{\frac{n}{\alpha+\beta+n}}_{=:c_1}\bar{X}_n + \underbrace{\frac{\alpha+\beta}{\alpha+\beta+n}}_{=:c_2}\cdot\frac{\alpha}{\alpha+\beta} = c_1\cdot\bar{X}_n + c_2\cdot\mathrm{E}_{apr}\tilde{\theta}\,,$$

where $c_1+c_2=1$. This means that the Bayes method is a middle ground between the estimator $\mathrm{E}_{apr}\tilde{\theta}$ (with no information about the sample $(X_1,\ldots,X_n)$) and the estimator $\bar{X}_n$ (with no information about the prior distribution of $\tilde{p}$) for p.

1.2.5 Resampling methods for obtaining point estimators

Let $(X_1,\ldots,X_n)$ be a random sample in a parametric model. The goal is to find an estimator $\hat{\theta}$ for the parameter θ. In order to construct this estimator, resampling methods will be applied, i.e. generating a new sample $(X_1^*,\ldots,X_n^*)$ by randomly drawing from the old random sample $(X_1,\ldots,X_n)$ independently with replacement. After resampling the sample mean, sample

variance and other estimators with respect to the new sample can be computed. In this case the dimension m of the parameter space Θ is arbitrary.

The following *resampling* methods are introduced:

1. *Jackknife*, which is supposed to imply its handiness in every situation
2. *Bootstrap*, which is supposed to imply its self-sufficiency

1. *Jackknife methods for estimating the variance or the bias of estimators*

   As an introductory example, consider $\theta = \mathrm{E}X = \mu$ or $\theta = \operatorname{Var} X = \sigma^2$ and the respective (unbiased) estimators $\hat{\mu} = \bar{X}_n$ or $\hat{\sigma}^2 = S_n^2$.

   It is already known that

$$\operatorname{Var} \hat{\mu} = \frac{\sigma^2}{n}\,, \qquad \operatorname{Var} \hat{\sigma}^2 = \frac{1}{n}\left(\mu_4' - \frac{n-3}{n-1}\sigma^4\right)\,.$$

   Now an estimator of the variance of $\hat{\mu}$ resp. $\hat{\sigma}^2$ is desired. In order to do so, the plug-in methods are useful:

$$\widehat{\operatorname{Var}}\, \hat{\mu} = \frac{S_n^2}{n}\,, \qquad \widehat{\operatorname{Var}}\, \hat{\sigma}^2 = \frac{1}{n}\left(\hat{\mu}_4' - \frac{n-3}{n-1}S_n^4\right)\,,$$

   where $\hat{\mu}_4'$ is the fourth centered empirical moment.

   In general there are *no* explicit formulas for $\operatorname{Var} \hat{\theta}$ known. That is where the *jackknife* method comes into play.

   - Let $X_{[i]}$ be the random sample given by $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$, $i = 1, \ldots, n$. Let

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

     and set

$$\hat{\theta}_{[i]} = \varphi_{n-1}(X_{[i]})\,, \quad \bar{\theta}_{[\cdot]} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{[i]}$$

$$\widehat{\operatorname{Var}}_{jn}(\hat{\theta}) \stackrel{\text{def.}}{=} \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{[i]} - \bar{\theta}_{[\cdot]}\right)^2\,.$$

   Definition 1.2.30. The estimator $\bar{\theta}_{[\cdot]}$ resp. $\widehat{\operatorname{Var}}_{jn}(\hat{\theta})$ is called a *jackknife estimator* for the expectation resp. variance of the estimator $\hat{\theta}$ of θ.

   Example 1.2.31. Let $\theta = \mu\,, \quad \hat{\theta} = \hat{\mu} = \bar{X}_n$. Then

$$\varphi_n(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i\,,$$

which implies that

$$\begin{aligned}\hat{\theta}_{[i]} = \frac{1}{n-1}\sum_{j\neq i} X_j &= \frac{1}{n-1}\left(-X_i + \sum_{j=1}^{n} X_j\right)\\ &= \frac{n}{n-1}\bar{X}_n - \frac{1}{n-1}X_i\,, \quad \forall i = 1,\dots,n\,,\\ \bar{\theta}_{[\cdot]} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{[i]} &= \frac{n}{n-1}\bar{X}_n - \frac{1}{n(n-1)}\sum_{i=1}^{n} X_i\\ &= \frac{n\cdot\bar{X}_n}{n-1} - \frac{\bar{X}_n}{n-1} = \frac{n-1}{n-1}\bar{X}_n = \bar{X}_n\,.\end{aligned}$$

Thus, a *jackknife estimator for* μ is equal to $\bar{X}_n$.
Construction of a *jackknife estimator for the variance*:

$$\begin{aligned}\widehat{\mathrm{Var}}_{jn}(\hat{\theta}) &= \frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{n}{n-1}\bar{X}_n - \frac{1}{n-1}X_i - \bar{X}_n\right)^2\\ &= \frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{1}{n-1}(\bar{X}_n - X_i)\right)^2\\ &= \frac{n-1}{n(n-1)^2}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2,\\ &= \frac{1}{n}S_n^2\,,\end{aligned}$$

which is exactly the plug-in estimator for the variance of $\hat{\mu}$.

- *jackknife for the bias of an estimator*
  Let $\hat{\theta}(X_1,\dots,X_n)$ be an estimator for θ. The bias of $\hat{\theta}$ given by $\mathrm{E}_\theta\hat{\theta} - \theta = \mathrm{Bias}(\hat{\theta})$.

  Definition 1.2.32. A *jackknife estimator of the bias* of $\hat{\theta}$ is given by

  $$\widehat{\mathrm{Bias}}_{jn}(\hat{\theta}) = (n-1)(\bar{\theta}_{[\cdot]} - \hat{\theta}).$$

  The following examples show that the procedure above leads to a decreasing bias: The estimator

  $$\tilde{\theta} = \hat{\theta} - \widehat{\mathrm{Bias}}_{jn}(\hat{\theta}) = n\hat{\theta} - (n-1)\bar{\theta}_{[\cdot]} \tag{1.10}$$

  generally has a *smaller bias* than $\hat{\theta}$. Here

  $$\hat{\theta}_{[i]} = \varphi_{n-1}(X_{[i]}) \quad \text{and} \quad \bar{\theta}_{[\cdot]} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{[i]}$$

  with

  $$\hat{\theta}(X_1,\dots,X_n) = \varphi_n(X_1,\dots,X_n)\ .$$

Example 1.2.33.

(a) Let $\theta = \mathrm{E}X_i = \mu$. Then $\hat{\theta} = \bar{X}_n$ is an unbiased estimator for μ. Now the question for the corrected bias estimator $\tilde{\mu}$ arises? (It is not supposed to be any worse!)
It holds that $\bar{\theta}_{[\cdot]} = \bar{X}_n$ thus, the bias estimator of jackknife $\widehat{\mathrm{Bias}}_{jn}(\hat{\theta}) = (n-1)(\bar{X}_n - \bar{X}_n) = 0$, and therefore $\tilde{\theta} = \hat{\theta} - 0 = \bar{X}_n$. Hence, the jackknife method does not (at least in this example) add additional bias.

(b) $\theta = \sigma^2 = \mathrm{Var}X_i$, $\hat{\theta} = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ is a biased moment estimator of the variance. The question of how $\tilde{\theta}$ looks like arises.
Exercise 1.2.34. Show that the bias corrected estimator $\tilde{\theta}$ is an unbiased estimator of the variance:

$$\tilde{\theta} = S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{n}{n-1}\hat{\sigma}^2$$

It follows that the bias of $\hat{\sigma}^2$ is completely removed by applying the jackknife method.
Idea of the proof: Show that

$$\widehat{\mathrm{Bias}}_{jn}(\hat{\theta}) = -\frac{1}{n(n-1)}\sum_{i=1}^n (X_i - \bar{X}_n)^2 .$$

Remark 1.2.35. The examples 1.2.33 a), b), which provided a jackknife estimator in analytic form are rather an exception. In most cases, the reduction of the bias is achieved by using Monte-Carlo methods on the basis of (1.10).

2. *Bootstrap estimator*

The bootstrap method draws a new random sample $(X_1^*, \ldots, X_n^*)$ from an approximate distribution $\hat{F}$ of the random sample variables X_i, $i = 1, \ldots, n$. Let E_* and Var_* be the expectation and variance with respect to the distribution P_* of $(X_1^*, \ldots, X_n^*)$. There are two possibilities for the construction of $\hat{F}$:

i) $\hat{F}(x) = \hat{F}_n(x)$, which is the empirical distribution of X_i, if X_i are i.i.d.

ii) $\hat{F}$, which is a parametric estimator of the parametric distribution F, of X_i. That means, if $X_i \sim F_\theta$, $i = 1, \ldots, n$ for a $\theta \in \Theta$ and $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ an estimator for θ, then $\hat{F} = F_{\hat{\theta}}$ (plug-in method).

Definition 1.2.36. A *bootstrap estimator* for the *expectation* (resp. *bias* or *variance*) of the estimator $\hat{\theta}(X_1, \ldots, X_n)$ is given by

(a) $\hat{\mathrm{E}}_{boot}(\hat{\theta}) = \mathrm{E}_*\hat{\theta}(X_1^*, \ldots, X_n^*)$.
(b) $\widehat{\mathrm{Bias}}_{boot}(\hat{\theta}) = \hat{\mathrm{E}}_{boot}\hat{\theta} - \hat{\theta}$.
(c) $\widehat{\mathrm{Var}}_{boot}(\hat{\theta}) = \mathrm{Var}_*(\hat{\theta}(X_1^*, \ldots, X_n^*))$.

Example 1.2.37. Let $\theta = \mu = \mathrm{E}X_i$ and $\hat{F} = \hat{F}_n$ be the empirical distribution function. How is a random sample $X_1^*, \ldots, X_n^*$ with $X_i^* \sim \hat{F}_n$ generated?

The empirical distribution function $\hat{F}_n$ weighs every observation x_i of the original sample with a weight $1/n$. As a consequence, it is sufficient to select one of the entries in $(x_1, \ldots, x_n)$ (with probability $1/n$, urn model "drawing with replacement"), in order to generate X_j^*, $j = 1, \ldots, n$.

Bootstrap estimator for the expectation $\hat{\mu} = \bar{X}_n$:

$$\begin{aligned}\hat{\mathrm{E}}_{boot}\hat{\mu} = \mathrm{E}_*\left(\frac{1}{n}\sum_{i=1}^n X_i^*\right) &\overset{X_i^* \text{ i.i.d.}}{=} \frac{1}{n}\cdot n\mathrm{E}_*(X_1^*) \\ &= \int x\, d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n\,.\end{aligned}$$

It follows that $\widehat{\mathrm{Bias}}_{boot}\hat{\mu} = 0$. Moreover,

$$\begin{aligned}\widehat{\mathrm{Var}}_{boot}(\hat{\mu}) = \mathrm{Var}_*\left(\frac{1}{n}\sum_{i=1}^n X_i^*\right) &\overset{X_i^* \text{ u.i.v.}}{=} \frac{1}{n^2}\cdot n\cdot \mathrm{Var}_*(X_1^*) \\ &= \frac{1}{n}\cdot\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{\hat{\sigma}^2}{n}\,,\end{aligned}$$

is a plug-in estimator for $\mathrm{Var}\bar{X}_n = \sigma^2/n$.

Monte-Carlo methods for constructing bootstrap estimators numerically:

What can be done, if there is no explicit expression of $\widehat{\mathrm{Var}}_{Boot}(\hat{\theta})$ (which is usually the case in statistics)?

Generate M independent random samples $(X_{i1}^*, \ldots, X_{in}^*)$, $i = 1, \ldots, M$ under i) or ii) by using Monte-Carlo simulation. Then,

$$\hat{\theta}_i = \hat{\theta}(X_{i1}^*, \ldots, X_{in}^*)\,, \quad i = 1, \ldots, M \quad \text{and set} \quad \hat{\mathrm{E}}_{boot}\hat{\theta} \approx \frac{1}{M}\sum_{i=1}^M \hat{\theta}_i\,.$$

Similarily a bootstrap estimator for $\mathrm{Bias}\,\hat{\theta}$ and $\mathrm{Var}\,\hat{\theta}$ is obtained:

$$\widehat{\mathrm{Bias}}_{boot}\hat{\theta} \approx \hat{\mathrm{E}}_{boot}\hat{\theta} - \hat{\theta}\,, \quad \widehat{\mathrm{Var}}_{boot}\hat{\theta} \approx \frac{1}{M-1}\sum_{i=1}^M \left(\hat{\theta}_i - \hat{\mathrm{E}}_{boot}\hat{\theta}\right)^2.$$

Furthermore, the distribution function of X_{ij}^* can be determined by the empirical distribution function, i.e.

$$\hat{F}_{boot}(x) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{n}\sum_{j=1}^{n} I(X_{ij}^* \le x)\,, \qquad x \in \mathbb{R}\,.$$

Using the methods above the *Bootstrap confidence intervals* for $\hat{\theta}$ can be constructed:
The quantiles $\hat{F}_{\hat{\theta}}^{-1}(\alpha_1)$ and $\hat{F}_{\hat{\theta}}^{-1}(1-\alpha_2)$ of the distribution of $\hat{\theta}(X_1^*, \ldots, X_n^*)$ originating from the sample $(\hat{\theta}_1, \ldots, \hat{\theta}_M)$ can be estimated empirically. Then

$$P\left(\hat{F}_{\hat{\theta}}^{-1}(\alpha_1) \le \hat{\theta}(X_1^*, \ldots, X_n^*) \le \hat{F}_{\hat{\theta}}^{-1}(1-\alpha_2)\right) \approx 1 - \alpha_1 - \alpha_2 = 1 - \alpha\,,$$

where $\alpha = \alpha_1 + \alpha_2$ is sufficiently small. Note that it is desired that X_i^* are similarily distributed as the X_i, and hence

$$P\left(\hat{F}_{\hat{\theta}}^{-1}(\alpha_1) \le \hat{\theta}(X_1, \ldots, X_n) \le \hat{F}_{\hat{\theta}}^{-1}(1-\alpha_2)\right) \approx 1 - \alpha$$

holds.

1.3 Further quality properties of point estimators

1.3.1 Cramér-Rao inequality

Let $(X_1, \ldots, X_n)$ be a random sample of i.i.d. random variables X_i with distribution function F_θ, $\theta \in \Theta \subset \mathbb{R}\,, i.e.\,, m = 1$. Let $\hat{\theta}(X_1, \ldots, X_n)$ be an estimator for θ. If $\hat{\theta}$ is unbiased, then the quality of another unbiased estimator $\tilde{\theta}$ of θ is determined by the its variance. That means, if $\mathrm{Var}_\theta\, \tilde{\theta} < \mathrm{Var}_\theta\, \hat{\theta}$ then $\tilde{\theta}$ is in a sense better. This section strives to answer the question, whether it is always possible to find a newer, better estimator $\tilde{\theta}$ with decreasing variance. Under certain conditions this is not possible. The lower bound for $\mathrm{Var}_\theta\, \hat{\theta}$ is given by the Cramér-Rao Theorem.
Let $L(x, \theta)$ be the likelihood function of X_i, i.e.

$$L(x,\theta) = \begin{cases} P_\theta(x)\,, & \text{in the discrete case}\,, \\ f_\theta(x)\,, & \text{in the absolutely continuous case,} \end{cases}$$

and $L(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} L(x_i, \theta)$ the likelihood function of the whole random sample $(X_1, \ldots, X_n)$. The conditions 1) to 5) for the asymptotically normal distribution on page 24 hold, where 5) holds for $k = 1$.

Theorem 1.3.1 (Inequality of Cramér-Rao)**.** Let $\hat{\theta}(X_1, \ldots, X_n)$ be an estimator for θ with the following properties:

1. $\mathrm{E}_\theta \hat{\theta}^2(X_1,\ldots,X_n) < \infty \quad \forall\, \theta \in \Theta$.
2. For all $\theta \in \Theta$ exists $\frac{d}{d\theta}\mathrm{E}_\theta \hat{\theta}(X_1,\ldots,X_n)$, given by

$$\begin{cases} \int_{\mathbb{R}^n} \hat{\theta}(x_1,\ldots,x_n)\frac{\partial}{\partial\theta}L(x_1,\ldots,x_n,\theta)dx_1\ldots dx_n, & \text{in the abs. cont. case ,} \\ \sum_{x_1,\ldots,x_n} \hat{\theta}(x_1,\ldots,x_n)\frac{\partial}{\partial\theta}L(x_1,\ldots,x_n,\theta), & \text{in the discrete case} \end{cases}$$

Then, a lower bound for the variance of $\hat{\theta}$ is attained, i.e.

$$\mathrm{Var}_\theta\, \hat{\theta}(X_1,\ldots,X_n) \geq \frac{\left(\frac{d}{d\theta}\mathrm{E}_\theta\, \hat{\theta}(X_1,\ldots,X_n)\right)^2}{n\cdot I(\theta)}\,, \qquad \theta \in \Theta\,,$$

where $I(\theta)$ is the Fisher information defined in (1.5).

Proof Let

$$\varphi_\theta(x_1,\ldots,x_n) = \frac{\partial}{\partial\theta}\log L(x_1,\ldots,x_n,\theta).$$

In Remark 1.2.20 it has been shown that

$$\mathrm{E}_\theta\varphi_\theta(X_1,\ldots,X_n) = 0\,, \quad \mathrm{Var}_\theta\, \varphi_\theta(X_1,\ldots,X_n) = n\cdot I(\theta)\,.$$

Applying the Cauchy-Schwartz inequality to

$$\mathrm{Cov}\,_\theta(\varphi_\theta(X_1,\ldots,X_n),\hat{\theta}(X_1,\ldots,X_n))$$

yields

$$\begin{aligned} &\mathrm{Cov}\,_\theta\left(\varphi_\theta(X_1,\ldots,X_n),\hat{\theta}(X_1,\ldots,X_n)\right) \\ &\qquad = \mathrm{E}_\theta\left(\varphi_\theta(X_1,\ldots,X_n)\cdot\hat{\theta}(X_1,\ldots,X_n)\right) - 0 \\ &\qquad \leq \sqrt{\mathrm{Var}_\theta\,\varphi_\theta(X_1,\ldots,X_n)}\sqrt{\mathrm{Var}_\theta\,\hat{\theta}(X_1,\ldots,X_n)} \end{aligned}$$

Thus,

$$\mathrm{Var}_\theta\,\hat{\theta}(X_1,\ldots,X_n) \geq \frac{\Big(\overbrace{\mathrm{E}_\theta\left(\varphi_\theta(X_1,\ldots,X_n)\cdot\hat{\theta}(X_1,\ldots,X_n)\right)}^{=:A}\Big)^2}{\mathrm{Var}_\theta\,\varphi_\theta(X_1,\ldots,X_n)} = \frac{A^2}{n\cdot I(\theta)}\,.$$

Now it suffices to show

$$A = \frac{d}{d\theta}\mathrm{E}_\theta\,\hat{\theta}(X_1,\ldots,X_n)\,.$$

Only the absolutely continuous case will be shown (in the discrete case, replace the integrals with sums):

$$\begin{aligned} A &= \int \frac{\partial}{\partial\theta}\log L(x_1,\ldots,x_n,\theta)\cdot\hat{\theta}(x_1,\ldots,x_n)\cdot L(x_1,\ldots,x_n,\theta)\,dx_1\ldots dx_n \\ &= \int \frac{\partial}{\partial\theta}L(x_1,\ldots,x_n,\theta)\cdot\hat{\theta}(x_1,\ldots,x_n)\,dx_1\ldots dx_n \\ &\overset{\text{Cond. 2)}}{=} \frac{d}{d\theta}\mathrm{E}_\theta\,\hat{\theta}(X_1,\ldots,X_n)\,. \end{aligned}$$

□

Corollary 1.3.2. If $\hat{\theta}$ is an unbiased estimator for θ and the conditions of Theorem 1.3.1 are fulfilled, then

$$Var_\theta\, \hat{\theta}(X_1, \ldots, X_n) \geq \frac{1}{n \cdot I(\theta)}\,.$$

Proof Apply the Cramér-Rao inequality with

$$\frac{d}{d\theta}\left(\mathrm{E}_\theta\, \hat{\theta}(X_1, \ldots, X_n)\right) = \frac{d}{d\theta}\theta = 1.$$

□

The following examples will show, that the estimator $\bar{X}_n$ of the expectation μ has the smallest variance within the class of all estimators μ which fulfill the conditions of Theorem 1.3.1. Hence, the sample mean $\bar{X}_n$ is the best unbiased estimator in this class for at least two families of distributions:

- Normal distribution and
- Poisson distribution.

Example 1.3.3.

1. Let $X_i \sim N(\mu, \sigma^2)$ and $\hat{\mu} = \bar{X}_n$ be an estimator for μ. Here, $\hat{\mu}$ is unbiased with $\mathrm{Var}\hat{\mu} = \sigma^2/n$. In the following it will be shown that the Cramér-Rao boundary for the variance of an unbiased estimator $\hat{\theta}$ for μ is also given by σ^2/n. In an initial step, the conditions of Theorem 1.3.1 will be validated. In order to show that

$$0 = \frac{d}{d\mu}\int_{\mathbb{R}} L(x,\mu)\, dx = \int_{\mathbb{R}} \frac{\partial}{\partial\mu} L(x,\mu)\, dx$$

with

$$L(x,\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

consider

$$\frac{\partial}{\partial\mu} L(x,\mu) = \frac{2(x-\mu)}{2\sigma^2} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{x-\mu}{\sigma^2} \cdot L(x,\mu)\,,$$

$$\int_{\mathbb{R}} \frac{\partial}{\partial\mu} L(x,\mu)\, dx = \mathrm{E}\left(\frac{X-\mu}{\sigma^2}\right) = 0\,.$$

For condition 2) in Theorem 1.3.1 it holds that

$$\begin{aligned}\frac{d}{d\mu}\mathrm{E}\bar{X}_n &= \frac{d}{d\mu}(\mu) = 1 \\ &\stackrel{?}{=} \frac{1}{n}\int_{\mathbb{R}^n} (x_1 + \ldots + x_n) \frac{\partial}{\partial\mu}\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right) dx_1 \ldots dx_n\,.\end{aligned}$$

Induction with repsect to n:

- Initial step $n = 1$:

$$\begin{aligned}\int_{\mathbb{R}} x \frac{\partial}{\partial \mu} L(x, \mu)\, dx &= \int_{\mathbb{R}} \frac{x(x-\mu)}{\sigma^2} L(x, \mu)\, dx \\ &= \frac{1}{\sigma^2} \left(\mathrm{E}_\mu\, X^2 - \mu^2 \right) = \frac{\mathrm{Var}_\mu\, X}{\sigma^2} = 1\,.\end{aligned}$$

- Induction hypothesis: For n it holds that

$$\int_{\mathbb{R}^n} (x_1 + \ldots + x_n) \cdot \frac{\partial}{\partial \mu} L(x_1, \ldots, x_n, \mu)\, dx_1 \ldots dx_n = n\,.$$

- Induction step $n \to n + 1$:

$$\begin{aligned}A &= \int_{\mathbb{R}^{n+1}} (x_1 + \ldots + x_{n+1}) \frac{\partial}{\partial \mu} \underbrace{L(x_1, \ldots, x_{n+1}, \mu)}_{= L(x_1, \ldots, x_n, \mu) \cdot L(x_{n+1}, \mu)}\, dx_1 \ldots dx_{n+1} \\ &\stackrel{?}{=} n + 1\,.\end{aligned}$$

  For A it holds that

$$\begin{aligned}A = \int_{\mathbb{R}^{n+1}} (x_1 + \ldots + x_n) \cdot \Bigg(& \frac{\partial}{\partial \mu} L(x_1, \ldots, x_n, \mu) \cdot L(x_{n+1}, \mu) \\ & + L(x_1, \ldots, x_n, \mu) \cdot \frac{\partial}{\partial \mu} L(x_{n+1}, \mu) \Bigg)\, dx_1 \ldots dx_n dx_{n+1} \\ & + \int_{\mathbb{R}^{n+1}} x_{n+1} \Bigg(\frac{\partial}{\partial \mu} L(x_1, \ldots, x_n, \mu) \cdot L(x_{n+1}, \mu) \\ & + L(x_1, \ldots, x_n, \mu) \cdot \frac{\partial}{\partial \mu} L(x_{n+1}, \mu) \Bigg)\, dx_1 \ldots dx_n dx_{n+1}\end{aligned}$$

$$
\begin{aligned}
&= n \cdot \underbrace{\int_{\mathbb{R}} L(x_{n+1}, \mu)\, dx_{n+1}}_{=1} \\
&+ \int_{\mathbb{R}^n} (x_1 + \ldots + x_n) \cdot L(x_1, \ldots, x_n, \mu)\, dx_1 \ldots dx_n \\
&\quad \cdot \underbrace{\int \frac{\partial}{\partial \mu} L(x_{n+1}, \mu)\, dx_{n+1}}_{=0} + \int_{\mathbb{R}} x_{n+1} L(x_{n+1}, \mu)\, dx_{n+1} \\
&\quad \cdot \underbrace{\int_{\mathbb{R}^n} \frac{\partial}{\partial \mu} L(x_1, \ldots, x_n, \mu)\, dx_1 \ldots dx_n}_{=0} \\
&+ \underbrace{\int_{\mathbb{R}} x_{n+1} \frac{\partial}{\partial \mu} L(x_{n+1}, \mu)\, dx_{n+1}}_{= \frac{d}{d\mu} \mathrm{E}_\mu X = \frac{d}{d\mu} \mu = 1} \\
&\quad \cdot \underbrace{\int_{\mathbb{R}^n} L(x_1, \ldots, x_n, \mu)\, dx_1 \ldots dx_n}_{=1} = n + 1 \,.
\end{aligned}
$$

Since all conditions are fulfilled, the bound can be computed by

$$\frac{1}{n \cdot I(\mu)}$$

with

$$I(\mu) = \mathrm{E}_\mu \left(\frac{\partial}{\partial \mu} \log L(X, \mu) \right)^2 .$$

Example 1.2.21 implies that

$$I(\mu) = \frac{1}{\sigma^2} \quad \Longrightarrow \quad n \cdot I(\mu) = \frac{n}{\sigma^2} \,.$$

In summary:

$$\mathrm{Var}_\mu\, \hat{\theta} \geq \frac{1}{\frac{n}{\sigma^2}} = \frac{\sigma^2}{n} = \mathrm{Var}_\mu\, \bar{X}_n$$

holds for an arbitrary estimator $\hat{\theta}$ for μ, which fulfills the conditions of Theorem 1.3.1.

2. The second example will be an exercise.

   Exercise 1.3.4. Let $X_i \sim Poisson(\lambda)\,, \quad i = 1, \ldots, n$. Show that the Cramér-Rao bound given by

   $$\frac{1}{n \cdot I(\lambda)} = \frac{\lambda}{n} = \mathrm{Var}_\lambda\, \bar{X}_n,$$

   which means that $\bar{X}_n$ is also the best unbiased estimator fulfilling the conditions of Theorem 1.3.1.

The following example will show that it is possible to construct an estimator with a smaller variance than the one provided by the Cramér-Rao bound, if the conditions of Theorem 1.3.1 are not fulfilled.

Example 1.3.5. Let $X_i \sim U[0,\theta]$, $\theta > 0$. Then the condition "supp $f_\theta(x) = [0,\theta]$ independent of θ" is not met. Additionally

$$0 \neq \int_{\mathbb{R}} \frac{\partial}{\partial\theta} L(x,\theta)\,dx = \int_0^\theta \left(\frac{1}{\theta}\right)' dx = -\frac{1}{\theta^2}\cdot\theta = -\frac{1}{\theta}$$

holds. Let $\hat{\theta}$ be an unbiased estimator of θ, then Cramér-Rao would imply that $\operatorname{Var}_\theta \hat{\theta} \geq (n\cdot I(\theta))^{-1}$, where

$$\begin{aligned} I(\theta) &= \mathrm{E}\left(\frac{\partial}{\partial\theta}\log L(X,\theta)\right)^2 = \int_0^\theta \frac{1}{\theta}\left(\frac{\partial}{\partial\theta}\log\left(\frac{1}{\theta}\right)\right)^2 dx \\ &= \frac{1}{\theta}\int_0^\theta dx\cdot\left(-\frac{1}{\theta}\right)^2 = \frac{1}{\theta^2}\,. \end{aligned}$$

Thus

$$\operatorname{Var}_\theta \hat{\theta} \geq \frac{\theta^2}{n}$$

would hold. Consider

$$\hat{\theta}(X_1,\ldots,X_n) = \frac{n+1}{n}\max\{X_1,\ldots,X_n\} = \frac{n+1}{n}X_{(n)}\,.$$

In order to show that

$$\mathrm{E}_\theta\,\hat{\theta}(X_1,\ldots,X_n) = \theta \quad \text{and} \quad \operatorname{Var}_\theta\,\hat{\theta}(X_1,\ldots,X_n) < \frac{\theta^2}{n}\,,$$

compute the k-th moments $\mathrm{E}_\theta X_{(n)}^k$, $k \in \mathbb{N}$.
It holds that

$$\begin{aligned} F_{X_{(n)}}(x) = F_{X_i}^n(x) &= \begin{cases} \frac{x^n}{\theta^n}\,, & x\in[0,\theta]\,, \\ 1\,, & x \geq \theta\,, \\ 0\,, & x < 0\,, \end{cases} \\ f_{X_{(n)}}(x) = F'_{X_{(n)}}(x) &= \frac{nx^{n-1}}{\theta^n}\cdot I(x\in[0,\theta])\,, \\ \mathrm{E}_\theta X_{(n)}^k &= \int_0^\theta x^k\frac{nx^{n-1}}{\theta^n}\,dx = \frac{n}{\theta^n}\int_0^\theta x^{n+k-1}\,dx = \frac{n\cdot\theta^{n+k}}{\theta^n\cdot(n+k)} = \frac{n\theta^k}{n+k}\,. \end{aligned}$$

Thus,

$$\mathrm{E}_\theta\,\hat{\theta} = \frac{n+1}{n}\cdot\mathrm{E}_\theta X_{(n)} = \frac{n+1}{n}\cdot\frac{n\theta}{n+1} = \theta\,,$$

which means that $\hat{\theta}$ is unbiased. Furthermore,

$$\begin{aligned}\operatorname{Var}_\theta \hat{\theta} &= \left(\frac{n+1}{n}\right)^2 \cdot \operatorname{Var}_\theta X_{(n)} = \left(\frac{n+1}{n}\right)^2 \cdot \left(\frac{n\theta^2}{n+2} - \frac{n^2\theta^2}{(n+1)^2}\right) \\ &= \frac{(n+1)^2}{n^2} \cdot \frac{n(n+1)^2 - n^2(n+2)}{(n+2)(n+1)^2} \cdot \theta^2 \\ &= \frac{\theta^2}{n(n+2)}(n^2 + 2n + 1 - n^2 - 2n) = \frac{\theta^2}{n(n+2)}.\end{aligned}$$

Ultimately, it follows that

$$\operatorname{Var}_\theta \hat{\theta} = \frac{\theta^2}{n(n+2)} < \frac{\theta^2}{n}\,.$$

1.3.2 Sufficiency

Let $(X_1, \ldots, X_n)$ be a random sample of i.i.d. random variables X_i with distribution function $F_\theta\,, \quad \theta \in \Theta \subseteq \mathbb{R}^m$. If the whole information $\{X_1 = x_1, \ldots, X_n = x_n\}$ passes to the estimator $\hat{\theta}(X_1, \ldots, X_n)$ of θ, then the function

$$\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^m\,, \qquad m \ll n^{4}$$

causes a loss of information, since $(X_1, \ldots, X_n)$ can (usually) not be reconstructed from $\hat{\theta}(X_1, \ldots, X_n)$. The class of so-called *sufficient* estimators minimize the loss of information in a stochastic sense:

Definition 1.3.6.

1. Let the random variables $X_1, \ldots, X_n$ and $\hat{\theta}(X_1, \ldots, X_n)$ be discrete. An estimator $\hat{\theta}$ of the parameter θ is called *sufficient*, if

$$P_\theta\left(X_1 = x_1, \ldots, X_n = x_n \,|\, \hat{\theta}(X_1, \ldots, X_n) = t\right)$$

   does not depend on θ, as long as $x_1, \ldots, x_n$ and t are in the support of $(X_1, \ldots, X_n)$ resp. $\hat{\theta}(X_1, \ldots, X_n)$.

2. Let $X_1, \ldots, X_n$ and $\hat{\theta}(X_1, \ldots, X_n)$ be absolutely continuous. Then, the estimator $\hat{\theta}$ is called *sufficient* for θ, if the probability

$$P\left((X_1, \ldots, X_n) \in B \,|\, \hat{\theta}(X_1, \ldots, X_n) = t\right)$$

   does not depend on $\theta \in \Theta$ for arbitrary $B \in \mathcal{B}_{\mathbb{R}^n}$ and $t \in \operatorname{supp} f_{\hat{\theta}}$, where $f_{\hat{\theta}}$ is the probability density function of $\hat{\theta}$.

Remark 1.3.7.

[4] in classical statistics

1. In Definition 1.3.6, 2. it holds that

$$P\left(\hat{\theta}(X_1,\dots,X_n)=t\right)=0,\quad \forall t,$$

because of the absolute continuity of $\hat{\theta}$. Therefore, the conditional probability (in contrary to Definition 1.3.6, 1.) is not understood in the classical sense, but as a conditional expectation. Conditional expectations were introduced in the lecture "Probability Theory and Stochastic Processes" (Section 1.1.4).

2. Consider the likelihood function

$$L_{\hat{\theta}}(x_1,\dots,x_n,\theta)=P_\theta\left(X_1=x_1,\dots,X_n=x_n\,|\,\hat{\theta}(X_1,\dots,X_n)=t\right)$$

for discrete $X_1,\dots,X_n$. Definition 1.3.6 implies that a new estimator for θ cannot be obtained from this conditional likelihood function $L_\theta(x_1,\dots,x_n,\theta)$, since it does not depend on θ. That means, the estimator $\hat{\theta}$ already provides all the information about θ obtainable from $(x_1,\dots,x_n)$.

3. Let $g:\mathbb{R}^m\to\mathbb{R}^m$ be a bijective Borel measurable function and $\hat{\theta}(X_1,\dots,X_n)$ a sufficient estimator of $\theta\in\Theta\subset\mathbb{R}^m$. Then $g(\hat{\theta}(X_1,\dots,X_n))$ is also a sufficient estimator for θ. This is due to the fact that

$$\left\{\omega\in\Omega:\,g\left(\hat{\theta}(X_1,\dots,X_n)\right)=t\right\}=\left\{\omega\in\Omega:\,\hat{\theta}(X_1,\dots,X_n)=g^{-1}(t)\right\},$$

for all $t\in\mathbb{R}^m$.

Lemma 1.3.8. Assume that the random variables $X_1,\dots,X_n$ and $\hat{\theta}(X_1,\dots,X_n)$ are either all discrete or absolutely continuous with likelihood functions

$$L(x_1,\dots,x_n,\theta)=\begin{cases}P_\theta(X_1=x_1,\dots,X_n=x_n), & \text{in the discrete case,}\\ f_{X_1,\dots,X_n}(x_1,\dots,x_n,\theta), & \text{in the abs. cont. case,}\end{cases}$$

$$L_{\hat{\theta}}(t,\theta)=\begin{cases}P_\theta(\hat{\theta}(X_1,\dots,X_n)=t), & \text{in the discrete case,}\\ f_{\hat{\theta}}(t,\theta), & \text{in the abs. cont, case.}\end{cases}$$

Denote the support of L by

$$\operatorname{supp} L=\{(x_1,\dots,x_n)\in\mathbb{R}^n:\quad L(x_1,\dots,x_n,\theta)>0\}.$$

Then, the estimator $\hat{\theta}$ is sufficient with respect to θ if and only if the ratio

$$\frac{L(x_1,\dots,x_n,\theta)}{L_{\hat{\theta}}(\hat{\theta}(x_1,\dots,x_n),\theta)} \tag{1.11}$$

does not depend on θ for all $(x_1,\dots,x_n)\in\operatorname{supp} L$ such that $\hat{\theta}(x_1,\dots,x_n)\in\operatorname{supp} L_{\hat{\theta}}$.

Proof Only the discrete case will be shown in the following.

"$\Longrightarrow$"
Let $\hat{\theta}$ be sufficient for θ. Then, it has to be verified that (1.11) does not depend on θ for all $(x_1, \dots, x_n) \in \mathbb{R}$, $t \in \mathbb{R}$ and $\theta \in \Theta$, such that $(x_1, \dots, x_n) \in \text{supp}L$. It holds that:

$$\begin{aligned} &P_\theta(X_1 = x_1, \dots, X_n = x_n \,|\, \hat{\theta}(X_1, \dots, X_n) = t) \\ &\qquad = \frac{P_\theta(X_1 = x_1, \dots, X_n = x_n, \hat{\theta}(X_1, \dots, X_n) = t)}{P_\theta(\hat{\theta}(X_1, \dots, X_n) = t)} \\ &\qquad = \begin{cases} 0\,, & \text{if } \hat{\theta}(x_1, \dots, x_n) \neq t \\ \frac{P_\theta(X_1 = x_1, \dots, X_n = x_n)}{P_\theta(\hat{\theta}(X_1, \dots, X_n) = \hat{\theta}(x_1, \dots, x_n))}\,, & \text{if } \hat{\theta}(x_1, \dots, x_n) = t\,. \end{cases} \end{aligned}$$

Thus (1.11) does not depend on θ.

"$\Longleftarrow$"
Can be done in the same spirit as the previous argument (backwards).

□

Example 1.3.9.

1. *Bernoulli distribution:* Let $X_i \sim Bernoulli(p)$, $p \in [0,1]$, $i = 1, \dots, n$, $\hat{p} = \bar{X}_n$ be an unbiased estimator for p. In the following, it will be shown that $\hat{p}$ is sufficient. It holds that

$$\hat{p} = \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} Y\,,$$

where $Y \sim Bin(n, p)$. By Remark 1.3.7 3. it is sufficient to show, that Y is a sufficient estimator for p. For $x_i \in \{0, 1\}$ $i = 1, \dots, n$ compute

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}\,.$$

Define the likelihood function L_Y by

$$L_Y(y, p) = \binom{n}{y} p^y (1-p)^{n-y}\,, \qquad y = 0, \dots, n\,.$$

Replacing y with the sum $\sum_{i=1}^n x_i$ yields

$$\frac{L(x_1, \dots, x_n, p)}{L_Y(\sum_{i=1}^n x_i, p)} = \frac{p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}}{\binom{n}{\sum_{i=1}^n x_i} p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}} = \frac{1}{\binom{n}{\sum_{i=1}^n x_i}}\,.$$

The term above obviously does not depend on p, thus Lemma 1.3.8 implies, that Y and therefore also $\hat{p}$ are sufficient.

2. *Normal distribution with known variance:* Let $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$, with known σ^2. Then, $\hat{\mu} = \bar{X}_n$ is an unbiased estimator for μ. In the following it will be shown, that $\hat{\mu}$ is sufficient: Considering

$$\begin{aligned} L(x_1, \ldots, x_n, \mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\right) \end{aligned}$$

and [33, Lemma 6.4.5] imply

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 + n(\bar{x}_n - \mu)^2}{2\sigma^2}\right).$$

Furthermore, note that $\hat{\mu} \sim N(\mu, {}^{\sigma^2}/_{n})$, hence

$$\begin{aligned} L_{\hat{\mu}}(x, \mu) &= \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{n}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right), \\ \frac{L(x_1, \ldots, x_n, \mu)}{L_{\hat{\mu}}(\bar{x}_n, \mu)} &= \frac{\frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 + n(\bar{x}_n - \mu)^2}{2\sigma^2}\right)}{\frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \cdot \exp\left(\frac{-n(\bar{x}_n - \mu)^2}{2\sigma^2}\right)} \\ &= \frac{1}{\sqrt{n}(2\pi\sigma^2)^{n/2-1}} \cdot \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x}_n)\right), \end{aligned}$$

which is independent of μ. Lemma 1.3.8 implies that $\hat{\mu} = \bar{X}_n$ is an sufficient estimator for μ.

The Neyman-Fisher factorization theorem, which will be introduced below, implies that the estimator $(\bar{X}_n, S_n^2)$ for (μ, σ^2) with unknown variance is sufficient.

Theorem 1.3.10 (Neyman-Fisher Factorization Theorem)**.** Under the conditions of Lemma 1.3.8 it holds that $\hat{\theta}(X_1, \ldots, X_n)$ is a sufficient estimator for θ if and only if there exist two measurable functions $g : \mathbb{R}^m \times \Theta \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}$, such that the following factorization of the likelihood function $L(x_1, \ldots, x_n, \theta)$ of the random sample $(X_1, \ldots, X_n)$ holds:

$$L(x_1, \ldots, x_n, \theta) = g\left(\hat{\theta}(x_1, \ldots, x_n), \theta\right) \cdot h(x_1, \ldots, x_n)$$

for $(x_1, \ldots, x_n) \in \operatorname{supp} L, \theta \in \Theta$.

Proof Only the discrete case will be shown.

1. If $\hat{\theta}$ is sufficient, then Lemma 1.3.8 implies that

$$\frac{L(x_1,\ldots,x_n,\theta)}{\underbrace{L_{\hat{\theta}}(\hat{\theta}(x_1,\ldots,x_n),\theta)}_{=g(\hat{\theta}(x_1,\ldots,x_n),\theta)}} = h(x_1,\ldots,x_n)$$

does not depend on θ. Thus, the factorization of Neyman-Fisher holds.

2. Let $L(x_1,\ldots,x_n,\theta) = g(\hat{\theta}(x_1,\ldots,x_n),\theta)\cdot h(x_1,\ldots,x_n)$ for all $(x_1,\ldots,x_n)\in \text{supp } L$, $\theta\in\Theta$. Furthermore, define

$$\begin{aligned} C &= \{(y_1,\ldots,y_n)\in\mathbb{R}^n : \hat{\theta}(y_1,\ldots,y_n) = \hat{\theta}(x_1,\ldots,x_n)\} \\ &= \hat{\theta}^{-1}\left(\hat{\theta}\,(x_1,\ldots,x_n)\right), \end{aligned}$$

then

$$\begin{aligned} \frac{P_\theta(X_1=x_1,\ldots,X_n=x_n)}{\underbrace{L_\theta(\hat{\theta}(x_1,\ldots,x_n),\theta)}_{=P_\theta(\hat{\theta}(X_1,\ldots,X_n)=\hat{\theta}(x_1,\ldots,x_n))}} &= \frac{g(\hat{\theta}(x_1,\ldots,x_n),\theta)\cdot h(x_1,\ldots,x_n)}{\sum_{(y_1,\ldots,y_n)\in C} P_\theta(X_1=y_1,\ldots,X_n=y_n)} \\ &= \frac{g(\hat{\theta}(x_1,\ldots,x_n),\theta)\cdot h(x_1,\ldots,x_n)}{\sum_{(y_1,\ldots,y_n)\in C} g(\underbrace{\hat{\theta}(y_1,\ldots,y_n)}_{=\hat{\theta}(x_1,\ldots,x_n)},\theta)\cdot h(y_1,\ldots,y_n)} \\ &= \frac{h(x_1,\ldots,x_n)}{\sum_{(y_1,\ldots,y_n)\in C} h(y_1,\ldots,y_n)}, \end{aligned}$$

does not depend on θ. Thus, $\hat{\theta}$ is sufficient by Lemma 1.3.8.

□

Example 1.3.11.

1. *Poisson distribution:* Let $X_i \sim Poisson(\lambda)$, $\lambda>0$, $\hat{\lambda}=\bar{X}_n$ be an unbiased estimator for λ. In the following it will be shown that $\hat{\lambda}$ is sufficient. For $x_i\in\{0,1,2,\ldots\}$, $i=1,\ldots,n$ it holds that

$$\begin{aligned} L(x_1,\ldots,x_n,\lambda) = \prod_{i=1}^n e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} &= \frac{e^{-\lambda n}\cdot\lambda^{\sum_{i=1}^n x_i}}{x_1!\cdot\ldots\cdot x_n!} = \frac{e^{-n\lambda}\lambda^{n\bar{x}_n}}{x_1!\cdot\ldots\cdot x_n!}, \\ &= g(\bar{x}_n,\lambda)\cdot h(x_1,\ldots,x_n), \end{aligned}$$

where $g(\bar{x}_n,\lambda) = e^{-n\lambda}\cdot\lambda^{n\bar{x}_n}$, $h(x_1,\ldots,x_n) = \frac{1}{x_1!\cdot\ldots\cdot x_n!}$.
Thus, $\hat{\lambda}=\bar{X}_n$ is sufficient by Theorem 1.3.10.

2. *Exponential distribution* Let $X_i \sim Exp(\lambda)$, $\lambda > 0$, $\hat{\lambda} = \bar{X}_n^{-1}$ be a moment estimator for λ, which is not unbiased but strongly consistent, since the strong law of large numbers implies that $\bar{X}_n \xrightarrow[n\to\infty]{\text{a.s.}} \mathrm{E}X_i = \frac{1}{\lambda}$. In the following it will be shown, that $\hat{\lambda}$ is sufficient. For $x_1 \geq 0, \ldots, x_n \geq 0$ it holds that

$$\begin{aligned} L(x_1, \ldots, x_n, \lambda) &= \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i} = \lambda^n e^{-\lambda n \bar{x}_n} \\ &= \lambda^n e^{-\frac{\lambda n}{\hat{\lambda}}} = g\left(\hat{\lambda}, \lambda\right) \cdot \underbrace{h(x_1, \ldots, x_n)}_{=1}, \end{aligned}$$

where $g(\hat{\lambda}, \lambda) = \lambda^n e^{-\frac{\lambda n}{\hat{\lambda}}}$ and $h(x_1, \ldots, x_n) \equiv 1$. Thus, $\hat{\lambda}$ is sufficient by Theorem 1.3.10.

Exercise 1.3.12. Using Theorem 1.3.10 show that the estimator $(\bar{X}_n, S_n^2)$ is sufficient for (μ, σ^2) if the random sample $(X_1, \ldots, X_n)$ is i.i.d. with distribution $X_i \sim N(\mu, \sigma^2)$ for all i.

Remark 1.3.13. An advantage of the Neyman-Fisher Theorem is, that if one wants to determine whether an estimator $\hat{\theta}$ is sufficient, the likelihood function of $\hat{\theta}$ does not need to be known explicitly. This is particularly important if the estimator $\hat{\theta}$ is rather complicated and the likelihood function cannot be computed.

1.3.3 Completeness

Definition 1.3.14. An estimator $\hat{\theta}(X_1, \ldots, X_n)$ of the parameter $\theta \in \Theta \subset \mathbb{R}^m$ is called *complete*, if for an arbitrary measurable function $g : \mathbb{R}^m \to \mathbb{R}$ with $\mathrm{E}_\theta g(\hat{\theta}(X_1, \ldots, X_n)) = 0$, $\theta \in \Theta$ it holds that

$$g\left(\hat{\theta}\left(X_1, \ldots, X_n\right)\right) \equiv 0\,. \quad P_\theta - \text{a.s. for all } \theta \in \Theta.$$

Remark 1.3.15.

1. Let $g_1, g_2 : \mathbb{R}^m \to \mathbb{R}$ be functions with

$$\mathrm{E}_\theta \left| g_i\left(\hat{\theta}\left(X_1, \ldots, X_n\right)\right)\right| < \infty$$

$\forall \theta \in \Theta$ and

$$\mathrm{E}_\theta g_1\left(\hat{\theta}\left(X_1, \ldots, X_n\right)\right) = \mathrm{E}_\theta g_2\left(\hat{\theta}\left(X_1, \ldots, X_n\right)\right),$$

where $\hat{\theta}$ is complete. Definition 1.3.14 then implies

$$g_1\left(\hat{\theta}\left(X_1, \ldots, X_n\right)\right) = g_2\left(\hat{\theta}(X_1, \ldots, X_n)\right), \quad \text{a.s.}$$

(Take $g = g_1 - g_2$).

Conclusion: The completeness as characteristic allows a comparison between the estimators $g_1(\hat{\theta})$ and $g_2(\hat{\theta})$ with respect to their almost surely equality.

2. If $\hat{\theta}$ is a complete estimator for θ, then $g(\hat{\theta})$ is also a complete estimator for θ for an arbitrary measurable function $g : \mathbb{R}^m \to \mathbb{R}^m$.

Example 1.3.16.

1. *Bernoulli distribution:* Let $X_i \sim Bernoulli(p)$, $p \in [0,1]$. In order to show that $\hat{p} = \bar{X}_n$ is complete, let g be an arbitrary real valued function. It is sufficient to show that $Y = \sum_{i=1}^n X_i$ is complete. It holds that $Y \sim Bin(n,p)$, which implies that

$$\mathrm{E}_p g(Y) = \sum_{k=0}^{n} g(k) \binom{n}{k} p^k (1-p)^{n-k} .$$

Furthermore, $\mathrm{E}_p g(Y) = 0$ if and only if

$$\sum_{k=0}^{n} g(k) \binom{n}{k} \Bigg(\underbrace{\frac{p}{1-p}}_{=t} \Bigg)^k = p_n(t) = 0$$

for $p \in (0,1)$, so $t \in (0,\infty)$. The polynomial $p_n(t)$ is of degree n, hence

$$\begin{aligned} & g(k)\binom{n}{k} = 0 \quad \text{for all } k \\ \Longrightarrow \quad & g(k) = 0, \quad k = 0, \dots, n \\ \Longrightarrow \quad & g(Y) = 0 \quad P_p\text{-a.s.}. \end{aligned}$$

Therefore, Y is complete and $\hat{p} = \bar{X}_n$ as well.

2. *Uniform distribution:* Let $X_i \sim U[0,\theta]\,, \quad i = 1, \dots, n$. It has already been shown that the estimator $\hat{\theta}(X_1, \dots, X_n) = \frac{n+1}{n} X_{(n)}$ is unbiased. In order to show its completeness, it is sufficient to show that $X_{(n)} = \max_{i=1,\dots,n} X_i$ is complete, i.e. all measurable functions $g : \mathbb{R} \to \mathbb{R}$ with $\mathrm{E}_\theta g(X_{(n)}) = 0$ need to fulfill $g(X_{(n)}) = 0$ almost surely.

The probability density function of $X_{(n)}$ is given by $f_{X_{(n)}}(x) = \frac{n x^{n-1}}{\theta^n} \cdot I_{[0,\theta]}(x)$ by Example 1.3.5. Hence, we can compute

$$\begin{aligned} 0 &= \frac{d}{d\theta} \mathrm{E}_\theta g(X_{(n)}) = \frac{d}{d\theta} \int_0^\theta g(x) f_{X_{(n)}}(x)\, dx = \frac{d}{d\theta} \frac{1}{\theta^n} \int_0^\theta n x^{n-1} g(x)\, dx \\ &= -n \frac{1}{\theta^{n+1}} \int_0^\theta g(x) n x^{n-1}\, dx + \frac{1}{\theta^n} n \theta^{n-1} g(\theta) = -\frac{n}{\theta} \underbrace{\mathrm{E}_\theta g(X_{(n)})}_{=0} + \frac{n}{\theta} g(\theta) \\ &= \frac{n}{\theta} g(\theta) = 0 \ \text{ for all } \theta > 0 \Longrightarrow g(x) = 0\,, \quad x > 0. \end{aligned}$$

It follows that $g(X_{(n)}) = 0$ holds almost surely.

1.3.4 Best unbiased estimator

Following [33, Definition 7.2.9.] note that for a random sample $(X_1, \ldots, X_n)$ with i.i.d. random variables $X_i \sim F_\theta$, $\theta \in \Theta \subset \mathbb{R}$ $(m = 1)$, the estimator $\hat{\theta}(X_1, \ldots, X_n)$ is called *best unbiased estimator*, if

$$\mathrm{E}_\theta \hat{\theta}^2(X_1, \ldots, X_n) < \infty \qquad \mathrm{E}_\theta \hat{\theta}(X_1, \ldots, X_n) = \theta, \quad \theta \in \Theta, \text{ and}$$

the estimator $\hat{\theta}$ has the smallest variance among all unbiased estimators.

Lemma 1.3.17 (Uniqueness of the best unbiased estimator)**.** If $\hat{\theta}$ is a best unbiased estimator for θ, then it is unique.

Proof Let $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ be a best unbiased estimator for θ and $\tilde{\theta}$ another best unbiased estimator for θ. In the following it will be shown that both estimators coincide, i.e. $\hat{\theta} = \tilde{\theta}$.
Ex adverso: Assume, that $\hat{\theta} \neq \tilde{\theta}$ and consider $\theta^* = 1/2(\hat{\theta} + \tilde{\theta})$. Obviously θ^* is unbiased and its variance is given by

$$\mathrm{Var}_\theta \theta^* = \frac{1}{4}\mathrm{Var}_\theta(\hat{\theta} + \tilde{\theta}) = \frac{1}{4}\mathrm{Var}_\theta \hat{\theta} + \frac{1}{4}\mathrm{Var}_\theta \tilde{\theta} + \frac{1}{2}\mathrm{Cov}_\theta(\hat{\theta}, \tilde{\theta})\,.$$

The Cauchy-Schwartz inequality implies $|\mathrm{Cov}_\theta(\hat{\theta}, \tilde{\theta})| \leq \sqrt{\mathrm{Var}_\theta \hat{\theta} \cdot \mathrm{Var}_\theta \tilde{\theta}} = \mathrm{Var}_\theta \hat{\theta}$ and therefore

$$\mathrm{Var}_\theta \theta^* \leq \frac{1}{2}\mathrm{Var}_\theta \hat{\theta} + \frac{1}{2}\mathrm{Var}_\theta \hat{\theta} = \mathrm{Var}_\theta \hat{\theta}\,.$$

Since $\hat{\theta}$ is a best unbiased estimator, it follows $\mathrm{Var}_\theta \theta^* = \mathrm{Var}_\theta \hat{\theta}$, and consequently $\varrho(\hat{\theta}, \tilde{\theta}) = 1$ implies that $\hat{\theta}$ and $\tilde{\theta}$ are linearly dependent, i.e. there exist some constants a and b, such that $\hat{\theta} = a\tilde{\theta} + b$. It holds that $a = 1$ since $\mathrm{Var}_\theta \hat{\theta} = a^2 \mathrm{Var} \tilde{\theta} = \mathrm{Var}_\theta \hat{\theta}$. Moreover, $b = 0$, because $\hat{\theta}$ and $\tilde{\theta}$ are unbiased: $\theta = \mathrm{E}_\theta \hat{\theta} = \mathrm{E}_\theta \tilde{\theta} + b = \theta + b$. Ultimately, $\hat{\theta} = \tilde{\theta}$, which completes the proof. □

Lemma 1.3.18. A unbiased estimator $\hat{\theta}$ with finite second moment is the best unbiased estimator for θ if and only if $\mathrm{Cov}_\theta(\hat{\theta}, \varphi) = 0\,, \quad \theta \in \Theta$ for an arbitrary sample function $\varphi : \mathbb{R}^n \to \mathbb{R}$ with $\mathrm{E}_\theta \varphi(X_1, \ldots, X_n) = 0\,, \forall \theta \in \Theta$.

Proof

"⇒" Let $\hat{\theta}$ be the best unbiased estimator for θ and $\varphi(X_1, \ldots, X_n)$ a sample function with $\mathrm{E}_\theta \varphi(X_1, \ldots, X_n) = 0\,, \forall \theta \in \Theta$. It is sufficient to show $\mathrm{Cov}_\theta(\hat{\theta}, \varphi) = \mathrm{E}_\theta(\hat{\theta}\varphi) = 0\,, \theta \in \Theta$.
Define $\tilde{\theta} = \hat{\theta} + a\varphi\,, a \in \mathbb{R}$. In order to compute

$$\mathrm{Var}_\theta \tilde{\theta} = \mathrm{Var}_\theta \hat{\theta} + a^2 \mathrm{Var}_\theta \varphi + 2a\mathrm{Cov}_\theta(\hat{\theta}, \varphi)$$

for $a \in \mathbb{R}$, let $g(a) = a^2 \mathrm{Var}_\theta \varphi + 2a\mathrm{Cov}_\theta(\varphi, \hat{\theta})$. For $\mathrm{Cov}_\theta(\varphi, \hat{\theta}) \neq 0$ there exists an $a \in \mathbb{R}$ with $g(a) < 0$. Since $\tilde{\theta}$ is an unbiased estimator for θ ($\mathrm{E}_\theta \tilde{\theta} = \mathrm{E}_\theta \hat{\theta} + a\mathrm{E}_\theta \varphi = \theta + 0 = \theta$) it holds that $\mathrm{Var}_\theta \tilde{\theta} \geq \mathrm{Var}_\theta \hat{\theta}$ for all $a \in \mathbb{R}$. This is a contradiction to $g(a) < 0$ for an $a \in \mathbb{R}$. Thus, $\mathrm{Cov}_\theta(\varphi, \hat{\theta}) = 0\,, \theta \in \Theta$.

"$\Leftarrow$" Let $\hat{\theta}$ be an unbiased estimator with $\mathrm{E}_\theta \hat{\theta}^2 < \infty, \theta \in \Theta$ and $\mathrm{Cov}_\theta(\varphi, \hat{\theta}) = 0\,, \theta \in \Theta$ if $\mathrm{E}_\theta \varphi = 0, \theta \in \Theta$. Let $\tilde{\theta}$ be another unbiased estimator for θ. In order to show that $\mathrm{Var}_\theta \tilde{\theta} \geq \mathrm{Var}_\theta \hat{\theta}$, consider

$$\tilde{\theta} = \hat{\theta} + (\underbrace{\tilde{\theta} - \hat{\theta}}_{=:\varphi})\,, \quad \mathrm{E}_\theta \varphi = \mathrm{E}_\theta \tilde{\theta} - \mathrm{E}_\theta \hat{\theta} = \theta - \theta = 0\,, \quad \forall \theta \in \Theta\,.$$

It follows that

$$Var_\theta \tilde{\theta} = \mathrm{Var}_\theta \hat{\theta} + \underbrace{\mathrm{Var}_\theta \varphi}_{\geq 0} + 2\underbrace{\mathrm{Cov}_\theta(\hat{\theta}, \varphi)}_{=0} \geq \mathrm{Var}_\theta \hat{\theta}\,,$$

which implies, that $\hat{\theta}$ is the best unbiased estimator for θ.

□

Theorem 1.3.19 (Lehmann-Scheffé)**.** Let $\hat{\theta}$ be an unbiased, complete and sufficient estimator for θ with $\mathrm{E}_\theta \hat{\theta}^2 < \infty$ for all $\theta \in \Theta$. Then, $\hat{\theta}$ is the best unbiased estimator for θ.

Proof In order to make use of Lemma 1.3.18 it has to be shown that $\mathrm{Cov}_\theta(\hat{\theta}, \varphi) = \mathrm{E}_\theta(\hat{\theta}\varphi) = 0\,, \theta \in \Theta$ for $\mathrm{E}_\theta \varphi = 0\,, \theta \in \Theta$. It holds that

$$\mathrm{E}_\theta(\hat{\theta}\varphi) = \mathrm{E}_\theta(\mathrm{E}(\hat{\theta}\varphi|\hat{\theta})) \overset{\hat{\theta}\ \sigma(\hat{\theta})\text{-measurable}}{=} \mathrm{E}_\theta(\hat{\theta} \cdot \mathrm{E}_\theta(\varphi|\hat{\theta})) = \mathrm{E}_\theta(\hat{\theta} \cdot g(\hat{\theta})) \overset{?}{=} 0\,,$$

for $g(\hat{\theta}) = 0$ almost surely. Since $\hat{\theta}$ is sufficient, $g(t) = \mathrm{E}_\theta(\varphi \,|\, \hat{\theta} = t)$ is independent of θ.
Consider $\mathrm{E}_\theta g(\hat{\theta})$. In order to show that $g(\hat{\theta}) = 0$ for all $\theta \in \Theta$, it has to be shown that $\mathrm{E}_\theta g(\hat{\theta}) = 0\, \theta \in \Theta$ since $\hat{\theta}$ is already assumed to be complete.

$$\mathrm{E}_\theta g(\hat{\theta}) = \mathrm{E}_\theta(\mathrm{E}_\theta(\varphi|\hat{\theta})) = \mathrm{E}_\theta \varphi = 0$$

is assumed to hold, thus $\mathrm{E}_\theta(\varphi\hat{\theta}) = 0$ and $\hat{\theta}$ is uncorrelated to $\varphi : \mathrm{E}_\theta \varphi = 0\,, \quad \theta \in \Theta$, which implies that $\hat{\theta}$ is the best unbiased estimator by Lemma 1.3.18 . □

Theorem 1.3.20. Let $\hat{\theta}$ be an unbiased estimator for θ and $\mathrm{E}_\theta \hat{\theta}^2 < \infty, \theta \in \Theta$. Let $\tilde{\theta}$ be a complete and sufficient estimator for θ. Then, the estimator $\theta^* = \mathrm{E}(\hat{\theta} \,|\, \tilde{\theta})$ is the best unbiased estimator for θ.

Proof

1. It has to be shown that $\mathrm{E}_\theta \theta^{*2} < \infty \, \forall \theta \in \Theta$. It holds that

$$\mathrm{E}_\theta \left(\theta^{*2} \right) = \mathrm{E}_\theta \left(\mathrm{E} \left(\hat{\theta} \,|\, \tilde{\theta} \right) \right)^2 \leq \mathrm{E}_\theta \left(\mathrm{E} \left(\hat{\theta}^2 \,|\, \tilde{\theta} \right) \right) = \mathrm{E}_\theta \hat{\theta}^2 < \infty \,,$$

by Jensen's inequality for the conditional expectation, which states

$$f(\mathrm{E}(X \,|\, \mathcal{B})) \overset{\text{f.s.}}{\leq} \mathrm{E}(f(X) \,|\, \mathcal{B})$$

for any random variable X, σ-algebra $\mathcal{B}$ and convex function f.

2. It has to be shown that θ^* is unbiased:
$\mathrm{E}_\theta \theta^* = \mathrm{E}_\theta(\mathrm{E}(\hat{\theta} \,|\, \tilde{\theta})) = \mathrm{E}_\theta \hat{\theta} = \theta$, $\theta \in \Theta$, since $\hat{\theta}$ is unbiased.

3. By Lemma 1.3.18, it is sufficient to show that
$\mathrm{E}_\theta(\theta^* \varphi) = 0$ for $\theta \in \Theta$, if $\mathrm{E}_\theta \varphi = 0$, $\theta \in \Theta$.

$$\mathrm{E}_\theta(\theta^* \varphi) = \mathrm{E}_\theta(\underbrace{\mathrm{E}(\hat{\theta} \,|\, \tilde{\theta})}_{= g(\tilde{\theta}),\, \tilde{\theta} \text{ suf.}} \varphi) = \mathrm{E}_\theta(g(\tilde{\theta})\varphi) = \mathrm{E}_\theta(\mathrm{E}(g(\tilde{\theta})\varphi \,|\, \tilde{\theta}))$$

$$\overset{g(\tilde{\theta}) \text{ } \tilde{\theta}\text{-measurable}}{=} \mathrm{E}_\theta(g(\tilde{\theta}) \cdot \underbrace{\mathrm{E}(\varphi \,|\, \tilde{\theta})}_{= g_1(\tilde{\theta})}) = 0 \,,$$

if $g_1(\tilde{\theta}) \overset{\text{a.s.}}{=} 0$, $\theta \in \Theta$. It needs to be shown that $\mathrm{E}_\theta g_1(\tilde{\theta}) = 0$. Now, $\mathrm{E}_\theta g_1(\tilde{\theta}) = \mathrm{E}_\theta(\mathrm{E}(\varphi \,|\, \tilde{\theta})) = \mathrm{E}_\theta \varphi = 0$ and the completeness of $\tilde{\theta}$ imply (similarly to the proof of Theorem 1.3.19) that $g_1(\tilde{\theta}) = 0$ almost surely.

□

Lemma 1.3.21 (Blackwell-Rao inequality)**.** Let $\hat{\theta}$ be an unbiased estimator for θ and $\mathrm{E}_\theta \hat{\theta}^2 < \infty \,, \theta \in \Theta$. Furthermore, let $\tilde{\theta}$ be a sufficient estimator for θ. Then, the unbiased estimator $\theta^* := \mathrm{E}_\theta(\hat{\theta} \,|\, \tilde{\theta})$ attains a variance which is smaller or equal to $\mathrm{Var}_\theta \hat{\theta}$.

Proof See proof of Theorem 1.3.20. Here, θ^* is unbiased, due to 2) in Theorem 1.3.20 and $\mathrm{Var}_\theta \theta^* = \mathrm{E}_\theta \theta^{*2} - \theta^2 \leq \mathrm{E}_\theta \hat{\theta}^2 - \theta^2 = \mathrm{Var}_\theta \hat{\theta}$ due to 1) in Theorem 1.3.20. □

Remark 1.3.22. The sufficiency of $\tilde{\theta}$ is not mentioned explicitly in the proof of Lemma 1.3.21. It is still necessary in order to assure that $\theta^* = \mathrm{E}_\theta(\hat{\theta} \,|\, \tilde{\theta}) = g(\tilde{\theta})$ does not depend on θ.

Corollary 1.3.23. If $\hat{\theta}$ is a complete and sufficient estimator for θ and there exists a function $g : \mathbb{R} \to \mathbb{R}$ such that $\mathrm{E}_\theta g(\hat{\theta}) = \theta$, $\forall \theta \in \Theta$, then $g(\hat{\theta})$ is the best unbiased estimator for θ.

Proof $g(\hat{\theta}) = \mathrm{E}(g(\hat{\theta}) \,|\, \hat{\theta})$, which is the best unbiased estimator by Theorem 1.3.20. □

1.3.5 δ–Method

Let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ be an estimator of a parameter $\theta \in \Theta \subseteq \mathbb{R}$ ($m = 1$), where $(X_1, \ldots, X_n)$ is a random sample of i.i.d. random variables X_j for $j = 1, \ldots, n$, $X_j \sim F_\theta$. Suppose that $\hat{\theta}_n$ is asymptotically normal distributed, i.e. there exists a sequence of functions $\{\sigma_n(\theta)\}_{n\in\mathbb{N}}$ with $\sigma_n(\theta) > 0$ and $\sigma_n(\theta) \underset{n\to\infty}{\longrightarrow} 0$, $\forall n \in \mathbb{N}$, $\theta \in \Theta$ such that

$$\frac{\hat{\theta}_n - \theta}{\sigma_n(\theta)} \xrightarrow{d} Y \sim N(0,1).$$

Let $g : \Theta \to \mathbb{R}$ be a Borel measurable function. What can be said about the asymptotic normality of $g(\hat{\theta}_n)$? In other words, this section aims to identify the sufficient conditions under which

$$\frac{g(\hat{\theta}_n) - g(\theta)}{\tilde{\sigma}_n(\theta)} \xrightarrow{d} Y \tag{1.12}$$

for another sequence $\{\tilde{\sigma}_n(\theta)\}_{n\in\mathbb{N}}$ with $\tilde{\sigma}_n(\theta) > 0$, $n \in \mathbb{N}$, and $\tilde{\sigma}_n(\theta) \underset{n\to\infty}{\longrightarrow} 0$, $\theta \in \Theta$. For linear $g(\theta) = a \cdot \theta + b$, $a, b \in \mathbb{R}$, relation (1.12) obviously holds. When does (1.12) hold for more general functions g? There may be multiple reasons for the consideration of functions $g(\hat{\theta}_n)$. One of those lies in the *variance stabilization* which will be discussed at the end of this section. There a function g is considered, such that $\tilde{\sigma}_n(\theta)$ does not depend on θ. This makes the construction of asymptotic confidence regions for θ much easier (cf. Section 2.2.3 for examples).
The following method of proving the asymptotic normality for $g(\hat{\theta}_n)$ makes use of the Taylor series decomposition of a sufficiently smooth function g. It has been known since the early 19$^{\text{th}}$ century and first asymptotically described by J. Doob [11]. The name "δ method" alludes to the differential or increment $dg(x) = g(x + \delta x) - g(x)$ which lies in the core of the method. Due to its very general nature, the results can be formulated for any asymptotically normal sequence of random variables $\{Y_n\}_{n\in\mathbb{N}}$, i.e. sequences with $\frac{Y_n - \mu}{\sigma_n} \xrightarrow{d} Y \sim N(0,1)$ for some $\mu \in \mathbb{R}$ and a normalizing sequence $\{\sigma_n\}_{n\in\mathbb{N}}$ with $\sigma_n > 0$ for all $n \in \mathbb{N}$ and $\sigma_n \underset{n\to\infty}{\longrightarrow} 0$.

Theorem 1.3.24. Suppose that

$$\frac{Y_n - \mu}{\sigma_n} \xrightarrow{d} Y \sim N(0,1) \tag{1.13}$$

for a sequence $\{Y_n\}_{n\in\mathbb{N}}$, μ and $\{\sigma_n\}_{n\in\mathbb{N}}$ as above. Let $g : \mathbb{R} \to \mathbb{R}$ be differentiable at $x = \mu$ with $g'(\mu) \neq 0$. Then,

$$\frac{g(Y_n) - g(\mu)}{g'(\mu)\sigma_n} \underset{n\to\infty}{\xrightarrow{d}} Y.$$

Proof First, show that (1.13) implies with $\sigma_n \xrightarrow[n\to\infty]{} 0$ that

$$Y_n \xrightarrow[n\to\infty]{P} \mu. \tag{1.14}$$

Indeed, Slutsky's Theorem (cf. [32, Theorem 3.4.3.]) yields

$$\begin{aligned}\frac{Y_n-\mu}{\sigma_n} \xrightarrow[n\to\infty]{d} Y,\ \sigma_n \xrightarrow{a.s.} 0 \implies Y_n-\mu = \sigma_n \cdot \frac{Y-\mu}{\sigma_n} \xrightarrow[n\to\infty]{d} 0\cdot Y = 0\\ \implies Y_n-\mu \xrightarrow[n\to\infty]{P} 0\end{aligned}$$

by [32, Theorem 3.3.4.]. Introduce the function

$$h(x) = \begin{cases}\frac{g(x)-g(\mu)}{x-\mu} - g'(\mu), & x\neq\mu,\\ 0, & x=\mu.\end{cases}$$

Since $g(x)$ is differentiable at $x=\mu$, $h(x)$ is continuous at $x=\mu$. The Continuous Mapping Theorem (cf [32, Theorem 3.4.4.] implies

$$h(Y_n) \xrightarrow[n\to\infty]{P} h(\mu) = 0,$$

i.e.

$$\frac{g(Y_n)-g(\mu)}{Y_n-\mu} - g'(\mu) \xrightarrow[n\to\infty]{P} 0.$$

Multiplying both sides by $\frac{Y_n-\mu}{\sigma_n}$ and using (1.13) in combination with Slutsky's Theorem implies that

$$\frac{h(Y_n)(Y_n-\mu)}{\sigma_n} = \frac{g(Y_n)-g(\mu)}{\sigma_n} - g'(\mu)\underbrace{\frac{Y_n-\mu}{\sigma_n}}_{\xrightarrow{d} Y\sim N(0,1)} \xrightarrow[n\to\infty]{P} 0.$$

Hence $\frac{g(Y_n)-g(\mu)}{\sigma_n} \xrightarrow[n\to\infty]{d} g'(\mu)\cdot Y$ as well and dividing by $g'(\mu)$ yields the desired result. □

Remark 1.3.25. If $g\in\mathcal{C}^1(B_\delta(\mu))$ for some $\delta>0$, where

$$B_\delta(\mu) = \{x\in\mathbb{R} : |x-\mu|\leq\delta\},$$

the proof above can be simplified by using the Mean Value Theorem

$$g(Y_n) = g(\mu) + g'(\xi)(Y_n-\mu),$$

where ξ lies between μ and Y_n. In addition, the Continuous Mapping Theorem together with (1.14) and the assumption $g\in\mathcal{C}^1(B_\delta(\mu))$ yield $g'(Y_n) \xrightarrow[n\to\infty]{P} g'(\mu)$. By Slutsky's Theorem a modified version of (1.13) holds:

$$\frac{g(Y_n)-g(\mu)}{g'(Y_n)\sigma_n} \xrightarrow[n\to\infty]{d} Y.$$

Example 1.3.26. By [33, Theorem 7.4.4, 2)], the sample variance

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

is an asymptotically normally distributed estimator of $\sigma^2 = \mathrm{Var}X_j > 0$:

$$\sqrt{n}\frac{S_n^2 - \sigma^2}{\sqrt{\mu_4' - \sigma_n^4}} \xrightarrow[n\to\infty]{d} Y \sim N(0,1),$$

where $\mu_4' = \mathrm{E}(X_j - \mathrm{E}X_j)^4$. One can show that the empirical standard deviation S_n is an asymptotically normal estimate of σ. Here,

$$\begin{aligned} g(x) &= \sqrt{x}, \\ g'(x) &= \frac{1}{2\sqrt{x}}, \\ \theta &= \sigma^2 > 0, \\ g'(\sigma^2) &= \frac{1}{2\sigma} > 0 \text{ and} \\ \sigma_n &= \sqrt{\frac{\mu_n' - \sigma^4}{n}} \end{aligned}$$

Following Theorem 1.3.24 it holds that

$$2\sigma\sqrt{n}\frac{S_n - \sigma}{\sqrt{\mu_4' - \sigma^4}} \xrightarrow[n\to\infty]{d} Y.$$

What happens if $g'(\mu) = 0$ in Theorem 1.3.24? In this case, a higher order Taylor approximation should be used, as the following result shows.

Theorem 1.3.27. Assume that a sequence of random variables $\{Y_n\}_{n\in\mathbb{N}}$ satisfies the conditions of Theorem 1.3.24. Let $g : \mathbb{R} \to \mathbb{R}$ be $m \geq 2$ times differentiable at μ with $g^{(j)}(\mu) = 0$, $j < m$ and $g^{(m)}(\mu) \neq 0$. Then,

$$m! \cdot \frac{g(Y_n) - g(\mu)}{g^{(m)}(\mu)\sigma_n^m} \xrightarrow[n\to\infty]{d} Y^m,$$

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

Proof Use the function

$$h(x) = \begin{cases} m!\dfrac{g(x) - g(\mu)}{(x-\mu)^m} - g^{(m)}(\mu), & x \neq \mu, \\ 0, & x = \mu. \end{cases}$$

in the proof of Theorem 1.3.24. □

Example 1.3.28. Suppose that $\{Y_n\}_{n\in\mathbb{N}}$ is a sequence of random variables with

$$\frac{Y_n}{\sigma_n} \xrightarrow[n\to\infty]{d} Y \sim N(0,1)$$

for $\sigma_n \xrightarrow[n\to\infty]{} 0$ with $\sigma_n > 0$ for all $n \in \mathbb{N}$. Apply Theorem 1.3.27 to $g(x) = \log^2(1+x)$ and $m = 2$, $\mu = 0$:

$$\begin{aligned} g'(x) &= \frac{2\log(1+x)}{1+x},\ g'(0) = 0, \\ g''(x) &= \frac{\frac{2}{1+x}(1+x) - 2\log(1+x)}{(1+x)^2} = 2\frac{1-\log(1+x)}{(1+x)^2} = 2\frac{\log\left(\frac{e}{1+x}\right)}{(1+x)^2}, \\ g''(0) &= 2 > 0. \end{aligned}$$

Then,

$$2\cdot\frac{\log^2(1+Y_n)}{2\sigma_n^2} = \frac{1}{\sigma_n^2}\log^2(1+Y_n) \xrightarrow{d} Y^2 \sim \chi_1^2.$$

As already mentioned above, it might be advantageous for some applications in the asymptotic theory of confidence intervals and statistical tests to eliminate the dependence of the asymptotic variance $\sigma_n(\theta)$ from the parameter θ. In other words, find a transformation g of the estimate $\hat{\theta}$ such that $\tilde{\sigma}_n(\theta)$ in (1.12) does not depend on θ anymore. This device is known as *variance stabilization*. By Theorem 1.3.24 a function $g : \mathbb{R} \to \mathbb{R}$ with $g'(\theta) \neq 0$ such that $g'(\theta)\cdot\sigma_n(\theta)$ depends only on $n \in \mathbb{N}$ has to be found. Let $\sigma_n(\theta) = \sigma(\theta)\cdot v_n$, with $v_n \to 0$. Then it suffices to solve the ordinary differential equation

$$g'(\theta) = \frac{c}{\sigma(\theta)},\ c \text{ constant.} \tag{1.15}$$

If $g'(\theta) = 0$, Theorem 1.3.27 can be applied here accordingly.

Example 1.3.29.

1. Consider a random sample $(X_1,\ldots,X_n)$ of centered i.i.d. random variables with $\mu_4 = \mathrm{E}X_j^4 < \infty$ and $\sigma^2 = \mathrm{Var}X_j > 0$. Since $\mu = \mathrm{E}X_j = 0$, consider the estimate $\tilde{S}_n^2 = \frac{1}{n}\sum_{j=1}^n X_j^2$ of σ^2. Assume that μ_4 is known. By [33, Theorem 7.4.4, 2)], it holds that

$$\sqrt{n}\frac{\tilde{S}_n^2 - \sigma^2}{\sqrt{\mu_4 - \sigma^4}} \xrightarrow[n\to\infty]{d} Y \sim N(0,1).$$

   By stabilizing the asymptotic variance in this case,

$$g'(\sigma^2) = \frac{1}{\sqrt{\mu_4 - (\sigma^2)^2}}$$

has to be solved. The solution is given by $g(x) = \arcsin\left(\frac{x}{\sqrt{\mu_4}}\right)$ and thus

$$\sqrt{n}\left(\arcsin\left(\frac{\tilde{S}_n^2}{\sqrt{\mu_4}}\right) - \arcsin\left(\frac{\sigma^2}{\sqrt{\mu_4}}\right)\right) \xrightarrow[n\to\infty]{d} Y \sim N(0,1).$$

2. Let $(X_1, \ldots, X_n)$ be a sample of i.i.d. random variables with $X_j \sim$Bernoulli(p), for $p \in (0,1)$. By [33, Theorem 7.3.2, a)], it holds that

$$\frac{\sqrt{n}(\hat{p}_n - p)}{\sqrt{p(1-p)}} \xrightarrow[n\to\infty]{d} Y \sim N(0,1),$$

where $\hat{p}_n = \bar{X}_n$. Similarly to 1), the variance stabilising transform g is given by $g(p) = 2\arcsin(\sqrt{p})$, since $g'(p) = \frac{1}{\sqrt{p(1-p)}}$. Applying Theorem 1.3.24 yields

$$2\sqrt{n}(\arcsin(\sqrt{\hat{p}_n}) - \arcsin(\sqrt{p})) \xrightarrow[n\to\infty]{d} Y \sim N(0,1) \tag{1.16}$$

3. Let $(X_1, \ldots, X_n)$ be a sample of i.i.d. Poisson(λ) distributed random variables with $\lambda > 0$. For $\theta = \lambda$, $\hat{\lambda}_n = \bar{X}_n$ it holds that

$$\sqrt{n}\frac{\hat{\lambda}_n - \lambda}{\sqrt{\lambda}} \xrightarrow[n\to\infty]{d} Y \sim N(0,1).$$

The variance stabilizing transform g is then given by $g(x) = 2\sqrt{x}$ because of $g'(\lambda) = \dfrac{1}{\sqrt{\lambda}}$. In summary, we get

$$2\sqrt{n}\left(\sqrt{\hat{\lambda}_n} - \sqrt{\lambda}\right) \xrightarrow[n\to\infty]{d} Y \sim N(0,1). \tag{1.17}$$

Remark 1.3.30. The δ–method can be extended to the asymptotic normality of (functions g of) d-dimensional random vectors $\{Y_n\}_{n\in\mathbb{N}}$, for $d \geq 2$. See [30, Section 3.3] for more details. It can be used to prove the asymptotic normality of the empirical Bravais-Pearson correlation coefficient

$$\rho_{XZ} = \frac{\sum_{j=1}^n X_j Z_j - n\bar{X}_n\bar{Z}_n}{S_{n,X}S_{n,Z}}$$

of i.i.d. random samples $(X_1, \ldots, X_n)$ and $(Z_1, \ldots, Z_n)$, where $S_{n,X}^2$ and $S_{n,Z}^2$ are their sample variances. Similarly, the *empirical coefficient of variation* $\frac{S_n}{\bar{X}_n}$ of one i.i.d. sample $(X_1, \ldots, X_n)$ can be shown to be asymptotically normal with

$$\sqrt{n}\left(\frac{S_n}{\bar{X}_n} - \frac{\sigma}{\mu}\right) \xrightarrow[n\to\infty]{d} Y \sim N\left(0, \frac{\sigma_*^2\mu^2}{4\sigma^2}\right),$$

where $\bar{X}_n$ is the sample mean and S_n^2 the sample variance, cf. [1]. Here

$$\sigma_*^2 = \frac{\mu_4}{\mu^4} - \left(\frac{\mu_2}{\mu^2}\right)^2 + 4\left(\frac{\mu_2}{\mu^2}\right)^3 - \frac{4\mu_2\mu_3}{\mu^5}$$

is a function of the first four moments of X_j which are assumed to be finite.

Chapter 2

Confidence Intervals

2.1 Introduction

This chapter will focus on the formal definition of confidence intervals. We will gain a deeper understanding of how they work and what they are used for. In particular, this chapter will cover *one-sample problems* and *two-sample problems*.
Recall the assumptions of parametric models: Let $(X_1, \ldots, X_n)$ be a random sample with $X_i \sim F_\theta$, $i = 1, \ldots, n$, and $F_\theta \in \{F_\theta : \theta \in \Theta\}$, where $\{F_\theta : \theta \in \Theta\}$ is some parametric family with $\Theta \subset \mathbb{R}$.
Each point estimator of θ provides a value for the parameter vector. It would also be beneficial to have information about the accuracy of the estimator, i.e., a neighborhood which contains θ with a certain probability $1-\alpha$. Here α denotes a *significance level*, which indicates the probability of θ being outside the predetermined neighborhood. Typical values are $\alpha = 0.01; 0.05; 0.1$. For $m = 1$ the neighborhood is an interval called *confidence interval* and the probability $1-\alpha$ is called *coverage probability* or *confidence level*. It is always desired to achieve a high confidence level, e.g., $1 - \alpha = 0,99; 0,95; 0,9$ are typical values.

Definition 2.1.1. Let $1 - \alpha$ be a confidence probability and
$\underline{\theta} : \mathbb{R}^n \to \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$, $\overline{\theta} : \mathbb{R}^n \to \overline{\mathbb{R}}$ be two measurable sample functions with the property

$$\underline{\theta}(x_1, \ldots x_n) \leq \overline{\theta}(x_1, \ldots, x_n) \quad \forall (x_1, \ldots x_n) \in \mathbb{R}^n.$$

If

1. $P_\theta\left(\theta \in \left[\underline{\theta}(X_1, \ldots, X_n), \overline{\theta}(X_1, \ldots X_n)\right]\right) \geq 1 - \alpha, \quad \theta \in \Theta,$
2. $\inf\limits_{\theta \in \Theta} P_\theta\left(\theta \in \left[\underline{\theta}(X_1, \ldots, X_n), \overline{\theta}(X_1, \ldots, X_n)\right]\right) = 1 - \alpha$ and
3. $\lim\limits_{n \to \infty} P_\theta\left(\theta \in \left[\underline{\theta}(X_1, \ldots, X_n), \overline{\theta}(X_1, \ldots, X_n)\right]\right) = 1 - \alpha, \quad \theta \in \Theta,$

then $I = \left[\underline{\theta}(X_1, \ldots, X_n), \overline{\theta}(X_1, \ldots X_n)\right]$ is called

1. *confidence interval*,
2. *minimal confidence interval*,
3. *asymptotic confidence interval*,

with confidence level $1-\alpha$. Here, $l_\theta(X_1, \ldots X_n) = \overline{\theta}(X_1, \ldots X_n) - \underline{\theta}(X_1, \ldots X_n)$ denotes the *length* of the confidence interval. It is desired to construct an interval, which has a relatively short length but a high confidence level, i.e., $1 - \alpha = .99$.

In Example 1.2.14, the construction of a confidence interval was introduced. This methodology can be generalized as follows.

1. Find a statistic $T(X_1, \ldots, X_n, \theta)$ which
   - depends on θ and
   - underlies a known (test) distribution F (possibly asymptotic as $n \to \infty$).
2. Determine the quantiles $F^{-1}(\alpha_1)$ and $F^{-1}(1-\alpha_2)$ of the distribution F for the niveaus α_1 and $1-\alpha_2$, such that $\alpha_1 + \alpha_2 = \alpha$.
3. Solve (if possible) the inequality $F^{-1}(\alpha_1) \leq T(X_1, \ldots, X_n, \theta) \leq F^{-1}(1-\alpha_2)$ w.r.t. θ. The respective solution (if the statistic T in θ is monotonically increasing) $I = [T^{-1}(F^{-1}(\alpha_1)), T^{-1}(F^{-1}(1-\alpha_2))]$ is a confidence interval for θ with confidence level $1-\alpha$, because

$$\begin{aligned} P_\theta(\theta \in I) &= P_\theta\left(T_\theta^{-1}(F^{-1}(\alpha_1)) \leq \theta \leq T^{-1}(F^{-1}(1-\alpha_2))\right) \\ &= P_\theta\left(F^{-1}(\alpha_1) \leq T_\theta(X_1, \ldots, X_n, \theta) \leq F^{-1}(1-\alpha_2)\right) \\ &= F(F^{-1}(1-\alpha_2)) - F(F^{-1}(\alpha_1)) \\ &= 1 - \alpha_2 - \alpha_1 \\ &= 1 - \alpha \text{ for all } \theta \in \Theta. \end{aligned}$$

For asymptotic confidence intervals the notation $\lim\limits_{n\to\infty}$ is introduced: $\lim\limits_{n\to\infty} P_\theta(\theta \in I) = \ldots = 1-\alpha$. Here T_θ^{-1} denotes the inverse of $T(X_1, \ldots, X_n, \theta)$ w.r.t. θ. A corresponding picture can be found in Figure 2.1.

Definition 2.1.2.

1. If $\alpha_1 = \alpha_2 = \alpha/2$, then the confidence interval given by

$$I = \left[T^{-1}\left(F^{-1}\left(\frac{\alpha}{2}\right)\right), T^{-1}\left(F^{-1}\left(1-\frac{\alpha}{2}\right)\right)\right]$$

   is called *symmetric*.

2. If $\alpha_1 = 0$, i.e., $\underline{\theta}(X_1, \ldots, X_n) = -\infty$, then the confidence interval $\left(-\infty, \overline{\theta}(X_1, \ldots, X_n)\right]$ is called *one sided*.

   Analogously, if $\alpha_2 = 0$ i.e., $\overline{\theta}(X_1, \ldots, X_n) = +\infty$, then the confidence interval is given by $[\underline{\theta}(X_1, \ldots, X_n), +\infty)$.

From now on, mostly symmetric confidence intervals will be constructed. More general, non symmetric confidence intervals can easily be constructed similarly.

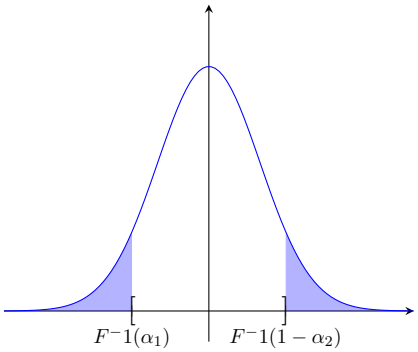


Figure 2.1: asymptotic confidence interval

Remark 2.1.3. One can observe, that the process of constructing a confidence interval is similar to constructing a test. In Definition 2.1.2, $T(X_1, \ldots, X_n)$ is called *test statistic*. Generally, a statistical test for every confidence interval can be constructed, but not the other way around.

2.2 One-sample problems

This section provides examples of confidence intervals for parameters of known distributions using the algorithm above.

2.2.1 Normal distribution

Let $X_1, \ldots, X_n$ be i.i.d. random sample with $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$.

Confidence interval for the expectation μ

- **Known variance σ^2:** Under the assumption that σ^2 is known, [33, Theorem 7.3.2] implies that an exact confidence interval for μ with confidence level $1 - \alpha$ can be constructed. Since $\overline{X}_n \sim N\left(\mu, \sigma^2/n\right)$,

$$T(X_1, \ldots, X_n, \mu) = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

Let z_{α_1} and $z_{1-\alpha_2}$ be quantiles of the $\mathcal{N}(0,1)$ distribution, such that $\alpha_1 + \alpha_2 = \alpha$. Then, $1-\alpha$ is the given confidence level and

$$\begin{aligned} 1-\alpha &= P\left(z_{\alpha_1} \leq T(X_1,\ldots,X_n,\mu) \leq z_{1-\alpha_2}\right) \\ &= P\left(z_{\alpha_1} \leq \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \leq z_{1-\alpha_2}\right) \\ &\overset{(-z_{\alpha_1}=z_{1-\alpha_1})}{=} P\left(\overline{X}_n - \frac{z_{1-\alpha_2}\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + \frac{z_{1-\alpha_1}\sigma}{\sqrt{n}}\right). \end{aligned}$$

Hence, $\left[\underline{\theta}(X_1,\ldots,X_n), \overline{\theta}(X_1,\ldots,X_n)\right]$ with

$$\underline{\theta}(X_1,\ldots,X_n) = \overline{X}_n - z_{1-\alpha_2}\frac{\sigma}{\sqrt{n}}$$

and

$$\overline{\theta}(X_1,\ldots,X_n) = \overline{X}_n + z_{1-\alpha_1}\frac{\sigma}{\sqrt{n}},$$

is a confidence interval for μ with confidence level $1-\alpha$. Its length is $l_\mu(X_1,\ldots,X_n) = \frac{\sigma}{\sqrt{n}}\left(z_{1-\alpha_2} + z_{1-\alpha_1}\right)$.
If $n \longrightarrow \infty$, then $l_\mu(X_1,\ldots,X_n) \longrightarrow 0$ which means, that if the amount of available information increases, i.e., $n \longrightarrow \infty$, the precision of the estimation also increases.
If the underlying distribution is symmetric i.e., $\alpha_1 = \alpha_2 = \alpha/2$, then

$$\begin{aligned} \underline{\theta}(X_1,\ldots,X_n) &= \overline{X}_n - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \\ \overline{\theta}(X_1,\ldots,X_n) &= \overline{X}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \end{aligned}$$

and

$$l_\mu(X_1,\ldots,X_n) = \frac{2\sigma}{\sqrt{n}} z_{1-\alpha/2}.$$

If the length $\varepsilon > 0$ is predetermined, the number of necessary observations n for achieving the desired precision can be calculated by solving

$$\frac{2\sigma}{\sqrt{n}} z_{1-\alpha/2} \leq \varepsilon \tag{2.1}$$

for n, which yields

$$n \geq \left(\frac{2\sigma z_{1-\alpha/2}}{\varepsilon}\right)^2.$$

For $\alpha_1 = 0$ or $\alpha_2 = 0$ one sided intervals like $\left(-\infty, \overline{X}_n + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right]$, resp. $\left[\overline{X}_n - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}, +\infty\right)$ can be constructed.

- **Unknown variance σ^2:** Using [33, Theorem 7.4.10.], the following confidence interval with confidence level $1-\alpha \in (0,1)$ for the expectation μ of a normally distributed random sample $(X_1, \dots, X_n)$ with unknown variance σ^2 can be constructed.

$$P\left(\mu \in \left[\bar{X}_n - \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}} S_n \,,\, \bar{X}_n + \frac{t_{n-1,1-\alpha/2}}{\sqrt{n}} S_n\right]\right) = 1-\alpha,$$

since

$$P\Bigg(\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \in \Big[\underbrace{\quad t_{n-1,\alpha/2} \quad}_{=-t_{n-1,1-\alpha/2} \text{ bc. of the sym. of } t \text{ dist.}}, t_{n-1,1-\alpha/2}\Big]\Bigg) = \tag{2.2}$$

$$\begin{aligned} &= F_{t_{n-1}}(t_{n-1,1-\alpha/2}) - F_{t_{n-1}}(t_{n-1,\alpha/2}) \\ &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha\,, \end{aligned}$$

where $t_{n-1,\alpha}$ is the α quantile of the t_{n-1} distribution. By solving for μ in (2.2) the remaining part can be shown.

Note that the length $l_\mu(X_1, \dots X_n) = \frac{2S_n}{\sqrt{n}} t_{n-1,1-\alpha/2}$ of the confidence interval is a random variable. Thus, the expected length

$$\mathbb{E}\, l_\mu(X_1, \dots X_n) = \frac{2}{\sqrt{n}} \mathbb{E}\, S_n t_{n-1,1-\alpha/2}$$

yields an answer to the question about the required number of observations n for a predetermined precision $\varepsilon > 0$ (cf. Equation (2.1)).

Confidence interval for the variance σ^2

- **Known expectation** μ: Consider the estimator $\tilde{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2$ for σ^2. [33, Theorem 7.4.8 1.] implies $\frac{n\tilde{S}_n^2}{\sigma^2} \sim \chi_n^2$.

  Define $T(X_1, \dots, X_n, \sigma^2) := \frac{n\tilde{S}_n^2}{\sigma^2}$, then

$$P\left(\chi^2_{n,\alpha_2} \le \frac{n\tilde{S}_n{}^2}{\sigma^2} \le \chi^2_{n,1-\alpha_1}\right) = P\left(\frac{n\tilde{S}_n^2}{\chi^2_{n,1-\alpha_1}} \le \sigma^2 \le \frac{n\tilde{S}_n^2}{\chi^2_{n,\alpha_2}}\right) = 1-\alpha.$$

  Thus, $\left[\frac{n\tilde{S}_n^2}{\chi^2_{n,1-\alpha_1}}, \frac{n\tilde{S}_n^2}{\chi^2_{n,\alpha_2}}\right]$ is a confidence interval for σ^2 with level $1-\alpha$, where $\alpha = \alpha_1 + \alpha_2$. The expected length is given by

$$\mathbb{E}\, l_{\sigma^2} = n\sigma^2 \left(\frac{1}{\chi^2_{n,\alpha_2}} - \frac{1}{\chi^2_{n,1-\alpha_1}}\right).$$

- **Unknown expectation μ**: Similarly to the construction above, [33, Theorem 7.4.8. 1.] implies that $\left[\frac{(n-1)S_n^2}{\chi^2_{n-1,1-\alpha_1}}, \frac{(n-1)S_n^2}{\chi^2_{n-1,\alpha_2}}\right]$ is a confidence interval for σ^2 with confidence level $1-\alpha$, where $\alpha = \alpha_1 + \alpha_2$. Note that $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$ for the sample variance $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$. The expected length is

$$\mathbb{E}\, l_{\sigma^2} = (n-1)\sigma^2\left(\frac{1}{\chi^2_{n-1,\alpha_2}} - \frac{1}{\chi^2_{n-1,1-\alpha_1}}\right).$$

2.2.2 Confidence intervals and stochastic inequalities

An alternative approach for obtaining confidence intervals is applying stochastic inequalities. Let, for example, $(X_1, \ldots, X_n)$ be a random sample of i.i.d. random variables with $\mathbb{E}\, X_i = \mu$, $\operatorname{Var} X_i = \sigma^2 \in (0, \infty)$, then the Tschebyschew inequality can be used to construct a simple confidence interval for μ:

$$P\left(|\overline{X}_n - \mu| > \varepsilon\right) \leq \frac{\operatorname{Var}\overline{X}_n}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} = \alpha.$$

Then, for $\varepsilon = \frac{\sigma}{\sqrt{n\alpha}}$

$$\begin{aligned} 1-\alpha &\leq P\left(|\overline{X}_n - \mu| \leq \varepsilon\right) \\ &= P\left(-\frac{\sigma}{\sqrt{n\alpha}} \leq -\overline{X}_n + \mu \leq \frac{\sigma}{\sqrt{n\alpha}}\right) \\ &= P\left(\overline{X}_n - \frac{\sigma}{\sqrt{n\alpha}} \leq \mu \leq \overline{X}_n + \frac{\sigma}{\sqrt{n\alpha}}\right) \end{aligned}$$

holds. The confidence interval $\left[\overline{X}_n - \frac{\sigma}{\sqrt{n\alpha}}, \overline{X}_n + \frac{\sigma}{\sqrt{n\alpha}}\right]$ for μ with known variance σ^2 is independent of the underlying distribution of X_i since no assumptions have been made.
More precise confidence intervals can be constructed by using the *Hoeffding inequality*:

Theorem 2.2.1 (Hoeffding inequality)**.** Let $Y_1, \ldots, Y_n$ be independent random variables with $\mathbb{E}\, Y_i = 0, a_i \leq Y_i \leq b_i$ a.s., $i = 1, \ldots, n$. For all $\varepsilon > 0$,

$$P\left(\sum_{i=1}^{n} Y_i \geq \varepsilon\right) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

holds.
(without proof)

Assume that $X_1, \ldots, X_n$ are i.i.d. with $X_i \sim \text{Bernoulli}(p)$, $p \in (0,1)$. Next, we will show how to construct a confidence interval for p.

Corollary 2.2.2. Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli (p) random variables. Then

$$P\left(|\overline{X}_n - p| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2}, \quad \varepsilon > 0.$$

Proof

$$\overline{X}_n - p = \frac{1}{n}\sum_{i=1}^{n} \underbrace{(X_i - p)}_{Y_i},\ Y_i \in [-p, 1-p],$$

holds, which means that $a_i = -p$, $b_i = 1-p$, $b_i - a_i = 1$, $i = 1, \ldots, n$, $\mathbb{E}\, Y_i = p - p = 0$. Then,

$$\begin{aligned} P_p\left(|\overline{X}_n - p| > \varepsilon\right) &= P_p\left(\left|\sum_{i=1}^{n} Y_i\right| \geq \varepsilon n\right) \\ &= P_p\left(\sum_{i=1}^{n} Y_i \geq \varepsilon n\right) + P_p\left(\sum_{i=1}^{n}(-Y_i) \geq \varepsilon n\right) \\ &\overset{\text{(Theorem 2.2.1)}}{\leq} 2e^{-\frac{2\varepsilon^2 n^2}{n}} = 2e^{-2\varepsilon^2 n}, \end{aligned}$$

where Theorem 2.2.1 is applied to $\{Y_i\}$ as well as $\{-Y_i\}$. □

Remark 2.2.3. Let $\alpha > 0$ and $\varepsilon_n = \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}$. Applying Corollary 2.2.2 with ε_n yields $P_p\left(|\overline{X}_n - p| > \varepsilon_n\right) \leq \alpha$, and thus $P_p\left(|\overline{X}_n - p| \leq \varepsilon_n\right) \geq 1-\alpha$. Hence,

$$\left[\overline{X}_n - \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}},\ \overline{X}_n + \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}\right]$$

is a confidence interval for p with level $1-\alpha$.

2.2.3 Asymptotic confidence intervals

The idea behind asymptotic confidence intervals is relatively simple, as it can be explained by using the example of an asymptotically normal distributed estimator $\hat{\theta}$ for a parameter θ. Let $(X_1, \ldots, X_n)$ be an i.i.d. random sample with $X_i \sim F_\theta$, $\theta \in \Theta \subseteq \mathbb{R}$. Let $\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n)$ be an estimator for θ, that is asymptotically normal distributed. If $\hat{\theta}_n$ is unbiased for every $n \in \mathbb{N}$, then

$$\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \xrightarrow{d} Y \sim \mathcal{N}(0,1),$$

where $\hat{\sigma}_n$ is a consistent estimator for the asymptotic variance of $\hat{\theta}_n$. Furthermore,

$$\begin{aligned}&\lim_{n\to\infty} P_\theta\left(z_{\alpha/2} \leq \frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \leq z_{1-\alpha/2}\right)\\ =&\lim_{n\to\infty} P_\theta\left(\theta \in \left[\hat{\theta}_n - z_{1-\alpha/2}\hat{\sigma}_n,\, \hat{\theta}_n + z_{1-\alpha/2}\hat{\sigma}_n\right]\right) = 1-\alpha.\end{aligned}$$

Thus,

$$\left[\hat{\theta}_n - z_{1-\alpha/2}\hat{\sigma}_n,\, \hat{\theta}_n + z_{1-\alpha/2}\hat{\sigma}_n\right]$$

is an asymptotic confidence interval for θ with level $1-\alpha$.
This approach can be applied to the following two examples:

- **Bernoulli distribution**

  Let $X_i \sim \text{Bernoulli}(p)$, $i = 1, \ldots, n$. Then $\theta = p$ and $\hat{\theta}_n = \hat{p}_n = \overline{X}_n$. Moreover, $\mathbb{E}_p\hat{p}_n = p, \text{Var}_p\hat{p}_n = \frac{p(1-p)}{n}$.
  Let $\hat{\sigma}^2 = \frac{1}{n}\hat{p}_n(1-\hat{p}_n) = \frac{\overline{X}_n}{n}(1-\overline{X}_n)$ be the Plug-In estimator for σ^2. Then the central limit theorem [33, Theorem 5.2.2.] and Slutsky's theorem [32, Theorem 3.4.1] imply

  $$\sqrt{n}\frac{\overline{X}_n - p}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \xrightarrow[n\to\infty]{d} Y \sim \mathcal{N}(0,1).$$

  Thus,

  $$\left[\overline{X}_n - z_{1-\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}},\, \overline{X}_n + z_{1-\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right]$$

  is an asymptotic confidence interval for p with confidence level $1-\alpha$. Since $p \in [0,1]$ is supposed to hold, one considers

  $$\underline{p}(X_1, \ldots, X_n) = \max\left\{0,\, \overline{X}_n - z_{1-\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right\}$$

  and

  $$\overline{p}(X_1, \ldots, X_n) = \min\left\{1,\, \overline{X}_n + z_{1-\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right\}.$$

  Remark 2.2.4.

1. Another confidence interval for the parameter p of the Bernoulli distribution can be obtained by considering an application of the central limit theorem

$$\lim_{n\to\infty} P_p\left(-z_{1-\alpha/2} \leq \sqrt{n}\frac{\overline{X}_n - p}{\sqrt{p(1-p)}} \leq z_{1-\alpha/2}\right) = 1-\alpha$$

and solving the quadratic inequality for p.
Exercise 2.2.5 Solve the inequality!

2. Using the variance stabilization from Example 1.3.29, 2., the relation (1.16) can be used to construct a confidence interval for p with sufficiently large n.

$$P\left(-z_{1-\frac{\alpha}{2}} \leq 2\sqrt{n}(\arcsin\sqrt{\bar{X}_n} - \arcsin\sqrt{p}) \leq z_{1-\frac{\alpha}{2}}\right) \approx 1-\alpha$$

holds, hence

$$\arcsin\sqrt{\bar{X}_n} - \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}} \leq \arcsin\sqrt{p} \leq \arcsin\sqrt{\bar{X}_n} + \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}.$$

With probability $1-\alpha$

$$\sqrt{p} \in \left[\sin\left(\arcsin\sqrt{\bar{X}_n} - \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right), \sin\left(\arcsin\sqrt{\bar{X}_n} + \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right)\right] \Rightarrow$$
$$p \in \left[\sin^2\left(\arcsin\sqrt{\bar{X}_n} - \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right), \sin^2\left(\arcsin\sqrt{\bar{X}_n} + \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right)\right]$$

holds. As $\sqrt{\bar{X}_n} \xrightarrow[n\to\infty]{a.s.} \sqrt{p} \in (0,1)$, $\frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \xrightarrow[n\to\infty]{} 0$ and since

$$\sin\left(\arcsin\sqrt{\bar{X}_n} \pm \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right) > 0$$

for sufficiently large n, the terms

$$\max\left\{0, \arcsin\sqrt{\bar{X}_n} - \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right\}$$

and

$$\min\left\{\frac{\pi}{2}, \arcsin\sqrt{\bar{X}_n} + \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right\}$$

do not need to be considered here.

- **Poisson distribution**:

Let $X_i \sim \text{Poisson}(\lambda)$, $i = 1, \ldots, n$, then $\theta = \lambda$, $\hat{\theta}_n = \hat{\lambda} = \overline{X}_n$. Since $\mathbb{E}_\lambda X_i = \text{Var}_\lambda X_i = \lambda$, the central limit theorem [33, Theorem 5.2.2.] can be applied:

$$\sqrt{n}\frac{\overline{X}_n - \lambda}{\sqrt{\lambda}} \xrightarrow[n\to\infty]{d} Y \sim \mathcal{N}(0,1)$$

Since $\overline{X}_n$ is strongly consistent for λ, Slutsky's theorem [32, Theorem 3.4.1] implies

$$\sqrt{n}\frac{\overline{X}_n - \lambda}{\sqrt{\overline{X}_n}} \xrightarrow[n\to\infty]{d} Y \sim \mathcal{N}(0,1).$$

Thus, a asymptotic confidence interval

$$\left[\overline{X}_n - z_{1-\alpha/2}\sqrt{\frac{\overline{X}_n}{n}},\ \overline{X}_n + z_{1-\alpha/2}\sqrt{\frac{\overline{X}_n}{n}}\right]$$

for the parameter λ with level $1-\alpha$ can be obtained.

Remark 2.2.6.

1. Similarly to Remark 2.2.4, solving the quadratic inequality

   $$\lim_{n\to\infty} P_\lambda\left(\sqrt{n}\frac{\overline{X}_n - \lambda}{\sqrt{\lambda}} \in [-z_{1-\alpha/2},\ z_{1-\alpha/2}]\right) = 1-\alpha$$

   for λ leads to an alternative asymptotic confidence interval for λ.

   Exercise 2.2.7. Solve this quadratic inequality.

2. Since $\lambda > 0$, the lower bound can be adjusted to

   $$\underline{\lambda}(X_1, \ldots, X_n) = \max\left\{0,\ \overline{X}_n - z_{1-\alpha/2}\sqrt{\frac{\overline{X}_n}{n}}\right\}$$

3. Using the variance stabilization transformation from Example 1.3.29, 3.

   $$P\left(-z_{1-\frac{\alpha}{2}} \le \sqrt{n}(\sqrt{\bar{x}_n} - \sqrt{\lambda}) \le z_{1-\frac{\alpha}{2}}\right) \xrightarrow[n\to\infty]{} 1-\alpha$$

   holds. For n sufficiently large, the asymptotic confidence interval for λ with confidence level $1-\alpha$ is given by

   $$\lambda \in \left[\left(\sqrt{\bar{X}_n} - \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right)^2, \left(\sqrt{\bar{X}_n} + \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right)^2\right].$$

Since $\sqrt{\bar{X}_n} \xrightarrow[n\to\infty]{a.s.} \sqrt{\lambda}$ by the strong law of large numbers and

$$\frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \xrightarrow[n\to\infty]{} 0,$$

for sufficiently large n

$$\sqrt{\bar{X}_n} - \frac{z_{1-\alpha}2}{2\sqrt{n}} > 0$$

holds. Hence $\max\left\{0, \sqrt{\bar{X}_n} - \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right\}$ does not need to be taken care of.

2.3 Two-sample problems

In this section, some characteristics or parameters of two different samples will be compared by constructing confidence intervals for simple functions of those parameters.
Consider two random samples $Y_1 = (X_{11}, \ldots, X_{1n_1})$ and $Y_2 = (X_{21}, \ldots, X_{2n_2})$ of random variables $X_{i1}, \ldots, X_{in_i}$, $i = 1, 2$, which are, within the sample Y_i i.i.d. with $X_{ij} \overset{d}{=} X_i$, $j = 1, \ldots n_i$, $i = 1, 2$. Assume for the prototype random variable $X_i \sim F_{\theta_i}$, $\theta_i \in \Theta \subset \mathbb{R}^m$. In general it will not be assumed that Y_1 and Y_2 are independent. If they are dependent, the random samples Y_1 and Y_2 are called *related samples*. Consider a function $g : \mathbb{R}^{2m} \to \mathbb{R}$ of the parameter vectors θ_1 and θ_2. In this lecture the cases $m = 1, 2$, $g(\theta_1, \theta_2) = \theta_{1j} - \theta_{2j}$ and $g(\theta_1, \theta_2) = \frac{\theta_{1j}}{\theta_{2j}}$ will mostly be covered, where $\theta_i = (\theta_{i1}, \ldots, \theta_{im})$, $i = 1, 2$. The goal is to construct a (possibly asymptotic) confidence interval for $g(\theta_1, \theta_2)$ by using (Y_1, Y_2).
As it turns out, the approach will be similar to Section 2.2. A statistic $T(Y_1, Y_2, g(\theta_1, \theta_2))$ is desired, that has a (possibly asymptotic) test distribution F and explicitly depends on $g(\theta_1, \theta_2)$.
By solving the inequality $F^{-1}_{\alpha_1} \leq T(Y_1, Y_2, g(\theta_1, \theta_2)) \leq F^{-1}_{1-\alpha_2}$ for $g(\theta_1, \theta_2)$ a (possibly asymptotic) confidence interval with level $1 - \alpha$, $\alpha = \alpha_1 + \alpha_2$ can be obtained.

2.3.1 Normally distributed samples

Assume, that $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2$.

Confidence interval for the difference $\mu_1 - \mu_2$ with known variance σ_1^2 and σ_2^2 and independent random samples

Let Y_1 and Y_2 be independent and σ_1^2, σ_2^2 known. Consider the parameter function $g(\mu_1, \mu_2) = \mu_1 - \mu_2$ and

$$\overline{X}_{in_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij},\ i = 1, 2$$

the sample mean of Y_1 and Y_2. Then, $\overline{X}_{in_i} \sim N(\mu_i, \frac{\sigma_i^2}{n_i})$, $i = 1, 2$. [33, Theorem 7.3.2, 4] implies that $\overline{X}_{1n_1}$ and $\overline{X}_{2n_2}$ are independent. The stability of the normal distribution implies

$$\overline{X}_{1n_1} - \overline{X}_{2n_2} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

and normalizing yields

$$T(Y_1, Y_2, \mu_1 - \mu_2) = \frac{\overline{X}_{1n_1} - \overline{X}_{2n_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1).$$

The confidence interval

$$\left[\overline{X}_{1n_1} - \overline{X}_{2n_2} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ \overline{X}_{1n_1} - \overline{X}_{2n_2} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

for $\mu_1 - \mu_2$ with level $1 - \alpha$ then results.

Confidence interval for the quotient σ_1^2/σ_2^2 with unknown expected values μ_1 and μ_2 and independent random samples

Let Y_1 and Y_2 be independent and $g(\sigma_1, \sigma_2) = \frac{\sigma_1^2}{\sigma_2^2}$. Construct a statistic $T(Y_1, Y_2, \frac{\sigma_1^2}{\sigma_2^2})$. Let

$$S_{in_i}^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_{in_i}\right)^2,\ i = 1, 2$$

be the sample variances of Y_1 and Y_2. Then, applying [33, Theorem 7.4.8.] yields $\frac{(n_i-1)S_{in_i}^2}{\sigma_i^2} \sim \chi_{n_i-1}^2$, $i = 1, 2$. Since $S_{in_i}^2$, $i = 1, 2$ are independent, the definition of the F distribution implies

$$T\left(Y_1, Y_2, \frac{\sigma_1^2}{\sigma_2^2}\right) = \frac{\frac{(n_2-1)S_{2n_2}^2}{(n_2-1)\sigma_2^2}}{\frac{(n_1-1)S_{1n_1}^2}{(n_1-1)\sigma_1^2}} = \frac{S_{2n_2}^2}{S_{1n_1}^2} \cdot \frac{\sigma_1^2}{\sigma_2^2} \sim F_{n_2-1, n_1-1}.$$

Thus, the confidence interval

$$\left[\frac{S_{1n_1}^2}{S_{2n_2}^2}F_{n_2-1,n_1-1,\alpha_1},\ \frac{S_{1n_1}^2}{S_{2n_2}^2}F_{n_2-1,n_1-1,1-\alpha_2}\right]$$

for $\frac{\sigma_1^2}{\sigma_2^2}$ with level $1-\alpha$ is obtained.

Confidence interval for the difference $\mu_1 - \mu_2$ of expected values with dependent samples

Let Y_1 and Y_2 be linked, i.e., $X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma^2)$ for an unknown $\sigma^2 > 0$, $n_1 = n_2 = n$. Since $X_{ij}, j = 1, \ldots, n$ are i.i.d.,
$Z_j = X_{1j} - X_{2j} \sim N(\mu_1 - \mu_2, \sigma^2)$, $j = 1, \ldots, n$ holds.
The goal is to construct a confidence interval for $\mu_1 - \mu_2$. Consider the random samples $(Z_1, \ldots, Z_n)$ and the results of Section 2.2, then the confidence interval

$$\left[\overline{Z}_n - t_{n-1,1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}},\ \overline{Z}_n + t_{n-1,1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}\right]$$

for $\mu_1 - \mu_2$ with level $1 - \frac{\alpha}{2}$ is obtained. Here,

$$\overline{Z}_n = \frac{1}{n}\sum_{j=1}^{n} Z_j = \frac{1}{n}\sum_{j=1}^{n}(X_{1j} - X_{2j}) = \overline{X}_{1n} - \overline{X}_{2n}$$

and

$$S_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(Z_j - \overline{Z}_n\right)^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_{1j} - X_{2j} - \overline{X}_{1n} + \overline{X}_{2n}\right)^2.$$

2.3.2 Poisson distributed random samples

Assume that the random samples Y_1 and Y_2 are independent and
$X_i \sim \text{Poisson}(\lambda_i)$, $i = 1, 2$. The goals is to construct confidence intervals for

$$g(\lambda_1, \lambda_2) = \lambda_1 - \lambda_2,$$
$$g(\lambda_1, \lambda_2) = \frac{n_2\lambda_2}{n_1\lambda_1 + n_2\lambda_2} = \frac{\lambda_2}{\rho\lambda_1 + \lambda_2},$$

where $\rho = \frac{n_1}{n_2} = \text{const}$ for $n_1, n_2 \to \infty$.

Asymptotic confidence interval for $\lambda_1 - \lambda_2$

In order to obtain an asymptotically $\mathcal{N}(0,1)$ distributed statistic
$T(Y_1, Y_2, \lambda_1 - \lambda_2)$, the central limit theorem of Ljapunow (cf. [33, Theorem 4.2.13]) will be used.

Lemma 2.3.1. For $n_1, n_2 \to \infty$ with $0 < c_1 \leq n_1/n_2 \leq c_2 < \infty$,

$$\frac{\overline{X}_{1n_1} - \overline{X}_{2n_2} - \lambda_1 + \lambda_2}{\sqrt{\frac{\lambda_1}{n_1} + \frac{\lambda_2}{n_2}}} \xrightarrow[n_1,n_2\to\infty]{d} Y \sim \mathcal{N}(0,1)$$

holds.

Proof Define the random variable

$$Z_{nk} = \begin{cases} \frac{X_{1k} - \lambda_1}{n_1\sqrt{\frac{\lambda_1}{n_1} + \frac{\lambda_2}{n_2}}}, & k = 1, \ldots, n_1 \\ -\frac{X_{2k-n_1} - \lambda_2}{n_2\sqrt{\frac{\lambda_1}{n_1} + \frac{\lambda_2}{n_2}}}, & k = n_1 + 1, \ldots, n_1 + n_2 \end{cases}$$

where $n = n_1 + n_2$. Then, $\mathbb{E}\, Z_{nk} = 0$ for all $k = 1, \ldots, n$, and

$$0 < \sigma_{nk}^2 = \operatorname{Var} Z_{nk} = \begin{cases} \frac{\operatorname{Var} X_{1k}}{n_1^2\left(\frac{\lambda_1}{n_1} + \frac{\lambda_2}{n_2}\right)} = \frac{\lambda_1}{n_1^2\left(\frac{\lambda_1}{n_1} + \frac{\lambda_2}{n_2}\right)}, & k = 1, \ldots, n_1, \\ \frac{\lambda_2}{n_2^2\left(\frac{\lambda_1}{n_1} + \frac{\lambda_2}{n_2}\right)}, & k = n_1 + 1, \ldots, n, \end{cases}$$

Thus,

$$\sum_{k=1}^{n} \sigma_{nk}^2 = \left(\frac{\lambda_1}{n_1^2} n_1 + \frac{\lambda_2}{n_2^2} n_2\right) \frac{1}{\frac{\lambda_1}{n_1} + \frac{\lambda_2}{n_2}} = 1.$$

Furthermore, for $\delta > 0$ and $n_1, n_2 \to \infty$

$$\begin{aligned} \lim_{n\to\infty} \sum_{k=1}^{n} \mathbb{E}\left(|Z_{nk}|\right)^{2+\delta} &= \lim_{n_1,n_2\to\infty} \left(\frac{\mathbb{E}\left(|X_{11} - \lambda_1|^{2+\delta}\right)}{n_1^{1+\delta}\left(\frac{\lambda_1}{n_1} + \frac{\lambda_2}{n_2}\right)^{(2+\delta)/2}} + \frac{\mathbb{E}\left(|X_{21} - \lambda_2|\right)^{2+\delta}}{n_2^{1+\delta}\left(\frac{\lambda_1}{n_1} + \frac{\lambda_2}{n_2}\right)^{(2+\delta)/2}} \right) \\ &= 0 \end{aligned}$$

holds. The Ljapunow condition is therefore met and [33, Theorem 4.2.13] implies

$$\sum_{k=1}^{n} Z_{nk} \xrightarrow[n_1,n_2\to\infty]{d} Y \sim \mathcal{N}(0,1).$$

Finally, $\sum_{k=1}^{n} Z_{nk} = \frac{\overline{X}_{1n_1} - \overline{X}_{2n_2} - \lambda_1 + \lambda_2}{\sqrt{\frac{\lambda_1}{n_1} + \frac{\lambda_2}{n_2}}}$, which completes the proof. □

The strong law of large numbers implies $\overline{X}_{in_i} \xrightarrow{f.s.} \lambda_i$, $i = 1, 2$ and using Slutskys theorem then yields

$$T(Y_1, Y_2, \lambda_1 - \lambda_2) = \frac{\overline{X}_{1n_1} - \overline{X}_{2n_2} - \lambda_1 + \lambda_2}{\sqrt{\overline{X}_{1n_1}/n_1 + \overline{X}_{n_2}/n_2}} \xrightarrow[n_1,n_2\to\infty]{d} Y \sim \mathcal{N}(0,1).$$

The asymptotic confidence interval for $\lambda_1 - \lambda_2$ with level $1-\alpha$ is thus given by

$$\left[\overline{X}_{1n_1} - \overline{X}_{2n_2} - z_{1-\alpha/2}\sqrt{\frac{\overline{X}_{1n_1}}{n_1} + \frac{\overline{X}_{2n_2}}{n_2}},\ \overline{X}_{1n_1} - \overline{X}_{2n_2} + z_{1-\alpha/2}\sqrt{\frac{\overline{X}_{1n_1}}{n_1} + \frac{\overline{X}_{2n_2}}{n_2}}\right]$$

Asymptotic confidence interval for $\frac{n_2\lambda_2}{n_1\lambda_1+n_2\lambda_2}$

Let ρ be some constant, $n_1/n_2 = \rho$ and
$g(\lambda_1, \lambda_2) = \frac{n_2\lambda_2}{n_1\lambda_1+n_2\lambda_2} = \frac{\lambda_2}{\rho\lambda_1+\lambda_2} \overset{\text{Def.}}{=} p$. The goal is to construct an asymptotic confidence interval for p. Consider the statistic

$$T(Y_1, Y_2, p) = \frac{S_{2n_2} - p(S_{1n_1} + S_{2n_2})}{\sqrt{\hat{p}(1-\hat{p})(S_{1n_1} + S_{2n_2}))}},$$

where $S_{in_i} = \sum_{j=1}^{n_i} X_{ij}$, $i = 1, 2$ and

$$\hat{p} = \frac{S_{2n_2}}{S_{1n_1} + S_{2n_2}} = \frac{n_2\overline{X}_{2n_2}}{n_1\overline{X}_{1n_1} + n_2\overline{X}_{2n_2}} \xrightarrow[n_1,n_2\to\infty]{f.s.} p$$

is a consistent estimator for p (by the strong law of large numbers). If it can be shown that $T(Y_1, Y_2, p) \xrightarrow[n_1,n_2\to\infty]{d} Y \sim \mathcal{N}(0,1)$, then

$$\lim_{n_1,n_2\to\infty} P\left(-z_{1-\alpha/2} \leq \frac{\frac{S_{2n_2}}{S_{1n_1}+S_{2n_2}} - p}{\sqrt{S_{1n_1} \cdot S_{2n_2}}} \cdot (S_{1n_1} + S_{2n_2})^{3/2} \leq z_{1-\alpha/2}\right) = 1-\alpha,$$

which yields the asymptotic confidence interval

$$\left[\underline{\theta}(Y_1, Y_2),\ \overline{\theta}(Y_1, Y_2),\right]$$

for p with level $1-\alpha$, where

$$\underline{\theta}(\lambda_1, \lambda_2) = \frac{S_{2n_2}}{S_{1n_1} + S_{2n_2}} - z_{1-\alpha/2} \cdot \sqrt{\frac{S_{1n_1} \cdot S_{2n_2}}{(S_{1n_1} + S_{2n_2})^3}}$$

and

$$\overline{\theta}(\lambda_1, \lambda_2) = \frac{S_{2n_2}}{S_{1n_1} + S_{2n_2}} + z_{1-\alpha/2} \cdot \sqrt{\frac{S_{1n_1} \cdot S_{2n_2}}{(S_{1n_1} + S_{2n_2})^3}}.$$

Since $0 < p < 1$, the boundaries can be adjusted as follows:

$$\underline{\theta}^*(Y_1, Y_2) = \max\{0,\ \underline{\theta}(Y_1, Y_2)\},$$
$$\overline{\theta}^*(Y_1, Y_2) = \min\{1,\ \overline{\theta}(Y_1, Y_2)\}.$$

Now the asymptotically normal distribution of $T(Y_1, Y_2, p)$ will be shown. It results from Slutskys theorem and the following Lemma:

Lemma 2.3.2. One has the following property

$$\frac{S_{2n_2} - p(S_{1n_1} + S_{2n_2})}{\sqrt{p(1-p)(S_{1n_1} + S_{2n_2})}} \xrightarrow[n_1 \to \infty]{d} Y \sim \mathcal{N}(0,1)$$

Proof In order to show the assertion, a version of the central limit theorem for sums of random variables with random numbers as sum limits (cf. [33, Theorem 4.2.2]) is used. Let $N_n = S_{1n_1} + S_{2n_2}$ be a sequence of non-negative random variables, then the sum is monotonically increasing. Let then $a_{n_2} = n_1\lambda_1 + n_2\lambda_2$. Obviously

$$\begin{aligned}
\frac{N_n}{a_{n_2}} &= \frac{S_{1n_1}}{n_1\lambda_1 + n_2\lambda_2} + \frac{S_{2n_2}}{n_1\lambda_1 + n_2\lambda_2} \\
&= \frac{\overline{X}_{1n_1}}{\lambda_1 + \rho^{-1}\lambda_2} + \frac{\overline{X}_{2n_2}}{\rho\lambda_1 + \lambda_2} \\
&\xrightarrow[n_1, n_2 \to \infty]{f.s.} \frac{\lambda_1}{\lambda_1 + \rho^{-1}\lambda_2} + \frac{\lambda_2}{\rho\lambda_1 + \lambda_2} \\
&= \frac{\rho\lambda_1}{\rho\lambda_1 + \lambda_2} + \frac{\lambda_2}{\rho\lambda_1 + \lambda_2} = 1
\end{aligned}$$

holds. Furthermore:

$$\begin{aligned}
P\left(S_{2n_2} = k \mid N_n = m\right) &= \frac{P\left(S_{2n_2} = k,\, S_{1n_1} + S_{2n_2} = m\right)}{P\left(S_{1n_1} + S_{2n_2} = m\right)} \\
&= \frac{P\left(S_{2n_2} = k,\, S_{1n_1} = m - k\right)}{P\left(S_{1n_1} + S_{2_{n-2}} = m\right)} \\
&= \frac{e^{-n_2\lambda_2}\frac{(\lambda_2 n_2)^k}{k!} \cdot e^{-n_1\lambda_1}\frac{(n_1\lambda_1)^{m-k}}{(m-k)!}}{e^{-n_1\lambda_1 - n_2\lambda_2}\frac{(n_1\lambda_1 + n_2\lambda_2)^m}{m!}} \\
&= \frac{m!}{(m-k)!k!}\left(\frac{n_2\lambda_2}{n_1\lambda_1 + n_2\lambda_2}\right)^m \left(\frac{n_1\lambda_1}{n_1\lambda_1 + n_2\lambda_2}\right)^{m-k} \\
&= \binom{m}{k} p^k (1-p)^{m-k}
\end{aligned}$$

which means that $S_{2n_2} \mid \{N_n = m\} \sim \mathrm{Bin}(m,p)$. Then, $\frac{S_{2n_2} - mp}{\sqrt{mp(1-p)}} \mid \{N_n = m\} \overset{d}{=} \frac{S_m - mp}{\sqrt{mp(1-p)}}$, where $S_m = \sum_{i=1}^{m} Z_i$ is a sum of identically distributed $Z_i \sim \mathrm{Bernoulli}(p)$. [33, Theorem 4.2.2] implies

$$\frac{S_{N_n} - N_n p}{\sqrt{N_n p(1-p)}} \xrightarrow{d} Y \sim \mathcal{N}(0,1) \iff \frac{S_{2n_2} - N_n p}{\sqrt{N_n p(1-p)}} \xrightarrow{d} Y \sim \mathcal{N}(0,1).$$

□

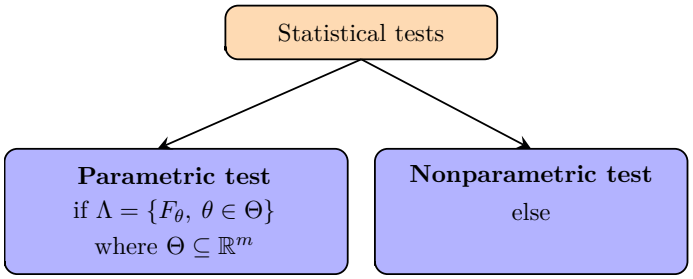
Chapter 3

Testing Statistical Hypotheses

In [33, Chapter 7], some tests like the Kolmogorow-Smironow test were introduced. This chapter, however, focuses on introducing tests for statistical significance formally.

3.1 General philosophy of testing

Let $(X_1, \ldots, X_n)$ be a random sample of i.i.d. random variables X_i with distribution function $F \in \Lambda$, where Λ is some class of distributions. Let $(x_1, \ldots, x_n)$ be a realization of the random sample $(X_1, \ldots, X_n)$. In statistical testing, hypotheses with respect to the nature of a (unknown) distribution F are posed and tested. Generally, two concepts are distinguished:



Parametric tests check, whether a parameter θ attains certain values (e.g. $\theta = 0$). Popular nonparametric tests are the so-called "goodness-of-fit tests", which check whether the distribution F is equal to a predetermined distribution F_0.
In an initial step, the term *Hypotheses* needs to be formalized. The set Λ of admissible distributions F is divided into two disjoint sets Λ_0 and Λ_1 , $\Lambda_0 \cup \Lambda_1 = \Lambda$. The assertion

"The *null hypothesis* $H_0 : F \in \Lambda_0$
is tested against the *alternative* $H_1 : F \in \Lambda_1$"

means, that we aim to assign the distribution function of the random variable X_i to Λ_0 or Λ_1, based on the explicit realization $(x_1, \ldots, x_n)$. The process of assigning the distribution of X_i involves a *decision rule*

$$\varphi : \mathbb{R}^n \to [0, 1],$$

which is a statistic with the following interpretation:
The sample space $\mathbb{R}^n$ is divided into 3 disjoint sets $K_0, K_{01}\ K_1$, such that $\mathbb{R}^n = K_0 \cup K_{01} \cup K_1$, where

$$\begin{aligned} K_0 &= \varphi^{-1}(\{0\}) &&= \{x \in \mathbb{R}^n : \varphi(x) = 0\}, \\ K_1 &= \varphi^{-1}(\{1\}) &&= \{x \in \mathbb{R}^n : \varphi(x) = 1\}, \\ K_{01} &= \varphi^{-1}((0, 1)) &&= \{x \in \mathbb{R}^n : 0 < \varphi(x) < 1\}. \end{aligned}$$

Thus $H_0 : F \in \Lambda_0$ is

- rejected, if $\varphi(x) = 1$, i.e., $x \in K_1$,
- not rejected, if $\varphi(x) = 0$, i.e., $x \in K_0$.

If $\varphi(x) \in (0, 1)$, i.e., $x \in K_{01}$, then $\varphi(x)$ is interpreted as a Bernoulli probability and a random variable $Y \sim \text{Bernoulli}(\varphi(x))$ is generated with

$$Y = \begin{cases} 1 & \Longrightarrow H_0 \text{ is rejected} \\ 0 & \Longrightarrow H_0 \text{ is not rejected} \end{cases}$$

If $K_{01} \neq \emptyset$, the decision rule is called *randomized*. If $K_{01} = \emptyset$, i.e., $\mathbb{R}^n = K_0 \cup K_1$ the tests are called *non-randomized*. K_0 and K_1 are called *acceptance region* and *rejection region* (*critical region*) of H_0 respectively. K_{01} is called *randomization region*.

Remark 3.1.1.

1. One deliberately says "H_0 *is not rejected*", instead of "H_0 *is accepted*", since statistical inference can generally not make positive decisions rather than negative decisions. The issue above is a general philosophical problem with respect to the falsifiability of hypotheses or scientific theories, which can generally not be at odds with the truth. (cf. *wissenschaftliche Erkenntnistheorie by Karl Popper (1902-1994)*).

2. Randomized tests are generally of a more theoretic nature (cf. Section 3.3). In practice, most non-randomized rules are used, which leads to a decision with respect to H_0 based on the explicit sample $(x_1, \ldots, x_n)$ alone. Here, $\varphi(x) = I_{K_1}, x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ holds.

In the following paragraph, non-randomized tests are considered in order to return to the more general approach in Section 3.3.

Definition 3.1.2. The non-randomized test rule $\varphi : \mathbb{R}^n \to \{0, 1\}$ provides a *(non-randomized) statistical test with significance level* α, if for $F \in \Lambda_0$

$$P_F\left(\varphi(X_1, \ldots, X_n) = 1\right) = P\left(H_0 \text{ reject } \mid H_0 \text{ true }\right) \leq \alpha.$$

Definition 3.1.3.

1. If H_0 is rejected, even though H_0 is correct, then a *type I error* has occurred. The probability

$$\alpha_n(F) = P_F\left(\varphi(x_1, \ldots, x_n) = 1\right), \quad F \in \Lambda_0,$$

is called *Probability of a type I error*. This probability is supposed to be lower than the significance level α.

2. A *type II error* occurs, if a wrong hypothesis H_0 is not rejected. Here

$$\beta_n(F) = P_F\left(\varphi(x_1, \ldots, x_n) = 0\right), \quad F \in \Lambda_1,$$

is called *Probability of a type II error*.

A summary of all possible errors can be found in the following matrix, which is called *confusion matrix*:

	H_0 true	H_0 false
reject H_0	Error of type I with probability $\alpha_n(F) \leq \alpha$	right decision
not rejecting H_0	right decision	Error of type II with probability $\beta_n(F)$

Here α_n and β_n are aimed to be small, which is contrary to the fact that a decreasing value of α increases the probability of mistakes of type II.

Definition 3.1.4.

1. The function

$$G_n(F) = P_F\left(\varphi(X_1, \ldots, X_n) = 1\right), \quad F \in \Lambda$$

is called *performance function* (or *power function*) of a test φ.

2. The constraint G_n on Λ_1 is called *power* of the test φ.
   With respect to the constraint it holds

$$\begin{cases} G_n(F) = \alpha_n(F) \leq \alpha, \ F \in \Lambda_0, \\ G_n(F) = 1 - \beta_n(F), \ F \in \Lambda_1. \end{cases}$$

Example 3.1.5. *Parametric tests.*
What does a parametric test look like? The parameter space Θ is given by $\Theta_0 \cup \Theta_1$, where $\Theta_0 \cap \Theta_1 = \emptyset$. Then

$$\Lambda_0 = \{F_\theta : \theta \in \Theta_0\},$$
$$\Lambda_1 = \{F_\theta : \theta \in \Theta_1\}.$$

P_F is replaced by P_θ. Furthermore α_n, G_n and β_n are defined on Θ instead of Λ.

Which hypotheses H_0 and H_1 are popular among parametric tests? The case $\Theta = \mathbb{R}$ is discussed below, but it should be noted that a more general choice of Θ is also possible.

1. $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$
2. $H_0 : \theta \geq \theta_0$ vs. $H_1 : \theta < \theta_0$
3. $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$
4. $H_0 : \theta \in [a, b]$ vs. $H_1 : \theta \notin [a, b]$

In the first case the, parametric test is called *two-sided* and in the second and third case *one-sided* (*right-* resp. *left-sided*). The fourth case is called *interval hypothesis* H_0.
Considering a one-sided or two-sided test, the power function may look like the one displayed in Figure 3.1 (a) or 3.1 (b), resp.

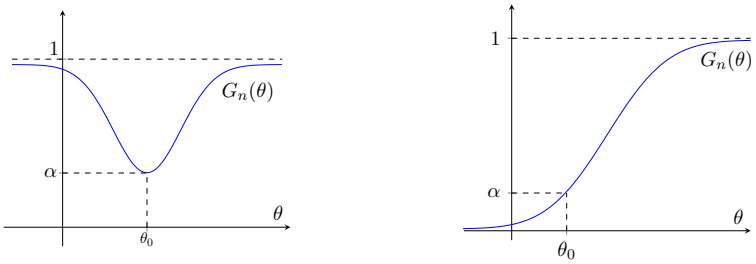


Figure 3.1: Performance function

In general models (not necessarily parametric), the ideal power function can be illustrated schematically, as in Figure 3.2.

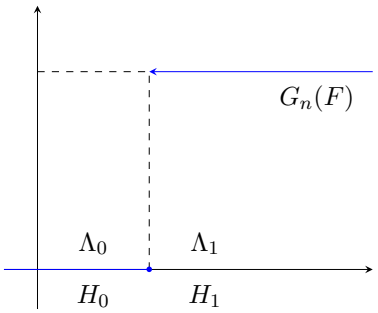


Figure 3.2: Schematic illustration of an ideal power function

- Definition 3.1.3, errors of type I and II and the rejection rule imply that the hypotheses H_0 and H_1 can not be treated symmetrically, since only the probability of errors of type I is controlled. That is the reason why statisticians mostly formulate the hypothesis of interest as H_1 instead of H_0, because if one decides that H_0 can be rejected, it can be assured that the probability of a false decision is below the significance level α.

- How is a statistical, non-randomized test constructed in practice? The construction of the rejection rule φ is very similar to constructing confidence intervals:

  1. Find a test statistic $T : \mathbb{R}^n \to \mathbb{R}$, which has a certain test distribution (perhaps asymptotically for $n \to \infty$) under H_0.
  2. Define $B_0 = [t_{\alpha_1}, t_{1-\alpha_2}]$, where t_{α_1} and $t_{1-\alpha_2}$ are quantiles of the test distribution of T with $\alpha_1 + \alpha_2 = \alpha \in [0, 1]$.
  3. If $T(X_1, \ldots, X_n) \in \mathbb{R} \setminus B_0 = B_1$, then set $\varphi(X_1, \ldots, X_n) = 1$ and reject H_0. Else, set $\varphi(X_1, \ldots, X_n) = 0$.

- If the distribution of T can only be determined asymptotically, then φ is called *asymptotic test*.

- Most of the times, even the asymptotic distribution of T is unknown. In this case, the so called *Monte-Carlo tests* come into play. In those tests the quantiles t_α are determined approximatively by conducting a large number of Monte-Carlo simulations of T (under H_0):

  If t^i, $i = 1, \ldots, m$ takes the values of T in m independent simulations, i.e. $t^i = T(x_1^i, \ldots, x_n^i)$, where x_j^i are independent realizations of $X_j \sim F \in \Lambda_0$ for $j = 1, \ldots, n$, $i = 1, \ldots, m$, then $t_\alpha \approx t^{(\lfloor \alpha \cdot m \rfloor)}$,[1] with $t^{(1)}, \ldots, t^{(m)}$ the order statistics and $\alpha \in [0, 1]$.

[1] set $t^{(0)} = -\infty$

Remark 3.1.6. It is easy to see that by using an arbitrary confidence interval

$$I_\theta = \left[I_1^\theta(X_1,\dots,X_n), I_2^\theta(X_1,\dots,X_n)\right]$$

with confidence level $1-\alpha$ for a parameter $\theta \in \mathbb{R}$, a test for θ can be constructed. The hypotheses $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ are tested under the following decision rule:

$$\varphi(X_1,\dots,X_n) = 1, \text{ if } \theta_0 \notin \left[I_1^{\theta_0}(X_1,\dots,X_n),\, I_2^{\theta_0}(X_1,\dots,X_n)\right].$$

The significance level of the test is α.

Example 3.1.7. *Normal distribution, testing the expected value with known variance.* Let

$$X_1,\dots,X_n \sim \mathcal{N}(\mu,\sigma^2)$$

with known variance σ^2. A confidence interval for μ is given by

$$I^\mu = [I_1^\mu(X_1,\dots,X_n),\, I_2^\mu(X_1,\dots,X_n)] = \left[\overline{X}_n - \frac{z_{1-\alpha/2}\cdot\sigma}{\sqrt{n}},\, \overline{X}_n + \frac{z_{1-\alpha/2}\cdot\sigma}{\sqrt{n}}\right]$$

(cf. Section 2.2.1). Hence, $H_0 : \mu = \mu_0$ (versus the alternative $H_1 : \mu \neq \mu_0$), is rejected, if

$$|\mu_0 - \overline{X}_n| > \frac{z_{1-\alpha/2}\cdot\sigma}{\sqrt{n}}.$$

In the language of testing, the above can be rewritten as

$$\varphi(x_1,\dots,x_n) = I\left((x_1,\dots x_n) \in K_1\right),$$

where

$$K_1 = \left\{(x_1,\dots,x_n) \in \mathbb{R}^n : |\mu_0 - \overline{x}_n| > \frac{z_{1-\alpha/2}\cdot\sigma}{\sqrt{n}}\right\}$$

is the rejection region. For the test statistic $T(X_1,\dots,X_n)$

$$T(X_1,\dots,X_n) = \frac{\overline{X}_n - \mu_0}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1)$$

under H_0 it holds $\alpha_n(\mu) = \alpha$.
The power function (cf. Figure 3.3) can be calculated as follows

$$\begin{aligned}
G_n(\mu) &= P_\mu\left(|\mu_0 - \overline{X}_n| > \frac{z_{1-\alpha/2}}{\sqrt{n}}\right) = 1 - P_\mu\left(\left|\overline{X}_n - \mu_0\right| \leq \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}}\right)\\
&= 1 - P_\mu\left(\left|\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} + \frac{\mu-\mu_0}{\sigma}\sqrt{n}\right| \leq z_{1-\alpha/2}\right)\\
&= 1 - P_\mu\left(-z_{1-\alpha/2} - \frac{\mu-\mu_0}{\sigma}\sqrt{n} \leq \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \leq z_{1-\alpha/2} - \frac{\mu-\mu_0}{\sigma}\sqrt{n}\right)\\
&= 1 - \Phi\left(z_{1-\alpha/2} - \frac{\mu-\mu_0}{\sigma}\sqrt{n}\right) + \Phi\left(-z_{1-\alpha/2} - \frac{\mu-\mu_0}{\sigma}\sqrt{n}\right)\\
&= \Phi\left(-z_{1-\alpha/2} + \frac{\mu-\mu_0}{\sigma}\sqrt{n}\right) + \Phi\left(-z_{1-\alpha/2} - \frac{\mu-\mu_0}{\sigma}\sqrt{n}\right).
\end{aligned}$$

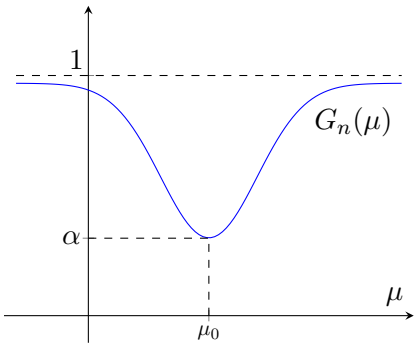


Figure 3.3: Performance function of a two-sided test of the expected value of a normal distribution with known variance.

The "yes-no" decision of tests is mainly viewed as too rough. That is why it is desirable to obtain a finer measure for the data with respect to the hypotheses H_0 and H_1. The so-called *p-value* solves the problem above and it is luckily included in most statistic software.

Definition 3.1.8. Let $(x_1, \ldots, x_n)$ be an explicit sample, i.e., a realization of $(X_1, \ldots, X_n)$ and $T(X_1, \ldots, X_n)$ the test statistic which was used to construct the decision rule φ. *The p-value of the test* φ is the smallest significance level to the value $t = T(x_1, \ldots, x_n)$ which leads to a rejection of H_0.
In the example of a one-sided test with rejection region $B_1 = (t, \infty)$, the rule of thumb for p is given by

$$p = \text{“ } P\left(T(X_1, \ldots, X_n) \geq t \mid H_0\right) \text{ ”},$$

where the quotation marks imply that the term is not a probability in the classical sense, rather than a conditional probability, which will be defined more precisely later.

Using the p-value, the rejection rule changes: The hypothesis H_0 is rejected with a significance level α, if $\alpha \geq p$. Previously, the significance of a test (rejection of H_0) was determined with respect to the following table:

p-value	interpretation
$p \leq 0,001$	very strongly significant
$0,001 < p \leq 0,01$	strongly significant
$0,01 < p \leq 0,05$	weakly significant
$p > 0,05$	not significant

Since the p-value can be calculated with ease nowadays, one can use the p-value directly to decide which significance level is sufficient for the underlying test.

Remark 3.1.9.

1. The significance level must not depend on p. By doing so, the general philosophy of testing is jeopardized!

2. The p-value is not a probability rather than a random variable since it depends on $(X_1, \ldots, X_n)$. The expression

$$p = P\left(T(X_1, \ldots, X_n) \geq t \mid H_0\right),$$

in Definition 3.1.8 for the p-value of an one-sided test with test statistic T can be interpreted as an *exceedance probability*. The exceedance probability is defined with respect to $t = T(x_1, \ldots, x_n)$ or more extreme values in order to be close to the hypothesis H_1 while repeating the random experiment under H_0:

$$p = P\left(T(X_1', \ldots, X_n') \geq T(x_1, \ldots, x_n) \mid H_0\right),$$

where $(X_1', \ldots, X_n') \stackrel{d}{=} (X_1, \ldots, X_n)$. If instead of the explicit sample $(x_1, \ldots, x_n)$ the random sample $(X_1, \ldots, X_n)$ is used, then

$$p = p(X_1, \ldots, X_n) = P\left(T(X_1', \ldots, X_n') \geq T(X_1, \ldots, X_n) \mid H_0, X_1, \ldots, X_n\right).$$

3. For other hypotheses H_0, the p-values may look different. For example:

   (a) A symmetric two-sided test has an acceptance region

$$B_0 = \left[-t_{1-\alpha/2},\, t_{1-\alpha/2}\right]$$

   for H_0. Therefore

$$p = P\left(|T(X_1', \ldots, X_n')| \geq T(X_1, \ldots X_n) \mid H_0, X_1, \ldots, X_n\right).$$

   (b) A left-sided test with $B_0 = [t_\alpha, \infty]$ results in

$$p = P(T(X_1', \ldots, X_n') \leq T(X_1, \ldots X_n) | H_0, X_1, \ldots, X_n).$$

4. The behavior of the p-value can be evaluated using the following lemma

   Lemma 3.1.10. If the distribution function F of T is continuous and monotonically increasing (e.g., the distribution T is absolutely continuous with continuous probability density function for example), then $p \sim U[0, 1]$.

Proof The result will be shown for right-sided tests only.

$$\begin{aligned} P(p \le \alpha \mid H_0) &= P\left(\overline{F}_T(T(X_1,\dots,X_n)) \le \alpha \mid H_0\right) \\ &= P\left(F_T(T(X_1,\dots,X_n)) \ge 1-\alpha \mid H_0\right) \\ &= P(U \ge 1-\alpha) = 1-(1-\alpha) = \alpha, \quad \alpha \in [0,1], \end{aligned}$$

since $F_T(T(X_1,\dots,X_n)) \stackrel{d}{=} U \sim U[0,1]$, and F_T is absolutely continuous. □

Exercise 3.1.11. Show that for an arbitrary random variable X with continuous and monotonically increasing distribution function F_X

$$F_X(X) \sim U[0,1]$$

holds.

If the distribution of T with domain $\{t_1,\dots,t_n\}$, $t_i < t_j$, is discrete for $i<j$, then the distribution of p is also discrete. In particular, it does not hold that $p \sim U[0,1]$. In this case $F_p(x)$ is a step function which touches the line $y=u$ at the points $u_k = \sum_{i=1}^{k} P(T(X_1,\dots,X_n)=t_i)$, $k=1,\dots,n$ (cf. Figure 4).

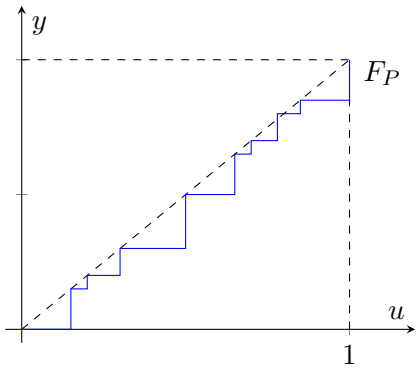


Figure 3.4: Distribution of p for discrete T

Definition 3.1.12.

1. If the power $G_n(\cdot)$ of a test φ with significance level α satisfies the inequality

$$G_n(F) \ge \alpha, \quad F \in \Lambda_1,$$

then the test is called *unbiased*.

2. Let φ and φ^* be two tests with significance level α and power functions $G_n(\cdot)$ and $G_n^*(\cdot)$. The test φ is said to be *more powerful* than φ^* if its power is larger:

$$G_n(F) \geq G_n^*(F) \quad \forall F \in \Lambda_1.$$

3. The test φ is called consistent, if $G_n(F) \underset{n\to\infty}{\longrightarrow} 1$ for all $F \in \Lambda_1$.

Remark 3.1.13.

1. The power of a one-sided test is mostly larger than the two-sided version:

   Example 3.1.14. Consider the Gauss-test for the expected value of the normal distribution if the variance is known. The two-sided test

   $$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu \neq \mu_0.$$

   implies that the power function is given by

   $$G_n(\mu) = \Phi\left(-z_{1-\alpha/2} + \sqrt{n}\frac{\mu - \mu_0}{\sigma}\right) + \Phi\left(-z_{1-\alpha/2} - \sqrt{n}\frac{\mu - \mu_0}{\sigma}\right).$$

   The one-sided test φ^* of the hypotheses

   $$H_0^* : \mu \leq \mu_0 \text{ vs. } H_1^* : \mu > \mu_0$$

   attains a power function given by

   $$G_n^*(\mu) = \Phi\left(-z_{1-\alpha} + \sqrt{n}\frac{\mu - \mu_0}{\sigma}\right).$$

   Since $G_n(\mu) \underset{n\to\infty}{\longrightarrow} 1$, $G_n^*(\mu) \underset{n\to\infty}{\longrightarrow} 1$ both tests are consistent. In the case above, φ^* is more powerful than φ. Moreover, both tests are unbiased (cf. Figure 3.1.14).

2. For testing interval hypotheses $H_0 : \theta \in [a, b]$ vs. $H_1 : \theta \notin [a, b]$ with confidence level α the following methodology can be used: Test

   (a) $H_0^a : \theta \geq a$ vs. $H_1^a : \theta < a$ with significance level $\alpha/2$,
   (b) $H_0^b : \theta \leq b$ vs. $H_1^b : \theta > b$ with significance level $\alpha/2$.

   H_0 is not rejected if H_0^a and H_0^b are not rejected. The probability for a type I error is α. The power of those tests is generally low.

3. As a rule of thumb, it holds that an increase in parameters that need to be estimated with respect to the test statistic leads to a decrease in power.

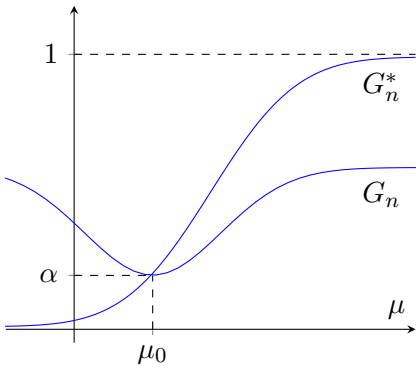


Figure 3.5: Power function of an one-sided and two-sided test for the expected value of a normal distribution.

3.2 Non-randomized tests

3.2.1 Parametric significance tests

This section provides examples of tests that can mostly be obtained from their corresponding confidence intervals for the parameters of distributions.

1. **Tests for the parameters of a normal distribution $\mathcal{N}(\mu, \sigma^2)$**

   (a) **Test of μ with unknown variance**

   - Hypotheses: $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$.
   - Test statistic:

   $$T(X_1, \ldots, X_n) = \frac{\overline{X}_n - \mu_0}{S_n} \sim t_{n-1} \quad | \, H_0$$

   - Decision rule:

   $$\varphi(X_1, \ldots, X_n) = 1, \text{ if } |T(X_1, \ldots, X_n)| > t_{n-1,1-\alpha/2}.$$

   (b) **Test of σ^2 with known μ**

   - Hypotheses: $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 \neq \sigma_0^2$.
   - Test statistic:

   $$T(X_1, \ldots, X_n) = \frac{n\tilde{S}_n^2}{\sigma_0^2} \sim \chi_n^2 \quad | \, H_0$$

   with $\tilde{S}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$.

   - Decision rule:

   $$\varphi(X_1, \ldots, X_n) = 1, \text{ if } T(X_1, \ldots, X_n) \notin \left[\chi_{n,\alpha/2}^2, \chi_{n,1-\alpha/2}^2 \right].$$

- Performance function:

$$\begin{aligned} G_n(\sigma^2) &= 1 - P_{\sigma^2}\left(\chi^2_{n,\alpha/2} \leq \frac{n\tilde{S}_n^2}{\sigma_0^2} \leq \chi^2_{n,1-\alpha/2}\right) \\ &= 1 - P_{\sigma^2}\left(\frac{\chi^2_{n,\alpha/2}\sigma_0^2}{\sigma^2} \leq \frac{n\tilde{S}_n^2}{\sigma^2} \leq \frac{\chi^2_{n,1-\alpha/2}\sigma_0^2}{\sigma^2}\right) \\ &= 1 - F_{\chi^2_n}\left(\chi^2_{n,1-\alpha/2}\frac{\sigma_0^2}{\sigma^2}\right) + F_{\chi^2_n}\left(\chi^2_{n,\alpha/2}\frac{\sigma_0^2}{\sigma_2}\right) \end{aligned}$$

(c) **Test for σ^2 with unknown μ**

- Hypotheses: $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 \neq \sigma_0^2$.
- Test statistic:

$$T(X_1, \ldots, X_n) = \frac{(n-1)S_n^2}{\sigma_0^2} \sim \chi^2_{n-1} \quad | H_0,$$

where $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$.

- Decision rule:

$\varphi(X_1, \ldots, X_n) = 1$, if $T(X_1, \ldots, X_n) \notin \left[\chi^2_{n-1,\alpha/2}, \chi^2_{n-1,1-\alpha/2}\right]$.

Exercise 3.2.1.

i. Find $G_n(\cdot)$ for the one-sided versions of the tests above.

ii. Show that the one-sided tests are unbiased, contrary to the two-sided tests being biased.

2. **Asymptotic tests**

Considering asymptotic tests, the test statistic distribution can only be estimated (for large n). In the same spirit, the confidence level α is obtained. Its construction is mostly based on limit theorems.

The general methodology is introduced via the *Wald test* (named after the statistician Abraham Wald (1902-1980)):

- Let $(X_1, \ldots, X_n)$ be a random sample and X_i be i.i.d. for $i = 1, \ldots, n$, with $X_i \sim F_\theta$, $\theta \in \Theta \subseteq \mathbb{R}$.
- $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ is tested. Let $\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n)$ be an asymptotically normal distributed estimator for θ. Let

$$\frac{\hat{\theta}_n - \theta_0}{\hat{\sigma}_n} \xrightarrow[n\to\infty]{d} Y \sim \mathcal{N}(0,1) \quad | H_0,$$

where $\hat{\sigma}_n^2$ is a consistent estimator for the variance of $\hat{\theta}_n$.

The test statistic is given by

$$T(X_1, \ldots, X_n) = \frac{\hat{\theta}_n(X_1, \ldots, X_n) - \theta_0}{\hat{\sigma}_n}.$$

- The decision rule is: H_0 is rejected, if

$$|T(X_1, \ldots, X_n)| > z_{1-\alpha/2}, \text{ where } z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2).$$

This decision rule should only be applied for large n. The probability of a type I error is equal to α, since $P(|T(X_1, \ldots, X_n)| > z_{1-\alpha/2} \mid H_0) \xrightarrow[n\to\infty]{} \alpha$ because of the asymptotically normal distribution of T.

The power function of the test is asymptotically given by

$$\lim_{n\to\infty} G_n(\theta) = 1 - \Phi\left(z_{1-\alpha/2} + \frac{\theta_0 - \theta}{\sigma}\right) + \Phi\left(-z_{1-\alpha/2} + \frac{\theta_0 - \theta}{\sigma}\right),$$

where $\hat{\sigma}_n^2 \xrightarrow[n\to\infty]{P} \sigma^2$.

Special cases of the Wald test are asymptotic tests for the expected value of Poisson or Bernoulli distributed random samples.

Example 3.2.2.

(a) **Bernoulli distribution**

Let $X_i \sim \text{Bernoulli}(p)$, $p \in (0,1)$ be i.i.d. random variables.

- Hypotheses: $H_0 : p = p_0$ vs. $H_1 : p \neq p_0$.
- Test statistic:

$$T(X_1, \ldots, X_n) = \begin{cases} \sqrt{n}\frac{\overline{X}_n - p_0}{\sqrt{\overline{X}_n(1-\overline{X}_n)}}, & \text{if } \overline{X}_n \neq 0, 1, \\ 0, & \text{otherwise.} \end{cases}$$

Under H_0, $T(X_1, \ldots, X_n) \xrightarrow[n\to\infty]{d} Y \sim \mathcal{N}(0,1)$ holds.

(b) **Poisson distribution**

Let $X_i \sim \text{Poisson}(\lambda)$, $\lambda > 0$ be i.i.d. random variables.

- Hypotheses: $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda \neq \lambda_0$.
- Test statistic:

$$T(X_1, \ldots, X_n) = \begin{cases} \sqrt{n}\frac{\overline{X}_n - \lambda_0}{\sqrt{\overline{X}_n}}, & \text{if } \overline{X}_n > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Under H_0, $T(X_1, \ldots, X_n) \xrightarrow[n\to\infty]{d} Y \sim N(0,1)$ holds.

3. **Two sample problems**

Let

$$Y_1 = (X_{11}, \ldots, X_{1n_1}),\ Y_2 = (X_{21}, \ldots, X_{2n_2}),\ n = \max\{n_1, n_2\}$$

be two random samples. Assume that X_{ij} are independent for $j = 1, \ldots, n_i$, $X_{ij} \sim F_{\theta_i}$, $i = 1, 2$.

(a) **Test for the equality of two expected values for normal distributed random samples**

- **with known variance**
  Let $X_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $i = 1, 2$, $j = 1, \ldots, n_i$. Here, σ_1^2, σ_2^2 are known and X_{ij} are independent for all i, j.
  The hypotheses are given by $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$. Consider the test statistic

$$T(Y_1, Y_2) = \frac{\overline{X}_{1n_1} - \overline{X}_{2n_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

  Under H_0, $T(Y_1, Y_2) \sim \mathcal{N}(0, 1)$ holds. The decision rule is given by: H_0 is rejected if $|T(Y_1, Y_2)| > z_{1-\alpha/2}$.

- **with unknown but equal variances**
  Let $X_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $i = 1, 2$, $j = 1, \ldots, n_i$. Here, σ_1^2, σ_2^2 are unknown, $\sigma_1^2 = \sigma_2^2$ and X_{ij} are independent for all i, j.
  The hypotheses are given by $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$. Consider the test statistic

$$T(Y_1, Y_2) = \frac{\overline{X}_{1n_1} - \overline{X}_{2n_2}}{S_{n_1 n_2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}},$$

  where $S_{n_1 n_2}^2$ is given by

$$\frac{1}{n_1 + n_2 - 2} \cdot \left(\sum_{j=1}^{n_1} \left(X_{1j} - \overline{X}_{1n_1} \right)^2 + \sum_{j=1}^{n_2} \left(X_{2j} - \overline{X}_{2n_2} \right)^2 \right).$$

  It can be shown that under H_0 $T(Y_1, Y_2) \sim t_{n_1+n_2-2}$ holds. The decision rule is then given by: H_0 is rejected if $|T(Y_1, Y_2)| > t_{n_1+n_2-2,1-\alpha/2}$.

(b) **Test for the equality of the expected value for linked random samples**

Let $Y_1 = (X_{11}, \ldots, X_{1n})$ and $Y_2 = (X_{21}, \ldots, X_{2n})$, $n_1 = n_2 = n$,

$$Z_j = X_{1j} - X_{2j} \sim \mathcal{N}(\mu_1 - \mu_2, \sigma^2),\ j = 1, \ldots, n,$$

be i.i.d. with $\mu_i = \mathbb{E}\, X_{ij}$, $i = 1, 2$. The hypotheses are given by: $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$ with unknown variance σ^2. The test statistic is given by

$$T(Z_1, \ldots, Z_n) = \sqrt{n} \frac{\overline{Z}_n}{S_n},$$

where

$$S_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(Z_j - \overline{Z}_n\right)^2.$$

Under H_0, $T(Z_1,\ldots,Z_n) \sim t_{n-1}$ holds. The decision rule is given by: H_0 is rejected, if $|T(Z_1,\ldots,Z_n)| > t_{n-1,1-\alpha/2}$.

(c) **Test for the equality of variances for independent Gaussian random samples**

Let $Y_1 = (X_{11},\ldots,X_{1n_1})$ and $Y_2 = (X_{21},\ldots,X_{2n_2})$ be i.i.d. with $X_{ij} \sim \mathcal{N}(\mu_i,\sigma_i^2)$, where μ_i and σ_i^2 are both unknown. The hypotheses are $H_0 : \sigma_1^2 = \sigma_2^2$ vs. $H_1 : \sigma_1^2 \neq \sigma_2^2$. The test statistic is given by

$$T(Y_1,Y_2) = \frac{S_{2n_2}^2}{S_{1n_1}^2},$$

where

$$S_{in_i}^2 = \frac{1}{n_i-1}\sum_{j=1}^{n}\left(X_{ij} - \overline{X}_{in_i}\right)^2,\ i = 1,2.$$

Under H_0, $T(Y_1,Y_2) \sim F_{n_2-1,n_1-1}$ holds. The decision rule is then given by: H_0 is rejected, if

$$T(Y_1,Y_2) \notin \left[F_{n_2-1,n_1-1,\alpha/2},\, F_{n_2-1,n_1-1,1-\alpha/2}\right].$$

(d) **Asymptotic two sample tests**

- **for Bernoulli distributed random samples**
  Let $X_{ij} \sim \text{Bernoulli}(p_i)$, $j = 1,\ldots,n_i$, $p_i \in (0,1)$, $i = 1,2$. The hypotheses are given by $H_0 : p_1 = p_2$ vs. $H_1 : p_1 \neq p_2$. The test statistic is then given by

$$T(Y_1,Y_2) = \begin{cases} \dfrac{\overline{X}_{1n_1} - \overline{X}_{2n_2}}{\sqrt{\frac{\overline{X}_{1n_1}(1-\overline{X}_{1n_1})}{n_1} + \frac{\overline{X}_{2n_2}(1-\overline{X}_{2n_2})}{n_2}}} & \\ 0, & \overline{X}_{1n_1} = \overline{X}_{2n_2} \in \{0,1\} \end{cases}$$

  Under H_0, $T(Y_1,Y_2) \xrightarrow[n_1,n_2\to\infty]{d} Y \sim \mathcal{N}(0,1)$ holds. The decision rule is then given by: H_0 is rejected if $|T(Y_1,Y_2)| > z_{1-\alpha/2}$. This is a test with asymptotic confidence level α.
- **for Poisson distributed random samples**

Let X_{ij} be independent, $X_{ij} \sim \text{Poisson}(\lambda_i)$, $\lambda_i > 0$, $i = 1, 2$. The hypotheses are: $H_0 : \lambda_1 = \lambda_2$ vs. $H_1 : \lambda_1 \neq \lambda_2$ and the test statistic is given by:

$$T(Y_1, Y_2) = \begin{cases} \frac{\overline{X}_{1n_1} - \overline{X}_{2n_2}}{\sqrt{\frac{\overline{X}_{1n_1}}{n_1} + \frac{\overline{X}_{2n_2}}{n_2}}} \\ 0, \qquad \overline{X}_{1n_1} = \overline{X}_{2n_2} = 0 \end{cases}$$

The decision rule is then given by: H_0 is rejected, if $|T(Y_1, Y_2)| > z_{1-\alpha/2}$. This is a test with asymptotic confidence level α.

Remark 3.2.3. Asymptotic tests must only be used for large samples, since for small samples the, asymptotic significance level can not be assured.

3.3 Randomized test

In this section, classical results of Neyman-Pearson with respect to the terminology of *most powerful tests* are presented. Here, randomized tests play a considerably important role.

3.3.1 Fundamentals

Let $(X_1, \ldots, X_n)$ be a random sample of i.i.d. random variables X_i and $(x_1, \ldots, x_n)$ a realization of $(X_1, \ldots, X_n)$. Assume that that the sample space $(B, \mathcal{B})$ is either given by $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ or $(\mathbb{N}_0^n, \mathcal{B}_{\mathbb{N}_0^n})$ depending on whether the distribution of X_i, $i = 1, \ldots, n$ is either absolutely continuous or discrete. If the random variables X_i are discrete, the domain is assumed to be $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. The domain is equipped with a measure μ, where

$$\mu = \begin{cases} \text{Lebesgue measure on } \mathbb{R}, & \text{if } B = \mathbb{R}^n, \\ \text{Counting measure on } \mathbb{N}_0, & \text{if } B = \mathbb{N}_0^n. \end{cases}$$

Thus

$$\int g(x)\mu(dx) = \begin{cases} \int_{\mathbb{R}} g(x)dx, & \text{in the absolutely continuous case,} \\ \sum_{x \in \mathbb{N}_0} g(x), & \text{in the discrete case,} \end{cases}$$

holds. Moreover, assume that $X_i \sim F_\theta$, $\theta \in \Theta \subseteq \mathbb{R}^m$, $i = 1, \ldots, n$ (parametric model). For $\Theta = \Theta_0 \cup \Theta_1$, $\Theta_0 \cap \Theta_1 = \emptyset$ the hypotheses are $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$, which are tested via the randomized test

$$\varphi(x) = \begin{cases} 1, & x \in K_1, \\ \gamma \in (0, 1), & x \in K_{01}, \qquad x = (x_1, \ldots, x_n), \\ 0, & x \in K_0. \end{cases}$$

If $x \in K_{01}$, then a random variable $Y \sim \text{Bernoulli}(\gamma)$ can be used in order to decide whether H_0 is rejected ($Y = 1$) or not ($Y = 0$).

Definition 3.3.1.

1. The *power (or performance) function* of a randomized test φ is given by

$$G_n(\theta) = G_n(\varphi, \theta) = \mathbb{E}_\theta \, \varphi(X_1, \ldots, X_n), \; \theta \in \Theta.$$

2. The test φ has the *significance level* $\alpha \in [0, 1]$ if $G_n(\varphi, \theta) \leq \alpha$, for all $\theta \in \Theta_0$. The number

$$\sup_{\theta \in \Theta_0} G_n(\varphi, \theta)$$

is called *scope* of the test φ. It obviously holds that the scope of an α confidence level test is smaller than or equal to α.

3. Let $\Psi(\alpha)$ be the set of all test with confidence level α. The test $\varphi_1 \in \Psi(\alpha)$ is called *(uniformly) more powerful* than the test $\varphi_2 \in \Psi(\alpha)$ if $G_n(\varphi_1, \theta) \geq G_n(\varphi_2, \theta)$, $\theta \in \Theta_1$, i.e., if φ_1 has a larger power.

4. A test $\varphi^* \in \Psi(\alpha)$ is called *(uniform) most powerful test* in $\Psi(\alpha)$ if

$$G_n(\varphi^*, \theta) \geq G_n(\varphi, \theta), \text{ for all tests } \varphi \in \Psi(\alpha), \; \theta \in \Theta_1.$$

Remark 3.3.2.

1. Definition 3.3.1 1. is a generalization of Definition 3.1.4, since for $\varphi(x) = I(x \in K_1)$,

$$\begin{aligned} G_n(\varphi, \theta) &= \mathbb{E}_\theta \, \varphi(X_1, \ldots, X_n) \\ &= P_\theta \left((X_1, \ldots, X_n) \in K_1 \right) \\ &= P_\theta \left(\text{reject } H_0 \right), \; \theta \in \Theta \end{aligned}$$

holds.

2. A most powerful test φ^* in $\Psi(\alpha)$ does not always exist. It only exists under certain conditions on $P_\theta, \Theta_0, \Theta_1$ and $\Psi(\alpha)$.

3.3.2 Neyman-Pearson test for simple hypotheses

In this section, simple hypotheses of the form

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta = \theta_1, \tag{3.1}$$

where $\theta_0, \theta_1 \in \Theta$, $\theta_1 \neq \theta_0$ are considered.

Therefore, $\Theta_0 = \{\theta_0\}$, $\Theta_1 = \{\theta_1\}$. Assume that F_{θ_i} has a probability density function $g_i(x)$ with respect to μ for $i = 0, 1$. From now on the notation $P_0 = P_{\theta_0}$, $P_1 = P_{\theta_1}$, $\mathbb{E}_0 = \mathbb{E}_{\theta_0}$, $\mathbb{E}_1 = \mathbb{E}_{\theta_1}$ will be used. Let $f_i(x) = \prod_{j=1}^n g_i(x_j)$, $x = (x_1, \ldots, x_n)$, $i = 0, 1$ be the density of the random sample under H_0 resp. H_1.

Definition 3.3.3. A *Neyman-Pearson test* (*NP test*) of simple hypotheses as in (3.1) is given by the rule

$$\varphi(x) = \varphi_K(x) = \begin{cases} 1, & \text{if } f_1(x) > K f_0(x), \\ \gamma, & \text{if } f_1(x) = K f_0(x), \\ 0, & \text{if } f_1(x) < K f_0(x), \end{cases} \tag{3.2}$$

for constants $K > 0$ and $\gamma \in [0, 1]$.

Remark 3.3.4.

1. Sometimes $K = K(x)$ and $\gamma = \gamma(x)$ are seen as functions of x and not as constants.

2. The *rejection region* of the Neyman-Pearson tests φ_K is

$$K_1 = \{x \in B : f_1(x) > K f_0(x)\}.$$

3. The *scope* of the Neyman-Pearson tests φ_K is given by

$$\begin{aligned} \mathbb{E}_0\, \varphi_K(X_1, \ldots, X_n) = P_0\big(&f_1(X_1, \ldots, X_n) \\ &> K f_0(X_1, \ldots X_n)\big) \\ &+ \gamma P_0\big(f_1(X_1, \ldots, X_n) = K f_0(X_1, \ldots, X_n)\big). \end{aligned}$$

4. Definition 3.3.3 can be given equivalently by defining the test statistic

$$T(x) = \begin{cases} \frac{f_1(x)}{f_0(x)}, & x \in B : f_0(x) > 0, \\ \infty, & x \in B : f_0(x) = 0. \end{cases}$$

Then the new test given by

$$\tilde{\varphi}_K(x) = \begin{cases} 1, & \text{if } T(x) > K, \\ \gamma, & \text{if } T(x) = K, \\ 0, & \text{if } T(x) < K, \end{cases}$$

can be introduced, which is for P_0- and P_1-almost all $x \in B$ equivalent to φ_k. $\varphi_K(x) = \tilde{\varphi}_K(x)\, \forall x \in B \setminus C$ holds, where $C = \{x \in B : f_0(x) = f_1(x) = 0\}$ has P_0- resp. P_1- measure zero.

Using this new formulation, the scope of φ resp. $\tilde{\varphi}_K$ is given by

$$\mathbb{E}_0\, \tilde{\varphi}_K = P_0(T(X_1, \ldots, X_n) > K) + \gamma \cdot P_0\left(T(X_1, \ldots, X_n) = K\right).$$

Theorem 3.3.5. (Optimality theorem) Let φ_K be a Neyman-Pearson test for a $K > 0$ and $\gamma \in [0,1]$, then φ_K is the most powerful test with confidence level $\alpha = \mathbb{E}_0\, \varphi_K$ of its scope.

Proof Let $\varphi \in \Psi(\alpha)$, i.e., $\mathbb{E}_0\,(\varphi(X_1,\ldots,X_n)) \leq \alpha$. In order to show that φ_K is more powerful than φ, it is sufficient to show for simple hypotheses H_0 and H_1, that $\mathbb{E}_1\,\varphi_K(X_1,\ldots,X_n) \geq \mathbb{E}_1\,\varphi(X_1,\ldots,X_n)$. Define the sets:

$$\begin{aligned} M^+ &:= \{x \in B : \varphi_K(x) > \varphi(x)\} \\ M^- &:= \{x \in B : \varphi_K(x) < \varphi(x)\} \\ M^= &:= \{x \in B : \varphi_K(x) = \varphi(x)\} \end{aligned}$$

It obviously holds that

$$\begin{aligned} &x \in M^+ \Rightarrow \varphi_K(x) > 0 \Rightarrow f_1(x) \geq K f_0(x), \\ &x \in M^- \Rightarrow \varphi_K(x) < 1 \Rightarrow f_1(x) \leq K f_0(x) \\ &B = M^+ \cup M^- \cup M^=. \end{aligned}$$

Hence

$$\begin{aligned} \mathbb{E}_1\,(\varphi_K(X_1,\ldots,X_n) - \varphi(X_1,\ldots,X_n)) &= \int_B (\varphi_K(x) - \varphi(x)) f_1(x)\mu(dx) \\ &= \left(\int_{M^+} + \int_{M^-} + \int_{M^=}\right)(\varphi_K(x) - \varphi(x)) f_1(x)\mu(dx) \\ &\geq \int_{M^+} (\varphi_K(x) - \varphi(x)) K f_0(x)\mu(dx) \\ &\quad + \int_{M^-} (\varphi_K(x) - \varphi(x)) K f_0(x)\mu(dx) \\ &= \int_B (\varphi_K(x) - \varphi(x)) K f_0(x)\mu(dx) \\ &= K\Big(\mathbb{E}_0\,\varphi_K(X_1,\ldots,X_n) - \mathbb{E}_0\,\varphi(X_1,\ldots,X_n)\Big) \\ &\geq K(\alpha - \alpha) = 0, \end{aligned}$$

since both tests obtain confidence level α. □

Remark 3.3.6.

1. As γ does not appear in the proof, the same result holds for $\gamma(x) \neq$ const.

2. The proof implies the inequality given by

$$\int_B (\varphi_K(x) - \varphi(x))\,(f_1(x) - K f_0(x))\,\mu(dx) \geq 0$$

if K is constant, resp.

$$\mathbb{E}_1\left(\varphi_K(X_1,\ldots,X_n)-\varphi(X_1,\ldots,X_n)\right)\geq \int_B\left(\varphi_K(x)-\varphi(x)\right)K(x)f_0(x)\mu(dx)$$

in the general case.

Theorem 3.3.7. (Fundamental lemma of Neyman-Pearson)

1. For an arbitrary $\alpha\in(0,1)$, there exists a Neyman-Pearson test φ_K with scope α, which is by Theorem 3.3.5 the most powerful α confidence level test.
2. If φ is also a most powerful test with confidence level α, then $\varphi(x)=\varphi_K(x)$ for μ-almost all $x\in K_0\cup K_1=\{x\in B: f_1(x)\neq Kf_0(x)\}$ and φ_K of part 1.

Proof 1. For $\varphi_K(x)$ it holds that

$$\varphi_K(x)=\begin{cases}1, & \text{if } x\in K_1=\{x: f_1(x)>K\cdot f_0(x)\},\\ \gamma, & \text{if } x\in K_{01}=\{x: f_1(x)=K\cdot f_0(x)\},\\ 0, & \text{if } x\in K_0=\{x: f_1(x)<K\cdot f_0(x)\}.\end{cases}$$

The scope of φ_K is given by

$$P_0\left(T(X_1,\ldots,X_n)>K\right)+\gamma P_0\left(T(X_1,\ldots,X_n)=K\right)=\alpha, \tag{3.3}$$

where

$$T(x_1,\ldots,x_n)=\begin{cases}\frac{f_1(x_1,\ldots,x_n)}{f_0(x_1,\ldots,x_n)}, & \text{if } f_0(x_1,\ldots,x_n)>0,\\ \infty, & \text{otherwise.}\end{cases}$$

The goal is to find a $K>0$ and a $\gamma\in[0,1]$, such that equation (3.3) holds. Let $\tilde{F}_0(x)=P_0(T(X_1,\ldots,X_n)\leq x)$, $x\in\mathbb{R}$ be the distribution function of T. Since $T\geq 0$, it holds that $\tilde{F}_0(x)=0$, if $x<0$. Furthermore, $P_0(T(X_1,\ldots,X_n)<\infty)=1$, which means that $\tilde{F}^{-1}(\alpha)\in[0,\infty)$, $\alpha\in(0,1)$. Equation (3.3) can then be rewritten as

$$1-\tilde{F}_0(K)+\gamma\left(\tilde{F}_0(K)-\tilde{F}_0(K-)\right)=\alpha, \tag{3.4}$$

where $\tilde{F}_0(K-)=\lim\limits_{x\nearrow K}\tilde{F}_0(x)$.

Let $K=\tilde{F}_0^{-1}(1-\alpha)$, then:

(a) If K is a point of continuity of $\tilde{F}_0$, then Equation (3.4) is satisfied for all $\gamma \in [0,1]$, for example for $\gamma = 0$.

(b) If K is not a point of continuity of $\tilde{F}_0$, then $\tilde{F}_0(K) - \tilde{F}_0(K-) > 0$, which implies that

$$\gamma = \frac{\alpha - 1 + \tilde{F}_0(K)}{\tilde{F}_0(K) - \tilde{F}_0(K-)}.$$

Therefore, a Neyman-Pearson test with confidence level α exists.

2. Define $M^{\neq} := \{x \in B : \varphi(x) \neq \varphi_K(x)\}$. It has to be shown that

$$\mu\left((K_0 \cup K_1) \cap M^{\neq}\right) = 0.$$

Consider

$$\begin{aligned}
&\mathbb{E}_1\,\varphi(X_1,\ldots,X_n) - \mathbb{E}_1\,\varphi_K(X_1,\ldots,X_n) = 0 \quad (\varphi \text{ and } \varphi_K \text{ are most powerful tests})\\
&\mathbb{E}_0\,\varphi(X_1,\ldots,X_n) - \mathbb{E}_0\,\varphi_K(X_1,\ldots,X_n) \leq 0 \quad (\varphi \text{ and } \varphi_K \text{ are } \alpha\text{-tests}\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{with scope } \varphi_K = \alpha)
\end{aligned}$$

$$\Rightarrow \int_B (\varphi - \varphi_K) \cdot (f_1 - K \cdot f_0)\, d\mu \geq 0.$$

In Remark 3.3.6 it has been shown, that

$$\begin{aligned}
&\int_B (\varphi - \varphi_K)(f_1 - K \cdot f_0) d\mu \leq 0\\
\Rightarrow &\int_B (\varphi - \varphi_K)(f_1 - K \cdot f_0) d\mu = 0 = \int_{M^{\neq} \cap (K_0 \cup K_1)} (\varphi - \varphi_K)(f_1 - K \cdot f_0) d\mu.
\end{aligned}$$

$\mu(M^{\neq} \cap (K_0 \cup K_1)) = 0$ holds if the integrand $(\varphi - \varphi_K)(f_1 - K \cdot f_0) > 0$ on $(K_0 \cup K_1) \cap M^{\neq}$. We need to show that

$$(\varphi_K - \varphi)(f_1 - K f_0) > 0 \text{ für } x \in (K_0 \cup K_1) \cap M^{\neq}. \tag{3.5}$$

Now,

$$\begin{aligned}
f_1 - K f_0 > 0 &\Rightarrow \varphi_K - \varphi > 0,\\
f_1 - K f_0 < 0 &\Rightarrow \varphi_K - \varphi < 0,
\end{aligned}$$

holds, since

$$\begin{aligned}
&f_1(x) > K f_0(x) \Rightarrow \varphi_K(x) = 1\\
&\text{and with } \varphi(x) < 1 \text{ we get } \varphi_K(x) - \varphi(x) > 0 \text{ on } M^{\neq}.\\
&f_1(x) < K f_0(x) \Rightarrow \varphi_K(x) = 0\\
&\text{and with } \varphi(x) > 0 \text{ we get } \varphi_K(x) - \varphi(x) < 0 \text{ on } M^{\neq}.
\end{aligned}$$

Thus inequality (3.5) holds and finally

$$\mu\left((K_0 \cup K_1) \cap M^{\neq}\right) = 0.$$

□

Remark 3.3.8. If φ and φ_K are most powerful α-tests, then they are P_0- resp. P_1- almost surely equal.

Example 3.3.9. (Neyman-Pearson test for the parameter of the Poisson distribution)
Let $(X_1, \ldots, X_n)$ be a random sample with $X_i \sim \text{Poisson}(\lambda)$, $\lambda > 0$, where X_i are i.i.d. for $i = 1, \ldots, n$. The hypotheses $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda = \lambda_1$ need to be tested. Here

$$g_i(x) = e^{-\lambda_i}\frac{\lambda_i^x}{x!},\ x \in \mathbb{N}_0,\ i = 0, 1,$$

$$f_i(x) = f_i(x_1, \ldots, x_n) = \prod_{j=1}^{n} g_i(x_j) = \prod_{j=1}^{n} e^{-\lambda_i}\frac{\lambda_i^{x_j}}{x_j!} = e^{-n\lambda_i} \cdot \frac{\lambda_i^{\sum_{j=1}^n x_j}}{(x_1! \cdot \ldots \cdot x_n!)},\ i = 0, 1.$$

The Neyman-Pearson test statistic is given by

$$T(x_1, \ldots, x_n) = \begin{cases} \frac{f_1(x)}{f_0(x)} = e^{-n(\lambda_1 - \lambda_0)} \cdot (\lambda_1/\lambda_0)^{\sum_{j=1}^n x_j}, & \text{if } x_1, \ldots, x_n \in \mathbb{N}_0, \\ \infty, & \text{otherwise.} \end{cases}$$

The Neyman-Pearson decision rule is given by

$$\varphi_K(x_1, \ldots, x_n) = \begin{cases} 1, & \text{if } T(x_1, \ldots, x_n) > K, \\ \gamma, & \text{if } T(x_1, \ldots, x_n) = K, \\ 0, & \text{if } T(x_1, \ldots, x_n) < K. \end{cases}$$

Choose $K > 0$, $\gamma \in [0, 1]$, such that φ_K has scope α. In order to do so, solve

$$\alpha = P_0(T(X_1, \ldots, X_n) > K) + \gamma P_0(T(X_1, \ldots, X_n) = K)$$

for γ and K.

$$\begin{aligned} P_0(T(X_1, \ldots, X_n) > K) &= P_0(\log T(X_1, \ldots, X_n) > \log K) \\ &= P_0\left(-n(\lambda_1 - \lambda_0) + \log\left(\frac{\lambda_1}{\lambda_0}\right)\sum_{j=1}^{n} X_j > \log K\right) \\ &= P_0\left(\sum_{j=1}^{n} X_j > A_K\right), \end{aligned}$$

where

$$A_K := \left\lfloor \frac{\log K + n \cdot (\lambda_1 - \lambda_0)}{\log \frac{\lambda_1}{\lambda_0}} \right\rfloor,$$

if for example $\lambda_1 > \lambda_0$. If $\lambda_1 < \lambda_0$ then replace $>$ with $<$ in the argument above.
Due to the stability of the Poisson distribution

$$\sum_{j=1}^{n} X_j \sim \text{Poisson}(n\lambda_0),$$

holds under H_0. Thus choose K as the smallest nonnegative number with

$$P_0\left(\sum_{j=1}^{n} X_j > A_K\right) \leq \alpha,$$

and set

$$\gamma = \frac{\alpha - P_0(\sum_{j=1}^{n} X_j > A_K)}{P_0(\sum_{j=1}^{n} X_j = A_K)},$$

where

$$P_0\left(\sum_{j=1}^{n} X_j > A_K\right) = 1 - \sum_{j=0}^{A_K} e^{-\lambda_0 n} \frac{(\lambda_0 n)^j}{j!},$$
$$P_0\left(\sum_{j=1}^{n} X_j = A_K\right) = e^{-\lambda_0 n} \frac{(\lambda_0 n)^{A_K}}{A_K!}.$$

Hence, the parameters K and γ have been found and a Neyman-Pearson test φ_K has been constructed.

3.3.3 One-sided Neyman-Pearson tests

So far Neyman-Pearson tests for simple hypotheses like $H_i : \theta = \theta_i$, $i = 0, 1$ have been considered. This section aims to introduce one-sided Neyman-Pearson tests for hypotheses of the form $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$.
In an initial step, a test for the following hypotheses is constructed: Let $(X_1, \ldots, X_n)$ be a random sample, X_i i.i.d. with

$$X_i \sim F_\theta \in \Lambda = \{F_\theta : \theta \in \Theta\},$$

where $\Theta \subset \mathbb{R}$ is open and Λ uniquely parameterized, i.e.,

$$\theta \neq \theta' \Rightarrow F_\theta \neq F_{\theta'}.$$

Furthermore, assume that F_θ has a density g_θ w.r.t. the Lebesgue measure (resp. counting measure) μ on $\mathbb{R}$ (resp. $\mathbb{N}_0$). Then

$$f_\theta(x) = \prod_{j=1}^{n} g_\theta(x_j), \quad x = (x_1, \ldots, x_n)$$

is a density of $(X_1, \ldots, X_n)$ with respect to μ^n on B.

Definition 3.3.10. A distribution on B with density f_θ is a member of the class of *distributions with monotone likelihood ratio* in T, if for all $\theta < \theta'$ exist a monotonically increasing function $h : \mathbb{R} \times \Theta^2 \to \mathbb{R} \cup \infty$ on $\mathbb{R}$and a statistic $T : B \to \mathbb{R}$ with the property

$$\frac{f_{\theta'}(x)}{f_\theta(x)} = h(T(x), \theta, \theta'),$$

where

$$h(T(x), \theta, \theta') = \infty, \text{ for all } x \in B : f_\theta(x) = 0 \text{ and } f_{\theta'}(x) > 0.$$

The case $f_\theta(x) = f_{\theta'}(x) = 0$ occurs with probability P_Θ- resp. $P_{\Theta'}$ zero.

Definition 3.3.11. Let Q_θ be a distribution on $(B, \mathcal{B})$ with probability density function f_θ w.r.t. μ. Q_θ is an element of the *one-parametric exponential family* ($\theta \in \Theta \subset \mathbb{R}$ open), if the density is given by:

$$f_\theta(x) = \exp\{c(\theta) \cdot T(x) + a(\theta)\} \cdot l(x), \quad x = (x_1, \ldots, x_n) \in B,$$

where $c(\theta)$ is a monotonically increasing function and $\mathrm{Var}_\theta\, T(X_1, \ldots, X_n) > 0$, $\theta \in \Theta$.

Lemma 3.3.12. Distributions of the one-parametric exponential family have a monotone likelihood ratio.

Proof Let Q_θ be in the one-parametric exponential family with probability density function

$$f_\theta(x) = \exp\{c(\theta) \cdot T(x) + a(\theta)\} \cdot l(x).$$

For $\theta < \theta'$

$$\frac{f_{\theta'}(x)}{f_\theta(x)} = \exp\{(c(\theta') - c(\theta)) \cdot T(x) + a(\theta') - a(\theta)\}$$

is monotone with respect to T, since $c(\theta') - c(\theta) > 0$ because of the monotonicity of $c(\theta)$. Thus f_θ has a monotone likelihood ratio. □

Example 3.3.13.

1. *Normal distributed random samples*

   Let $X_i \sim \mathcal{N}(\mu, \sigma_0^2)$, $i = 1, \ldots, n$ be i.i.d. random variables, with unknown $\mu \in \mathbb{R}$ and known variance $\sigma_0^2 > 0$ (here μ denotes the expected value of X_i and not a measure on $\mathbb{R}$). The probability density function of the vector $X = (X_1, \ldots, X_n)^\top$ is given by

$$\begin{aligned} f_\mu(x) &= \prod_{i=1}^n g_\mu(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma_0^2}} \\ &= \frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (x_i-\mu)^2\right\} \\ &= \frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma_0^2}\left(\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + \mu^2 n\right)\right\} \\ &= \exp\Big(\underbrace{\frac{\mu}{\sigma_0^2}}_{c(\mu)} \cdot \underbrace{\sum_{i=1}^n x_i}_{T(x)} - \underbrace{\frac{\mu^2 n}{2\sigma_0^2}}_{a(\mu)}\Big) \cdot \underbrace{\frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n x_i^2}{2\sigma_0^2}\right)}_{l(x)}. \end{aligned}$$

   Thus $\mathcal{N}(\mu, \sigma_0^2)$ is a member of the one-parametric exponential family with $c(\mu) = \frac{\mu}{\sigma_0^2}$ and $T(x) = \sum_{i=1}^n x_i$.

2. *Binomial distributed random samples*

   Let $X_i \sim \mathrm{Bin}(k, p)$ be i.i.d., $i = 1, \ldots, n$. The parameter $p \in (0, 1)$ is assumed to be unknown. The probability mass function of $X = (X_1, \ldots, X_n)^\top$ is given by

$$\begin{aligned} f_p(x) &= P_p\left(X_i = x_i, i = 1, \ldots, n\right) \\ &= \prod_{i=1}^n \binom{k}{x_i} p^{x_i}(1-p)^{k-x_i} = p^{\sum_{i=1}^n x_i} \cdot \frac{(1-p)^{nk}}{(1-p)^{\sum_{i=1}^n x_i}} \cdot \prod_{i=1}^n \binom{k}{x_i} \\ &= \exp\Big\{\Big(\underbrace{\sum_{i=1}^n x_i}_{T(x)}\Big) \cdot \underbrace{\log\left(\frac{p}{1-p}\right)}_{c(p)} + \underbrace{nk \cdot \log(1-p)}_{a(p)}\Big\} \cdot \underbrace{\prod_{i=1}^n \binom{k}{x_i}}_{l(x)}, \end{aligned}$$

   thus $\mathrm{Bin}(n, p)$ is a member of the one-parametric exponential family with

$$c(p) = \log\left(\frac{p}{1-p}\right)$$

   and

$$T(x) = \sum_{i=1}^n x_i.$$

Lemma 3.3.14. If φ_K is the Neyman-Pearson test of the hypotheses $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$, then

$$\mu^n(K_0 \cup K_1) = \mu^n(\{x \in B : f_1(x) \neq K f_0(x)\}) > 0.$$

Proof Since $\theta_0 \neq \theta_1$ and because of the unique parametrization it holds that $f_0 \neq f_1$ on a set with μ-measure greater than 0.
Assume that $\mu(K_0 \cup K_1) = 0$. Then $f_1(x) = K \cdot f_0(x)$ μ-almost surely, which means that

$$1 = \int_B f_1(x)dx = K \cdot \int_B f_0(x)dx.$$

This yields $K = 1$ and $f_1(x) = f_0(x)$ μ-almost surely, which is a contradiction to the unique parametrization. □

Assume that $(X_1, \ldots, X_n)$ is an i.i.d. random sample, where X_i have the density g_θ, $i = 1, \ldots, n$ and $(X_1, \ldots, X_n)$ has the density $f_\theta(x) = \prod_{i=1}^n g_\theta(x_i)$ from the class of distributions with monotone likelihood ratio and a statistic $T(x_1, \ldots, x_n)$.
Consider the hypotheses $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$ and the Neyman-Pearson test:

$$\varphi^*_{K^*}(x) = \begin{cases} 1, & \text{if } T(x) > K^*, \\ \gamma^*, & \text{if } T(x) = K^*, \\ 0, & \text{if } T(x) < K^* \end{cases} \tag{3.6}$$

for $K^* \in \mathbb{R}$ and $\gamma^* \in [0, 1]$. The power function of $\varphi^*_{K^*}$ at θ_0 is given by

$$G_n(\theta_0) = \mathbb{E}_0\, \varphi^*_{K^*} = P_0\left(T(X_1, \ldots, X_n) > K^*\right) + \gamma^* \cdot P_0\left(T(X_1, \ldots, X_n) = K^*\right).$$

Theorem 3.3.15.

1. If $\alpha = \mathbb{E}_0\, \varphi^*_{K^*}(X_1, \ldots, X_n) > 0$, then the defined test is a most powerful test of the one-sided hypotheses H_0 vs. H_1 with confidence level α.

2. For every confidence level $\alpha \in (0, 1)$ exists a $K^* \in \mathbb{R}$ and $\gamma^* \in [0, 1]$, such that $\varphi^*_{K^*}$ is a most powerful test with scope α.

3. The power function $G_n(\theta)$ of $\varphi^*_{K^*}(\theta)$ is monotonically nondecreasing in θ. If $0 < G_n(\theta) < 1$, then G_n is even monotonically increasing.

Proof

1. Assume $\theta_1 > \theta_0$ and consider the simple hypotheses $H_0' : \theta = \theta_0$ and $H_1' : \theta = \theta_1$. Let

$$\varphi_K(x) = \begin{cases} 1, & f_1(x) > K f_0(x), \\ \gamma, & f_1(x) = K f_0(x), \\ 0, & f_1(x) < K f_0(x), \end{cases}$$

be the Neyman-Pearson test for H_0', H_1' with $K > 0$. Since f_θ has the monotone likelihood ratio with statistic T, i.e.,

$$\frac{f_1(x)}{f_0(x)} = h(T(x), \theta_0, \theta_1),$$

there exists $K > 0$, such that

$$\left\{ x : f_1(x)/f_0(x) \begin{array}{l} > K \\ < K \end{array} \right\} \subset \left\{ T(x) \begin{array}{l} > K^* \\ < K^* \end{array} \right\} \quad \text{with } K = h(K^*, \theta_0, \theta_1).$$

φ_K is a most powerful Neyman-Pearson test with confidence level $\alpha = \mathbb{E}_0 \varphi_K = \mathbb{E}_0 \varphi^*_{K^*}$.

$\alpha > 0$ implies that $K < \infty$, since $K = \infty$ would yield

$$\begin{aligned} 0 < \alpha = \mathbb{E}_0 \varphi_K &\le P_0 \left(T(X_1, \ldots, X_n) \ge K^* \right) \\ &\le P_0 \left(\frac{f_1(X_1, \ldots, X_n)}{f_0(X_1, \ldots, X_n)} = \infty \right) \\ &= P_0 \left(f_1(X_1, \ldots, X_n) > 0, f_0(X_1, \ldots, X_n) = 0 \right) \\ &= \int_B I \left(f_1(x) > 0, f_0(x) = 0 \right) \cdot f_0(x) \mu(dx) \\ &= 0. \end{aligned}$$

For the test $\varphi^*_{K^*}$ in (3.6) it holds that

$$\varphi^*_{K^*}(x) = \begin{cases} 1, & \text{if } f_1(x)/f_0(x) > K, \\ \gamma^*(x), & \text{if } f_1(x)/f_0(x) = K, \\ 0, & \text{if } f_1(x)/f_0(x) < K, \end{cases}$$

where $\gamma^*(x) \in \{\gamma^*, 0, 1\}$. Thus $\varphi^*_{K^*}$ is a most powerful Neyman-Pearson test for H_0' vs. H_1' (cf. Remark 3.3.4, 1. and Remark 3.3.6) for an arbitrary $\theta_1 > \theta_0$. That is why $\varphi^*_{K^*}$ is a most powerful Neyman-Pearson test for $H_0'' : \theta = \theta_0$ vs. $H_1'' : \theta > \theta_0$.

The same assertion is obtained by part 3. of the theorem for $H_0 : \theta \le \theta_0$ vs. $H_1 : \theta > \theta_0$, since then $G_n(\theta) \le G_n(\theta_0) = \alpha$ for all $\theta < \theta_0$.

2. See proof of Theorem 3.3.7, 1.).

3. It has to be shown that $G_n(\theta)$ is monotone. In order to do so, let $\theta_1 < \theta_2$ and show that $\alpha_1 = G_n(\theta_1) \leq G_n(\theta_2)$. Consider the new, simple hypotheses $H_0'' : \theta = \theta_1$ vs. $H_1'' : \theta = \theta_2$. The test $\varphi^*_{K^*}$ can similarly to 1. be stated as a Neyman-Pearson test (for the hypotheses H_0'' and H_1''), which is a most powerful test with confidence level α_1. Consider another constant test $\varphi(x) = \alpha_1$. Then $\alpha_1 = \mathbb{E}_{\theta_2} \varphi \leq \mathbb{E}_{\theta_2} \varphi^*_{K^*} = G_n(\theta_2)$. This implies that $G_n(\theta_1) \leq G_n(\theta_2)$.

   It is now to be shown that for $G_n(\theta) \in (0,1)$ it holds that $G_n(\theta_1) < G_n(\theta_2)$. Assume that $\alpha_1 = G_n(\theta_1) = G_n(\theta_2)$ and $\theta_1 < \theta_2$ for $\alpha_1 \in (0,1)$. Then $\varphi(x) = \alpha_1$ is also a most powerful test for H_0'' and H_1''. Theorem 3.3.7, 2. implies

$$\mu^n(\{x \in K_0 \cup K_1 : \underbrace{\varphi(x)}_{=\alpha_1} \neq \varphi^*_{K^*}(x)\}) = 0$$

   which is a contradiction to the construction of the test $\varphi^*_{K^*}$. This test can not be equal to $\alpha_1 \in (0,1)$ on $K_0 \cup K_1$.

□

Remark 3.3.16.

1. Theorem 3.3.15 can be applied to the Neyman-Pearson tests of the one-sided hypotheses

$$H_0 : \theta \geq \theta_0 \text{ vs. } H_1 : \theta < \theta_0,$$

   with the corresponding difference

$$\begin{aligned} \theta &\mapsto -\theta \\ T &\mapsto -T. \end{aligned}$$

   Thus the most powerful α test also exists in that case.

2. It can be shown that the power function $G_n(\varphi^*_{K^*}, \theta)$ of the most powerful Neyman-Pearson tests on $\Theta_0 = (-\infty, \theta_0)$ attains the following minimization property:

$$G_n(\varphi^*_{K^*}, \theta) \leq G_n(\varphi, \theta) \quad \forall \varphi \in \Psi(\alpha),\ \theta \leq \theta_0.$$

Example 3.3.17. Consider a normally distributed random sample $(X_1, \ldots, X_n)$ of i.i.d. random variables X_i with $X_i \sim \mathcal{N}(\mu, \sigma_0^2)$ and known $\sigma_0^2 > 0$. The hypotheses

$$H_0 : \mu \leq \mu_0 \text{ vs. } H_1 : \mu > \mu_0$$

are tested. Example 3.1.7 provides the test statistic

$$T(X_1, \ldots, X_n) = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma_0},$$

where under H_0 it holds that $T(X_1, \ldots, X_n) \sim \mathcal{N}(0, 1)$. H_0 is rejected, if

$$T(X_1, \ldots, X_n) > z_{1-\alpha}, \quad \text{with } \alpha \in (0, 1).$$

It will be shown that this test is the most powerful Neyman-Pearson test with confidence level α. Example 3.3.13 implies that the probability density function f_n of $(X_1, \ldots, X_n)$ is a member of the one-parametric exponential family with

$$\tilde{T}(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i.$$

Then f_μ of $(x_1, \ldots, x_n)$ is also a member of the one-parametric exponential family with respect to the statistic

$$T(X_1, \ldots, X_n) = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma_0},$$

since it holds that

$$\begin{aligned} f_\mu(x) &= \exp\Big(\underbrace{\frac{\mu}{\sigma_0^2}}_{\tilde{c}(\mu)} \cdot \underbrace{\sum_{i=1}^{n} x_i}_{\tilde{T}} - \underbrace{\frac{\mu^2 n}{2\sigma_0^2}}_{\tilde{a}(\mu)} \Big) \cdot l(x) \\ &= \exp\Big(\underbrace{\frac{\mu\sqrt{n}}{\sigma_0}}_{c(\mu)} \cdot \underbrace{\sqrt{n}\frac{\overline{x}_n - \mu}{\sigma_0}}_{T} + \underbrace{\frac{\mu^2 n}{2\sigma_0^2}}_{a(\mu)} \Big) \cdot l(x). \end{aligned}$$

The statistic T can be used in the construction of the Neyman-Pearson tests (cf. Equation (3.6)):

$$\varphi^*_{K^*}(x) = \begin{cases} 1, & \text{if } T(x) > z_{1-\alpha}, \\ 0, & \text{if } T(x) = z_{1-\alpha}, \\ 0, & \text{if } T(x) < z_{1-\alpha} \end{cases}$$

(with $K^* = z_{1-\alpha}$ and $\gamma^* = 0$). Theorem 3.3.15 implies that this test is the most powerful Neyman-Pearson test with confidence level α for our hypotheses:

$$\begin{aligned} G_n(\varphi_{K^*}, \mu_0) &= P_0\left(T(X_1, \ldots, X_n) > z_{1-\alpha}\right) + 0 \cdot P_0\left(T(X_1, \ldots, X_n) \leq z_{1-\alpha}\right) \\ &= 1 - \Phi(z_{1-\alpha}) = 1 - (1 - \alpha) = \alpha. \end{aligned}$$

3.3.4 Unbiased two-sided tests

Let $(X_1, \ldots, X_n)$ be a random sample of i.i.d. random variables with probability density function

$$f_\theta(x) = \prod_{i=1}^{n} g_\theta(x_i).$$

In the following, a two-sided test of the hypotheses

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0$$

is considered. There cannot be a most powerful test φ with confidence level α for all $\alpha \in (0, 1)$. Assume that φ is the most powerful with confidence level α for H_0 vs. H_1, then φ would be the most powerful test for the hypotheses

1. $H_0' : \theta = \theta_0$ vs. $H_1' : \theta > \theta_0$
2. $H_0'' : \theta = \theta_0$ vs. $H_1'' : \theta < \theta_0$.

By Theorem 3.3.15, 3. the power function would then be given by

1. $G_n(\varphi, \theta) < \alpha$ on $\theta < \theta_0$, resp.
2. $G_n(\varphi, \theta) > \alpha$ on $\theta < \theta_0$,

which is a contradiction!
That is why the class of all possible tests is reduced to the class of unbiased tests (cf. Definition 3.1.12). The test φ is unbiased if and only if

$$G_n(\varphi, \theta) \leq \alpha \text{ for } \theta \in \Theta_0 \text{ and}$$
$$G_n(\varphi, \theta) \geq \alpha \text{ for } \theta \in \Theta_1.$$

Example 3.3.18.

1. $\varphi(x) \equiv \alpha$ is unbiased.
2. The two-sided Gauss test is unbiased, (cf. Example 3.1.7): $G_n(\varphi, \mu) \geq \alpha$ for all $\mu \in \mathbb{R}$.

Assume that X_i are i.i.d. The probability density function f_θ of $(X_1, \ldots, X_n)$ is assumed to be a member of the one-parametric exponential family

$$f_\theta(x) = \exp\{c(\theta) \cdot T(x) + a(\theta)\} \cdot l(x), \tag{3.7}$$

where $c(\theta)$ and $a(\theta)$ are continuously differentiable on Θ with

$$c'(\theta) > 0 \quad \text{and} \quad \operatorname{Var}_\theta T(X_1, \ldots, X_n) > 0$$

for all $\theta \in \Theta$. Let $f_\theta(x)$ be continuous in (x, θ) on $B \times \Theta$.

Exercise 3.3.19. Show that the following relation holds:

$$a'(\theta) = -c'(\theta)\mathbb{E}_\theta\, T(X_1, \ldots, X_n).$$

Lemma 3.3.20. Let φ be an unbiased test with confidence level α for

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0.$$

Then

1. $\alpha = \mathbb{E}_0\, \varphi(X_1, \ldots, X_n) = G_n(\varphi, \theta_0)$,
2. $\mathbb{E}_0\, [T(X_1, \ldots, X_n)\varphi(X_1, \ldots, X_n)] = \alpha \cdot \mathbb{E}_0\, T(X_1, \ldots, X_n)$,

Proof

1. The power function of φ is given by

$$G_n(\varphi, \theta) = \int_B \varphi(x) f_\theta(x) \mu n(dx).$$

Since f_θ is in the one-parametric exponential family, $G_n(\varphi, \theta)$ is differentiable (under the integral) with respect to θ and hence continuous in θ. Since φ is unbiased, it holds that

$$G_n(\varphi, \theta_0) \leq \alpha,\ G_n(\varphi, \theta) \geq \alpha, \quad \theta \neq \theta_0$$

Thus $G_n(\varphi, \theta_0) = \alpha$ and θ_0 minimizes G_n, which proves 1.

2. Since θ_0 minimizes G_n, it holds that

$$\begin{aligned} 0 = G_n'(\varphi, \theta_0) &= \int_B \varphi(x)(c'(\theta_0)T(x) + a'(\theta_0)) f_0(x)\mu(dx) \\ &= c'(\theta_0) \cdot \mathbb{E}_0\, [\varphi(X_1, \ldots X_n)T(X_1, \ldots, X_n)] + a'(\theta_0) \cdot G_n(\varphi, \theta_0) \\ &= c'(\theta_0) \cdot \mathbb{E}_0\, [\varphi(X_1, \ldots, X_n)T(X_1, \ldots, X_n)] + \alpha a'(\theta_0) \\ &\overset{\text{(Exerc. 3.3.19)}}{=} c'(\theta_0)\, (\mathbb{E}_0\, (\varphi \cdot T) - \alpha \mathbb{E}_0\, T) \end{aligned}$$

Therefore, $\mathbb{E}_0\, (\varphi T) = \alpha \mathbb{E}_0\, T$.

□

In the following paragraph, a modification of the Neyman-Pearson test for simple hypotheses of the form

$$H_0 : \theta = \theta_0 \text{ vs. } H_1' : \theta = \theta_1, \quad \theta_1 \neq \theta_0,$$

is introduced. For $\lambda, K \in \mathbb{R}$, $\gamma : B \to [0, 1]$, define

$$\varphi_{K,\lambda}(x) = \begin{cases} 1, & \text{if } f_1(x) > (K + \lambda T(x)) f_0(x), \\ \gamma(x), & \text{if } f_1(x) = (K + \lambda T(x)) f_0(x), \\ 0, & \text{if } f_1(x) < (K + \lambda T(x)) f_0(x), \end{cases} \tag{3.8}$$

where $T(x)$ is the statistic in (3.7).
Let $\tilde{\Psi}(\alpha)$ be the class of all tests that satisfy the conditions 1. and 2. of Lemma 3.3.20. Lemma 3.3.20 implies that the set of unbiased tests with confidence level α is a subset of $\tilde{\Psi}(\alpha)$.

Theorem 3.3.21. The modified Neyman-Pearson test $\varphi_{K,\lambda}$ is the most powerful α test in $\tilde{\Psi}(\alpha)$ for the hypotheses H_0 vs. H_1' with confidence level $\alpha = \mathbb{E}_0\,\varphi_{K,\lambda}$, if $\varphi_{K,\lambda} \in \tilde{\Psi}(\alpha)$.

Proof It has to be shown that $\mathbb{E}_1\,\varphi_{K,\lambda} \geq \mathbb{E}_1\,\varphi$ for all $\varphi \in \tilde{\Psi}(\alpha)$, resp. $\mathbb{E}_1\,(\varphi_{K,\lambda} - \varphi) \geq 0$. It holds that

$$\begin{aligned}
\mathbb{E}_1\,(\varphi_{K,\lambda} - \varphi) &= \int_B (\varphi_{K,\lambda}(x) - \varphi(x)) f_1(x)\mu(dx) \\
&\overset{\text{(Rem. 3.3.6, 2.))}}{\geq} \int_B (\varphi_{K,\lambda}(x) - \varphi(x))(K + \lambda T(x)) f_0(x)\mu(dx) \\
&= K\Big(\underbrace{\mathbb{E}_0\,\varphi_{K,\lambda}}_{=\alpha} - \underbrace{\mathbb{E}_0\,\varphi}_{=\alpha}\Big) + \lambda\Big(\underbrace{\mathbb{E}_0\,(\varphi_{K,\lambda}\cdot T)}_{\alpha\mathbb{E}_0\,T} - \underbrace{\mathbb{E}_0\,(\varphi\cdot T)}_{=\alpha\cdot\mathbb{E}_0\,T}\Big) \\
&= 0,
\end{aligned}$$

since $\varphi, \varphi_{K,\lambda} \in \tilde{\Psi}(\alpha)$. □

Consider the following decision rule, which will later be used in testing two-sided hypotheses given by

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0,$$

$$\varphi_c(x) = \begin{cases} 1, & \text{if } T(x) \notin (c_1, c_2), \\ \gamma_1, & \text{if } T(x) = c_1, \\ \gamma_2, & \text{if } T(x) = c_2, \\ 0, & \text{if } T(x) \in (c_1, c_2), \end{cases} \tag{3.9}$$

for $c_1 \leq c_2 \in \mathbb{R}$, $\gamma_1, \gamma_2 \in [0,1]$ and the statistic $T(x)$, $x = (x_1, \ldots, x_n) \in B$, which is in the density (3.7). In the following it is shown that φ_c can be rewritten as a Neyman-Pearson test.
For the density

$$f_\theta(x) = \exp\{c(\theta)T(x) + a(\theta)\} \cdot l(x)$$

assume that $l(x) > 0$, $c'(x) > 0$, and $a'(x)$ exists for $\theta \in \Theta$.

Lemma 3.3.22. Let $(X_1, \ldots, X_n)$ be a random sample of i.i.d. random variables with probability density function $f_\theta(x), x \in B$, which is a member of the one-parametric exponential family. Let $T(x)$ be the respective statistic in the exponent of the density f_θ. For arbitrary real numbers $c_1 \leq c_2$, $\gamma_1, \gamma_2 \in [0,1]$ and parameters $\theta_0, \theta_1 \in \Theta : \theta_0 \neq \theta_1$ the test φ_c in (3.9) can be rewritten as a modified Neyman-Pearson test $\varphi_{K,\lambda}$ as in (3.8) with $K, \lambda \in \mathbb{R}$, $\gamma(x) \in [0,1]$.

Proof If the notation

$$f_{\theta_i}(x) = f_i(x), \quad i = 0, 1,$$

is used then

$$\frac{f_1(x)}{f_0(x)} = \exp\Big\{ \underbrace{(c(\theta_1) - c(\theta_0))}_{c} T(x) + \underbrace{a(\theta_1) - a(\theta_0)}_{a} \Big\},$$

and therefore

$$\{x \in B : f_1(x) > (K + \lambda T(x))\, f_0(x)\} = \{x \in B : \exp(cT(x) + a) > K + \lambda T(x)\}.$$

Can one find such K and λ in $\mathbb{R}$ for the line $K + \lambda t$, $t \in \mathbb{R}$, which intersects or touches the convex curve $\exp(ct + a)$ exactly in c_1 and c_2 (if $c_1 \neq c_2$) or in $t = c_1$ (if $c_1 = c_2$) resp.? As it turns out, such K and λ can always be found (cf. Figure 3.6).

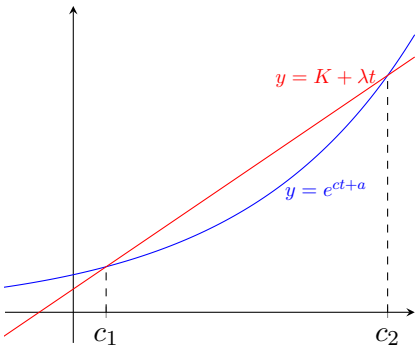


Figure 3.6: Intersection of a line with a convex curve

Let $\gamma(x) = \gamma_i$ for $\{x \in B : T(x) = c_i\}$. Then

$$\{x : \exp(cT(x) + a) > K + \lambda T(x)\} = \{x : T(x) \notin [c_1, c_2]\}$$

and

$$\{x : \exp(cT(x) + a) < K + \lambda T(x)\} = \{x : T(x) \in (c_1, c_2)\}.$$

□

Remark 3.3.23.

1. The inversion of Lemma 3.3.22 does not hold, since for given curves $y = K + \lambda t$ and $y = \exp(ct + a)$ the intersections c_1 and c_2 do not have to exist. The line can be underneath the curve $y = \exp(ct + a)$.

2. The test φ_c does not explicitly use the parameters θ_0 and θ_1, which makes it different from $\varphi_{K,\lambda}$, since it uses the densities f_0 and f_1.

In the following, the fundamental theorem for two-sided tests for the hypotheses

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0$$

will be presented.

Theorem 3.3.24. Fundamental theorem for two-sided tests
Under the conditions of Lemma 3.3.22, let φ_c be a test as in (3.9), for which $\varphi_c \in \tilde{\Psi}(\alpha)$ holds. Then φ_c is the most powerful unbiased test with confidence level α (and thus most powerful test in $\tilde{\Psi}(\alpha)$) for the hypotheses

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0.$$

Proof Let $\theta_1 \in \Theta$, $\theta_1 \neq \theta_0$ be arbitrary. By Lemma 3.3.22, φ_c is a modified Neyman-Pearson test $\varphi_{K,\lambda}$ for a specific choice of K and $\lambda \in \mathbb{R}$, but $\varphi_{K,\lambda}$ is a most powerful test in $\tilde{\Psi}(\alpha)$ by Theorem 3.3.21 for $H_0 : \theta = \theta_0$ vs. $H_1' : \theta = \theta_1$. Since φ_c does not depend on θ_1, it is the most powerful test in $\tilde{\Psi}(\alpha)$ for $H_1 : \theta \neq \theta_0$. Since unbiased tests with confidence level α are in $\tilde{\Psi}(\alpha)$ it only has to be shown that φ_c is unbiased. φ_c is the most powerful test and thus not worse than the constant unbiased test $\varphi = \alpha$, i.e.

$$G_n(\varphi_c, \theta) \geq G_n(\varphi, \theta) = \alpha, \quad \theta \neq \theta_0.$$

Thus φ_c is also unbiased. □

Remark 3.3.25. It has been shown that φ_c is the most powerful test within its scope. It should still be shown that for arbitrary $\alpha \in (0, 1)$ constants $c_1, c_2, \gamma_1, \gamma_2$ can be found, which satisfy $\mathbb{E}_0 \varphi_c = \alpha$. The proof is rather technical and will thus be omitted here. The following example shows how $c_1, c_2, \gamma_1, \gamma_2$ have to be chosen.

Example 3.3.26. Two-sided-Gauss-test
Example 3.1.7 considers the following test for the expectation of a normally distributed random sample $X = (X_1, \ldots, X_n)$ with i.i.d. X_i and $X_i \sim \mathcal{N}(\mu, \sigma_0^2)$ where σ_0^2 is known. The hypotheses

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu \neq \mu_0$$

are tested. The test $\varphi(x)$ is given by

$$\varphi(x) = I\left(x \in \mathbb{R}^n : |T(x)| > z_{1-\alpha/2}\right),$$

where

$$T(x) = \sqrt{n}\frac{\overline{x}_n - \mu_0}{\sigma_0}.$$

It has to be shown that φ is the most powerful test with confidence level α in $\tilde{\Psi}(\alpha)$ (and thus the most powerful unbiased test). By Theorem 3.3.24 it has to be shown that φ can be rewritten as φ_c with (3.9), since the n-dimensional Normal distribution with probability density function f_μ (cf. example 3.3.17) is a member of the one-parametric exponential family with statistic

$$T(x) = \sqrt{n}\frac{\overline{x}_n - \mu}{\sigma_0}.$$

Let $c_1 = -z_{1-\alpha/2}$, $c_2 = z_{1-\alpha/2}$, $\gamma_1 = \gamma_2 = 0$. Then

$$\varphi(x) = \varphi_c(x) = \begin{cases} 1, & \text{if } |T(x)| > z_{1-\alpha/2}, \\ 0, & \text{if } |T(x)| \leq z_{1-\alpha/2}. \end{cases}$$

The assertion is thus proven, since the power function $G_n(\varphi, \theta)$ of φ as in Example 3.1.7 implies, that φ is an unbiased test with confidence level α (and therefore $\varphi \in \tilde{\Psi}(\alpha)$).

Remark 3.3.27. So far, we only assumed that *one* parameter of the distribution of the random sample $(X_1, \ldots, X_n)$ is unknown. This has been necessary in order to be able to introduce the above theory of most powerful (Neyman-Pearson) tests for one–parametric exponential families. In order to consider the case with more unknown parameters (as in the example of two-sided tests for the expected value of a normally distributed random sample with unknown variance), a deeper understanding of randomized tests is needed. If one is interested, the theory can be found in [26].

3.4 Goodness-of-fit tests

Let $(X_1, \ldots, X_n)$ be a random sample of i.i.d. random variables with $X_i \sim F$ for $i = 1, \ldots, n$. Goodness-of-fit testing tests the hypotheses

$$H_0 : F = F_0 \text{ vs. } H_1 : F \neq F_0,$$

where F_0 is a given distribution function.

A goodness-of-fit test has already been introduced in the lecture "Elementary probability theory". The Kolmogorow-Smirnov test can be found in Remark 7.6.8.
In this section, further non-parametric goodness-of-fit tests are introduced. The first one, namely the χ^2-goodness-of-fit test, was introduced by K. Pearson.

3.4.1 χ^2-goodness-of-fit test

The Kolmogorov-Smirnov test is based on the distance

$$D_n = \sup_{x \in \mathbb{R}} \mid \hat{F}_n(x) - F_0(x) \mid$$

between the empirical distribution function of the random sample $(X_1, \ldots, X_n)$ and the distribution function F_0. In practice this test is usually too sensitive, since irregularities in the random samples might lead to an unjustified rejection of H_0. A solution to this problem is a test which coarsens the null hypothesis H_0 and is based on the χ^2-goodness-of-fit statistic.
Partition the domain of X_i into r classes $(a_j, b_j]$, $j = 1, \ldots, r$ with the property

$$-\infty \leq a_1 < b_1 = a_2 < b_2 = \ldots = a_r < b_r \leq \infty.$$

Instead of $X_i, i = 1, \ldots, n$, consider the so-called *class sizes* Z_j, $j = 1, \ldots, r$, where

$$Z_j = \#\{i : a_j < X_i \leq b_j,\, 1 \leq i \leq n\}.$$

Lemma 3.4.1. The random vector $Z = (Z_1, \ldots, Z_r)^\top$ is *multinomial distributed* with parameter vector

$$p = (p_1, \ldots, p_{r-1})^\top \in [0,1]^{r-1},$$

where

$$p_j = P(a_j < X_1 \leq b_j) = F(b_j) - F(a_j),\ j = 1, \ldots, r-1, \quad p_r = 1 - \sum_{j=1}^{r-1} p_j.$$

Notation:

$$Z \sim M_{r-1}(n, p).$$

Proof We show that for all numbers $k_1, \ldots k_r \in \mathbb{N}_0$ with $k_1 + \ldots + k_r = n$

$$P(Z_i = k_i,\, i = 1, \ldots, r) = \frac{n!}{k_1! \cdot \ldots \cdot k_r!} p_1^{k_1} \cdot \ldots \cdot p_r^{k_r} \tag{3.10}$$

holds. Since X_i are i.i.d. it holds that

$$P\left(X_j \in (a_{i_j}, b_{i_j}],\, j = 1, \ldots, n\right) = \prod_{j=1}^{n} P\left(a_{i_j} < X_1 \leq b_{i_j}\right) = p_1^{k_1} \cdot \ldots \cdot p_r^{k_r},$$

if the sequence of intervals $(a_{i_j}, b_{i_j}]_{j=1,\ldots,n}$ contains the interval $(a_i, b_i]$ k_i times, $i = 1, \ldots, r$. The formula (3.10) results from the law of total probability as a sum over all permutations of sequences $(a_{i_j}, b_{i_j}]_{j=1,\ldots,n}$. □

In the sense of Lemma 3.4.1 new hypotheses w.r.t. the nature of F are tested:

$$H_0 : p = p_0 \text{ vs. } H_1 : p \neq p_0,$$

where $p = (p_1, \ldots, p_{r-1})^\top$ is the parameter vector of Z, and $p_0 = (p_{01}, \ldots, p_{0,r-1})^\top \in (0,1)^{r-1}$ with $\sum_{i=1}^{r-1} p_{0i} < 1$. In this case,

$$\Lambda_0 = \{F \in \Lambda : F(b_j) - F(a_j) = p_{0j}, \quad j = 1, \ldots, r-1\}$$

and $\Lambda_1 = \Lambda \setminus \Lambda_0$ holds, where Λ is the set of all distribution functions. In order to test H_0 vs. H_1, the *Pearson test statistic*

$$T_n(x) = \sum_{j=1}^{r} \frac{(z_j - np_{0j})^2}{np_{0j}},$$

where $x = (x_1, \ldots, x_n)$ is an explicit sample and z_j, $j = 1, \ldots, r$ the corresponding class sizes. Under H_0,

$$\mathbb{E}\, Z_j = np_{0j}, \quad j = 1, \ldots, r,$$

holds and thus H_0 is rejected, if $T_n(X)$ attains higher values than expected. The following theorem shows that $T(X_1, \ldots, X_n)$ is asymptotically (for $n \to \infty$) χ^2_{r-1}-distributed, which leads to the following goodness-of-fit test (χ^2 goodness-of-fit test):

$$H_0 \text{ is rejected, if } T_n(x_1, \ldots, x_n) > \chi^2_{r-1,1-\alpha}.$$

This test is named after its inventor *Karl Pearson* (1857-1936).

Theorem 3.4.2. Under H_0,

$$\lim_{n\to\infty} P_{p_0}\left(T_n(X_1, \ldots, X_n) > \chi^2_{r-1,1-\alpha}\right) = \alpha,\ \alpha \in (0,1),$$

holds, which means the χ^2-Pearson test is an asymptotic test with confidence level α.

Proof Denote by $Z_{nj} = Z_j(X_1, \ldots, X_n)$ the class sizes of the random samples $(X_1, \ldots, X_n)$. By Lemma 3.4.1

$$Z_n = (Z_{n1}, \ldots, Z_{nr}) \sim M_{r-1}(n, p_0) \text{ under } H_0$$

holds. Moreover, $\mathbb{E}\, Z_{nj} = np_{0j}$ and

$$\operatorname{Cov}(Z_{ni}, Z_{nj}) = \begin{cases} np_{0j}(1 - p_{0j}), & i = j, \\ -np_{0i}p_{0j}, & i \neq j \end{cases}$$

should hold for all $i, j = 1, \ldots, r$. Since

$$Z_{nj} = \sum_{i=1}^{n} I(a_j < X_i \leq b_j), \quad j = 1, \ldots, r,$$

it holds that $Z_n = (Z_{n1}, \ldots, Z_{n,r-1})$ is the sum of n i.i.d. random vectors with $Y_i \in \mathbb{R}^{r-1}$ with coordinates $Y_{ij} = I(a_j < X_i \leq b_j)$, $j = 1, \ldots, r-1$. Thus, the multivariate limit theorem (which is proven in Lemma 3.4.3) yields

$$Z_n' = \frac{Z_n - \mathbb{E}\, Z_n}{\sqrt{n}} = \frac{\sum\limits_{i=1}^{n} Y_i - n\mathbb{E}\, Y_1}{\sqrt{n}} \xrightarrow[n\to\infty]{d} Y \sim \mathcal{N}(0, K),$$

with $\mathcal{N}(0, K)$ a $(r-1)$ dimensional multivariate normal distribution (cf. [33, Example 3.4.5.3.] with expectation vector 0 and covariance matrix $K = (\sigma_{ij}^2)$, where

$$\sigma_{ij}^2 = \begin{cases} -p_{0i}p_{0j}, & i \neq j, \\ p_{0i}(1 - p_{0j}), & i = j \end{cases}$$

for $i, j = 1, \ldots, r-1$. This matrix K is invertible with $K^{-1} = A = (a_{ij})$,

$$a_{ij} = \begin{cases} \frac{1}{p_{0r}}, & i \neq j, \\ \frac{1}{p_{0i}} + \frac{1}{p_{0r}}, & i = j. \end{cases}$$

Moreover, K (as a covariance matrix) is symmetric and positive semi-definite. Results from Linear Algebra ensure the existence of an invertible $(r-1) \times (r-1)$ matrix $A^{1/2}$, with $A = A^{1/2}(A^{1/2})^\top$. Thus,

$$K = A^{-1} = ((A^{1/2})^\top)^{-1} \cdot (A^{1/2})^{-1}.$$

If $(A^{1/2})^\top$ is applied to Z_n', we get

$$(A^{1/2})^\top \cdot Z_n' \xrightarrow[n\to\infty]{d} (A^{1/2})^\top \cdot Y,$$

where

$$(A^{1/2})^\top \cdot Y \sim \mathcal{N}\left(0,\, (A^{1/2})^\top \cdot K \cdot A^{1/2}\right) = \mathcal{N}\left(0,\, \mathcal{I}_{r-1}\right)$$

by the properties of the multivariate normal distribution. Furthermore, the continuous mapping theorem, which has been introduced in [32, Theorem 3.4.4.], implies that

$$Y_n \xrightarrow[n\to\infty]{d} Y \Longrightarrow \varphi(Y_n) \xrightarrow[n\to\infty]{d} \varphi(Y)$$

for random variables $\{Y_n\}$, $Y \in \mathbb{R}^m$, and continuous mappings $\varphi : \mathbb{R} \to \mathbb{R}$. Repeatedly applying the continuous mapping theorem implies that

$$\left|(A^{1/2})^\top Z_n'\right|^2 \xrightarrow[n\to\infty]{d} \left|(A^{1/2})^\top Y\right|^2 = R \sim \chi^2_{r-1}.$$

It needs to be shown that

$$T_n(X_1, \ldots, X_n) = \left|(A^{1/2})^\top Z_n'\right|^2.$$

Now,

$$\begin{aligned}
\left|(A^{1/2})^\top Z_n'\right|^2 &= ((A^{1/2})^\top Z_n')^\top ((A^{1/2})^\top Z_n') \\
&= Z_n'^\top \cdot \underbrace{A^{1/2} \cdot (A^{1/2})^\top}_{A} Z_n' = Z_n'^\top A Z_n' \\
&= n \sum_{j=1}^{r-1} \frac{1}{p_{0j}} \left(\frac{Z_{nj}}{n} - p_{0j}\right)^2 + \frac{n}{p_{0r}} \sum_{i=1}^{r-1} \sum_{j=1}^{r-1} \left(\frac{Z_{ni}}{n} - p_{0i}\right)\left(\frac{Z_{nj}}{n} - p_{0j}\right) \\
&= \sum_{j=1}^{r-1} \frac{(Z_{nj} - np_{0j})^2}{np_{0j}} + \frac{n}{p_{0r}} \left(\sum_{j=1}^{r-1} \left(\frac{Z_{nj}}{n} - p_{0j}\right)\right)^2 \\
&= \sum_{j=1}^{r-1} \frac{(Z_{nj-np_{0j}})^2}{np_{0j}} + \frac{n}{p_{0r}} \left(\frac{Z_{nr}}{n} - p_{0r}\right)^2 \\
&= \sum_{j=1}^{r} \frac{(Z_{nj} - np_{0j})^2}{np_{0j}} = T_n(X_1, \ldots, X_n),
\end{aligned}$$

since

$$\begin{aligned}
\sum_{j=1}^{r-1} Z_{nj} &= n - Z_{nr}, \\
\sum_{j=1}^{r-1} p_{0j} &= 1 - p_{0r}.
\end{aligned}$$

□

Lemma 3.4.3. Multivariate central limit theorem
Let $\{Y_n\}_{n\in\mathbb{N}}$ be a sequence of i.i.d. random vectors with $\mathbb{E}\,Y_1 = \mu \in \mathbb{R}^m$ and covariance matrix $K \in \mathbb{R}^{m\times m}$. Then

$$\frac{\sum\limits_{i=1}^{n} Y_i - n\mu}{\sqrt{n}} \xrightarrow[n\to\infty]{d} Y \sim N(0, K). \tag{3.11}$$

Proof Let $Y_j = (Y_{j1}, \ldots, Y_{jm})^\top$. By the continuous mapping theorem for characteristic functions the convergence in (3.11) is equivalent to

$$\varphi_n(t) \underset{n\to\infty}{\longrightarrow} \varphi(t), \quad t \in \mathbb{R}^m, \tag{3.12}$$

where

$$\varphi_n(t) = \mathbb{E}\, e^{itS_n} = \mathbb{E}\, \exp\left\{ i \sum_{j=1}^{m} t_j \frac{Y_{1j} + \ldots + Y_{nj} - n\mu_j}{\sqrt{n}} \right\},$$

is the characteristic function of the random vector

$$S_n = \frac{\sum\limits_{i=1}^{n} Y_i - n\mu}{\sqrt{n}},$$

and

$$\varphi(t) = e^{-t^\top K t/2}$$

is the characteristic function of the $\mathcal{N}(0, K)$ distribution. The function $\varphi_n(t)$ can be rewritten as

$$\varphi_n(t) = \mathbb{E}\, \exp\left\{ i \sum_{i=1}^{n} \frac{\sum\limits_{i=1}^{m} t_j (Y_{ij} - \mu_j)}{\sqrt{n}} \right\}, \quad t = (t_1, \ldots, t_m)^\top \in \mathbb{R}^m,$$

where

$$L_i := \sum_{j=1}^{m} t_j (Y_{ij} - \mu_j)$$

is a random variable with

$$\mathbb{E}\, L_i = 0,$$
$$\operatorname{Var} L_i = \mathbb{E}\left[\sum_{k,j=1}^{m} t_j (Y_{ij} - \mu_j)(Y_{ik-\mu_k}) t_k \right] = t^\top K t, \quad i \in \mathbb{N}.$$

If $t^\top Kt = 0$, then $L_i = 0$ almost surely for all $i \in \mathbb{N}$, which implies $\varphi_n(t) = \varphi(t) = 1$. Thus the convergence in (3.11) holds.
If $t^\top Kt > 0$, then $\varphi_n(t)$ is the characteristic function of a random variable

$$\sum_{i=1}^{n} L_i/\sqrt{n}$$

evaluated at 1, and $\varphi(t)$ is the characteristic function of a one-dimensional normal distribution $\mathcal{N}(0, t^\top Kt)$ evaluated at 1. The central limit theorem for one-dimensional random variables then implies (cf. [33, Theorem 5.2.2.])

$$\sum_{i=1}^{n} \frac{L_i}{\sqrt{n}} \xrightarrow[n\to\infty]{d} L \sim \mathcal{N}(0, t^\top Kt)$$

and thus

$$\varphi_n(t) = \varphi_{\left(\sum_{i=1}^{n} L_i/\sqrt{n}\right)}(1) \xrightarrow[n\to\infty]{} \varphi_L(1) = \varphi(t),$$

which proves the convergence in (3.11). □

Remark 3.4.4.

1. The method of reducing a multidimensional convergence to a one-dimensional convergence, using linear combinations of random variables, as in the proof above is called *Cramér-Wold* device.

2. The χ^2-Pearson test works asymptotically for large random samples. Naturally the question of how big n is arises. In this case, the "rule of thumb" is given by: np_{0j} should be larger or equal to a, with $a \in (2, \infty)$. For a larger class number, i.e., $r \geq 10$, even $a = 1$ is sufficient. In the following, it is shown that the χ^2 goodness-of-fit test is consistent.

Lemma 3.4.5. The χ^2-Pearson test is consistent, i.e., for all $p \in [0,1]^{r-1}$, $p \neq p_0$

$$\lim_{n\to\infty} P_p\left(T_n(X_1, \ldots, X_n) > \chi^2_{r-1,1-\alpha}\right) = 1$$

holds.

Proof Under H_1, the strong law of large numbers implies

$$\frac{Z_{nj}}{n} = \frac{\sum_{i=1}^{n} I(a_j < X_i \leq b_j)}{n} \xrightarrow[n\to\infty]{a.s.} \underbrace{\mathbb{E}\, I(a_j < X_1 \leq b_j)}_{=p_j}.$$

Choose j such that $p_j \neq p_{0j}$. Then

$$T_n(X_1, \ldots, X_n) \geq \frac{(Z_{nj} - np_{0j})^2}{np_{0j}} \geq \underbrace{n\left(\frac{Z_{nj}}{n} - p_{0j}\right)^2}_{\sim n(p_j - p_{0j})^2} \xrightarrow[n\to\infty]{a.s.} \infty,$$

and thus

$$P_p\left(T_n(X_1, \ldots, X_n) > \chi^2_{r-1,1-\alpha}\right) \xrightarrow[n\to\infty]{f.s.} 1.$$

□

3.4.2 χ^2-goodness-of-fit test of Pearson-Fisher

Let $(X_1, \ldots, X_n)$ be a random sample of i.i.d. random variables X_i, $i = 1, \ldots, n$. The goal is to test whether the distribution function F of X_i is an element of a given parametric family

$$\Lambda_0 = \{F_\theta : \theta \in \Theta\}, \quad \Theta \subset \mathbb{R}^m.$$

Let a_i, b_i, $i = 1, \ldots, r$ be given with $m < r$,

$$-\infty \leq a_1 < b_1 = a_2 < b_2 = \ldots = a_r < b_r \leq \infty$$

and

$$Z_{nj} = \#\{X_i,\, i = 1, \ldots, n :\, a_j < X_i \leq b_j\}, \quad j = 1, \ldots, r,$$
$$Z_n = (Z_{n1}, \ldots, Z_{nr})^\top.$$

Lemma 3.4.1 implies $Z \sim M_{r-1}(n, p)$, $p = (p_0, \ldots, p_{r-1})^\top \in [0,1]^{r-1}$. Under $H_0 : F \in \Lambda_0$, $p = p(\theta)$, $\theta \in \Theta \subset \mathbb{R}^m$ holds. Presume $p \in C(\Theta)$. By coarsening the hypothesis H_0 the new hypotheses:

$$H_0 : p \in \{p(\theta) : \theta \in \Theta\} \text{ vs. } H_1 : p \notin \{p(\theta) : \theta \in \Theta\}$$

are to be tested. In order to test these hypotheses, the χ^2-*Pearson-Fisher test* is constructed as follows:

1. Find a maximum-likelihood estimator $\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n)$ (weakly consistent) for θ, such that $\hat{\theta}_n \xrightarrow[n\to\infty]{P} \theta$. Here, $\{\hat{\theta}_n\}_{n\in\mathbb{N}}$ is asymptotically normal distributed.

2. Construct the plug-in estimator $p(\hat{\theta}_n)$ for $p(\theta)$.

3. For the test statistic

$$\hat{T}_n(X_1, \ldots, X_n) = \sum_{j=1}^{r} \frac{\left(Z_{nj} - np_j(\hat{\theta})\right)^2}{np_j(\hat{\theta})} \xrightarrow[n\to\infty]{d} \eta \sim \chi^2_{r-m-1}$$

holds under H_0 and certain assumptions.

4. H_0 is rejected, if $\hat{T}_n(X_1, \ldots, X_n) > \chi^2_{r-m-1,1-\alpha}$. This is an asymptotic test with confidence level α.

Remark 3.4.6.

1. The χ^2-Pearson-Fisher test assumes that the function $p(\theta)$ can be stated explicitly, but θ is unknown. That means for every class of distributions Λ_0, the function $p(\cdot)$ has to be calculated.

2. Why is $\hat{T}_n$ able to discriminate between the hypotheses H_0 and H_1? The strong law of large numbers implies

$$\frac{1}{n}Z_{nj} - p_j(\hat{\theta}_n) = \underbrace{\frac{1}{n}Z_{nj} - p_j(\theta)}_{\xrightarrow{P}0} - \underbrace{(p_j(\hat{\theta}_n) - p_j(\theta))}_{\xrightarrow{P}0} \xrightarrow[n\to\infty]{P} 0,$$

   if $\hat{\theta}_n$ is weakly consistent and $p_j(\cdot)$ a continuous function for all $j = 1, \ldots, r$.

   Thus, under H_0 $\hat{T}_n(X_1, \ldots, X_n)$ is supposed to take relatively small values. A significant deviation of this behavior is is supposed to lead to the rejection of H_0.

For the distribution F_θ of X_i the following regularity properties are assumed to hold (cf. Theorem 1.2.22).

1. The distribution function F_θ is either absolutely continuous or discrete for all $\theta \in \Theta$.

2. The parametrization is unique, i.e. $\theta \neq \theta_1 \Leftrightarrow F_\theta \neq F_{\theta_1}$.

3. The support supp $L(x, \theta) = \{x \in \mathbb{R} : L(x, \theta) > 0\}$ of the likelihood function given by

$$L(x, \theta) = \begin{cases} P_\theta(X_1 = x), & \text{in case of discrete } F_\theta, \\ f_\theta(x), & \text{in the absolutely continuous case,} \end{cases}$$

   does not depend on θ.

4. $L(x, \theta)$ is assumed to be three times continuously differentiable and for $k = 1, \ldots, 3$ and $i_1, \ldots, i_k \in \{1, \ldots, m\}$,

$$\sum_{k=1}^{3} \int \frac{\partial^k L(x, \theta)}{\partial\theta_{i_1} \cdot \ldots \cdot \partial\theta_{i_k}} dx = \frac{\partial^k}{\partial\theta_{i_1} \cdot \ldots \cdot \partial\theta_{i_k}} \sum_{k=1}^{3} \int L(x, \theta) dx = 0$$

   holds.

5. For all $\theta_0 \in \Theta$ there exist a constant c_{θ_0} and a measurable function $g_{\theta_0} : \operatorname{supp} L \to \mathbb{R}_+$, such that

$$\left| \frac{\partial^3 \log L(x,\theta)}{\partial\theta_{i_1}\partial\theta_{i_2}\partial\theta_{i_3}} \right| \leq g_{\theta_0}(x), \quad |\theta - \theta_0| < c_{\theta_0},$$

and

$$\mathbb{E}_{\theta_0}\, g_{\theta_0}(X_1) < \infty.$$

Define the *Fisher information matrix* by

$$I(\theta) = \left(\mathbb{E}\left[\frac{\partial \log L(X_1,\theta)}{\partial\theta_i} \frac{\partial \log L(X_1,\theta)}{\partial\theta_j} \right] \right)_{i,j=1,\ldots,m}. \tag{3.13}$$

Theorem 3.4.7. Asymptotical normal distribution of consistent maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$, multivariate case $\boldsymbol{m > 1}$
Let $X_1, \ldots, X_n$ be i.i.d. with likelihood function L, which satisfies the regularity assumptions 1.-5. Let $I(\theta)$ be positive definite for all $\theta \in \Theta \subset \mathbb{R}^m$ and $\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n)$ be a sequence of weakly consistent maximum likelihood estimators for θ. Then

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{d} N(0, I^{-1}(\theta)).$$

Without proof (cf. proof of Theorem 1.2.22).

For the coarsed hypothesis $H_0 : p \in \{p(\theta), \theta \in \Theta\}$ construct the piecewise constant likelihood function

$$L(x,\theta) = p_j(\theta), \text{ if } x \in (a_j, b_j].$$

Then, the likelihood function of the random sample $(x_1, \ldots, x_n)$ is given by

$$L(x_1, \ldots, x_n, \theta) = \prod_{j=1}^{r} p_j(\theta)^{Z_j(x_1,\ldots,x_n)}$$

$$\Rightarrow \log L(x_1, \ldots, x_n, \theta) = \sum_{j=1}^{r} Z_j(x_1, \ldots, x_n) \cdot \log p_j(\theta).$$

For the maximum likelihood estimator, we get

$$\hat{\theta}_n = \hat{\theta}(x_1, \ldots, x_n) = \operatorname*{argmax}_{\theta\in\Theta} \log L(x_1, \ldots, x_n, \theta)$$

$$\Rightarrow \sum_{j=1}^{r} Z_j(x_1, \ldots, x_n) \frac{\partial p_j(\theta)}{\partial\theta_i} \cdot \frac{1}{p_j(\theta)} = 0, \quad i = 1, \ldots, m.$$

Furthermore, the property $\sum_{j=1}^{r} p_j(\theta) = 1$ implies

$$\sum_{j=1}^{r} \frac{\partial p_j(\theta)}{\partial\theta_i} = 0 \Rightarrow \sum_{j=1}^{r} \frac{Z_j(x_1, \ldots, x_n) - n p_j(\theta)}{p_j(\theta)} \cdot \frac{\partial p_j(\theta)}{\partial\theta_i} = 0, \quad i = 1, \ldots, m.$$

Lemma 3.4.8. In the case above, $I(\theta) = C^\top(\theta) \cdot C(\theta)$ holds, where $C(\theta)$ is a $(r \times m)$-matrix with elements

$$c_{ij}(\theta) = \frac{\partial p_i(\theta)}{\partial \theta_j} \cdot \frac{1}{\sqrt{p_i(\theta)}}.$$

Proof

$$\begin{aligned}
\mathbb{E}_0\left[\frac{\partial \log L(X_1,\theta)}{\partial \theta_i} \cdot \frac{\partial \log L(X_1,\theta)}{\partial \theta_j}\right] &= \sum_{k=1}^{r} \frac{\partial \log p_k(\theta)}{\partial \theta_i} \cdot \frac{\partial \log p_k(\theta)}{\partial \theta_j} \cdot p_k(\theta) \\
&= \sum_{k=1}^{r} \frac{\partial p_k(\theta)}{\partial \theta_i} \frac{1}{p_k(\theta)} \cdot \frac{\partial p_k(\theta)}{\partial \theta_j} \cdot \frac{1}{p_k(\theta)} \cdot p_k(\theta) \\
&= \left(C^\top(\theta) \cdot C(\theta)\right)_{ij},
\end{aligned}$$

since

$$\log L(X_1,\theta) = \sum_{i=1}^{r} \log p_j(\theta) \cdot I\left(x \in (a_j, b_j]\right).$$

□

Theorem 3.4.7 implies

Corollary 3.4.9. Let $\hat{\theta}_n = \hat{\theta}(X_1,\ldots,X_n)$ be a weakly consistent maximum likelihood estimator of θ in the coarsened model, which satisfies the regularity assumptions 1.-5. Assume that the Fisher information matrix $I(\theta) = C^\top(\theta) \cdot C(\theta)$ is positive definite for all $\theta \in \Theta$. Then, $\hat{\theta}$ is asymptotically normal distributed

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{d} Y \sim \mathcal{N}\left(0, I^{-1}(\theta)\right).$$

Theorem 3.4.10. Let $\hat{\theta}_n$ be a maximum likelihood estimator in the coarsed model for θ, which satisfies all assumptions of Corollary 3.4.9. The test statistic

$$\hat{T}_n(X_1,\ldots,X_n) = \sum_{j=1}^{r} \frac{(Z_j(X_1,\ldots,X_n) - np_j(\hat{\theta}_n))^2}{np_j(\hat{\theta}_n)}$$

is asymptotically χ^2_{r-m-1}-distributed under H_0:

$$\lim_{n\to\infty} P_\theta\left(\hat{T}_n(X_1,\ldots,X_n) > \chi^2_{r-m-1,1-\alpha}\right) = \alpha.$$

Without proof (cf. [27]).

This theorem implies that the χ^2-Pearson-Fisher test is an asymptotic test with confidence level α.

Example 3.4.11.

1. *χ^2-Pearson-Fisher test of the normal distribution*

   Let $(X_1, \ldots, X_n)$ be a random sample. We test, whether $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Define

   $$\theta := (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+.$$

   Let $\{(a_j, b_j]\}_{j=1,\ldots,r}$ be an arbitrary partition of $\mathbb{R}$ in r disjoint intervals. Recall that density of the one-dimensional $\mathcal{N}(\mu, \sigma^2)$-distribution is given by

   $$f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

   and define

   $$p_j(\theta) := P_0\left(X_1 \in (a_j, b_j]\right) = \int_{a_j}^{b_j} f_\theta(x)dx, \quad j = 1, \ldots, r$$

   with class sizes

   $$Z_j = \#\left\{i : X_i \in (a_j, b_j]\right\}.$$

   The goal is to find the maximum-likelihood estimator in the coarsed model

   $$\begin{aligned}
   \frac{\partial p_j(\theta)}{\partial \mu} &= \int_{a_j}^{b_j} \frac{\partial}{\partial \mu} f_\theta(x)dx = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \int_{a_j}^{b_j} \frac{x-\mu}{\sigma^2} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx, \\
   \frac{\partial p_j(\theta)}{\partial \sigma^2} &= \int_{a_j}^{b_j} \frac{\partial}{\partial \sigma^2} f_\theta(x)dx \\
   &= \frac{1}{\sqrt{2\pi}} \int_{a_j}^{b_j} \left[-\frac{1}{2} \cdot \frac{1}{(\sigma^2)^{3/2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \right. \\
   &\qquad \left. + \frac{1}{\sqrt{\sigma^2}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \cdot \left(\frac{(x-\mu)^2}{2(\sigma^2)^2} \right) \right] dx \\
   &= -\frac{1}{2}\frac{1}{\sigma^2} \int_{a_j}^{b_j} f_\theta(x)dx + \frac{1}{2(\sigma^2)^2} \int_{a_j}^{b_j} (x-\mu)^2 f_\theta(x)dx.
   \end{aligned}$$

   The necessary conditions for a maximum are

   $$\sum_{i=1}^{r} Z_j \frac{\int_{a_j}^{b_j} x f_\theta(x)dx}{\int_{a_j}^{b_j} f_\theta(x)dx} - \mu \underbrace{\sum_{j=1}^{r} Z_j}_{=n} = 0,$$

   $$\frac{1}{\sigma^2} \sum_{j=1}^{r} Z_j \frac{\int_{a_j}^{b_j} (x-\mu)^2 f_\theta(x)dx}{\int_{a_j}^{b_j} f_\theta(x)dx} - \underbrace{\sum_{j=1}^{r} Z_j}_{=n} = 0,$$

which results in the maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}^2$ for μ and σ^2

$$\hat{\mu} = \frac{1}{n}\sum_{j=1}^{r} Z_j \frac{\int_{a_j}^{b_j} x f_\theta(x)dx}{\int_{a_j}^{b_j} f_\theta(x)dx}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{r} Z_j \frac{\int_{a_j}^{b_j} (x-\mu)^2 f_\theta(x)dx}{\int_{a_j}^{b_j} f_\theta(x)dx}.$$

Construct an approximation of $\hat{\mu}$ and $\hat{\sigma}^2$ for $r \to \infty$ as follows: If $r \to \infty$ (and thus $n \to \infty$), then $b_j - a_j$ is small and by the simple quadratic rule

$$\int_{a_j}^{b_j} x f_\theta(x)dx \approx (b_j - a_j)\, y_j f_\theta(y_j),$$
$$\int_{a_j}^{b_j} f_\theta(x)dx \approx (b_j - a_j)\, f_\theta(y_j),$$

holds, where $y_1 = b_1$, $y_r = b_{r-1} = a_r$, and

$$y_j = (b_{j+1} + b_j)/2, \quad j = 2, \ldots, r-1.$$

Thus, for the maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}^2$

$$\hat{\mu} \approx \frac{1}{n}\sum_{j=1}^{r} y_j \cdot Z_j = \tilde{\mu}$$
$$\hat{\sigma}^2 \approx \frac{1}{n}\sum_{j=1}^{r} (y_j - \tilde{\mu})^2 Z_j = \tilde{\sigma}^2$$

and

$$\tilde{\theta} = \left(\tilde{\mu}, \tilde{\sigma}^2\right)$$

holds. In the χ^2-Pearson-Fisher test H_0 is rejected if

$$\hat{T}_n = \frac{\sum_{j=1}^{r} \left(Z_j - np_j(\tilde{\theta})\right)^2}{np_j(\tilde{\theta})} > \chi^2_{r-3,1-\alpha}.$$

2. *χ^2-Pearson-Fisher test for the Poisson distribution*

Let $(X_1, \ldots, X_n)$ be a random sample of i.i.d. random variables. We aim to test, whether $X_i \sim \text{Poisson}(\lambda)$, $\lambda > 0$. Set $\theta = \lambda$ and $\Theta = (0, +\infty)$. Coarsing Θ leads to

$$-\infty = a_1 < \underbrace{b_1}_{=0} = a_2 < \underbrace{b_2}_{=1} = a_3 < \ldots < \underbrace{b_{r-1}}_{=r-2} = a_r < b_r = +\infty.$$

Then,

$$
\begin{aligned}
p_j(\lambda) &= P_\lambda\left(X_1 = j-1\right) = e^{-\lambda}\frac{\lambda^{j-1}}{(j-1)!}, \quad j = 1, \ldots, r-1,\\
p_r(\lambda) &= \sum_{i=r-1}^{\infty} e^{-\lambda}\frac{\lambda^i}{i!},\\
\frac{dp_j(\lambda)}{d\lambda} &= -e^{-\lambda}\frac{\lambda^{j-1}}{(j-1)!} + (j-1)\frac{\lambda^{j-2}}{(j-1)!}e^{-\lambda} = e^{-\lambda}\frac{\lambda^{j-1}}{(j-1)!}\left(\frac{j-1}{\lambda} - 1\right)\\
&= p_j(\lambda)\cdot\left(\frac{j-1}{\lambda} - 1\right), \quad j = 1, \ldots, r-1,\\
\frac{dp_r(\lambda)}{d\lambda} &= \sum_{i\geq r-1} p_i(\lambda)\left(\frac{i-1}{\lambda} - 1\right).
\end{aligned}
$$

Next, we have to solve the maximum likelihood equation

$$
0 = \sum_{j=1}^{r-1} Z_j \cdot \left(\frac{j-1}{\lambda} - 1\right) + Z_r \frac{\sum\limits_{i\geq r-1} p_i(\lambda)\left(\frac{i-1}{\lambda} - 1\right)}{p_r(\lambda)}.
$$

If $r \longrightarrow \infty$, then $r(n)$ exists for every n with $Z_{r(n)} = 0$. Thus, for $r > r(n)$

$$
\sum_{j=1}^{r-1}(j-1)Z_j - \lambda\underbrace{\sum_{j=1}^{r} Z_j}_{=n} = 0
$$

holds, which yields the maximum likelihood estimator

$$
\frac{1}{n}\sum_{j=1}^{r-1}(j-1)Z_j = \frac{1}{n}\sum_{j=1}^{n} X_j = \overline{X}_n.
$$

Hence, the χ^2-Pearson-Fisher test rejects H_0, if

$$
\hat{T}_n = \sum_{j=1}^{r} \frac{\left(Z_j - np_\lambda(\overline{X}_n)\right)^2}{\left(np_j(\overline{X}_n)\right)^2} > \chi^2_{r-2,1-\alpha}.
$$

3.4.3 Shapiros goodness-of-fit test

Let $(X_1, \ldots, X_n)$ be a random sample of i.i.d. random variables $X_i \sim F$. The hypotheses

$$
\begin{aligned}
&H_0 : F \in \{N(\mu, \sigma^2) : \mu \in \mathbb{R},\, \sigma^2 > 0\} \text{ vs.}\\
&H_1 : F \notin \{N(\mu, \sigma^2), \mu \in \mathbb{R},\, \sigma^2 > 0\}
\end{aligned}
$$

are to be tested. The χ^2-tests of sections 3.4.1 - 3.4.2 are asymptotic, which makes them impractical for small sample sizes.
The following test is more suitable for testing H_0, if only a small sample is available.
Consider the order statistic $X_{(1)}, \ldots, X_{(n)}$, i.e., $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ and compare their correlation to the mean of the corresponding order statistic of a $\mathcal{N}(0,1)$-distribution. Let $(Y_1, \ldots, Y_n)$ be a random sample of i.i.d. random variables with $Y_1 \sim \mathcal{N}(0,1)$. Define $a_i := \mathbb{E}\, Y_{(i)}$, $i = 1, \ldots, n$. If the empirical correlation coefficient ρ_{aX} between $(a_1, \ldots, a_n)$ and $(X_{(1)}, \ldots, X_{(n)})$ is close to 1, the random sample is normally distributed. In the following, the approach above will be formalized.
Let b_i be the expected value of the i-th order statistic of a $\mathcal{N}(\mu, \sigma^2)$-distributed random sample of i.i.d. random variables Z_i, with $b_i = \mathbb{E}\, Z_{(i)}$, $i = 1, \ldots, n$. It holds $b_i = \mu + \sigma a_i$, $i = 1, \ldots, n$ and considering the correlation coefficient yields

$$\rho_{bX} = \frac{\sum\limits_{i=1}^{n} \left(b_i - \overline{b}_n\right)\left(X_{(i)} - \overline{X}_n\right)}{\sqrt{\sum\limits_{i=1}^{n} \left(b_i - \overline{b}_n\right)^2 \sum\limits_{i=1}^{n} \left(X_{(i)} - \overline{X}_n\right)^2}}. \tag{3.14}$$

Since ρ is invariant with respect to linear transformations and

$$\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} \mathbb{E}\, Y_i = \mathbb{E}\left(\sum_{i=1}^{n} Y_i\right) = 0,$$

$$\rho_{bX} = \rho_{aX} = \frac{\sum\limits_{i=1}^{n} a_i \left(X_{(i)} - \overline{X}_n\right)}{\sqrt{\sum\limits_{i=1}^{n} a_i^2 \sum\limits_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2}} = \frac{\sum\limits_{i=1}^{n} a_i X_{(i)} - \overline{X}_n \overbrace{\sum\limits_{i=1}^{n} a_i}^{=0}}{\sqrt{\sum\limits_{i=1}^{n} a_i^2 \sum\limits_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2}}$$

$$= \frac{\sum\limits_{i=1}^{n} a_i X_{(i)}}{\sqrt{\sum\limits_{i=1}^{n} a_i^2 \cdot \sum\limits_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2}}$$

holds.
The test statistic is then given by

$$T_n = \frac{\sum\limits_{i=1}^{n} a_i X_{(i)}}{\sqrt{\sum\limits_{i=1}^{n} a_i^2 \sum\limits_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2}} \quad \text{(Shapiro-Francia test)}$$

The values a_i are known and can be found in tables or by using statistic software. Note that $|T_n| \leq 1$.
H_0 is rejected if $T_n \leq q_{n,\alpha}$, where $q_{n,\alpha}$ is the α-quantile of the distribution of T_n. Those quantiles can also be found in tables or by using Monte-Carlo-Simulations.

Remark 3.4.12. Another famous test of this kind is obtained by replacing the linear transformation $b_i = \mu + \sigma a_i$ with another linear transformation given by

$$(a'_1, \ldots, a'_n)^\top = K^{-1} \cdot (a_1, \ldots, a_n),$$

where $K = (k_{ij})_{j=1}^n$ is the covariance matrix of $\left(Y_{(1)}, \ldots, Y_{(n)}\right)$ with

$$k_{ij} = \mathbb{E}\left(Y_{(i)} - a_i\right)\left(Y_{(j)} - a_j\right), \quad i, j = 1, \ldots, n.$$

The constructed test is called *Shapiro-Wilk test*.

3.5 More nonparametric tests

3.5.1 Binomial test

Let $(X_1, \ldots, X_n)$ be a random sample of i.i.d random variables with $X_i \sim$ Bernoulli(p), where $p \in [0, 1]$. We want to test the hypotheses

$$H_0 : p = p_0 \text{ vs. } H_1 : p \neq p_0$$

The test statistic is given by

$$T_n = \sum_{i=1}^{n} X_i \underset{H_0}{\sim} \text{Bin}(n, p_0),$$

and H_0 is rejected if

$$T_n \notin [\text{Bin}(n, p_0)_{\alpha/2}, \text{Bin}(n, p_0)_{1-\alpha/2}],$$

where $\text{Bin}(n, p)_\alpha$ is the α quantile of the $\text{Bin}(n, p)$ distribution
For different H_0, like $p \leq p_0$ ($p \geq p_0$) the rejection region has to be adjusted. The quantiles $\text{Bin}(n, p)_\alpha$ can also be found in tables or by using Monte-Carlo simulations. If n is sufficiently large, the quantiles can be approximated using the central limit theorem of DeMoivre-Laplace:

$$P\left(T_n \leq x\right) = P\left(\frac{T_n - np_0}{\sqrt{np_0(1-p_0)}} \leq \frac{x - np_0}{\sqrt{np_0(1-p_0)}}\right) \underset{n\to\infty}{\approx} \Phi\left(\frac{x - np_0}{\sqrt{np_0(1-p_0)}}\right).$$

This yields

$$z_\alpha \approx \frac{\text{Bin}(n,p_0)_\alpha - np_0}{\sqrt{np_0(1-p_0)}}$$
$$\Rightarrow \text{Bin}(n,p_0)_\alpha \approx \sqrt{np_0(1-p_0)} \cdot z_\alpha + np_0$$

Using Poisson approximation (for $n \to \infty, np_0 \to \lambda_0$)

$$\text{Bin}(n,p_0)_{\alpha/2} \approx \text{Poisson}(\lambda_0)_{\alpha/2},$$
$$\text{Bin}(n,p_0)_{1-\alpha/2} \approx \text{Poisson}(\lambda_0)_{1-\alpha/2},$$

holds if $\lambda_0 = np_0$.
Question: Can the symmetry of a distribution be tested by using the binomial test?
Let $(Y_1, \ldots, Y_n)$ be a random sample of i.i.d. random variables with distribution function F. The hypotheses are

$$H_0 : F \text{ is symmetric vs. } H_1 : F \text{ is not symmetric.}$$

A symmetric distributions median is around 0. Thus the hypothesis H_0 is coarsed, and

$$H_0' : F^{-1}(0,5) = 0 \text{ vs. } H_1' : F^{-1}(0,5) \neq 0$$

is tested instead. More generally, for $\beta \in [0,1]$ we consider

$$H_0'' : F^{-1}(\beta) = \gamma_\beta \text{ vs. } H_1'' : F^{-1}(\beta) \neq \gamma_\beta.$$

H_0'' vs. H_1'' is tested by using the binomial test: Define $X_i := I(Y_i \leq \gamma_\beta)$. Under H_0''

$$X_i \sim \text{Bernoulli}(F(\gamma_\beta)) = \text{Bernoulli}(\beta).$$

holds. For $a_1 = -\infty$, $b_1 = \gamma_\alpha$, $a_2 = b_1$, $b_2 = +\infty$ define two disjoint classes $(a_1, b_1]$, $(a_2, b_2]$ in the sense of the χ^2-Pearson test. The test statistic is given by

$$T_n = \sum_{i=1}^{n} X_i = \#\{Y_i : Y_i \leq \gamma_\beta\} \sim Bin(n,\beta), \quad p = F(\gamma_\beta),$$

and the hypothesis $F^{-1}(\beta) = \gamma_\beta$ is equivalent to $H_0''' : p = \beta$. In this case, the decision rule states that H_0''' is rejected, if

$$T_n \notin \left[\text{Bin}(n,\beta)_{\alpha/2}, \text{Bin}(n,\beta)_{1-\alpha/2}\right].$$

This is a test with confidence level α.

3.5.2 Randomness iteration tests

Sometimes in biological research a sequence of 0s and 1s is tested for its randomness or the existence of bigger clusters within those numbers. This hypothesis can be tested statistically by using the so-called *iteration tests* . Let $(X_1, ..., X_n)$ be a random sample, $X_i \in \{0, 1\}$, $\sum_{i=1}^{n} X_i = n_1$ the total number of ones, $n_2 = n - n_1$ the total number of zeroes and n_1, n_2 predetermined. An exemplary realization of $(X_1, \ldots, X_n)$ with $n = 18$, $n_1 = 12$ could be

$$x = (0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1).$$

The following hypotheses are to be tested:

$$H_0 : \text{every sequence } x \text{ is equally likely vs.}$$
$$H_1 : \text{There are preferred sequences (clustering).}$$

Let

$$\Omega = \left\{ x = (x_1, \ldots, x_n) : \quad x_i \in \{0, 1\},\, i = 1, \ldots, n,\, \sum_{i=1}^{n} x_i = n_1 \right\}$$

be the sample space. Then, the space $(\Omega, \mathcal{F}, P)$ with $\mathcal{F} = \mathcal{P}(\Omega)$,

$$P(x) = \frac{1}{|\Omega|} = \frac{1}{\binom{n}{n_1}}$$

is a Laplace space.
Let

$$\begin{aligned} T_n(X) &= \#\{\text{Iterations in } X\} = \#\{\text{Subsequences of zeros or ones}\} \\ &= \#\{\text{Change spots from 0 to 1 or from 1 to 0}\} + 1. \end{aligned}$$

For $x = (0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0)$, $T_n(x) = 7 = 6 + 1$ holds.
$T_n(X)$ is used as a test statistic for H_0 vs. H_1 as follows. H_0 is rejected, if $T(x)$ is small, i.e. $T_n(x) < F_{T_n}^{-1}(\alpha)$. This is a test with confidence level α. The question arises, how the quantiles $F_{T_n}^{-1}$ can be calculated?

Theorem 3.5.1. Under H_0

1.

$$P(T_n = k) = \begin{cases} \frac{2\binom{n_1-1}{i-1}\binom{n_2-1}{i-1}}{\binom{n}{n_1}}, & \text{if } k = 2i, \\ \frac{\binom{n_1-1}{i}\binom{n_2-1}{i-1}+\binom{n_1-1}{i-1}\binom{n_2-1}{i}}{\binom{n}{n_1}}, & \text{if } k = 2i+1, \end{cases}$$

2.

$$\mathbb{E}\, T_n = 1 + \frac{2n_1n_2}{n},$$

3.

$$\mathrm{Var}\,(T_n) = \frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}.$$

holds.

Proof

1. Assume that $k = 2i$ (the uneven case works analogously). How can i clusters of ones be selected? The number of those possibilities is equal to the number of ways, how n_1 particles can be distributed to i classes.

$$0|00|\ldots|0|\,(n_1).$$

This is the number of possibilities, how $i-1$ partitions can be distributed on $n_1 - 1$ positions, which is equal to $\binom{n_1-1}{i-1}$. The same holds for the zeroes.

2. Let $Y_j = I\left\{X_{j-1} \neq X_j\right\}_{j=2,\ldots,n}$. Then,

$$\mathbb{E}\, T_n(X) = 1 + \sum_{j=2}^{n} \mathbb{E}\, Y_j = 1 + \sum_{j=2}^{n} P\left(X_{j-1} \neq X_j\right)$$

and the probabilities can be rewritten as

$$\begin{aligned} P\left(X_{j-1} \neq X_j\right) &= \frac{2\binom{n-2}{n_1-1}}{\binom{n}{n_1}} = 2 \cdot \frac{\frac{(n-2)!}{(n-2-(n_1-1))!(n_1-1)!}}{\frac{n!}{(n-n_1)!n_1!}} \\ &= \frac{2n_1(n-n_1)}{(n-1)n} \\ &= \frac{2n_1n_2}{n(n-1)}. \end{aligned}$$

Hence,

$$\mathbb{E}\, T_n = 1 + (n-1)\frac{2n_1n_2}{n(n-1)} = 1 + 2\frac{n_1n_2}{n}.$$

Exercise 3.5.2. Proof the third assertion.

□

Example 3.5.3 (*Wald-Wolfowitz test*)**.** Let $Y = (Y_1, \ldots, Y_{n_1})$, $Z = (Z_1, \ldots, Z_{n_2})$ be two independent random samples of i.i.d. random variables, $Y_i \sim F$, $Z_i \sim G$.

$$H_0 : F = G \text{ vs. } H_1 : F \neq G.$$

is to be tested. Define $(Y, Z) \coloneqq (Y_1, \ldots, Y_{n_1}, Z_1, \ldots, Z_{n_2})$ and let X_i' be the sample variables of (Y, Z), $i = 1, \ldots, n$, $n = n_1 + n_2$. Consider the order statistic $X'_{(i)}$, $i = 1, \ldots, n$ and set

$$X_i = \begin{cases} 1, & \text{if } X'_{(i)} = Y_j \text{ for a } j = 1, \ldots, n_1, \\ 0, & \text{if } X'_{(i)} = Z_j \text{ for a } j = 1, \ldots, n_2. \end{cases}$$

Under H_0, the sample values in (Y, Z) are well distributed, i.e., every combination of 0 and 1 in $(X_1, \ldots, X_n)$ is equally likely. Thus, the randomness iteration test can be applied to test H_0 vs. H_1. H_0 is rejected if $T_n(x) \leq F_{T_n}^{-1}(\alpha)$, $x = (x_1, \ldots, x_n)$.
The quantiles of F_{T_n} can be calculated directly if n is sufficiently large, since

$$\frac{n_1}{n_1 + n_2} \underset{n\to\infty}{\longrightarrow} p \in (0, 1)$$

implies that T_n is asymptotically normal distributed.

Theorem 3.5.4. Under the assumptions above

$$\begin{aligned} \lim_{n\to\infty} \frac{\mathbb{E}\, T_n}{n} &= 2p(1-p), \\ \lim_{n\to\infty} \frac{1}{n} \operatorname{Var} T_n &= 4p^2(1-p)^2, \\ \frac{T_n - 2p(1-p)}{2\sqrt{n}p(1-p)} &\xrightarrow[n\to\infty]{d} Y \sim \mathcal{N}(0,1), \quad \text{if } \frac{n_1}{n_1+n_2} \longrightarrow p \in (0,1) \end{aligned}$$

holds. Thus, the quantiles for T_n can be approximated for large n by

$$\begin{aligned} \alpha &= P\left(T_n \leq F_{T_n}^{-1}(\alpha)\right) = P\left(\frac{T_n - 2np(1-p)}{2\sqrt{n}p(1-p)} \leq \frac{x - 2np(1-p)}{2\sqrt{n}p(1-p)}\right)\Big|_{x=F_{T_n}^{-1}(\alpha)} \\ &\approx \Phi\left(\frac{F_{T_n}^{-1}(\alpha) - 2np(1-p)}{2\sqrt{n}p(1-p)}\right) \\ &\Rightarrow z_\alpha \approx \frac{F_{T_n}^{-1}(\alpha) - 2np(1-p)}{2\sqrt{n}p(1-p)}, \end{aligned}$$

which implies

$$F_{T_n}^{-1}(\alpha) \approx 2np(1-p) + 2\sqrt{n}p(1-p) \cdot z_\alpha$$

In practice $\hat{p} = \frac{n_1}{n_1+n_2}$ is used for p.

Chapter 4

Linear Regression

In Section 6.7.3 of the lecture "Elementary probability theory and statistics", (cf. [33]) a simple form of linear regression was introduced via

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

Using matrix notation, this can be rewritten as $Y = X\beta + \varepsilon$, where $Y = (Y_1, \ldots, Y_n)^\top$ is a random vector and

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

is a $n \times 2$ matrix, which contains the so-called *predictor variables* $x_i, i = 1, \ldots, n$ and is called *design matrix*. Further, $\beta = (\beta_0, \beta_1)^\top$ resp. $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ is the so-called parameter resp. error vector. With respect to the error vector, we assume that $\varepsilon \sim \mathcal{N}(0, \mathcal{I} \cdot \sigma^2)$ is multivariate normally distributed.
In multivariate linear regression, i.e. not simple simple linear regression, an arbitrary $(n \times m)$ design matrix

$$X = (x_{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,m}}$$

and a m-dimensional parameter vector $\beta = (\beta_1, \ldots, \beta_m)^\top$ are permissible for $m \geq 2$. That means we consider

$$Y = X\beta + \varepsilon, \tag{4.1}$$

where $\varepsilon \sim \mathcal{N}(0, K)$ and K an arbitrary covariance matrix. In general, this choice of K can result in the errors not being independent which means that $K \neq \text{diag } (\sigma_1^2, \ldots, \sigma_n^2)$.

The goal of this chapter is to construct estimators and tests for β. But before getting into detail about this, properties of the multivaritate normal distribution need to be discussed.

4.1 Multivariate normal distribution

In the lecture notes of "Elementary probability theory and statistics" (cf. [33]) the multivariate normal distribution was introduced in example 3.4.5 as follows:

Definition 4.1.1. Let $X = (X_1, \ldots, X_n)^\top$ be a n-dimensional random vector, $\mu \in \mathbb{R}^n$, K a symmetric positive definite $(n \times n)$ matrix. X is *multivariate normal distribution* with parameters μ and K ($X \sim \mathcal{N}(\mu, K)$), if X is absolutely continuous distributed with probability density function

$$f_X(x) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sqrt{\det(K)}} \exp\left\{-\frac{1}{2}(x-\mu)^\top K^{-1}(x-\mu)\right\},$$

where $x = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$.

However, this is not the only way to define the multivariate normal distribution. Thus, let us discuss three more definitions of $\mathcal{N}(\mu, K)$.

Definition 4.1.2. The random vector $X = (X_1, \ldots, X_n)^\top$ is multivariate normally distributed ($X \sim \mathcal{N}(\mu, K)$) with mean vector $\mu \in \mathbb{R}^n$ and covariance matrix K, if the characteristic function $\varphi_X(t) = \mathbb{E}\, e^{i(t,X)}$, $t \in \mathbb{R}^n$, is given by

$$\varphi_X(t) = \exp\left\{it^\top \mu - \frac{1}{2} t^\top K t\right\}, \quad t \in \mathbb{R}^n.$$

Definition 4.1.3. The random vector $X = (X_1, \ldots, X_n)^\top$ is multivariate normally distributed ($X \sim \mathcal{N}(\mu, K)$) with mean vector $\mu \in \mathbb{R}^n$ and covariance matrix K, if

$$\text{for all } a \in \mathbb{R}^n: \text{ the random variable } (a, X) = a^\top X \sim \mathcal{N}(a^\top \mu, a^\top K a)$$

is a one-dimensional normally distributed random variable.

Definition 4.1.4. Let $\mu \in \mathbb{R}^n$ and K be a covariance matrix. A random vector $X = (X_1, \ldots, X_n)^\top$ is multivariate normally distributed with mean vector μ and covariance matrix K ($X \sim \mathcal{N}(\mu, K)$), if

$$X \stackrel{d}{=} \mu + C \cdot Y,$$

where C is a $n \times m$ matrix with $\operatorname{rank}(C) = m$, $K = C \cdot C^\top$ and $Y \sim \mathcal{N}(0, \mathcal{I})$ is an m-dimensional random vector with i.i.d.$\sim \mathcal{N}(0,1)$ coordinates.

Remark 4.1.5. Definition 4.1.4 is an analogue to the one-dimensional case where we know that $Y \sim \mathcal{N}(\mu, \sigma^2)$ if and only if $Y \overset{d}{=} \mu + \sigma X$ with $X \sim \mathcal{N}(0, 1)$.

Exercise 4.1.6. Show that the function

$$f_X(x) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sqrt{\det(K)}} \exp\left\{-\frac{1}{2}(x-\mu)^\top K^{-1}(x-\mu)\right\}, \; x \in \mathbb{R}^n$$

from Definition 4.1.1 is indeed a probability density function.

Lemma 4.1.7. Let X and Y be two n-dimensional random vectors with characteristic functions

$$\varphi_X(t) = \mathbb{E}\, e^{i(t,X)} = \mathbb{E}\, e^{it^\top X}$$
$$\varphi_Y(t) = \mathbb{E}\, e^{i(t,Y)} = \mathbb{E}\, e^{it^\top Y}$$

for $t \in \mathbb{R}^n$. Then, it holds

1. *Uniqueness theorem:*

$$X \overset{d}{=} Y \Leftrightarrow \varphi_X(t) = \varphi_Y(t), \quad t \in \mathbb{R}^n$$

2. If X and Y are independent, then:

$$\varphi_{X+Y}(t) = \varphi_X(t) \cdot \varphi_Y(t), \quad t \in \mathbb{R}^n.$$

without proof (cf. proof of Theorem 2.1.4 (5), [32, Corollary 2.1.10].

Theorem 4.1.8.

1. The definitions 4.1.2 - 4.1.4 of the multivariate normal distribution are equivalent.

2. The definition 4.1.1 and 4.1.4 are equivalent for $n = m$.

Remark 4.1.9.

1. If the matrix K in Definition 4.1.4 has full rank n, then X's probability density function is given as in Definition 4.1.1. In this case it is called *regular*.

2. If $\operatorname{rank}(K) = m < n$, then the distribution $\mathcal{N}(\mu, K)$ is concentrated on the m-dimensional subspace

$$\{y \in \mathbb{R}^n : y = \mu + Cx, \, x \in \mathbb{R}^m\}$$

by Definition 4.1.4. In this case, $\mathcal{N}(\mu, K)$ is obviously not absolutely continuous distributed and is thus called *singular*.

Proof In an initial step, Definition 4.1.3 $\Leftrightarrow$ 4.1.2 $\Leftrightarrow$ 4.1.4 is proven.

1. (a) First, we show that Definitions 4.1.2 and 4.1.3 are equivalent, i.e., for the random variable X with characteristic function φ_X it holds that

$$\begin{aligned} & \varphi_X(t) = \exp\{it^\top \mu - \frac{1}{2} t^\top K t\} \\ \Leftrightarrow \quad & \text{for all } a \in \mathbb{R}^n : a^\top X \sim \mathcal{N}(a^\top \mu, a^\top K a). \end{aligned}$$

   Simple calculations yield

$$\varphi_{t^\top X}(1) = \mathbb{E}\, e^{it^\top X \cdot 1} \overset{\varphi_{\mathcal{N}(\mu,\sigma^2)}}{=} \exp\{it^\top \mu - \frac{1}{2} t^\top K t\} = \varphi_X(t),$$

   for all $t \in \mathbb{R}$. (This is called the *Procedure of Cramér-Wold*, cf. multivariate central limit theorem).

   (b) Next, we show that Definitions 4.1.3 and 4.1.4 are equivalent. Using the notation $y = C^\top t$ we compute

$$\begin{aligned} \varphi_{\mu+CY}(t) = \mathbb{E}\, e^{i(t,\mu+CY)} &= \mathbb{E}\, e^{it^\top \mu + it^\top CY} = e^{it^\top \mu} \cdot \mathbb{E}\, e^{i(C^\top t, Y)} \\ &\overset{Y \sim \mathcal{N}(0,\mathcal{I})}{=} e^{it^\top \mu} \cdot \exp\left(-\frac{1}{2} y^\top \cdot y\right) \\ &= \exp\left\{it^\top \mu - \frac{1}{2} t^\top C \cdot C^\top t\right\} \\ &= \exp\left\{it^\top \mu - \frac{1}{2} t^\top K t\right\}, \; t \in \mathbb{R}^n. \end{aligned}$$

2. It needs to be shown that for $X \sim \mathcal{N}(\mu, K)$ in the sense of Definition 4.1.4 and $Y \sim \mathcal{N}(\mu, K)$ in the sense of Definition 4.1.1 the relation $\operatorname{rank}(K) = n$ implies that $\varphi_X = \varphi_Y$.

   Definition 4.1.2 (which is equivalent to Definition 4.1.4) implies, that

$$\begin{aligned} \varphi_X(t) &= \exp\left\{it^\top \mu - \frac{1}{2} t^\top K t\right\}, \quad t \in \mathbb{R}^n, \\ \varphi_Y(t) &= \mathbb{E}\, e^{it^\top Y} \\ &= \int_{\mathbb{R}^n} e^{it^\top y} \frac{1}{(2\pi)^{n/2}\sqrt{\det K}} \cdot \exp\Big\{-\frac{1}{2} \overbrace{(y-\mu)}^{x}{}^\top K^{-1} \overbrace{(y-\mu)}^{x} \Big\} dy \\ &= e^{it^\top \mu} \cdot \int_{\mathbb{R}^n} \frac{1}{(2\pi)^{n/2}\sqrt{\det K}} \cdot \exp\left\{it^\top x - \frac{1}{2} x^\top K^{-1} x\right\} dx \end{aligned}$$

   Diagonalising K : $\exists$ orthogonal $(n \times n)$ matrix V : $V^\top = V^{-1}$ and $V^\top K V = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, where $\lambda_i > 0$, $i = 1, \ldots, n$. By applying

the substitution $x = Vz$, $t = Vs$ it holds that

$$\begin{aligned}
\varphi_Y(t) &= \frac{e^{it^\top \mu}}{(2\pi)^{n/2}\sqrt{\det K}} \cdot \int_{\mathbb{R}^n} \exp\left\{is^\top V^\top V z - \frac{1}{2} z^\top V^\top K^{-1} V z\right\} dz \\
&= \frac{e^{it^\top \mu}}{\sqrt{(2\pi)^n \lambda_1 \cdot \ldots \cdot \lambda_n}} \cdot \int_{\mathbb{R}} \ldots \int_{\mathbb{R}} \exp\left\{is^\top z - \frac{1}{2}\sum_{i=1}^n \frac{z_i^2}{\lambda_i}\right\} dz_1 \ldots dz_n \\
&= e^{it^\top \mu} \prod_{i=1}^n \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\lambda_i}} e^{is_i z_i - \frac{z_i^2}{(2\lambda_i)}} dz_i \\
&= e^{it^\top \mu} \cdot \prod_{i=1}^n \varphi_{\mathcal{N}(0,\lambda_i)}(s_i) = e^{it^\top \mu} \prod_{i=1}^n e^{\frac{-s_i^2 \lambda_i}{2}} \\
&= \exp\left\{it^\top \mu - \frac{1}{2} s^\top \text{diag } (\lambda_1, \ldots, \lambda_n) s\right\} \\
&= \exp\left\{it^\top \mu - \frac{1}{2}(V^\top t)^\top V^\top K V V^\top t\right\} \\
&= \exp\left\{it^\top \mu - \frac{1}{2} t^\top \underbrace{VV^\top}_{\mathcal{I}} K \underbrace{VV^\top}_{\mathcal{I}} t\right\} \\
&= \exp\left\{it^\top \mu - \frac{1}{2} t^\top K t\right\}, \; t \in \mathbb{R}^n.
\end{aligned}$$

□

4.1.1 Properties of the multivariate normal distribution

Theorem 4.1.10. Let $X = (X_1, \ldots, X_n) \sim \mathcal{N}(\mu, K)$, $\mu \in \mathbb{R}^n$, K symmetric and positive semidefinite. Then the following properties hold:

1. μ is the *vector of expectations* of X:

$$\mathbb{E}\, X = \mu, \quad \text{that means } \mathbb{E}\, X_i = \mu_i,\ i = 1, \ldots, n.$$

   K is the *covariance matrix* of X:

$$K = (k_{ij}), \text{ with } k_{ij} = \text{Cov } (X_i, X_j).$$

2. Every partial vector $X' = (X_{i_1}, \ldots, X_{i_k})^\top$ $(1 \le i_1 < \ldots < i_k \le n)$ of X is also multivariate normally distributed, $X' \sim \mathcal{N}(\mu', K')$, where $\mu' = (\mu_{i_1}, \ldots, \mu_{i_k})^\top$, $K' = (k'_{jl}) = (\text{Cov}(X_{i_j}, X_{i_l}))$, $j, l = 1, \ldots, k$. In particular it holds that $X_i \sim \mathcal{N}(\mu_i, k_{ii})$, where $k_{ii} = \text{Var}\, X_i$, $i = 1, \ldots, n$.

3. Two partial vectors of X are independent if and only if the corresponding elements k_{ij} of K, which represent the cross covariances, are zero, i.e. $X' = (X_1, \ldots, X_k)^\top$, $X'' = (X_{k+1}, \ldots, X_n)$ are independent

(where the given order is chosen for the sake of simplicity) if and only if $k_{ij} = 0$ for $1 \leq i \leq k$, $j > k$ or $i > k$, $1 \leq j \leq k$, i.e.

$$K = \left(\begin{array}{c|c} K' & 0 \\ \hline 0 & K'' \end{array}\right)$$

where K' and K'' are covariance matrices of X' resp. X''.

4. *Conclusion stability:* If X and Y are independent, n-dimensional random vectors with $X \sim \mathcal{N}(\mu_1, K_1)$ and $Y \sim \mathcal{N}(\mu_2, K_2)$, then

$$X + Y \sim \mathcal{N}(\mu_1 + \mu_2, K_1 + K_2).$$

Exercise 4.1.11. Prove Theorem 4.1.10.

Theorem 4.1.12 (*Linear transformation of* $\mathcal{N}(\mu, K)$)**.** Let $X \sim \mathcal{N}(\mu, K)$ be an n-dimensional random vector and A an $(m \times n)$ matrix with $\operatorname{rank}(A) = m \leq n$, $b \in \mathbb{R}^m$. Then the random vector $Y = AX + b$ is multivariate normally distributed with

$$Y \sim \mathcal{N}(A\mu + b, AKA^\top).$$

Proof Without loss of generality assume $\mu = 0$ and $b = 0$, since $\varphi_{Y-a}(t) = e^{-it^\top a} \cdot \varphi_Y(t)$, for $a = A\mu + b$. It has to be shown that:

$$Y = AX,\ X \sim \mathcal{N}(0, K) \Rightarrow Y \sim \mathcal{N}(0, AKA^\top).$$

This can be done by calculation as follows.

$$\begin{aligned} \varphi_Y(t) = \varphi_{AX}(t) &= \mathbb{E}\, e^{it^\top AX} = \mathbb{E}\, e^{i(X, \overbrace{A^\top t}^{:=s})} \\ &\overset{\text{(Def. 4.1.2)}}{=} \exp\left\{-\frac{1}{2} s^\top K s\right\} = \exp\left\{-\frac{1}{2} t^\top AKA^\top t\right\},\ t \in \mathbb{R}^n \\ &\Rightarrow Y \sim N\left(0, AKA^\top\right). \end{aligned}$$

□

4.1.2 Linear and quadratic forms of normally distributed random variables

Definition 4.1.13. Let $X = (X_1, \ldots, X_n)^\top$, $Y = (Y_1, \ldots, Y_n)^\top$ be two random vectors on $(\Omega, \mathcal{F}, P)$ and A be a symmetric, real-valued $(n \times n)$ matrix.

1. $Z = AX$ is called *linear form* of X with matrix A.

2. $Z = Y^\top AX$ is called *bilinear form* of X and Y with matrix A. Furthermore, rewriting the vector notation yields
$$Z = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}X_jY_i.$$

3. The random variable $Z = X^\top AX$ (which is a bilinear form X with itself) is called *quadratic form* of X with matrix A.

Theorem 4.1.14. Let $Z = Y^\top AX$ be a bilinear form of random vectors $X, Y \in \mathbb{R}^n$ with respect to the symmetric matrix A. If $\mu_X = \mathbb{E}\,X$, $\mu_Y = \mathbb{E}\,Y$ and $K_{XY} = (\mathrm{Cov}(X_i, Y_j))_{i,j=1,\dots,n}$ the cross covariance matrix of X and Y, then
$$\mathbb{E}\,Z = \mu_Y^\top A\mu_X + \mathrm{trace}(AK_{XY}).$$

Proof
$$\begin{aligned}
\mathbb{E}\,Z &= \mathbb{E}\,\mathrm{trace}(Z) = \mathbb{E}\,\mathrm{trace}(Y^\top AX) \quad (\text{since } \mathrm{trace}(AB) = \mathrm{trace}(BA))\\
&= \mathbb{E}\,\mathrm{trace}(AXY^\top) = \mathrm{trace}(A\mathbb{E}\,(XY^\top))\\
&= \mathrm{trace}\left(A\mathbb{E}\left((X-\mu_X)\cdot(Y-\mu_Y)^\top + \mu_XY^\top + X\mu_Y^\top - \mu_X\mu_Y^\top\right)\right)\\
&= \mathrm{trace}\left(A(K_{XY} + \mu_X\mu_Y^\top + \mu_X\mu_Y^\top - \mu_X\mu_Y^\top)\right)\\
&= \mathrm{trace}\left(AK_{XY} + A\mu_X\mu_Y^\top\right)\\
&= \mathrm{trace}(AK_{XY}) + \mathrm{trace}\left(A\mu_X\cdot\mu_Y^\top\right)\\
&= \mathrm{trace}\left(\mu_Y^\top A\mu_X\right) + \mathrm{trace}\,(AK_{XY}) = \mu_Y^\top A\mu_X + \mathrm{trace}\,(AK_{XY})\,.
\end{aligned}$$

□

Corollary 4.1.15. For quadratic forms it holds that
$$\mathbb{E}\,(X^\top AX) = \mu_X^\top A\mu_X + \mathrm{trace}(A\cdot K),$$
where $\mu_X = \mathbb{E}\,X$ and K is the covariance matrix of X.

Theorem 4.1.16 (*Covariance of quadratic forms*)**.** Let $X \sim \mathcal{N}(\mu, K)$ be an n-dimensional random vector and $A, B \in \mathbb{R}^{n\times n}$ two symmetric matrices. Then
$$\mathrm{Cov}\left(X^\top AX, X^\top BX\right) = 4\mu^\top AKB\mu + 2\cdot\mathrm{trace}(AKBK).$$

Lemma 4.1.17 (mixed moments)**.** Let $Y = (Y_1, \dots, Y_n)^\top \sim \mathcal{N}(0, K)$ be a random vector. Then
$$\begin{aligned}
\mathbb{E}\,(Y_iY_jY_k) &= 0,\\
\mathbb{E}\,(Y_iY_jY_kY_l) &= k_{ij}\cdot k_{kl} + k_{ik}\cdot k_{jl} + k_{jk}\cdot k_{il}, \quad 1 \le i,j,k,l \le n,
\end{aligned}$$
where $K = (k_{ij})_{i,j=1,\dots,n}$ is the covariance matrix of Y.

Exercise 4.1.18. Prove the Lemma.

Proof of Theorem 4.1.16.

$$
\begin{aligned}
\operatorname{Cov}\left(X^\top AX, X^\top BX\right) &= \mathbb{E}\left(X^\top AX \cdot X^\top BX\right) \\
&\quad - \mathbb{E}\left(X^\top AX\right) \cdot \mathbb{E}\left(X^\top BX\right) \\
&\overset{\text{(Corollary 4.1.15)}}{=} \mathbb{E}\left((\overbrace{X-\mu}^{:=Y}+\mu)^\top A(\overbrace{X-\mu}^{=Y}+\mu)\cdot(\overbrace{X-\mu}^{=Y}+\mu)^\top B(\overbrace{X-\mu}^{=Y}+\mu)\right) \\
&\quad - \left(\mu^\top A\mu + \operatorname{trace}(AK)\right)\left(\mu^\top B\mu + \operatorname{trace}(BK)\right) \\
&= \mathbb{E}\left[\left(Y^\top AY + 2\mu^\top AY + \mu^\top A\mu\right)\left(Y^\top BY + 2\mu^\top BY + \mu^\top B\mu\right)\right] \\
&\quad - \mu^\top A\mu \cdot \mu^\top B\mu - \mu^\top A\mu \cdot \operatorname{trace}(BK) - \mu^\top B\mu \cdot \operatorname{trace}(AK) \\
&\quad - \operatorname{trace}(AK) \cdot \operatorname{trace}(BK) \\
&= \mathbb{E}\left(Y^\top AY \cdot Y^\top BY\right) + 2\mathbb{E}\left(Y^\top AY \cdot \mu^\top BY\right) + \mathbb{E}\left(Y^\top AY\right) \cdot \mu^\top B\mu \\
&\quad + 2\mathbb{E}\left(\mu^\top AY \cdot Y^\top BY\right) + 4\mathbb{E}\left(\mu^\top AY \cdot \mu^\top BY\right) + 2\underbrace{\mathbb{E}\left(\mu^\top AY\right)}_{=0}\mu^\top B\mu \\
&\quad + \mu^\top A\mu \cdot \mathbb{E}\left(Y^\top BY\right) + 2\mu^\top A\mu \cdot \underbrace{\mathbb{E}\,\mu^\top BY}_{=0} + \mu^\top A\mu \cdot \mu^\top B\mu \\
&\quad - \mu^\top A\mu \cdot \mu^\top B\mu - \mu^\top A\mu \cdot \operatorname{trace}(BK) - \mu^\top B\mu \cdot \operatorname{trace}(AK) \\
&\quad - \operatorname{trace}(AK) \cdot \operatorname{trace}(BK) \\
&= \mathbb{E}\left(Y^\top AY \cdot Y^\top BY\right) + 2\mu^\top B\overbrace{\mathbb{E}\left(Y \cdot Y^\top AY\right)}^{=0 \text{ (Lemma 4.1.17)}} + \mu^\top B\mu \cdot \operatorname{trace}(AK) \\
&\quad + 2\mu^\top A\overbrace{\mathbb{E}\left(Y \cdot Y^\top BY\right)}^{=0} + 4\mu^\top A\overbrace{\mathbb{E}\left(YY^\top\right)}^{=K} B\mu + \mu^\top A\mu \cdot \operatorname{trace}(BK) \\
&\quad - \mu^\top A\mu \cdot \operatorname{trace}(BK) - \mu^\top B\mu \cdot \operatorname{trace}(AK) - \operatorname{trace}(AK)\operatorname{trace}(BK) \\
&= \mathbb{E}\left(Y^\top AY \cdot Y^\top BY\right) + 4\mu^\top AKB\mu - \operatorname{trace}(AK) \cdot \operatorname{trace}(BK).
\end{aligned}
$$

Since

$$
\begin{aligned}
\mathbb{E}\left(Y^\top AY \cdot Y^\top BY\right) &= \mathbb{E}\left(\sum_{i,j=1}^n a_{ij}Y_iY_j \cdot \sum_{k,l=1}^n b_{kl}Y_kY_l\right) \\
&= \sum_{i,j,k,l=1}^n a_{ij}b_{kl}\mathbb{E}\left(Y_iY_jY_kY_l\right) \\
&\overset{\text{(Lemma 4.1.17)}}{=} \sum_{i,j,k,l=1}^n a_{ij}b_{kl}\left(k_{ij}\cdot k_{kl} + k_{ik}\cdot k_{jl} + k_{jk}\cdot k_{il}\right)
\end{aligned}
$$

$$= \sum_{i,j=1}^{n} a_{ij}k_{ij} \cdot \sum_{k,l=1}^{n} b_{kl} \cdot k_{kl} + 2 \sum_{i,j,k,l=1}^{n} a_{ij} \cdot k_{jl} \cdot b_{lk} \cdot k_{ki}$$
$$= 2 \cdot \operatorname{trace}(AKBK) + \operatorname{trace}(AK) \cdot \operatorname{trace}(BK)$$

it holds that

$$\begin{aligned}\operatorname{Cov}&\left(X^\top AX, X^\top BX\right)\\ &= 2 \cdot \operatorname{trace}(AKBK) + \operatorname{trace}(AK) \cdot \operatorname{trace}(BK) + 4\mu^\top AKB\mu\\ &\quad - \operatorname{trace}(AK) \cdot \operatorname{trace}(BK) = 4\mu^\top AKB\mu + 2 \cdot \operatorname{trace}(AKBK).\end{aligned}$$

□

Corollary 4.1.19.

$$\operatorname{Var}\left(X^\top AX\right) = 4\mu^\top AKA\mu + 2 \cdot \operatorname{trace}\left((AK)^2\right)$$

Theorem 4.1.20. Let $X \sim \mathcal{N}(\mu, K)$ and $A, B \in \mathbb{R}^{n\times n}$ be two symmetric matrices. Then

$$\operatorname{Cov}\left(BX, X^\top AX\right) = 2BKA\mu$$

Proof

$$\begin{aligned}\operatorname{Cov}&\left(BX, X^\top AX\right) =\\ &\overset{\text{(Folgerung 4.1.15)}}{=} \mathbb{E}\left[(BX - B\mu)(X^\top AX - \mu^\top A\mu - \operatorname{trace}(AK))\right]\\ &= \mathbb{E}\left[B(X-\mu)\left((X-\mu)^\top A(X-\mu)\right.\right.\\ &\qquad \left.\left.+2\mu^\top AX - 2\mu^\top A\mu - \operatorname{trace}(AK)\right)\right],\end{aligned}$$

Since

$$(X-\mu)^\top A(X-\mu) = X^\top AX - \mu^\top AX - X^\top A\mu + \mu^\top A\mu$$

and by substituting $Z = X - \mu$ (which implies $\mathbb{E}\,Z = 0$)

$$\begin{aligned}\operatorname{Cov}\left(BX, X^\top AX\right) &= \mathbb{E}\left[BZ(Z^\top AZ + 2\mu^\top AZ - \operatorname{trace}(AK))\right]\\ &= \mathbb{E}\left(BZ \cdot Z^\top AZ\right) + 2\mathbb{E}\left(BZ \cdot \mu^\top AZ\right)\\ &\qquad - \operatorname{trace}(AK) \cdot \overbrace{\mathbb{E}\left(BZ\right)}^{=B\mathbb{E}\,Z=0}\\ &= 2\mathbb{E}\left(BZ \cdot Z^\top A\mu\right) + \mathbb{E}\left(BZZ^\top AZ\right)\\ &= 2B\,\underbrace{\mathbb{E}\left(ZZ^\top\right)}_{\operatorname{Cov} X = K} A\mu + B \cdot \underbrace{\mathbb{E}\left(ZZ^\top AZ\right)}_{=0}\\ &= 2BKA\mu,\end{aligned}$$

since $Z \sim \mathcal{N}(0, K)$ and Lemma 4.1.17 and the proof of Theorem 4.1.16. □

Definition 4.1.21. Let $X_i \sim \mathcal{N}(\mu_i, 1)$, $i = 1, \ldots, n$ be independent. Then the random variable

$$Y = X_1^2 + \ldots + X_n^2$$

is *non-centered* $\chi^2_{n,\mu}$ distributed with n *degrees of freedom* and the *non-centrality parameter*

$$\mu = \sum_{i=1}^{n} \mu_i^2.$$

In Remark 2.2.6, WT&SP (cf. [32]), the moment generating function of random variables were introduced. For the proof of Theorem 4.1.23 the following uniqueness theorem will be used:

Lemma 4.1.22 (*Uniqueness theorem for moment generating functions*)**.** Let X_1 and X_2 be two absolutely continuous random variables with moment generating functions

$$M_{X_i}(t) = \mathbb{E}\, e^{tX_i}, \quad i = 1, 2,$$

which are defined on the interval (a, b). If f_1 and f_2 are the probability density functions of the distributions of X_1 and X_2, then

$$f_1(x) = f_2(x) \text{ for almost all } x \in \mathbb{R} \Leftrightarrow M_{X_1}(t) = M_{X_2}(t),\ t \in (a, b).$$

Without proof.

Theorem 4.1.23. The probability density function of a $\chi^2_{n,\mu}$ distributed random variable X (with $n \in \mathbb{N}$ and $\mu > 0$) is given by the mixture function of the density of a χ^2_{n+2J} distribution with mixture variable $J \sim$ Poisson$(\mu/2)$:

$$f_X(x) = \begin{cases} \sum\limits_{j=0}^{\infty} e^{-\mu/2} \frac{(\mu/2)^j}{j!} \cdot \frac{e^{-x/2} x^{\frac{n+2j}{2}-1}}{\Gamma\left(\frac{n+2j}{2}\right) \cdot 2^{\frac{n+2j}{2}}}, & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{4.2}$$

Proof

1. First, calculate $M_X(t)$, $X \sim \chi^2_{n,\mu}$:

$$\begin{aligned} M_X(t) &= \mathbb{E}\,(e^{tX}) = \mathbb{E}\, \exp\left\{ t \sum_{i=1}^{n} X_i^2 \right\} \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \cdot \int\limits_{-\infty}^{\infty} e^{tx_i^2} \cdot e^{-\frac{(x_i - \mu_i)^2}{2}} dx_i \quad \left(t < \frac{1}{2},\ X_i \sim \mathcal{N}(\mu_i, 1) \right) \end{aligned}$$

It holds that

$$\begin{aligned} tx_i^2 - \frac{(x_i-\mu_i)^2}{2} &= \frac{1}{2}(2tx_i^2 - x_i^2 + 2x_i\mu_i - \mu_i^2) \\ &= -\frac{1}{2}\left(x_i^2(1-2t) - 2x_i\mu_i + \frac{\mu_i^2}{(1-2t)} - \frac{\mu_i^2}{(1-2t)} + \mu_i^2\right) \\ &= -\frac{1}{2}\left(\left(x_i\cdot\sqrt{1-2t} - \frac{\mu_i}{\sqrt{1-2t}}\right)^2 + \mu_i^2\left(1-\frac{1}{1-2t}\right)\right) \\ &= -\frac{1}{2}\left(\frac{(x_i(1-2t)-\mu_i)^2}{1-2t} - \mu_i^2\cdot\frac{2t}{1-2t}\right) \end{aligned}$$

Substituting

$$y_i = \frac{(x_i\cdot(1-2t)-\mu_i)}{\sqrt{1-2t}}$$

yields

$$\begin{aligned} M_X(t) &= (1-2t)^{-\frac{n}{2}}\prod_{i=1}^n \exp\left\{\mu_i^2\cdot\left(\frac{t}{1-2t}\right)\right\}\cdot\underbrace{\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{y_i^2}{2}}dy_i}_{=1} \\ &= (1-2t)^{-\frac{n}{2}}\cdot\exp\left\{\frac{t}{1-2t}\cdot\sum_{i=1}^n\mu_i^2\right\} \\ &= \frac{1}{(1-2t)^{n/2}}\cdot\exp\left\{\frac{\mu t}{1-2t}\right\}, \quad t<\frac{1}{2}. \end{aligned}$$

2. Let Y be a random variable with probability density function (4.2). Calculating $M_Y(t)$ yields

$$\begin{aligned} M_Y(t) &= \sum_{j=0}^{\infty} e^{-\frac{\mu}{2}}\frac{(\mu/2)^j}{j!}\underbrace{\int_0^{\infty} e^{xt}\cdot\frac{e^{-\frac{x}{2}}\cdot x^{\frac{n+2j}{2}-1}}{\Gamma\left(\frac{n+2j}{2}\right)\cdot\frac{n+2j}{2}}\,dx}_{=M_{\chi^2_{n+2j}}(t)=\frac{1}{(1-2t)^{(n+2j)/2}}\text{Satz 1.1.4}} \\ &= \frac{e^{-\frac{\mu}{2}}}{(1-2t)^{\frac{n}{2}}}\cdot\sum_{j=1}^{\infty}\left(\frac{\mu}{2(1-2t)}\right)^j\cdot\frac{1}{j!} \\ &= \frac{1}{(1-2t)^{\frac{n}{2}}}\cdot\exp\left\{-\frac{\mu}{2}+\frac{\mu}{2(1-2t)}\right\} \\ &= \frac{1}{(1-2t)^{\frac{n}{2}}}\cdot\exp\left\{\frac{\mu\cdot(1-(1-2t))}{2\cdot(1-2t)}\right\} \\ &= (1-2t)^{-\frac{n}{2}}\cdot\exp\left\{\frac{\mu t}{1-2t}\right\} \\ &\Longrightarrow \quad M_X(t) = M_Y(t), \quad t<\frac{1}{2} \end{aligned}$$

Using Lemma 4.1.22 implies $f_X(x) = f_Y(x)$ for almost all $x \in \mathbb{R}$.

□

Remark 4.1.24.

1. Definition 4.1.21 can be rewritten as:
   If $X \sim \mathcal{N}(\vec{\mu}, \mathcal{I})$, $\vec{\mu} = (\mu_1, \ldots, \mu_n)^\top$, then $|X|^2 = X^\top X \sim \chi^2_{n,\mu}$, where $\mu = |\vec{\mu}|^2$.

2. The property above can be generalized for $X \sim \mathcal{N}(\vec{\mu}, K)$, with a symmetric, positive definite $(n \times n)$ matrix K:
   $$X^\top K^{-1} X \sim \chi^2_{n,\tilde{\mu}}, \quad \text{where } \tilde{\mu} = \vec{\mu}^\top K^{-1} \vec{\mu},$$
   and since K is positive definite, there exists a $K^{\frac{1}{2}}$, such that $K = K^{\frac{1}{2}} K^{\frac{1}{2}\top}$. Then
   $$Y = K^{-\frac{1}{2}} X \sim \mathcal{N}(K^{-\frac{1}{2}} \mu, \mathcal{I}),$$
   since
   $$K^{-\frac{1}{2}} K K^{-\frac{1}{2}\top} = K^{-\frac{1}{2}} \cdot K^{\frac{1}{2}} \cdot K^{\frac{1}{2}\top} \cdot K^{-\frac{1}{2}\top} = \mathcal{I}$$
   and thus $Y^\top Y \overset{1.}{\sim} \chi^2_{n,\tilde{\mu}}$, with
   $$\tilde{\mu} = \left(K^{-\frac{1}{2}} \vec{\mu}\right)^\top K^{-\frac{1}{2}} \vec{\mu} = \vec{\mu}^\top K^{-\frac{1}{2}\top} K^{-\frac{1}{2}} \vec{\mu} = \vec{\mu}^\top K^{-1} \vec{\mu}.$$

Theorem 4.1.25. Let $X \sim \mathcal{N}(\mu, K)$, where K is a symmetric, positive definite $(n \times n)$ matrix and let A be another symmetric $(n \times n)$ matrix with the property $AK = (AK)^2$ (idempotence) and $\operatorname{rank}(A) = r \leq n$. Then:
$$X^\top A X \sim \chi^2_{r,\tilde{\mu}}, \text{ where } \tilde{\mu} = \mu^\top A \mu.$$

Proof A is positive semidefinite since

$$\begin{aligned} AK = (AK)^2 = AK \cdot AK \quad & | \, K^{-1} \\ \Longrightarrow A = AKA \Rightarrow \forall x \in \mathbb{R}^n : \; & x^\top A x = x^\top AKAx \\ = (\underbrace{Ax}_{=y})^\top K (\underbrace{Ax}_{=y}) \geq 0 \; & \text{because of the positive definiteness of } K. \\ \Longrightarrow \exists H : \; & \text{a } (n \times r) \text{ matrix with rank } (H) = r : \; A = HH^\top \end{aligned}$$

Thus it holds that
$$X^\top A X = X^\top H \cdot H^\top X = (\underbrace{H^\top X}_{=Y})^\top \cdot H^\top X = Y^\top Y$$

Further, $Y \sim \mathcal{N}(H^\top \mu, \mathcal{I}_r)$, since by Theorem 4.1.12 $Y \sim \mathcal{N}(H^\top \mu, H^\top K H)$ and rank $(H) = r$. Consequently, $H^\top H$ is a regular $(r \times r)$ matrix and

$$\begin{aligned} H^\top K H &= (H^\top H)^{-1}(H^\top \underbrace{H \cdot H^\top K H \cdot (H^\top}_{=AKA=A} H)(H^\top H)^{-1} \\ &= (H^\top H)^{-1} H^\top \cdot \underbrace{A}_{=HH^T} \cdot H (H^\top H)^{-1} \\ &= \mathcal{I}_r \end{aligned}$$

Then

$$X^\top A X = \mid Y \mid^2 \sim \chi^2_{r,\tilde{\mu}} \text{ with } \tilde{\mu} = (H^\top \mu)^2 = \mu^\top H \cdot H^\top \mu = \mu^\top A \mu.$$

□

Theorem 4.1.26 (Independence)**.** Let $X \sim \mathcal{N}(\mu, K)$ and K be a symmetric, positive semidefinite $(n \times n)$ matrix.

1. Let A, B be $(r_1 \times n)$ resp. $(r_2 \times n)$ matrices, $r_1, r_2 \leq n$ with $AKB^\top = 0$. Then the vectors AX and BX are independent.
2. Furthermore, let C be a symmetric, positive semidefinite $(n \times n)$ matrix with the property $AKC = 0$. Then AX and $X^\top C X$ are independent.

Proof

1. By theorem 4.1.10, 3) it holds that AX and BX are independent, if and only if $\varphi_{(AX,BX)}(t) = \varphi_{AX}(t) \cdot \varphi_{BX}(t)$, $t = (t_1, t_2)^\top \in \mathbb{R}^{r_1 + r_2}$, $t_1 \in \mathbb{R}^{r_1}$, $t_2 \in \mathbb{R}^{r_2}$. It has to be shown that:

$$\varphi_{(AX,BX)}(t) = \mathbb{E}\, e^{\left(i t_1^\top A + t_2^\top B\right) \cdot X} \stackrel{!}{=} \mathbb{E}\, e^{i t_1^\top A X} \cdot \mathbb{E}\, e^{i t_2^\top B X}.$$

It holds that

$$\begin{aligned} \varphi_{(AX,BX)}(t) &= \mathbb{E} e^{i\left(t_1^\top A + t_2^\top B\right) \cdot X} \\ &\stackrel{(Def. 4.1.2)}{=} e^{i\left(t_1^\top A + t_2^\top B\right) \cdot \mu - \frac{1}{2} \cdot \left(t_1^\top A + t_2^\top B\right) \cdot K \cdot \left(t_1^\top A + t_2^\top B\right)^\top}, \end{aligned}$$

and with

$$\begin{aligned} &\left(t_1^\top A + t_2^\top B\right) \cdot K \cdot \left(t_1^\top A + t_2^\top B\right)^\top \\ &= \left(t_1^\top A\right) K \left(t_1^\top A\right)^\top + \left(t_1^\top A\right)^\top K \left(t_2^\top B\right) \\ &\qquad + \left(t_2^\top B\right) K \left(t_1^\top A\right)^\top + \left(t_2^\top B\right) K \left(t_2^\top B\right)^\top \\ &= t_1^\top A K A^\top t_1 + t_1^\top \cdot \underbrace{A K B^\top}_{=0} \cdot t_2 + t_2^\top \cdot \underbrace{B K A^\top}_{=(AKB^\top)^\top = 0} \cdot t_1 + t_2^\top B K B^\top t_2 \end{aligned}$$

we get

$$\begin{aligned}\varphi_{(AX,BX)}(t) &= e^{it_1^\top A - \frac{1}{2}t_1^\top AKA^\top t_1} \cdot e^{it_2^\top B - \frac{1}{2}t_2^\top BKB^\top t_2} \\ &= \varphi_{AX}(t_1) \cdot \varphi_{BX}(t_2), \quad t_1 \in \mathbb{R}^{r_1},\, t_2 \in \mathbb{R}^{r_2}\end{aligned}$$

2. C is symmetric, positive semidefinite $\Longrightarrow$ There exists a $(n \times r)$ matrix H with rank $(H) = r \leq n$ and $C = HH^\top$, $\Longrightarrow H^\top H$ has rank r and is thus invertible. Then

$$X^\top CX = X^\top HH^\top X = (H^\top X)^\top \cdot H^\top X = |H^\top X|^2.$$

If AX and $H^\top X$ are independent, then AX and $X^\top CX = |H^\top X|^2$ are independent by the transformation theorem for random vectors. By 1) AX and $H^\top X$ are independent, if $AK(H^\top)^\top = AKH = 0$. By assumption

$$AKC = AKH \cdot H^\top = 0 \Longrightarrow AKH \cdot H^\top H = 0,$$

since $\exists (H^\top H)^{-1}$, it holds that

$$\begin{aligned}&0 = AKH \cdot H^\top H \cdot (H^\top H)^{-1} = AKH \Longrightarrow AKH = 0 \\ &\Longrightarrow AX \text{ and } H^\top X \text{ are independent} \\ &\Longrightarrow AX \text{ and } X^\top CX \text{ are independent.}\end{aligned}$$

□

4.2 Multivariate linear regression models with full rank

Multivariate linear regression has the form

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

where $Y = (Y_1, \ldots, Y_n)^\top$ is the random vector of the so-called *response variables*, the *design matrix*

$$X = (x_{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,m}}$$

is deterministic and has full rank, i.e., rank $(X) = r = m \leq n$, $\beta = (\beta_1, \ldots, \beta_m)^\top$ is the *parameter vector* and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ is the random vector of the *error terms*. In our setting, the error terms fulfill $\mathbb{E}\,\varepsilon_i = 0$, $\operatorname{Var}\varepsilon_i = \sigma^2 > 0$, $i \in \{1, \ldots, n\}$. The goal of this section is to estimate β and σ^2.

4.2.1 Method of least squares

Let the design matrix $X = (X_1, \ldots, X_m)$ be defined by deterministic vectors $X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^\top$, $j = 1, \ldots, m$, which generate the m-dimensional linear subspace $L_X = \langle X_1, \ldots, X_m \rangle$. Further, define the mean squared deviation between Y and $X\beta$ via

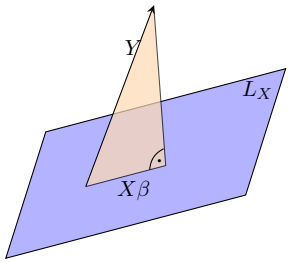
$$e(\beta) = \frac{1}{n}|Y - X\beta|^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - x_{i1}\beta_1 - \ldots - x_{im}\beta_m)^2 .$$

Then, the ordinary least squares estimator, or *OLS estimator* for short, $\hat{\beta}$ of β is defined as

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(e(\beta)). \tag{4.3}$$

Why does a solution $\beta \in \mathbb{R}^m$ of the quadratic optimisation (4.3) exist? Geometrically, $X\hat{\beta}$ can be interpreted as the orthogonal projection of the data vector Y on the linear subvector L_X as depicted in Figure 4.1. Formally, the existence of the solution will be shown by using the following theorem.

Figure 4.1: Projection on the linear subspace L_X



Theorem 4.2.1. Under the above conditions, there exists an unique OLS estimator $\hat{\beta}$, which solves the so-called *normal equation*

$$X^\top X\beta = X^\top Y. \tag{4.4}$$

Thus, it holds that

$$\hat{\beta} = \left(X^\top X\right)^{-1} X^\top Y.$$

Proof The necessary condition for the existence of the minimum is $e'(\beta) = 0$, that means

$$e'(\beta) = \left(\frac{\partial e(\beta)}{\partial \beta_1}, \ldots, \frac{\partial e(\beta)}{\partial \beta_m}\right)^\top = 0.$$

It holds that

$$e'(\beta) = \frac{2}{n}\left(X^\top X\beta - X^\top Y\right)$$

$\Longrightarrow \hat{\beta}$ is a solution of the normal equation $X^\top X\beta = X^\top Y$. Sufficient conditions for a minimum are given, since

$$e''(\beta) = \left(\frac{\partial^2 e(\beta)}{\partial\beta_i\partial\beta_j}\right)_{i,j=1,\ldots,m} = \frac{2}{n}X^\top X.$$

$X^\top X$ is symmetric and positive definite, since X has full rank:

$$\forall y \neq 0,\, y \in \mathbb{R}^m: \quad y^\top X^\top Xy = (Xy)^\top Xy = |Xy|^2 > 0$$

and $y \neq 0 \Longrightarrow Xy \neq 0$ implies that $e''(\beta)$ is positive definite. Thus $X^\top X$ is invertible. That means, $\hat{\beta}$ minimizes $e(\beta)$. The estimator $\hat{\beta} = \left(X^\top X\right)^{-1} X^\top Y$ can be obtained, by multiplying $\left(X^\top X\right)^{-1}$ to the left of the normal equation $X^\top X\beta = X^\top Y$. □

Example 4.2.2.

1. *Ordinary least squares*

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad m = 2,\, \beta = (\beta_1, \beta_2)^\top,\, Y = X\beta + \varepsilon$$

$\hat{\beta} = \left(\hat{\beta}_1, \hat{\beta}_2\right)$ yields the OLS estimator from [33].

$$\hat{\beta}_2 = \frac{S_{XY}^2}{S_{XX}^2}, \quad \hat{\beta}_1 = \overline{Y}_n - \overline{X}_n\hat{\beta}_2,$$

where

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i, \quad \overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$$

$$S_{XY}^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)\left(Y_i - \overline{Y}_n\right)$$

$$S_{XX}^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$$

Exercise 4.2.3. Prove that!

2. *Multiple linear regression*

   $Y = X\beta + \varepsilon$ with design matrix

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix} \text{ for } \beta = (\beta_0, \beta_1, \ldots, \beta_m)^\top .$$

The OLS estimator $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ is obviously a linear estimator with respect to Y.
Next, let us show that $\hat{\beta}$ is the *best linear, unbiased estimator* of β *(BLUE)* in the class

$$\mathcal{L} = \left\{ \tilde{\beta} = AY + b : \mathbb{E}\,\tilde{\beta} = \beta \right\}$$

of all linear unbiased estimators.

Theorem 4.2.4 (*Properties of the OLS estimator* $\hat{\beta}$)**.** Let $Y = X\beta + \varepsilon$ be a multivariate linear regression model with full rank m and error terms $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, which satisfy the following conditions:

$$\mathbb{E}\,\varepsilon = 0, \quad \text{Cov}\,(\varepsilon_i, \varepsilon_j) = \sigma^2 \delta_{ij},\ i, j = 1, \ldots, n \text{ for a } \sigma^2 \in (0, \infty).$$

Then,

1. the OLS estimator $\hat{\beta} = \left(X^\top X\right)^{-1} X^\top Y$ is unbiased: $\mathbb{E}\,\hat{\beta} = \beta$.
2. $\text{Cov}\,(\hat{\beta}) = \sigma^2 \left(X^\top X\right)^{-1}$
3. $\hat{\beta}$ has minimal variance among the estimators from $\mathcal{L}$, i.e.,

$$\forall \tilde{\beta} \in \mathcal{L} : \quad \text{Var}\,\tilde{\beta}_j \geq \text{Var}\,\hat{\beta}_j, \quad j = 1, \ldots, m.$$

Proof 1. Let us compute

$$\begin{aligned} \mathbb{E}\,\hat{\beta} &= \mathbb{E}\left[\left(X^\top X\right)^{-1} X^\top (X\beta + \varepsilon) \right] \\ &= \left(X^\top X\right)^{-1} \cdot X^\top X \cdot \beta + \left(X^\top X\right)^{-1} X^\top \cdot \underbrace{\mathbb{E}\,\varepsilon}_{=0} \\ &= \beta \quad \forall \beta \in \mathbb{R}^m. \end{aligned}$$

2. For all $\tilde{\beta} = AY + b \in \mathcal{L}$ it holds that

$$\begin{aligned}
&\beta = \mathbb{E}\,\tilde{\beta} = A\mathbb{E}\,Y + b = AX\beta + b \quad \forall \beta \in \mathbb{R}^m. \\
&\Longrightarrow b = 0, \quad AX = \mathcal{I}. \\
&\Longrightarrow \tilde{\beta} = AY = A\left(X\beta + \varepsilon\right) = AX\beta + A\varepsilon \\
&= \beta + A\varepsilon.
\end{aligned}$$

For

$$\hat{\beta} = \underbrace{\left(X^\top X\right)^{-1} X^\top}_{=A} Y$$

it holds that

$$\begin{aligned}
\text{Cov}\ \hat{\beta} &= \left(\mathbb{E}\left(\left(\hat{\beta}_i - \beta_i\right)\left(\hat{\beta}_j - \beta_j\right)\right)\right)_{i,j=1,\ldots,m} \\
&= \mathbb{E}\left(A\varepsilon \cdot (A\varepsilon)^\top\right) = \mathbb{E}\left(A\varepsilon\varepsilon^\top A^\top\right) = A\mathbb{E}\left(\varepsilon\varepsilon^\top\right) \cdot A^\top \\
&= A \cdot \sigma^2 \mathcal{I} A^\top = \sigma^2 AA^\top = \sigma^2 \left(X^\top X\right)^{-1} X^\top \left(\left(X^\top X\right)^{-1} X^\top\right)^\top \\
&= \sigma^2 \left(X^\top X\right)^{-1} X^\top X \left(X^\top X\right)^{-1} = \sigma^2 \left(X^\top X\right)^{-1}.
\end{aligned}$$

3. Let $\tilde{\beta} \in \mathcal{L}$, $\tilde{\beta} = \beta + A\varepsilon$. It has to be shown that

$$\left(\text{Cov}\ (\tilde{\beta})\right)_{ii} = \sigma^2 (AA^\top)_{ii} \geq \left(\text{Cov}\ (\hat{\beta})\right)_{ii} = \sigma^2 (X^\top X)^{-1}_{ii},$$

for $i = 1, \ldots, m$.
Let $D = A - (X^\top X)^{-1} X^\top$, then $A = D + (X^\top X)^{-1} X^\top$,

$$\begin{aligned}
AA^\top &= \left(D + \left(X^\top X\right)^{-1} X^\top\right)\left(D^\top + X\left(X^\top X\right)^{-1\top}\right) \\
&= DD^\top + \left(X^\top X\right)^{-1},
\end{aligned}$$

since

$$\begin{aligned}
DX\left(X^\top X\right)^{-1} &= \left(\underbrace{AX}_{=\mathcal{I}} - \underbrace{\left(X^\top X\right)^{-1} X^\top X}_{=\mathcal{I}}\right)\left(X^\top X\right)^{-1} \\
&= 0
\end{aligned}$$

and

$$\begin{aligned}
\left(X^\top X\right)^{-1} X^\top D^\top &= \left(X^\top X\right)^{-1} X^\top \left(A^\top - X\left(X^\top X\right)^{-1\top}\right) \\
&= \left(X^\top X\right)^{-1} \left(\underbrace{(AX)^\top}_{=\mathcal{I}} - \underbrace{X^\top X\left(X^\top X\right)^{-1}}_{=\mathcal{I}}\right) \\
&= 0.
\end{aligned}$$

$$\begin{aligned}
&\Longrightarrow \left(AA^\top\right)_{ii} = \underbrace{\left(DD^\top\right)_{ii}}_{\geq 0} + \left(X^\top X\right)^{-1}_{ii} \geq \left(X^\top X\right)^{-1}_{ii} \\
&\Longrightarrow \operatorname{Var}\hat{\beta}_i \leq \operatorname{Var}\tilde{\beta}_i, \quad i = 1, \ldots, m.
\end{aligned}$$

□

Theorem 4.2.5. Let $\hat{\beta}_n$ be the OLS estimator of the linear regression model from above and $\{a_n\}_{n\in\mathbb{N}}$ be a sequence with $a_n \neq 0$, $n \in \mathbb{N}$, $a_n \to 0$ $(n \to \infty)$. Additionally, assume that there exists a regular $(m \times m)$ matrix Q with

$$Q = \lim_{n\to\infty} a_n \left(X_n^\top X_n\right).$$

Then, $\hat{\beta}_n$ is weakly consistent, i.e.,

$$\hat{\beta}_n \xrightarrow[n\to\infty]{p} \beta.$$

Proof

$$\hat{\beta}_n \xrightarrow[n\to\infty]{p} \beta \Longleftrightarrow P\left(\left|\hat{\beta}_n - \beta\right| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0 \quad \forall \varepsilon > 0.$$

$$\begin{aligned}
P\left(\left|\hat{\beta}_n - \beta\right| > \varepsilon\right) &= P\left(\left|\hat{\beta}_n - \beta\right|^2 > \varepsilon^2\right) \\
&= P\left(\sum_{i=1}^{m} \left|\hat{\beta}_{in} - \beta_i\right|^2 > \varepsilon^2\right) \\
&\leq P\left(\bigcup_{i=1}^{m} \left\{\left|\hat{\beta}_{in} - \beta_i\right|^2 > \frac{\varepsilon^2}{m}\right\}\right) \\
&\leq \sum_{i=1}^{m} P\left(\left|\hat{\beta}_{in} - \beta_i\right| > \frac{\varepsilon}{\sqrt{m}}\right) \\
&\overset{\text{Tschebyschew}}{\leq} m \sum_{i=1}^{m} \frac{\operatorname{Var}\hat{\beta}_{in}}{\varepsilon^2} \xrightarrow[n\to\infty]{} 0 \\
&\text{if } \operatorname{Var}\hat{\beta}_{in} \xrightarrow[n\to\infty]{} 0, \quad i = 1, \ldots, m.
\end{aligned}$$

$\operatorname{Var}\hat{\beta}_{in}$ is a diagonal element of the matrix

$$\operatorname{Cov}\hat{\beta}_n \overset{(Satz 4.2.4)}{=} \sigma^2 \left(X_n^\top X_n\right)^{-1}.$$

If $\operatorname{Cov}\hat{\beta}_n \xrightarrow[n\to\infty]{} 0$ is true, then so is the theorem. Thus, let us show this convergence.

Since there exists a matrix

$$Q^{-1} = \lim_{n\to\infty} \frac{1}{a_n} \left(X_n^\top X_n\right)^{-1}$$

we can show that

$$\begin{aligned}\lim_{n\to\infty} \operatorname{Cov} \hat{\beta}_n &= \sigma^2 \lim_{n\to\infty} \left(X_n^\top X_n\right)^{-1} = \sigma^2 \lim_{n\to\infty} a_n \cdot \frac{1}{a_n} \left(X_n^\top X_n\right)^{-1} \\ &= 0 \cdot Q^{-1} \cdot \sigma^2 = 0.\end{aligned}$$

□

4.2.2 Estimator of the variance σ^2

Introduce the estimator $\hat{\sigma}^2$ for the variance σ^2 of the error terms ε_i as follows:

$$\hat{\sigma}^2 = \frac{1}{n-m} \left|Y - X\hat{\beta}\right|^2. \tag{4.5}$$

This is a generalized version of the variance estimator from the simple linear regression, which has already been introduced in [33].

Theorem 4.2.6 (*Expectation*)**.** The variance estimator

$$\hat{\sigma}^2 = \frac{1}{n-m} \left|Y - X\hat{\beta}\right|^2$$

is unbiased. That means,

$$\mathbb{E}\,\hat{\sigma}^2 = \sigma^2.$$

Proof

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-m} \left(Y - X\hat{\beta}\right)^\top \left(Y - X\hat{\beta}\right) \\ &= \frac{1}{n-m} \left(Y - X(X^\top X)^{-1} X^T Y\right)^\top \left(Y - X\left(X^\top X\right)^{-1} X^\top Y\right) \\ &= \frac{1}{n-m} (DY)^\top DY\end{aligned}$$

where $D = \mathcal{I} - X(X^\top X)^{-1}X^\top$ is a $(n \times n)$ matrix. Then,

$$\hat{\sigma}^2 = \frac{1}{n-m} Y^\top D^\top DY = \frac{1}{n-m} Y^\top D^2 Y = \frac{1}{n-m} Y^\top DY,$$

if $D^\top = D$ and $D^2 = D$, i.e. D is symmetric and idempotent. Indeed, it holds that:

$$\begin{aligned}D^\top &= \mathcal{I} - \left(X^\top\right)^\top \left(X^\top X\right)^{\top-1} X^\top = \mathcal{I} - X\left(X^\top X\right)^{-1} X^\top = D. \\ D^2 &= \left(\mathcal{I} - X(X^\top X)^{-1}X^T\right)\left(\mathcal{I} - X\left(X^\top X\right)^{-1} X^\top\right) \\ &= \mathcal{I} - 2X\left(X^\top X\right)^{-1} X^\top + X\left(X^\top X\right)^{-1} X^\top X\left(X^\top X\right)^{-1} X^\top \\ &= \mathcal{I} - X\left(X^\top X\right)^{-1} X^\top = D.\end{aligned}$$

Furthermore it holds that

$$\hat{\sigma}^2 = \frac{1}{n-m} \cdot \operatorname{trace}\left(Y^\top D Y\right) = \frac{1}{n-m} \cdot \operatorname{trace}\left(D Y Y^\top\right)$$
$$\Longrightarrow \mathbb{E}\,\hat{\sigma}^2 = \frac{1}{n-m} \cdot \operatorname{trace}\left(D \mathbb{E}\left(Y Y^\top\right)\right) = \frac{\sigma^2}{n-m} \cdot \operatorname{trace}(D),$$

since

$$\operatorname{trace}\left(D \cdot \mathbb{E}\left(Y Y^\top\right)\right) =$$
$$= \operatorname{trace}\Big(D(X\beta)(X\beta)^\top + DX\beta \underbrace{\mathbb{E}\,\varepsilon^\top}_{=0} + D \underbrace{\mathbb{E}\,\varepsilon}_{=0}(X\beta)^\top + D \cdot \underbrace{\mathbb{E}\,\varepsilon\varepsilon^\top}_{=\operatorname{Cov}\varepsilon = \sigma^2 \cdot \mathcal{I}}\Big)$$

and

$$DX = \left(\mathcal{I} - X\left(X^\top X\right)^{-1} X^T\right) X$$
$$= X - X\left(X^\top X\right)^{-1} X^\top X = X - X = 0.$$

Now it needs to be shown that $\operatorname{trace}(D) = n - m$:

$$\operatorname{trace}(D) = \operatorname{trace}\left(\mathcal{I} - X\left(X^\top X\right)^{-1} X^\top\right)$$
$$= \operatorname{trace}(\mathcal{I}) - \operatorname{trace}\left(X\left(X^\top X\right)^{-1} X^\top\right)$$
$$= n - \operatorname{trace}\Big(\underbrace{X^\top X \cdot \left(X^\top X\right)^{-1}}_{\text{eine } (m \times m)\text{-Matrix}}\Big) = n - m.$$

□

4.2.3 Maximum likelihood estimator for β and σ^2

In order to construct a maximum likelihood estimator for β and σ^2 resp. the distributional properties of the OLS estimators $\hat{\beta}$ and $\hat{\sigma}^2$, the distribution of ε resp. Y has to be specified. In the following, normally distributed i.i.d error terms are assumed, i.e.,

$$\varepsilon \sim N\left(0, \sigma^2 \mathcal{I}\right), \quad \sigma^2 > 0.$$

Clearly, this implies

$$Y \sim N\left(X\beta, \sigma^2 \mathcal{I}\right).$$

What do the distributions of the OLS estimators $\hat{\beta}$ and $\hat{\sigma}^2$ look like? Since $\hat{\beta}$ is linearly dependent of Y, unbiased and $\operatorname{Cov}\hat{\beta} = \hat{\sigma}^2\left(X^\top X\right)^{-1}$, simple calculations yield

$$\hat{\beta} \sim N\left(\beta, \sigma^2\left(X^\top X\right)^{-1}\right).$$

In the following, the maximum likelihood estimator for β and σ^2, namely $\tilde{\beta}$ and $\tilde{\sigma}^2$ are calculated. Once they have been calculated, it can be seen that they are closely related to the OLE estimator.

$$\begin{aligned}\tilde{\beta} &= \hat{\beta},\\ \tilde{\sigma}^2 &= \frac{n-m}{n}\hat{\sigma}^2.\end{aligned}$$

Consider the Likelihood function of Y:

$$L(y,\beta,\sigma^2) = f_Y(y) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n}\cdot\exp\left\{-\frac{1}{2\sigma^2}\left(y-X\beta\right)^\top\left(y-X\beta\right)\right\}$$

and the Log likelihood function

$$\log L(y,\beta,\sigma^2) = -\frac{n}{2}\log\left(2\pi\right)\underbrace{-\frac{n}{2}\log\left(\sigma^2\right)-\frac{1}{2\sigma^2}\left|y-X\beta\right|^2}_{:=g}.$$

The maximum likelihood estimators are then given by

$$\left(\tilde{\beta},\tilde{\sigma}^2\right) = \underset{\beta\in\mathbb{R}^m,\,\sigma^2>0}{\operatorname{argmax}}\ \log L(y,\beta,\sigma^2),$$

if they exist.

Theorem 4.2.7 (*Maximum likelihood estimation of $\tilde{\beta}$ and $\tilde{\sigma}^2$*)**.** There exist unique maximum likelihood estimators for β and σ^2 which are given by

$$\begin{aligned}\tilde{\beta} &= \hat{\beta} = \left(X^\top X\right)^{-1}X^\top Y\\ \tilde{\sigma}^2 &= \frac{n-m}{n}\hat{\sigma}^2 = \frac{1}{n}\left|Y-X\tilde{\beta}\right|^2.\end{aligned}$$

Proof Fix $\sigma^2>0$ and find

$$\tilde{\beta} = \underset{\beta\in\mathbb{R}^m}{\operatorname{argmax}}\log L(Y,\beta,\sigma^2) = \underset{\beta\in\mathbb{R}^m}{\operatorname{argmin}}\left|Y-X\beta\right|^2,$$

which implies that $\tilde{\beta}$ coincides with the known OLS estimator $\hat{\beta} = \left(X^\top X\right)^{-1}X^\top Y$ and does not depend on σ^2. Therefore, we can compute

$$\tilde{\sigma}^2 = \underset{\sigma^2>0}{\operatorname{argmax}}\log L\left(Y,\tilde{\beta},\sigma^2\right) = \underset{\sigma^2>0}{\operatorname{argmax}}\, g(\sigma^2).$$

It holds that

$$g\left(\sigma^2\right)\underset{\sigma^2\to+\infty}{\longrightarrow}-\infty,\quad g\left(\sigma^2\right)\underset{\sigma^2\to 0}{\longrightarrow}-\infty,$$

since $|Y - X\beta|^2 \neq 0$ because $Y \sim N(X\beta, \sigma^2\mathcal{I}) \in \{Xy : y \in \mathbb{R}^m\}$ with probability zero. Since

$$g'(\sigma^2) = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{|Y - X\beta|}{2(\sigma^2)^2} = 0, \quad \tilde{\sigma}^2 = \frac{1}{n}\left|Y - X\tilde{\beta}\right|^2$$

maximizes $g(\sigma^2)$, that means $\tilde{\sigma}^2$ is a maximum likelihood estimator for σ^2. □

Theorem 4.2.8. Under the assumptions above it holds that

1. $\mathbb{E}\,\tilde{\sigma}^2 = \frac{n-m}{n}\sigma^2$, that means $\tilde{\sigma}^2$ is biased; but it is asymptotically unbiased.
2. $\frac{n}{\sigma^2}\tilde{\sigma}^2 \sim \chi^2_{n-m}$, $\frac{n-m}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-m}$.

Proof

1. Trivial (similar to the proof of Theorem 4.2.6)
2. Only the assertion for $\hat{\sigma}^2$ is shown:

$$\begin{aligned}
\frac{n-m}{\sigma^2}\hat{\sigma}^2 &= \frac{1}{\sigma^2}\left|Y - X\hat{\beta}\right|^2 \\
&= \frac{1}{\sigma^2}Y^\top \underbrace{D}_{=D^2} Y \quad \text{(by the proof of Theorem 4.2.6)} \\
&= \frac{1}{\sigma^2}(DY)^\top DY = \frac{1}{\sigma^2}\big(\underbrace{D(X}_{=0}\beta + \varepsilon)\big)^\top \cdot \underbrace{D(X}_{=0}\beta + \varepsilon) \\
&= \frac{1}{\sigma^2}(D\varepsilon)^\top D\varepsilon = \left(\frac{\varepsilon}{\sigma}^\top\right) D \left(\frac{\varepsilon}{\sigma}\right),
\end{aligned}$$

where

$$\left(\frac{\varepsilon}{\sigma}\right) \sim N(0, \mathcal{I}).$$

By Theorem 4.1.25 it holds that

$$\frac{\varepsilon}{\sigma}^\top D \frac{\varepsilon}{\sigma} \sim \chi^2_r,$$

where $r = \text{rank}\,(D)$, since $D\mathcal{I} = D$ is idempotent. If $r = n - m$, then $(n-m)\hat{\sigma}^2 \sim \chi^2_{n-m}$. It needs to be shown that rank $(D) = r = n - m$. From linear algebra it is known that rank $(D) = n - \dim(\text{Kern}\,(D))$. Now Kern $(D) = \{Xx : x \in \mathbb{R}^m\}$ and thus $\dim(\text{Kern}\,(D)) = m$, since rank $(X) = m$. It holds that $\{Xx : x \in \mathbb{R}^n\} \subseteq \text{Kern}\,(D)$, since

$$DX = (\mathcal{I} - X(X^\top X)^{-1}X^\top)X = X - (X^\top X)^{-1}X^\top X = 0.$$

and Kern $(D) \subseteq \{Xx : x \in \mathbb{R}^m\}$, since

$$\forall y \in \text{Kern } (D): \quad Dy = 0 \iff (\mathcal{I} - X(X^\top X)^{-1}X^\top)y = 0$$
$$\iff y = X \cdot \underbrace{(X^\top X)^{-1}X^\top y}_{x} = Xx \in \{Xx : x \in \mathbb{R}^m\}.$$

□

Theorem 4.2.9. Let $Y = X\beta + \varepsilon$ be a multivariate regression model with $Y = (Y_1, \ldots, Y_n)^\top$, design matrix X with rank $(X) = m$, $\beta = (\beta_1, \ldots, \beta_m)^\top$, $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathcal{I})$. Then, the estimators $\hat{\beta} = (X^\top X)^{-1}X^\top Y$ for β resp. $\hat{\sigma}^2 = \frac{1}{n-m}|Y - X\hat{\beta}|^2$ for σ^2 are independent.

Proof In this proof Theorem 4.1.26 is used. In order to do so, $\hat{\beta}$ has to be expressed as a linear and $\hat{\sigma}^2$ as quadratic form of ε. It has been shown in the proofs of Theorem 4.2.4 and 4.2.8 that

$$\hat{\beta} = \beta + \underbrace{(X^\top X)^{-1}X^\top}_{=A}\varepsilon,$$
$$\hat{\sigma}^2 = \frac{1}{n-m}\varepsilon^\top D\varepsilon, \text{ where } D = \mathcal{I} - X(X^\top X)^{-1}X^\top.$$

Moreover it holds that $AD = 0$, by the proof of Theorem 4.2.6

$$(AD)^\top = D^\top A^\top = \underbrace{D \cdot X}_{=0}((X^\top X)^{-1})^\top = 0.$$

Since $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathcal{I})$, it holds that

$$A\sigma^2\mathcal{I}D = 0.$$

Thus, the assumptions of Theorem 4.1.26 are satisfied and $\hat{\beta}$ and $\hat{\sigma}^2$ are independent. □

4.2.4 Tests for regression parameters

In this section the hypotheses

$$H_0 : \beta = \beta_0 \text{ vs. } H_1 : \beta \neq \beta_0$$

are tested for a $\beta_0 \in \mathbb{R}^m$. In order to do so, define the test statistic

$$T = \frac{\left(\hat{\beta} - \beta_0\right)^\top X^\top X \left(\hat{\beta} - \beta_0\right)}{m\hat{\sigma}^2}.$$

Theorem 4.2.11 implies that under H_0

$$T \sim F_{m,n-m}.$$

Thus, H_0 is rejected, if $T > F_{m,n-m,1-\alpha}$, where $F_{m,n-m,1-\alpha}$ is the $(1-\alpha)$ quantile of the $F_{m,n-m}$ distribution. This is a test with confidence level $\alpha \in (0,1)$.
Special case: The case $\beta_0 = 0$ describes *test for connectivity* ; that means it is tested, whether $\beta_1, \ldots, \beta_m$ are relevant for describing the data Y.

Remark 4.2.10.

1. How can we test that the test statistic T can actually distinguish H_0 from H_1? Introduce

$$\tilde{Y} = Y - X\hat{\beta} =: Y - \hat{Y}.$$

Then,

$$\hat{\sigma}^2 = \frac{1}{n-m}\left|\tilde{Y}\right|^2$$

and $\tilde{Y}$ is the vector of *residuals*.

Without loss of generality assume $\beta_0 = 0$. If H_0 is false, then $\beta \neq 0$, and

$$|X\beta|^2 = (X\beta)^\top X\beta = \beta^\top X^\top X\beta > 0,$$

since X has full rank. This implies that H_0 has to be rejected, if

$$\left|\hat{Y}\right|^2 = \left|X\hat{\beta}\right|^2 = \hat{\beta}^\top X^\top X\hat{\beta} \gg 0.$$

In the test statistic $|X\hat{\beta}|^2$ the variation of the estimation of β is not considered. In order to do so, a new test statistic T can be defined by dividing $|X\hat{\beta}|^2$ by $\hat{\sigma}^2$, i.e.

$$T = \frac{\hat{\beta}^\top X^\top X\hat{\beta}}{m\cdot\hat{\sigma}^2} = \frac{\left|\hat{Y}\right|^2}{\frac{m}{n-m}\left|Y-\hat{Y}\right|^2}.$$

The Pythagorean theorem implies

$$|Y|^2 = \left|\tilde{Y}\right|^2 + \left|\hat{Y}\right|^2.$$

Then, under H_0

$$\mathbb{E}\,|\hat{Y}|^2 = \mathbb{E}\,|Y|^2 - \mathbb{E}\,|Y-\hat{Y}|^2 = n\sigma^2 - \mathbb{E}\,|\tilde{Y}|^2$$

holds and thus

$$\frac{\mathbb{E}\,|\hat{Y}|^2}{\mathbb{E}\left(\frac{m}{n-m}\left|\tilde{Y}\right|^2\right)} \overset{(H_0)}{=} \frac{n\sigma^2 - \mathbb{E}\,|\tilde{Y}|^2}{\frac{m}{n-m}\mathbb{E}\,|\tilde{Y}|^2} = \frac{n-m}{m}\left(\frac{n\sigma^2}{\mathbb{E}\,|\tilde{Y}|^2} - 1\right),$$

where we have used that $\mathbb{E}\,|Y|^2 = \mathbb{E}\left(Y^\top Y\right) = \sigma^2 n$ and $Y \sim \mathcal{N}(0, \sigma^2\mathcal{I})$. Consequently, the test statistic T is sensible with respect to variations of H_0.

2. The term

$$\left|\tilde{Y}\right|^2 = \left|Y - \hat{Y}\right|^2$$

is called *residual distribution*. Now the *coefficient of determination* R^2 as introduced in [33] can be generalized as

$$R^2 = 1 - \frac{|\tilde{Y}|^2}{\left|Y - \overline{Y}_n \cdot e\right|^2},$$

where $e = (1, \ldots, 1)^\top$, $\overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$.

Theorem 4.2.11. Under $H_0 : \beta = \beta_0$ it holds that

$$T = \frac{\left(\hat{\beta} - \beta_0\right)^\top X^\top X \left(\hat{\beta} - \beta_0\right)}{m\hat{\sigma}^2} \sim F_{m,n-m}.$$

Proof It holds that

$$\begin{aligned} &\hat{\beta} \sim N\left(\beta_0, \sigma^2 \left(X^\top X\right)^{-1}\right) \\ &\implies \hat{\beta} - \beta_0 \sim N\Big(0, \underbrace{\sigma^2 (X^\top X)^{-1}}_{:=K}\Big). \end{aligned}$$

If $A = \frac{X^\top X}{\sigma^2}$, then $AK = \mathcal{I}$ is idempotent. Then by Theorem 4.1.25

$$\left(\hat{\beta} - \beta_0\right)^\top A \left(\hat{\beta} - \beta_0\right) \overset{H_0}{\sim} \chi^2_m.$$

Note that under H_1 the distribution of $(\hat{\beta} - \beta_0)^\top A(\hat{\beta} - \beta_0)$ does not follow a centered χ^2 distribution.
Furthermore, it holds that

$$\frac{n-m}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-m}.$$

Also, Theorem 4.2.9 implies the independence of $(\hat{\beta} - \beta_0)^\top A(\hat{\beta} - \beta_0)$ and $\frac{n-m}{\sigma^2}\hat{\sigma}^2$. Therefore,

$$T = \frac{(\hat{\beta} - \beta_0)^\top (X^\top X)(\hat{\beta} - \beta_0)/m}{(n-m)\hat{\sigma}^2/(n-m)} \sim F_{m,n-m}$$

by definition of the F distribution.

□

Now, let us test the relevance of the parameters β_j, i.e., we test

$$H_0 : \beta_j = \beta_{0j} \text{ vs. } H_1 : \beta_j \neq \beta_{0j}.$$

Theorem 4.2.12. Under $H_0 : \beta_j = \beta_{0j}$ it holds that

$$T_j = \frac{\hat{\beta}_j - \beta_{0j}}{\hat{\sigma}\sqrt{x^{jj}}} \sim t_{n-m},$$

where x^{jj} is the j-th diagonal entry of the matrix $(X^\top X)^{-1}$.

Proof $\hat{\beta} \overset{H_0}{\sim} \mathcal{N}(\beta_0, \sigma^2(X^\top X)^{-1})$ implies $\hat{\beta}_j \overset{H_0}{\sim} \mathcal{N}(\beta_{0j}, \sigma^2 x^{jj})$ and thus $\hat{\beta}_j - \beta_{0j} \sim \mathcal{N}(0, \sigma^2 x^{jj})$. Consequently, $A := \frac{\hat{\beta}_j - \beta_{0j}}{\sigma\sqrt{x^{jj}}} \sim \mathcal{N}(0,1)$. Furthermore, it holds that $B := \frac{(n-m)\hat{\sigma}^2}{\sigma^2} \overset{H_0}{\sim} \chi^2_{n-m}$, and by Theorem 4.2.9 the statistics A and B are independent which implies that

$$T_j = \frac{\frac{\hat{\beta}_j - \beta_{0j}}{\sigma\sqrt{x^{jj}}}}{\sqrt{\frac{(n-m)\hat{\sigma}^2}{(n-m)\sigma^2}}} \sim t_{n-m}.$$

□

Thus, a test for $H_0 : \beta_j = \beta_{j0}$ vs. $H_1 :$ not H_0, with confidence level α can be constructed using test statistic T by rejected the null hypothesis, if $|T| > t_{n-m,1-\alpha/2}$.
Next, let us test the hypothesis

$$H_0 : \beta_{j_1} = \beta_{0j_1}, \ldots, \beta_{j_l} = \beta_{0j_l} \text{ vs. } H_1 : \exists i \in \{1, \ldots, l\} : \beta_{j_i} \neq \beta_{0j_i}.$$

Exercise 4.2.13. Show that under H_0 the following assertion holds:

$$T = \frac{(\hat{\beta}' - \beta_0')^\top K' (\hat{\beta}' - \beta_0')}{l\hat{\sigma}^2} \sim F_{l,n-m},$$

where

$$\begin{aligned}
\hat{\beta}' &= (\hat{\beta}_{j_1}, \ldots, \hat{\beta}_{j_l}), \\
\beta_0' &= (\beta_{0j_1}, \ldots, \beta_{0j_l}), \\
K' &= \begin{pmatrix} x^{j_1 j_1} & \ldots & x^{j_1 j_l} \\ \vdots & \vdots & \vdots \\ x^{j_l j_1} & \ldots & x^{j_l j_l} \end{pmatrix}^{-1}.
\end{aligned}$$

Construct the corresponding F test!

Test for linear combinations of parameters

Let us consider

$$H_0 : H\beta = c \text{ vs. } H_1 : H\beta \neq c,$$

where H is a $(r \times m)$ matrix and $c \in \mathbb{R}^r$.

Theorem 4.2.14. Under H_0 it holds that

$$T = \frac{(H\hat{\beta} - c)^\top (H(X^\top X)^{-1}H^\top)^{-1}(H\hat{\beta} - c)}{r\hat{\sigma^2}} \sim F_{r,n-m}.$$

Thus $H_0 : H\beta = c$ is rejected, if $T > F_{r,n-m,1-\alpha}$.

Exercise 4.2.15. Prove Theorem 4.2.14!

4.2.5 Confidence region

1. *Confidence interval for β_j*

   In Theorem 4.2.12 it has been shown that

   $$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \cdot \sqrt{x^{jj}}} \sim t_{n-m},$$

   where $(X^\top X)^{-1} = (x^{ij})_{i,j=1,\ldots,m}$. By using the standard methodology, the following confidence interval for β_j with confidence level $1-\alpha$ can be constructed as follows

   $$P\left(\hat{\beta}_j - t_{n-m,1-\alpha/2} \cdot \hat{\sigma}\sqrt{x^{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{n-m,1-\alpha/2} \cdot \hat{\sigma}\sqrt{x^{jj}}\right) = 1-\alpha.$$

2. *Simultaneous confidence region for $\beta = (\beta_1, \ldots, \beta_m)^\top$*

   From the *Bonferroni inequality* it is known that

   $$P\left(\bigcap_{j=1}^{m} A_j\right) \geq \sum_{j=1}^{m} P(A_j) - (m-1),$$

   for sets $A_1, \ldots, A_m$. Now, using the sets

   $$A_j := \left\{\beta_j \in \left[\hat{\beta}_j - t_{n-m,1-\alpha/(2m)} \cdot \hat{\sigma}\sqrt{x^{jj}}, \hat{\beta}_j + t_{n-m,1-\alpha/(2m)} \cdot \hat{\sigma}\sqrt{x^{jj}}\right]\right\}$$

   yields

   $$\begin{aligned} &P\Big(A_j,\ j = 1, \ldots, m\Big) \\ &\geq \sum_{j=1}^{m} P(A_j) - (m-1) = m \cdot \left(1 - \frac{\alpha}{m}\right) - m + 1 = 1 - \alpha. \end{aligned}$$

   This implies that

   $$\left\{\beta = (\beta_1, \ldots, \beta_m)^\top : \beta_j \in A_j\right\}$$

   is a simultaneous confidence region for β with confidence level $1-\alpha$.

3. *Confidence ellipsoid for β.*

   In Theorem 4.2.11 it has been shown that

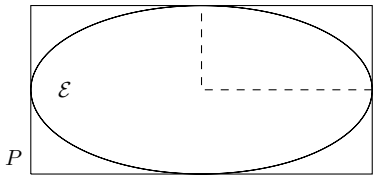
$$T = \frac{(\hat{\beta} - \beta)^\top (X^\top X)(\hat{\beta} - \beta)}{m\hat{\sigma}^2} \sim F_{m,n-m}.$$

   This implies that

$$P\left(T \leq F_{m,n-m,1-\alpha}\right) = 1 - \alpha \quad \text{and}$$
$$\mathcal{E} = \left\{ \beta \in \mathbb{R}^m : \frac{(\hat{\beta} - \beta)^\top (X^\top X)(\hat{\beta} - \beta)}{m\hat{\sigma}^2} \leq F_{m,n-m,1-\alpha} \right\}$$

   is a confidence ellipsoid with confidence level $1 - \alpha$ (see Figure 4.2).

Figure 4.2: Confidence ellipsoid



   Since an ellipsoid can be embedded in the minimal parallelepiped P, such that the length of the sides of P are $2\times$ length of the half-axes of $\mathcal{E}$, the following simultaneous confidence region for $\beta = (\beta_1, \ldots, \beta_m)^\top$ can be constructed:

$$P = \left\{ \beta : \hat{\beta}_j - \hat{\sigma}\sqrt{m x^{jj} F_{m,n-m,1-\alpha}} \leq \beta_j \leq \hat{\beta}_j + \hat{\sigma}\sqrt{m x^{jj} F_{m,n-m,1-\alpha}} \right\}$$
$$j = 1, \ldots, m$$

4. *Confidence interval for the expected target value $x_{01}\beta_1 + \ldots + x_{0m}\beta_m$.*

   Let $Y_0 = x_{01}\beta_1 + \ldots + x_{0m}\beta_m + \varepsilon_0$ be a new target value with $\mathbb{E}\,\varepsilon_0 = 0$. Then

$$\mathbb{E}\, Y_0 = \sum_{i=1}^{n} x_{0i}\beta_i.$$

   In the following, a confidence interval for $\mathbb{E}\, Y_0$ is constructed. In oder to do so, the proof idea of Theorem 4.2.12 combined with Theorem 4.2.14 with $H = (x_{01}, \ldots, x_{0m}) = x_0^\top$, $r = 1$ is used. Then

$$T = \frac{\sum_{i=1}^{m} \hat{\beta}_i x_{0i} - \sum_{i=1}^{m} \beta_i x_{0i}}{\hat{\sigma}\sqrt{x_0^\top (X^\top X)^{-1} x_0}} \sim t_{n-m}.$$

Thus

$$\Big\{\beta = (\beta_1, \ldots, \beta_m)^\top : \sum_{i=1}^{m} x_{0i}\hat{\beta}_i - \hat{\sigma}\sqrt{x_0^\top (X^\top X)^{-1} x_0} \cdot t_{n-m,1-\alpha/2} \leq \sum_{i=1}^{m} x_{0i}\beta_i \leq \sum_{i=1}^{m} x_{0i}\hat{\beta}_i + \hat{\sigma}\sqrt{x_0^\top (X^\top X)^{-1} x_0} \cdot t_{n-m,1-\alpha/2}\Big\}$$

is a confidence interval $\sum_{i=1}^{m} x_{0i}\beta_i$ with confidence level $1-\alpha$.

5. *Forecast interval for the target variable Y_0.*

   For $Y_0 = \sum_{i=1}^{m} x_{0i}\beta_i + \varepsilon_0$ with $\varepsilon_0 \sim \mathcal{N}(0, \sigma^2)$ independent of $\varepsilon_1, \ldots, \varepsilon_n$, it holds that

$$\begin{aligned} & x_0^\top \hat{\beta} - Y_0 \sim \mathcal{N}(0, \sigma^2(1 + x_0^\top (X^\top X)^{-1} x_0)) \\ \Longrightarrow & \frac{x_0^\top \hat{\beta} - Y_0}{\sigma\sqrt{1 + x_0^\top (X^\top X)^{-1} x_0}} \sim \mathcal{N}(0,1) \\ \Longrightarrow & \frac{x_0^\top \hat{\beta} - Y_0}{\hat{\sigma}\sqrt{1 + x_0^\top (X^\top X)^{-1} x_0}} \sim t_{n-m}, \end{aligned}$$

   Thus,

$$\left(x_0^\top \hat{\beta} - c,\, x_0^\top \hat{\beta} + c\right)$$

$$\text{with } c = \hat{\sigma}\sqrt{1 + x_0^\top (X^\top X)^{-1} \cdot x_0} \cdot t_{n-m,1-\alpha/2}$$

   is a forecast interval for the target variable Y_0 with confidence level $1-\alpha$.

6. *Confidence band for the regression plane $y = \beta_1 + \sum_{i=2}^{m} x_i\beta_i$ in the multiple regression model.*

   Let $Y = X\beta + \varepsilon$, where

$$X = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1m} \\ 1 & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \cdots & x_{nm} \end{pmatrix} \quad \text{and } \varepsilon \sim \mathcal{N}(0,\, \sigma^2 \cdot \mathcal{I}).$$

   The goal is to construct a confidence band $B(x)$ for y. It holds that

$$P\left(y = \beta_1 + \sum_{i=2}^{m} \beta_i x_i \in B(x)\right) = 1 - \alpha \quad \forall x \in \mathbb{R}_1^{m-1},$$

   where $R_1^{m-1} = \left\{(1, x_2, \ldots, x_m)^\top \in \mathbb{R}^m\right\}$.

Theorem 4.2.16. Furthermore,

$$P\Bigg(\max_{x\in\mathbb{R}_1^{m-1}} \frac{\Big(x^T\hat{\beta} - \overbrace{\Big(\beta_1 + \sum_{i=2}^{m}\beta_i x_i\Big)}^{=y}\Big)^2}{\hat{\sigma}^2 x^\top (X^\top X)^{-1}x} \leq m\cdot F_{m,n-m,1-\alpha}\Bigg) = 1-\alpha$$

holds. Without proof.

4.3 Multivariate linear regression with rank(X) < m

Let $Y = X\beta + \varepsilon$, $Y \in \mathbb{R}^n$, where X is a $(n \times m)$ matrix with rank $(X) = r < m$, $\beta = (\beta_1, \ldots, \beta_m)^\top$, $\varepsilon \in \mathbb{R}^n$, $\mathbb{E}\,\varepsilon = 0$, $\mathbb{E}\,(\varepsilon_i\varepsilon_j) = \delta_{ij}\sigma^2$, $i, j = 1, \ldots, n$, $\sigma^2 > 0$.
Even though the rank of the matrix is not full anymore, the OLS estimator $\hat{\beta}$ is still a solution to the normal equation

$$\left(X^\top X\right)\beta = X^\top Y.$$

However, $X^\top X$ is not invertible, since

$$\text{rank }(X^\top X) \leq \min\left\{\text{rank }(X), \text{rank }(X^\top)\right\} = r < m.$$

Consequently, in order to obtain $\hat{\beta}$ from the normal equation, both sides of the equations are multiplied with the *generalized inverse* of $X^\top X$.

4.3.1 Generalized inverse

Definition 4.3.1. Let A be a $(n \times m)$ matrix. A $(m \times n)$ matrix A^- is called *generalized inverse* of A, if

$$AA^-A = A.$$

The matrix A^- is not unique, which is shown in the following lemmas.

Lemma 4.3.2. Let A be a $(n \times m)$ matrix, $m \leq n$ with rank $(A) = r \leq m$. There exist invertible matrices P $(n \times n)$ and Q $(m \times m)$, such that

$$PAQ = \begin{pmatrix} \mathcal{I}_r & 0 \\ 0 & 0 \end{pmatrix}, \text{ where } I_r = \text{diag}(\underbrace{1, \ldots, 1}_{r \text{ times}}). \tag{4.6}$$

Corollary 4.3.3. For an arbitrary $(n \times m)$ matrix A with $n \geq m$, rank (A) $r \leq m$ it holds that

$$A^- = Q \begin{pmatrix} \mathcal{I}_r & A_2 \\ A_1 & A_3 \end{pmatrix} P, \tag{4.7}$$

where P and Q are matrices as in (4.6), $\mathcal{I}_r = \text{diag}\,(\overbrace{1, \ldots, 1}^{r \text{ times}})$, and A_1, A_2 resp. A_3 are arbitrary $((m-r) \times r)$, $(r \times (n-r))$ resp. $((m-r) \times (n-r))$ matrices.
In particular

$$\begin{aligned} A_1 &= 0, \\ A_2 &= 0, \\ A_3 &= \text{diag}\,(\underbrace{1, \ldots, 1}_{s-r \text{ times}}, 0, \ldots, 0), \\ s &\in \{r, \ldots, m\} \end{aligned}$$

can be chosen, which means rank $(A^-) = s \in \{r, \ldots, m\}$ for

$$A^- = Q \begin{pmatrix} \mathcal{I}_s & 0 \\ 0 & 0 \end{pmatrix} P.$$

Proof In the following it is shown that for A^- as in (4.7), it holds that $AA^-A = A$. Lemma 4.3.2 implies that

$$\begin{aligned} A &= P^{-1} \cdot \text{diag}\,(1, \ldots, 1, 0, \ldots, 0) \cdot Q^{-1} \quad \text{and thus} \\ AA^-A &= P^{-1} \begin{pmatrix} \mathcal{I}_r & 0 \\ 0 & 0 \end{pmatrix} Q^{-1}Q \cdot \begin{pmatrix} \mathcal{I}_r & A_2 \\ A_1 & A_3 \end{pmatrix} PP^{-1} \begin{pmatrix} \mathcal{I}_r & 0 \\ 0 & 0 \end{pmatrix} Q^{-1} \\ &= P^{-1} \begin{pmatrix} \mathcal{I}_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \mathcal{I}_r & A_2 \\ A_1 & A_3 \end{pmatrix} \begin{pmatrix} \mathcal{I}_r & 0 \\ 0 & 0 \end{pmatrix} Q^{-1} \\ &= P^{-1} \begin{pmatrix} \mathcal{I}_r & 0 \\ 0 & 0 \end{pmatrix} Q^{-1} = A. \end{aligned}$$

□

Lemma 4.3.4. Let A be an arbitrary $(n \times m)$ matrix with rank $(A) = r \leq m$, $m \leq n$.

1. If $(A^\top A)^-$ is a generalized inverse of an $(m \times m)$ matrix $A^\top A$, then $\left((A^\top A)^-\right)^\top$ is also a generalized inverse of $A^\top A$.

2. It holds that

$$(A^\top A)(A^\top A)^- A^\top = A^\top \quad \text{resp.}$$
$$A(A^\top A)^-(A^\top A) = A.$$

Proof

1. $A^\top A$ is symmetric, i.e.

$$\underbrace{\left(A^\top A(A^\top A)^- A^\top A\right)^\top}_{=A^\top A\left((A^\top A)^-\right)^\top A^\top A} = \left(A^\top A\right)^\top = A^\top A.$$

   Thus $\left((A^\top A)^-\right)^\top$ is a generalized inverse of $A^\top A$.

2. Let $B = (A^\top A)(A^\top A)^- A^\top - A^\top$. In the following it is shown that $B = 0$ by proving that $BB^\top = 0$.

$$\begin{aligned} BB^\top &= \left((A^\top A)(A^\top A)^- A^\top - A^\top\right)\left(A\left((A^\top A)^-\right)^\top A^\top A - A\right) \\ &= A^\top A(A^\top A)^- A^\top A\left((A^\top A)^-\right)^\top A^\top A - \underbrace{A^\top A(A^\top A)^- A^\top A}_{=A^\top A} \\ &\quad - \underbrace{A^\top A\left((A^\top A)^-\right)^\top \cdot A^\top A}_{=A^\top A} + A^\top A \\ &= A^\top A - 2A^\top A + A^\top A = 0. \end{aligned}$$

   The assertion $A(A^\top A)^- A^\top A = A$ can be shown, by transposing the matrices on both sides of the equation $A^\top A(A^\top A)^- A^\top = A^\top$.

□

4.3.2 OLS estimator for β

Theorem 4.3.5. Let X be a $(n \times m)$ design matrix with rank $(X) = r < m$ in the linear regression model $Y = X\beta + \varepsilon$. The generalized solution of the normal equation

$$\left(X^\top X\right)\beta = X^\top Y$$

is given by

$$\beta = \left(X^\top X\right)^- X^\top Y + \left(\mathcal{I}_m - \left(X^\top X\right)^- X^\top X\right) z, \quad z \in \mathbb{R}^m. \tag{4.8}$$

Proof

1. In the following it is shown that β as in (4.8) is a solution of the normal equation.

$$\begin{aligned} X^\top X\beta &= \underbrace{(X^\top X)(X^\top X)^- X^\top}_{=X^\top \text{(Lemma 4.3.4, 2.))}} Y + \Big(X^\top X - \underbrace{X^\top X(X^\top X)^- X^\top X}_{=X^\top X}\Big) z \\ &= X^\top Y \end{aligned}$$

2. Let us show that an arbitrary solution β' of the normal equation can be written as (4.8). Let β be the solution (4.8). Calculating the difference of the equations yields

$$\begin{array}{rl} (X^\top X)\beta' & = X^\top Y \\ - \quad (X^\top X)\beta & = X^\top Y \\ \hline (X^\top X)(\beta' - \beta) & = 0 \end{array}$$

$$\begin{aligned} \beta' &= (\beta' - \beta) + \beta \\ &= \beta' - \beta + (X^\top X)^- X^\top Y + \Big(\mathcal{I}_m - (X^\top X)^- X^\top X\Big) z \\ &= (X^\top X)^- X^\top Y + \Big(\mathcal{I}_m - (X^\top X)^- X^\top X\Big) z + (\beta' - \beta) \\ &\qquad - \underbrace{(X^\top X)^- X^\top X(\beta' - \beta)}_{=0} \\ &= (X^\top X)^- X^\top Y + \Big(\mathcal{I}_m - (X^\top X)^- X^\top X\Big) \Big(\underbrace{z + \beta' - \beta}_{=z_0}\Big) \\ &\Longrightarrow \beta' \text{ can be rewritten as (4.8).} \end{aligned}$$

□

Remark 4.3.6. Theorem 4.3.5 yields the set of all extreme points of the OLS minimization problem

$$e(\beta) = \frac{1}{n} |Y - X\beta|^2 \longrightarrow \min_\beta .$$

Thus, the set of all OLS estimators of β in (4.8) should satisfy additional conditions.

Theorem 4.3.7.

1. All OLS estimators of β can be written as

$$\overline{\beta} = \Big(X^\top X\Big)^- X^\top Y,$$

where $(X^\top X)^-$ is an arbitrary generalized inverse of $X^\top X$.

2. $\overline{\beta}$ is not unbiased, since
$$\mathbb{E}\,\overline{\beta} = \left(X^\top X\right)^- X^\top X\beta.$$
3. It holds that
$$\operatorname{Cov}\overline{\beta} = \sigma^2 \left(X^\top X\right)^- \left(X^\top X\right) \left(\left(X^\top X\right)^-\right)^\top.$$

Proof

1. We show that $e(\beta) \geq e(\overline{\beta}) \quad \forall \beta \in \mathbb{R}^m$.
$$\begin{aligned} n \cdot e(\beta) &= |Y - X\beta|^2 = (Y - X\overline{\beta} + X(\overline{\beta} - \beta))^\top (Y - X\overline{\beta} + X(\overline{\beta} - \beta)) \\ &= (Y - X\overline{\beta})^\top (Y - X\overline{\beta}) + \left(X(\overline{\beta} - \beta)\right)^\top \left(X(\overline{\beta} - \beta)\right) \\ &\quad + 2(\overline{\beta} - \beta)^\top X^\top (Y - X\overline{\beta}) \\ &= n \cdot e(\overline{\beta}) + \underbrace{2 \cdot (\overline{\beta} - \beta)^\top (X^\top Y - (X^\top X\overline{\beta}))}_{=0} + \left|X(\overline{\beta} - \beta)\right|^2 \\ &\geq n \cdot e(\overline{\beta}) + 0 = n \cdot e(\overline{\beta}), \quad \text{since} \end{aligned}$$
$\overline{\beta}$ can be rewritten as in (4.8) with $z = 0$ and is thus a solution to the normal equation.

2. It holds that
$$\begin{aligned} \mathbb{E}\,\overline{\beta} = \mathbb{E}\left((X^\top X)^- X^\top Y\right) &= \left(X^\top X\right)^- X^\top \mathbb{E}\,Y \\ = (X^\top X)^- X^\top X\beta, \\ Y = X\beta + \varepsilon, \quad \mathbb{E}\,\varepsilon = 0. \end{aligned}$$
Note that this implies $\mathbb{E}\,Y = X\beta$. Why is $\overline{\beta}$ not unbiased, i.e. $(X^\top X)^- X^\top X\beta \neq \beta$, $\beta \in \mathbb{R}^m$?
Since rank $(X) = r < m$, rank $(X^\top X) < m$ and rank $((X^\top X)^- X^\top X) < m$. Thus, there exists a $\beta \neq 0$, for which it holds that
$$\left(X^\top X\right)^- X^\top X\beta = 0 \neq \beta,$$
so $\overline{\beta}$ is not unbiased. Furthermore it holds that all solutions of (4.8) are not unbiased estimators. Applying the expectation on (4.8) yields the following in the case of unbiasedness
$$\begin{aligned} &\forall \beta \in \mathbb{R}^m: \quad \beta = (X^\top X)^- X^\top X\beta + \left(\mathcal{I}_m - (X^\top X)^-(X^\top X)\right) z, \quad z \in \mathbb{R}^m. \\ &\implies \left(\mathcal{I}_m - (X^\top X)^-(X^\top X)\right)(z - \beta) = 0 \quad \forall z, \beta \in \mathbb{R}^m \\ &\implies (X^\top X)^-(X^\top X)(\beta - z) = \beta - z, \quad \forall z, \beta \in \mathbb{R}^m. \end{aligned}$$
Since this equation can't hold for all $\beta \in \mathbb{R}^m$, the assumption leads to a contradiction.

3. It holds that

$$\begin{aligned}\operatorname{Cov}\left(\overline{\beta}_i, \overline{\beta}_j\right) &= \operatorname{Cov}\left(\Big(\underbrace{(X^\top X)^- X^\top}_{:=A=(a_{kl})} Y\Big)_i, \left((X^\top X)^- X^\top Y\right)_j\right) \\ &= \operatorname{Cov}\left(\sum_{k=1}^n a_{ik} Y_k, \sum_{l=1}^n a_{jl} Y_l\right) \\ &= \sum_{k,l=1}^n a_{ik} a_{jl} \underbrace{\operatorname{Cov}\left(Y_k, Y_l\right)}_{=\sigma^2 \cdot \delta_{kl}} = \sigma^2 \sum_{k=1}^n a_{ik} a_{jk} = \left(\sigma^2 A A^\top\right)_{i,j} \\ &= \left(\sigma^2 (X^\top X)^- X^\top X \left((X^\top X)^-\right)^\top\right)_{i,j.}\end{aligned}$$

□

4.3.3 Functions that can be estimated without bias

Definition 4.3.8. A linear combination $a^\top \beta$ of $\beta_1, \ldots, \beta_m$, $a \in \mathbb{R}^m$ is called *estimable without bias*, if

$$\exists\, c \in \mathbb{R}^n : \quad \mathbb{E}\left(c^\top Y\right) = a^\top \beta,$$

i.e. if a linear unbiased estimator $c^\top Y$ for $a^\top \beta$ exists.

Theorem 4.3.9. The function $a^\top \beta$, $a \in \mathbb{R}^m$ is estimable without bias if and only if one of the following conditions is satisfied:

1. $\exists\, c \in \mathbb{R}^n : \quad a^\top = c^\top X$.
2. a fulfills the equation

$$a^\top \left(X^\top X\right)^- X^\top X = a^\top. \tag{4.9}$$

Proof

1. „$\Longrightarrow$“: If $a^\top \beta$ is estimable, then there exists a vector $d \in \mathbb{R}^n$ with $\mathbb{E}\,(d^\top Y) = a^\top \beta \quad \forall \beta \in \mathbb{R}^m$. So

$$\begin{aligned}&a^\top \beta = d^\top \mathbb{E}\, Y = d^\top X \beta \Rightarrow \left(a^\top - d^\top X\right)\beta = 0, \quad \forall \beta \in \mathbb{R}^m \\ &\Longrightarrow a^\top = d^\top X,\end{aligned}$$

   Finally, setting $c = d$ proves this implication.

   „$\Longleftarrow$“: We can easily compute $\mathbb{E}\,(c^\top Y) = c^\top \mathbb{E}\, Y = c^\top X \beta \ = a^\top \beta$. Therefore, $a^\top \beta$ is estimable without bias.

2. „$\Longrightarrow$“: If $a^\top\beta$ is estimable without bias, then

$$a^\top (X^\top X)^- X^\top X \overset{1.}{=} c^\top \underbrace{X \cdot (X^\top X)^- X^\top X}_{=X \text{ (Lemma 4.3.4)}} = c^\top X \overset{(1.)}{=} a^\top.$$

Thus, relation (4.9) is satisfied.

„$\Longleftarrow$“: If $a^\top (X^\top X)^- X^\top X = a^\top$, then $c = (a^\top (X^\top X)^- X^\top)^\top$ and the first assertion imply that $a^\top\beta$ is estimable.

□

Remark 4.3.10. In case of a regression with rank $(X) = m$ the equation (4.9) is always satisfied since $(X^\top X)^- = (X^\top X)^{-1}$ and thus $a^\top\beta$ is estimable for all $a \in \mathbb{R}^m$.

Theorem 4.3.11 (Examples of estimable functions)**.** If rank $(X) = r < m$, then the following linear combinations of β are estimable:

1. The coordinates $\sum_{j=1}^{m} x_{ij}\beta_j$, $i = 1, \ldots, n$ of the vector of expectations $\mathbb{E}\, Y = X\beta$.
2. Arbitrary linear combinations of estimable functions.

Proof

1. Set $\tilde{x}_i = (x_{i1}, \ldots, x_{im})$, $i = 1, \ldots, n$. Then

$$\sum_{j=1}^{m} x_{ij\beta_j} = \tilde{x}_i^\top \beta \quad \forall i = 1, \ldots, n,$$

$$X\beta = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)^\top \beta.$$

$\tilde{x}_i\beta$ is estimable, if $\tilde{x}_i$ satisfies (4.9), which can be expressed in matrices for all $i = 1, \ldots, n$ as follows:

$$X \left(X^\top X\right)^- X^\top X = X.$$

By Lemma 4.3.4 this is valid.

2. For $a_1, \ldots, a_k \in \mathbb{R}^m$ let $a_1^\top\beta, \ldots, a_k^\top\beta$ be estimable functions. For all $\lambda = (\lambda_1, \ldots, \lambda_k)^\top \in \mathbb{R}^k$ show that $\sum_{i=1}^{k} \lambda_i \cdot a_i^\top\beta = \lambda^\top A\beta$ is estimable, where $A = (a_1, \ldots, a_k)^\top$. It needs to be shown that $b = (\lambda^\top A)^\top$ satisfies (4.9), i.e.

$$\lambda^\top A \left(X^\top X\right)^- X^\top X = \lambda^\top A.$$

This equation is satisfied, since $a_i^\top (X^\top X)^- X^\top X = a_i^\top$, $i = 1, \ldots, k$. By Theorem 4.3.9, 2.) $\lambda^\top A\beta$ is estimable.

□

Theorem 4.3.12 (*Gauß-Markov*)**.** Let $a^\top\beta$ be an estimable function, $a \in \mathbb{R}^m$ in the linear regression model $Y = X\beta + \varepsilon$ with rank $(X) \leq m$.

1. The best linear unbiased estimator of $a^\top\beta$ is given by $a^\top\overline{\beta}$, where
$$\overline{\beta} = \left(X^\top X\right)^- X^\top Y$$
is an OLS estimator for β.
2. $\operatorname{Var}(a^\top\overline{\beta}) = \sigma^2 a^\top (X^\top X)^- a$.

Proof The linearity of $a^\top\overline{\beta} = a^\top(X^\top X)^- X^\top Y$ as a function of Y is clear. For the unbiasedness it holds that

$$\begin{aligned}\mathbb{E}(a^\top\overline{\beta}) &= a^\top \mathbb{E}\,\overline{\beta} = a^\top (X^\top X)^- X^\top X\beta \\ &= c^\top \underbrace{X(X^\top X)^- X^\top X}_{=X \text{ (Lemma 4.3.4)}}\beta = \underbrace{c^\top X}_{=a^\top}\beta = a^\top\beta \quad \forall \beta \in \mathbb{R}^m.\end{aligned}$$

First, calculate $\operatorname{Var}(a^\top\overline{\beta})$ (i.e. prove the second assertion) and show that it is minimal:

$$\begin{aligned}\operatorname{Var}(a^\top\overline{\beta}) &= \operatorname{Var}\left(\sum_{i=1}^m a_i\overline{\beta}_i\right) = \sum_{i,j=1}^m a_i a_j \cdot \operatorname{Cov}\left(\overline{\beta}_i, \overline{\beta}_j\right) \\ &= a^\top \operatorname{Cov}\left(\overline{\beta}\right) a \overset{\text{(Satz 4.3.7)}}{=} a^\top \sigma^2 \left((X^\top X)^- X^\top X (X^\top X)^-\right)^\top a \\ &= \sigma^2 \cdot a^\top \underbrace{\left((X^\top X)^-\right)^\top}_{=(X^\top X)^-} X^\top X \underbrace{\left((X^\top X)^-\right)^\top}_{(X^\top X)^-} a \\ &\overset{\text{Lemma 4.3.4, 1.)}}{=} \sigma^2 a^\top (X^\top X)^- X^\top X (X^\top X)^- a \\ &\overset{\text{Theorem 4.3.9, 1.)}}{=} \sigma^2 \cdot c^\top \underbrace{X \cdot (X^\top X) X^\top X}_{=X} (X^\top X)^- X^\top c \\ &= \sigma^2 \underbrace{c^\top X}_{=a^\top} (X^\top X)^- \underbrace{X^\top c}_{=a} = \sigma^2 a^\top (X^\top X)^- a.\end{aligned}$$

Now it is shown that for an arbitrary linear unbiased estimator $b^\top Y$ of $a^\top\beta$ it holds that $\operatorname{Var}(b^\top Y) \geq \operatorname{Var}(a^\top\overline{\beta})$. Since $b^\top Y$ is unbiased, it holds that $\mathbb{E}(b^\top Y) = a^\top\beta$. Using Theorem 4.3.9 it holds that $a^\top = b^\top X$. Now, consider

$$\begin{aligned}0 &\leq \operatorname{Var}\left(b^\top Y - a^\top\overline{\beta}\right) \\ &= \operatorname{Var}\left(b^\top Y\right) - 2\operatorname{Cov}\left(b^\top Y, a^\top\overline{\beta}\right) + \operatorname{Var}\left(a^\top\overline{\beta}\right) \\ &= \operatorname{Var}(b^\top Y) - 2\sigma^2 a^\top (X^\top X)^- a + \sigma^2 a^\top (X^\top X)^- a \\ &= \operatorname{Var}(b^\top Y) - \operatorname{Var}\left(a^\top\overline{\beta}\right)\end{aligned}$$

with

$$\begin{aligned}\text{Cov}\left(b^\top Y, a^\top\overline{\beta}\right) &= \text{Cov}\left(b^\top Y, a^\top(X^\top X)^- X^\top Y\right) = \sigma^2 a^\top (X^\top X)^- \underbrace{X^\top b}_{=a} \\ &= \sigma^2 a^\top (X^\top X)^- a.\end{aligned}$$

Thus, $\text{Var}\left(b^\top Y\right) \geq \text{Var}\left(a^\top\overline{\beta}\right)$ and $a^\top\overline{\beta}$ is a best linear unbiased estimator for $a^\top\beta$. □

Remark 4.3.13.

1. If rank $(X) = m$, then $a^\top\hat{\beta}$ is the best linear unbiased estimator for $a^\top\beta$, $a \in \mathbb{R}^m$.
2. The estimator $a^\top\overline{\beta} = a^\top(X^\top X)^- X^\top Y$ does not depend on the choice of the generalized inverse as is shown in the following theorem.

Theorem 4.3.14. The best linear unbiased estimator $a^\top\overline{\beta}$ for $a^\top\beta$ is uniquely determined.

Proof

$$a^\top\overline{\beta} = a^\top(X^\top X)^- X^\top Y \overset{\text{Theorem 4.3.9, 1.)}}{=} c^\top X(X^\top X)^- X^\top Y.$$

In order to show that $X(X^\top X)^- X^\top$ does not depend on $(X^\top X)^-$, we prove that for arbitrary generalized inverses A_1 and A_2 of $(X^\top X)$ it holds that $XA_1X^\top = XA_2X^\top$. By Lemma 4.3.4, 2.) it holds that

$$XA_1X^\top X = X = XA_2X^\top X.$$

Multiplying all parts of the equation with $A_1X^\top$ on the right yields

$$XA_1\underbrace{X^\top XA_1X^\top}_{=X^\top} = XA_1X^\top = XA_2\underbrace{X^\top XA_1X^\top}_{=X^\top}$$

Thus, $XA_1X^\top = XA_2X^\top$. □

4.3.4 Normally distributed error terms

Let $Y = X\beta + \varepsilon$ be a linear regression model with rank $(X) = r < m$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathcal{I})$. As in Section 4.2.3 the maximum likelihood estimator $\tilde{\beta}$ and $\tilde{\sigma}^2$ can be derived for β and σ^2. Exactly as in Theorem 4.2.7 it can be shown that

$$\begin{aligned}\tilde{\beta} &= \overline{\beta} = (X^\top X)^- X^\top Y \quad \text{and} \\ \tilde{\sigma}^2 &= \frac{1}{n}\left|Y - X\overline{\beta}\right|^2.\end{aligned}$$

Now the distributional properties of $\overline{\beta}$ and $\tilde{\sigma}^2$ are discussed. First the unbiasedness of $\tilde{\sigma}^2$ is discussed. $\tilde{\sigma}^2$ is not unbiased but, the corrected estimator

$$\overline{\sigma}^2 = \frac{1}{n-r}|Y - X\beta|^2 = \frac{n}{n-r}\tilde{\sigma}^2$$

is unbiased.

Theorem 4.3.15. The estimator $\overline{\sigma}^2$ is unbiased for σ^2.

The proof of Theorem 4.3.15 is similar to the proof of Theorem 4.2.6 in which $\hat{\beta} = (X^\top X)^{-1}X^\top Y$ and $\hat{\sigma}^2 = \frac{1}{n-m}|Y - X\beta|^2$ are considered for the case rank $(X) = m$. Thus, Theorem 4.2.6 is a special case of Theorem 4.3.15. Define $D := \mathcal{I} - X(X^\top X)^- X^\top$.

Lemma 4.3.16. For D it holds that

1. $D^\top = D$ (symmetry),
2. $D^2 = D$ (idempotence),
3. $DX = 0$,
4. $\operatorname{trace}(D) = n - r$.

Proof

1. It holds that

$$\begin{aligned} D^\top &= \left(\mathcal{I} - X(X^\top X)^- X^\top\right)^\top = \mathcal{I} - X\left((X^\top X)^-\right)^\top X^\top \\ &= \mathcal{I} - X(X^\top X)^- X^\top = D, \end{aligned}$$

   since $\left((X^\top X)^-\right)^\top$ is also a generalized inverse of $X^\top X$ (cf. Lemma 4.3.4, 1.)).

2. It holds that

$$\begin{aligned} D^2 &= \left(\mathcal{I} - X(X^\top X)^- X^\top\right)^2 \\ &= \mathcal{I} - 2X(X^\top X)^- X^\top + \underbrace{X(X^\top X)^- X^\top X}_{=X \text{ (Lemma 4.3.4, 2.)}} (X^\top X)^- X^\top \\ &= \mathcal{I} - X(X^\top X)^- X^\top = D. \end{aligned}$$

3. $DX = X - \underbrace{X(X^\top X)^- X^\top X}_{=X \text{ (Lemma 4.3.4, 2.)}} = X - X = 0.$

4. It holds that

$$\begin{aligned}\operatorname{trace}(D) &= \operatorname{trace}(I) - \operatorname{trace}\left(X(X^\top X)^- X^\top\right)\\ &= n - \operatorname{trace}\left(X(X^\top X)^- X^\top\right).\end{aligned}$$

The symmetry and idempotence of the matrix A imply $\operatorname{trace}(A) = \operatorname{rank}(A)$ as is known from linear algebra. Since $X(X^\top X)^- X^\top$ is symmetric and idempotent, it is sufficient to show $\operatorname{rank}(X(X^\top X)^- X^\top) = r$. By Lemma 4.3.4 2.) it holds that

$$\begin{aligned}\operatorname{rank}(X) = r = \operatorname{rank}&\left(X(X^\top X)^- X^\top X\right)\\ &\le \min\left\{\operatorname{rank}\left(X(X^\top X)^- X^\top\right), \underbrace{\operatorname{rank}(X)}_{=r}\right\}\\ &\le \operatorname{rank}\left(X(X^\top X)^- X^\top\right) \le \operatorname{rank}(X) = r\\ &\Longrightarrow \operatorname{rank}\left(X(X^\top X)^- X^\top\right) = r\\ &\Longrightarrow \operatorname{trace}\left(X(X^\top X)^- X^\top\right) = r.\end{aligned}$$

□

Proof of Theorem 4.3.15 By using Lemma 4.3.16 it can be shown that

$$\begin{aligned}\overline{\sigma}^2 &= \frac{1}{n-r}\left|Y - X\overline{\beta}\right|^2 = \frac{1}{n-r}\left|Y - X(X^\top X)^- X^\top Y\right|^2 = \frac{1}{n-r}|DY|^2\\ &= \frac{1}{n-r}\left|\underbrace{DX}_{=0}\beta + D\varepsilon\right|^2 = \frac{1}{n-r}|D\varepsilon|^2 = \frac{1}{n-r}\varepsilon^\top \underbrace{D^\top D}_{=D^2=D}\varepsilon = \frac{1}{n-r}\varepsilon^\top D\varepsilon.\end{aligned}$$

Thus

$$\begin{aligned}\mathbb{E}\,\overline{\sigma}^2 &= \frac{1}{n-r}\mathbb{E}\left(\varepsilon^\top D\varepsilon\right) = \frac{1}{n-r}\mathbb{E}\operatorname{trace}\left(\varepsilon^\top D\varepsilon\right) = \frac{1}{n-r}\operatorname{trace}\Big(D\cdot\mathbb{E}\left(\underbrace{\varepsilon\varepsilon^\top}_{\sigma^2\mathcal{I}}\right)\Big)\\ &= \frac{\sigma^2}{n-r}\cdot\operatorname{trace}(D) = \sigma^2\end{aligned}$$

by Lemma 4.3.16, 4.), because $\mathbb{E}\,\varepsilon\varepsilon^\top = \sigma^2\mathcal{I}$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathcal{I})$ □

Theorem 4.3.17. The following distributional properties hold

1. $\overline{\beta} \sim N\left((X^\top X)^- X^\top X\beta,\ \sigma^2(X^\top X)^-(X^\top X)\left((X^\top X)^-\right)^\top\right)$,
2. $\frac{(n-r)\overline{\sigma}^2}{\sigma^2} \sim \chi^2_{n-r}$,
3. $\overline{\beta}$ and $\overline{\sigma}^2$ are independent.

Proof

1. It holds that

$$\begin{aligned}\overline{\beta} &= (X^\top X)^- X^\top Y = (X^\top X)^- X^\top (X\beta + \varepsilon) \\ &= \underbrace{(X^\top X)^- X^\top X\beta}_{=\mu} + \underbrace{(X^\top X)^- X^\top}_{=A}\varepsilon.\end{aligned}$$

   Consequently,

$$\begin{aligned}\overline{\beta} &\sim N\left(\mu,\, \sigma^2 AA^\top\right) \\ &= N\left((X^\top X)^- X^\top X\beta,\, \sigma^2 (X^\top X)^- X^\top X((X^\top X)^-)^\top\right)\end{aligned}$$

   with $AA^\top = (X^\top X)^- X^\top X((X^\top X)^-)^\top$.

2. It holds that $\overline{\sigma}^2 = \frac{1}{n-r}\varepsilon^\top D\varepsilon$ by the proof of Theorem 4.3.15. Thus,

$$\frac{(n-r)\overline{\sigma}^2}{\sigma^2} = \underbrace{\left(\frac{\varepsilon}{\sigma}\right)^\top}_{\sim\mathcal{N}(0,\mathcal{I})} D\left(\frac{\varepsilon}{\sigma}\right) \overset{\text{(Satz 4.1.25)}}{\sim} \chi^2_{n-r}.$$

3. Consider $A\varepsilon$ and $\varepsilon^\top D\varepsilon$. It is sufficient to show that they are independent in order to show the independence of $\overline{\beta}$ and $\overline{\sigma}^2$, since $\overline{\beta} = \mu + A\varepsilon$, $\overline{\sigma}^2 = \frac{1}{n-r}\varepsilon^\top D\varepsilon$. It holds that $A \cdot \sigma^2\mathcal{I} \cdot D = 0$. By Theorem 4.1.26 $A\varepsilon$ and $\varepsilon^\top D\varepsilon$ are independent.

□

4.3.5 Hypothesis testing

Consider the hypothesis test $H_0 : H\beta = d$ vs. $H_1 : H\beta \neq d$, where H is an $(s \times m)$ matrix $(s \leq m)$ with rank $(H) = s$ and $d \in \mathbb{R}^s$.
In Theorem 4.2.14 for the case rank $(X) = r = m$ the following test statistic was considered

$$T = \frac{(H\hat{\beta} - d)^\top (H(X^\top X)^{-1}H^\top)^{-1}(H\hat{\beta} - d)}{s\hat{\sigma}^2} \overset{(H_0)}{\sim} F_{s,n-m}.$$

In general, we may consider

$$T = \frac{(H\overline{\beta} - d)^\top (H(X^\top X)^- H^\top)^{-1}(H\overline{\beta} - d)}{s\overline{\sigma}^2}. \tag{4.10}$$

We will show that $T \overset{(H_0)}{\sim} F_{s,n-r}$. Then, a test with confidence level $\alpha \in (0,1)$ can be constructed by rejecting H_0, if $T > F_{s,n-r,1-\alpha}$.

Definition 4.3.18. The hypothesis $H_0 : H\beta = d$ is called *testable*, if all coordinates of the vector $H\beta$ are estimable functions.

Theorem 4.3.9 provides conditions for H, under which $H_0 : H\beta = d$ is testable. They are formulated in the following Lemma

Lemma 4.3.19. The hypothesis $H_0 : H\beta = d$ is testable if and only if

1. There exists an $(s \times n)$ matrix C such that $H = CX$, or
2. $H(X^\top X)^- X^\top X = H$.

First, show that the test statistic T in (4.10) is well defined, i.e. the $(s \times s)$ matrix $H(X^\top X)^- H^\top$ is positive definite and thus invertible. Corollary 4.3.3 implies

$$X^\top X = P^{-1} \begin{pmatrix} \mathcal{I}_r & 0 \\ 0 & 0 \end{pmatrix} P^{-1} \tag{4.11}$$

for an $(m \times m)$ matrix P, which is symmetric and invertible. Thus,

$$(X^\top X)^- = P \cdot \begin{pmatrix} \mathcal{I}_r & 0 \\ 0 & \mathcal{I}_{m-r} \end{pmatrix} P = P \cdot P,$$

holds, i.e. there exists a unique generalized inverse $X^\top X$ with this representation. This implies that the $(s \times s)$ matrix $HPPH^\top = (PH^\top)^\top \cdot PH^\top$ is positive definite because rank $(PH^\top) = s$. Let now $(X^\top X)^-$ be an arbitrary generalized inverse of $X^\top X$. Then, Lemma 4.3.19 implies

$$H(X^\top X)^- H^\top = CX(X^\top X)^- X^\top C^\top = CXPPX^\top C^\top = HPPH^\top,$$

because $X(X^\top X)^- X^\top$ is invariant with respect to the choice $(X^\top X)^-$ by the proof of Theorem 4.3.14. Thus, $H\left(X^\top X\right)^- H^\top$ is positive definite for an arbitrary generalized inverse $\left(X^\top X\right)^-$ and the test statistic T is well defined.

Theorem 4.3.20. If $H_0 : H\beta = d$ is testable, then $T \overset{(H_0)}{\sim} F_{s,n-r}$.

Proof This proof is similar to the proof of Theorem 4.2.14. First, we compute

$$\begin{aligned} H\overline{\beta} - d &= H(X^\top X)^- X^\top (X\beta + \varepsilon) - d \\ &= \underbrace{H(X^\top X)^- X^\top X\beta - d}_{=\mu} + \underbrace{H(X^\top X)^- X^\top}_{=B} \varepsilon. \end{aligned}$$

Show that $\mu \overset{(H_0)}{=} 0$.

$$\mu \overset{\text{(Lemma 4.3.19)}}{=} C \cdot \underbrace{X(X^\top X)^- X^\top X}_{=X \text{ (Lemma 4.3.4, 2.)}} \cdot \beta - d = CX\beta - d = H\beta - d \overset{(H_0)}{=} 0.$$

Using Theorem 4.3.17 yields $(H\overline{\beta} - d)^\top \left(H(X^\top X)^- H^\top\right)^{-1} \left(H\overline{\beta} - d\right)$ and $s \cdot \overline{\sigma}^2$ are independent and $\frac{(n-r)\overline{\sigma}^2}{\sigma^2} \sim \chi^2_{n-r}$. It remains to show

$$\left(\underbrace{H\overline{\beta} - d}_{=\varepsilon^\top B^\top}\right)^\top \left(H(X^\top X)^- H^\top\right)^{-1} \left(\underbrace{H\overline{\beta} - d}_{=B\varepsilon}\right) \overset{(H_0)}{\sim} \chi^2_s.$$

It holds that

$$\begin{aligned} &\varepsilon^\top B^\top \left(H(X^\top X)^- H^\top\right)^{-1} B\varepsilon \\ &= \varepsilon^\top \underbrace{X \left((X^\top X)^-\right)^\top H^\top \left(H(X^\top X)^- H^\top\right)^{-1} H(X^\top X)^- X^\top}_{A} \varepsilon \end{aligned}$$

It can be shown that A is symmetric, idempotent and rank $(A) = s$. The idempotence can be shown as follows

$$\begin{aligned} A^2 &= X\left((X^\top X)^-\right)^\top H^\top \left(H(X^\top X)^- H^\top\right)^{-1} \underbrace{H(X^\top X)^- X^\top X}_{H \text{ (Lemma 4.3.19, 2.)}} \left((X^\top X)^-\right)^T H^\top \cdot \\ &\quad \cdot \left(H(X^\top X)^- H^\top\right)^{-1} H(X^\top X)^- X^\top \\ &= X \left((X^\top X)^-\right)^\top H^\top \left(H(X^\top X)^- H^\top\right)^{-1} H(X^\top X)^- X^\top = A, \end{aligned}$$

since $\left((X^\top X)^-\right)^\top$ is also a generalized inverse of $X^\top X$ (by Lemma 4.3.4). Thus, $H(X^\top X)^- H^\top = CX(X^\top X)^- X^\top C^\top$ does not depend on the choice of $(X^\top X)^-$, cf. proof of Theorem 4.3.14. Using Theorem 4.1.25 yields $\frac{\varepsilon^\top}{\sigma} A \frac{\varepsilon}{\sigma} \sim \chi^2_s$, because $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathcal{I})$ and thus $T \overset{H_0}{\sim} F_{s,n-r}$. □

4.3.6 Confidence regions

Similar to Section 4.2.5, confidence regions for different functions of the parameter vector β can be found. Theorem 4.3.20 directly yields the following confidence region with confidence level $1 - \alpha \in (0, 1)$:

Corollary 4.3.21. Let $Y = X\beta + \varepsilon$ be a multivariate regression model with rank $(X) = r < m$, H an $(s \times m)$ matrix with rank $(H) = s$, $s \in \{1, \ldots, m\}$ and $H_0 : H\beta = d$ testable $\forall d \in \mathbb{R}^s$. Then,

$$\left\{ d \in \mathbb{R}^s : \frac{\left(H\overline{\beta} - d\right)^\top \left(H(X^\top X)^- H^\top\right)^{-1} \left(H\overline{\beta} - d\right)}{s \cdot \overline{\sigma}^2} \leq F_{s,\, n-r,\, 1-\alpha} \right\}$$

is a confidence region for $H\beta$ with confidence level $1 - \alpha$.

Corollary 4.3.22. Let $h^\top \beta$ be an estimable linear function of β, $h \in \mathbb{R}^m$. Then,

$$\left(h^\top \overline{\beta} - t_{n-r,\,1-\alpha/2} \cdot \overline{\sigma}\sqrt{h^\top (X^\top X)^- h},\; h^\top \overline{\beta} + t_{n-r,\,1-\alpha/2} \cdot \overline{\sigma}\sqrt{h^\top (X^\top X)^- h}\right)$$

is a confidence interval for $h^\top \beta$ with confidence level $1 - \alpha$.

Proof Set $s = 1$ and $H = h^\top$. Theorem 4.3.20 implies

$$\begin{aligned} T &= \frac{\left(h^\top \overline{\beta} - d\right)^\top \left(h^\top (X^\top X)^- h\right)^{-1} \left(h^\top \overline{\beta} - d\right)}{\overline{\sigma}^2} = \frac{\left(h^\top \overline{\beta} - d\right)\left(h^\top \overline{\beta} - d\right)}{\overline{\sigma}^2 \left(h^\top (X^\top X)^- h\right)} \\ &= \frac{\left(h^\top \overline{\beta} - d\right)^2}{\overline{\sigma}^2 \left(h^\top (X^\top X)^- h\right)} \sim F_{1,\,n-r} \end{aligned}$$

under the condition $h^\top \beta = d$, since $h^\top \left(X^\top X\right)^- h$ is one-dimensional. Thus, it holds that

$$\sqrt{T} = \frac{h^\top \beta - h^\top \overline{\beta}}{\overline{\sigma}\sqrt{h^\top (X^\top X)^- h}} \sim t_{n-r}.$$

Therefore,

$$P\left(-t_{n-r,1-\alpha/2} \le \sqrt{T} \le t_{n-r,1-\alpha/2}\right) = 1 - \alpha.$$

This implies the confidence interval above. □

An even stronger version of Corollary 4.3.22 can be proven which holds for all h of a linear subspace:

Theorem 4.3.23 (Confidence band of Scheffé)**.** Let $H = (h_1, \ldots, h_s)^\top$ where $h_1, \ldots, h_s \in \mathbb{R}^m$, $1 \le s \le m$ and $H_0 : H\beta = d$ testable $\forall d \in \mathbb{R}^s$. Let rank $(H) = s$ and $\mathcal{L} =< h_1, \ldots, h_s >$ the linear subspace with span $\{h_1, \ldots, h_s\}$. Then

$$P\left(\max_{h \in \mathcal{L}} \left\{ \frac{\left(h^\top \beta - h^\top \overline{\beta}\right)^2}{\overline{\sigma}^2 h^\top (X^\top X)^- h} \right\} \le s F_{s,\,n-r,\,1-\alpha}\right) = 1 - \alpha$$

Thus,

$$\left[h^\top \overline{\beta} \pm \sqrt{s F_{s,\,n-r,\,1-\alpha}} \cdot \overline{\sigma}\sqrt{h^\top (X^\top X)^- h}\right]$$

is a (uniform with respect to $h \in \mathcal{L}$) confidence band for $h^\top \beta$.

Proof Set

$$T_1 := \left(H\overline{\beta} - H\beta\right)^\top \left(H(X^\top X)^- H^\top\right)^{-1} \left(H\overline{\beta} - H\beta\right).$$

Then, Theorem 4.3.20 implies

$$P\Big(T_1 \leq s \cdot \overline{\sigma}^2 F_{s,\,n-r,\,1-\alpha}\Big) = 1 - \alpha$$

for all $\alpha \in (0,1)$. If it can be shown that

$$T_1 = \max_{x \in \mathbb{R}^s,\, x \neq 0} \left\{ \frac{\left(x^\top \left(H\overline{\beta} - H\beta\right)\right)^2}{x^\top \left(H(X^\top X)^- H^\top\right) x} \right\}, \tag{4.12}$$

then the proof is concluded, since

$$\begin{aligned}
1 - \alpha &= P\Big(T_1 \leq \underbrace{s\overline{\sigma}^2 F_{s,\,n-r,\,1-\alpha}}_{t}\Big) \\
&= P\left(\max_{x \in \mathbb{R}^s,\, x \neq 0} \left\{ \frac{\left(x^\top \left(H\overline{\beta} - H\beta\right)\right)^2}{x^\top \left(H(X^\top X)^- H^\top\right) x} \right\} \leq t \right) \\
&= P\left(\max_{x \in \mathbb{R}^s,\, x \neq 0} \left\{ \frac{\left((H^\top x)^\top \overline{\beta} - (H^\top x)^\top \beta\right)^2}{(H^\top x)^\top (X^\top X)^- (H^\top x)} \right\} \leq t \right) \\
&\overset{H^\top x = h \in \mathcal{L}}{=} P\left(\max_{h \in \mathcal{L}} \left\{ \frac{\left(h^\top \overline{\beta} - h^\top \beta\right)^2}{h^\top (X^\top X)^- h} \right\} \leq s\overline{\sigma}^2 F_{s,\,n-r,\,1-\alpha} \right).
\end{aligned}$$

In order to prove the validity of (4.12) it is sufficient to show that T_1 is an upper bound of

$$\frac{\left(x^\top (H\overline{\beta} - H\beta)\right)^2}{x^\top \left(H(X^\top X)^- H^\top\right) x},$$

which is also a maximum. Since $H(X^\top X)^- H^\top$ is positive definite and invertible, there exists an invertible $(s \times s)$ matrix B with the property $BB^\top = H(X^\top X)^- H^\top$. Then,

$$\begin{aligned}
\left(x^\top (H\overline{\beta} - H\beta)\right)^2 &= \Big(\underbrace{x^\top B}_{(B^\top x)^\top} \cdot B^{-1}(H\overline{\beta} - H\beta)\Big)^2 \\
&\overset{C.S.}{\leq} |B^\top x|^2 \cdot |B^{-1}(H\overline{\beta} - H\beta)|^2 \\
&= x^\top BB^\top x \left(H\overline{\beta} - H\beta\right)^\top \cdot \underbrace{(B^{-1})^\top B^{-1}}_{= (B^\top)^{-1} B^{-1} = (BB^\top)^{-1}} (H\overline{\beta} - H\beta) \\
&= x^\top H(X^\top X)^- H^\top x \cdot \left(H\overline{\beta} - H\beta\right)^\top \left(H(X^\top X)^- H^\top\right)^{-1} (H\overline{\beta} - H\beta).
\end{aligned}$$

Thus, it holds that

$$\frac{\left(x^\top(H\overline{\beta}-H\beta)\right)^2}{x^\top\left(H(X^\top X)^- H^\top\right)x} \leq \left(H\overline{\beta}-H\beta\right)^\top\left(H(X^\top X)^- H^\top\right)^{-1}\left(H\overline{\beta}-H\beta\right)$$
$$= T_1.$$

For $x = \left(H(X^\top X)^- H^\top\right)^{-1}\left(H\overline{\beta}-H\beta\right)$ it can be shown that it actually is a maximum. □

4.3.7 Introduction to variance analysis

In this section we discuss an example for the application of linear models with a design matrix that doesn't have full rank. It is the assertion of the *variability of the expected values* in the random sample $Y = (Y_1, \ldots, Y_n)^\top$ in short *ANOVA*(analysis of variance).
First, consider the *single factor variance analysis*, in which it is assumed, that the random sample $(Y_1, \ldots, Y_n)$ can be partitioned in k homogeneous subclasses $(Y_{ij},\ j = 1, \ldots, n_i)$, $i = 1, \ldots, k$ with

1. $\mathbb{E}(Y_{ij}) = \mu_i = \mu + \alpha_i, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, k.$

2. $n_i > 1, \quad i = 1, \ldots, k, \quad \sum_{i=1}^{k} n_i = n, \quad \sum_{i=1}^{k} n_i\alpha_i = 0.$

Here μ is a factor, which is equal in all classes and $\alpha_i \in \mathbb{R}$ are the *class specific differences* between the expected values $\mu_1, \ldots, \mu_k$. The number $i = 1, \ldots, k$ of the classes are called *levels of the influencing factor* (e.g. the doses of a drug in a clinical trial) and α_i, $i = 1, \ldots, k$ can be interpreted as the *effect* of the i-th level. The constraint $\sum_{i=1}^{k} n_i\alpha_i = 0$ causes that the conversion $(\mu_1, \ldots, \mu_k) \longleftrightarrow (\mu, \alpha_1, \ldots, \alpha_k)$ is unique and that $\mu = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mathbb{E}\,Y_{ij}$.
Furter, it is assumed that μ_i can be measured with uncorrelated measurement errors ε_{ij}, i.e.

$$Y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, \ldots, k,\ j = 1, \ldots, n_i \tag{4.13}$$
$$\mathbb{E}\,\varepsilon_{ij} = 0, \quad \operatorname{Var}\varepsilon_{ij} = \sigma^2, \quad \varepsilon_{ij} \text{ uncorrelated},\ i = 1, \ldots, k,\ j = 1, \ldots, n_i. \tag{4.14}$$

Here, we want to test the *classical ANOVA hypothesis* that *no* variability in the expected values μ_i can be found, i.e.

$$H_0: \quad \mu_1 = \mu_2 = \ldots = \mu_k,$$

which means, that

$$H_0: \quad \alpha_1 = \alpha_2 = \ldots = \alpha_k.$$

The constraint

$$\sum_{i=1}^{k} n_i \alpha_i = 0.$$

implies $\alpha_i = 0$.
The problem (4.13) can be rewritten in terms of multivariate linear regression as follows:

$$Y = X\beta + \varepsilon, \text{ where } Y = (Y_{11}, \ldots, Y_{1n_1}, Y_{21}, \ldots, Y_{2n_2}, \ldots, Y_{k1}, \ldots, Y_{kn_k})^\top,$$
$$\beta = (\mu, \alpha_1, \ldots, \alpha_k)^\top,$$
$$\varepsilon = (\varepsilon_{11}, \ldots, \varepsilon_{1n_1}, \ldots, \varepsilon_{k1}, \ldots, \varepsilon_{kn_k})^\top,$$

$$X = \begin{pmatrix} 1 & 1 & 0 & \ldots & \ldots & 0 \\ 1 & 1 & 0 & \ldots & \ldots & 0 \\ \vdots & & & & & \\ 1 & 1 & 0 & \ldots & \ldots & 0 \\ 1 & 0 & 1 & 0 & \ldots & 0 \\ \vdots & & & & & \\ 1 & 0 & 1 & 0 & \ldots & 0 \\ \vdots & & & & & \\ 1 & 0 & \ldots & \ldots & 0 & 1 \\ \vdots & & & & & \\ 1 & 0 & \ldots & \ldots & 0 & 1 \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \\ a \\ a \end{matrix}}\right\} n_1 \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \\ a \end{matrix}}\right\} n_2 \\ \vdots \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \\ a \end{matrix}}\right\} n_k \end{matrix}$$

The $(n \times (k+1))$ matrix X has rank $k < m = k+1$; thus the theory of section 4.3 can be applied to this model.

Exercise 4.3.24. Show that the ANOVA hypothesis

$$H_0: \quad \alpha_i = 0, \quad \forall i = 1, \ldots, k$$

is not testable!

In order to consider an equivalent but testable hypothesis

$$H_0: \quad \alpha_1 - \alpha_2 = 0, \ldots, \alpha_1 - \alpha_k = 0 \quad \text{resp.} \quad H_0: \quad H\beta = 0,$$

we construct a $(k-1) \times (k+1)$ matrix

$$H = \begin{pmatrix} 0 & 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & 0 & -1 & \dots & 0 \\ \vdots & & & & & \\ 0 & 1 & 0 & \dots & -1 & 0 \\ 0 & 1 & 0 & \dots & 0 & -1 \end{pmatrix}$$

which we can use as part of our hypothesis test. (Show that!)

In the *two factor variance analysis* the random sample $(Y_1, \dots, Y_n)$ is divided in $k_1 \cdot k_2$ homogeneous groups depending on two factors

$$Y_{i_1 i_2 j}, \quad j = 1, \dots, n_{i_1 i_2}$$

for $i_1 = 1, \dots, k_1$, $i_2 = 1, \dots, k_2$, such that

$$\sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} n_{i_1 i_2} = n.$$

Here it is assumed that

$$\mathbb{E}\, Y_{i_1 i_2 j} = \mu_{i_1 i_2} = \mu + \alpha_{i_1} + \beta_{i_2} + \gamma_{i_1 i_2}, \quad i_1 = 1, \dots, k_1,\, i_2 = 1, \dots, k_2,$$

thus the following linear model is constructed:

$$\begin{aligned} Y_{i_1 i_2 j} = \mu_{i_1 i_2} + \varepsilon_{i_1 i_2 j} = \mu + \alpha_{i_1} + \beta_{i_2} + \gamma_{i_1 i_2} + \varepsilon_{i_1 i_2 j}, \\ j = 1, \dots, n_{i_1 i_2},\, i_1 = 1, \dots, k_1,\, i_2 = 1, \dots, k_2. \end{aligned}$$

Exercise 4.3.25. Write down the design matrix X for this case explicitly and show that it also doesn't have full rank.

Chapter 5

Generalized linear models

Another class of regression models usually allows for an arbitrary functional connection g between the mean of the goal variable $\mathrm{E}Y_i$ and the linear part $X\beta$, which is a linear combinations of the entries of the design matrix $X = (x_{ij})$ and the parameter vector $\beta = (\beta_1, \ldots, \beta_m)^\top$. On the other hand it allows distributions of Y_i, which are not necessarily based on the normal distribution (and functions of those). Thus it is possible to consider data Y_i that has a finite number of characteristics (e.g. "yes" and "no" in economic surveys). The class of all possible distributions is bounded by the *Exponential family*.
Let $Y_1, \ldots, Y_n$ be a random sample of the goal variable of the model and let

$$X = (x_{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,m}}$$

the design matrix of the output variables, which are not random.

Definition 5.0.1. The *generalized linear model* is given by

$$\left(g(\mathbb{E}Y_1), \ldots, g(\mathbb{E}Y_n)\right)^\top = X\beta \quad \text{with } \beta = (\beta_1, \ldots, \beta_m)^\top \tag{5.1}$$

where $g : G \subset \mathbb{R} \to \mathbb{R}$ is the so called *link function* with domain G. The rank is given by $\operatorname{rank}(X) = m$.

Under the assumption that g is known explicitly, the parameter vector β is desired to be estimated using $(Y_1, \ldots, Y_n)$. Here it is assumed that Y_i , $i = 1, \ldots, n$ are independent but not necessarily identically distributed. But their distribution is a member of the following family of distributions.

5.1 Exponential family of distributions

Definition 5.1.1. The distribution of a random variable Y is a member of the *exponential family*, if the functions $a : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ and $b : \Theta \to \mathbb{R}$ exist, such that

- in the *absolutely continuous case* the probability density function of Y is given by

$$f_\theta(y) = \exp\Big\{\frac{1}{\tau^2}(y\theta + a(y,\tau) - b(\theta))\Big\}, \quad y \in \mathbb{R} \tag{5.2}$$

- in the *discrete case* the probability mass function of Y is given by

$$P_\theta(Y = y) = \exp\Big\{\frac{1}{\tau^2}(y\theta + a(y,\tau) - b(\theta))\Big\}, y \in C \tag{5.3}$$

where C is the (at most) countable *domain* of Y, τ^2 the so called *error term*, $\theta \in \Theta \subset \mathbb{R}$ a parameter and

$$\Theta = \Big\{\theta \in \mathbb{R} : \int_{\mathbb{R}} \exp\Big\{\frac{y\theta + a(y,\tau)}{\tau^2}\Big\} dy < \infty\Big\}$$

respectively in the discrete case:

$$\Theta = \Big\{\theta \in \mathbb{R} : \sum_{y \in C} \exp\Big\{\frac{y\theta + a(y,\tau)}{\tau^2}\Big\} < \infty\Big\}$$

which is the natural parameter space with at least two different elements.

Lemma 5.1.2. Θ is an interval.

Proof Show that $\Theta \subset \mathbb{R}$ is convex. It is then (possibly an infinite) interval. For arbitrary θ_1, $\theta_2 \in \Theta$ (at least one pair exists by Definition 5.1.1) it holds that $\alpha\theta_1 + (1-\alpha)\theta_2 \in \Theta$ for all $\alpha \in (0,1)$. In order show that the statement above holds, suppose that the distribution of Y is absolutely continuous. Since $\theta_i \in \Theta$, it holds that

$$\int_{\mathbb{R}} \exp\Big\{\frac{1}{\tau^2}\Big(y\theta_i + a(y,\tau)\Big)\Big\} dy < \infty, \quad i = 1, 2.$$

The inequality

$$\alpha x_1 + (1-\alpha)x_2 \leq \max\{x_1, x_2\}, \quad x_1, x_2 \in \mathbb{R}, \quad \alpha \in (0,1),$$

implies

$$\begin{aligned}
&\exp\Big\{\frac{1}{\tau^2}\Big(y(\alpha\theta_1 + (1-\alpha)\theta_2) + a(y,\tau)\Big)\Big\} \\
&= \exp\Big\{\alpha\frac{1}{\tau^2}\Big(y\theta_1 + a(y,\tau)\Big) + (1-\alpha)\frac{1}{\tau^2}\Big(y\theta_2 + a(y,\tau)\Big)\Big\} \\
&\leq \max_{i=1,2} \exp\Big\{\frac{1}{\tau^2}\Big(y\theta_i + a(y,\tau)\Big)\Big\} \\
&\leq \exp\Big\{\frac{1}{\tau^2}\Big(y\theta_1 + a(y,\tau)\Big)\Big\} + \exp\Big\{\frac{1}{\tau^2}\Big(y\theta_2 + a(y,\tau)\Big)\Big\},
\end{aligned}$$

and thus

$$\int_{\mathbb{R}} \exp\left\{\frac{1}{\tau^2}\Big(y(\alpha\theta_1 + (1-\alpha)\theta_2) + a(y,\tau)\Big)\right\} dy$$
$$\leq \sum_{i=1}^{2} \int_{\mathbb{R}} \exp\left\{\frac{1}{\tau^2}\Big(y\theta_i + a(y,\tau)\Big)\right\} dy < \infty$$

by the assumptions of the lemma. In summary

$$\alpha\theta_1 + (1-\alpha)\theta_2 \in \Theta,$$

and Θ is a interval. □

Example 5.1.3. Which distributions belong to the exponential family?

1. **Normal distribution**: If $Y \sim \mathcal{N}(\mu, \sigma^2)$, then μ is the parameter of interest and σ^2 the error term. It holds that

$$\begin{aligned} f_\mu(y) &= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(y-\mu)^2}{2\sigma^2}} \\ &= \exp\left\{\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\left(\frac{y^2}{\sigma^2} - \frac{2y\mu}{\sigma^2} + \frac{\mu^2}{\sigma^2}\right)\right\} \\ &= \exp\left\{\frac{1}{\sigma^2}\left(y\mu - \frac{y^2}{2} - \left(\frac{\mu^2}{2} + \frac{\sigma^2}{2}\log(2\pi\sigma^2)\right)\right)\right\} \end{aligned}$$

   and setting $\theta = \mu$, $\tau = \sigma$,

$$a(y,\tau) = -\frac{y^2}{2} - \frac{\sigma^2}{2}\log(2\pi\sigma^2) \quad \text{and} \quad b(\mu) = b(\theta) = \frac{\mu^2}{2}$$

   satisfies Equation (5.2)

2. **Bernoulli distribution**: $Y \sim \text{Bernoulli}(p)$, $p \in [0;1]$
   The Bernoulli distribution is usually used in surveys in market research where

$$Y = \begin{cases} 1, & \text{if the answer is “yes”} \\ 0, & \text{if the answer is “no”} \end{cases}$$

   for a given question in the respective survey.
   Here the probabilities are given by $P(Y=1) = p$, $P(Y=0) = 1-p$. Then for $y \in \{0,1\}$ it holds that:

$$\begin{aligned} P_\theta(Y=y) = p^y(1-p)^{1-y} &= e^{y\log p + (1-y)\log(1-p)} \\ &= e^{y\log\frac{p}{1-p} - (-\log(1-p))} \end{aligned}.$$

   Thus the Bernoulli distribution is a member of the exponential family with $\theta = \log\frac{p}{1-p}$, $\tau = 1$,

$$a(y,\tau) = 0, \quad b(\theta) = -\log(1-p) = \log(1+e^\theta).$$

3. **Poisson distribution**: If $Y \sim$ Poisson (λ), $\lambda > 0$, then for $y \in \mathbb{N}_0$

$$P_\theta(Y = y) = e^{-\lambda} \cdot \frac{\lambda^y}{y!} = e^{y \log \lambda - \log(y!) - \lambda}.$$

Thus the Poisson distribution is a member of the exponential family with $\theta = \log \lambda$, $\tau = 1$,

$$a(y, \tau) = -\log(y!), \quad b(\theta) = \lambda = e^\theta \ .$$

Lemma 5.1.4. If the distribution of a random variable Y is a member of the exponential family, $\mathbb{E}Y^2 < \infty$ and $b : \Theta \to \mathbb{R}$ is two times continuously differentiable with $b''(\theta) > 0$ for all $\theta \in \Theta$, then

$$\mathbb{E}Y = b'(\theta), \quad \mathrm{Var}Y = \tau^2 b''(\theta).$$

Proof

1. Only the case for absolutely continuous distributions is discussed below. The discrete case can be handled simultaneously by replacing the $\int$ with $\sum$. It holds that

$$\begin{aligned}
\mathrm{E}Y &= \int_{\mathbb{R}} y f_\theta(y) dy = \int_{\mathbb{R}} y \exp\left\{ \frac{1}{\tau^2} \Big(y\theta + a(y,\tau) - b(\theta) \Big) \right\} dy \\
&= e^{-\frac{b(\theta)}{\tau^2}} \cdot \tau^2 \int_{\mathbb{R}} \frac{\partial}{\partial \theta} \exp\left\{ \frac{1}{\tau^2} \Big(y\theta + a(y,\tau) \Big) \right\} dy \\
&= e^{-\frac{b(\theta)}{\tau^2}} \cdot \tau^2 \frac{\partial}{\partial \theta} \int_{\mathbb{R}} \exp\left\{ \frac{1}{\tau^2} \Big(y\theta + a(y,\tau) \Big) \right\} dy \\
&= e^{-\frac{b(\theta)}{\tau^2}} \cdot \tau^2 \frac{\partial}{\partial \theta} \left(e^{\frac{b(\theta)}{\tau^2}} \underbrace{\int_{\mathbb{R}} \exp\left\{ \frac{1}{\tau^2} \Big(y\theta + a(y,\tau) - b(\theta) \Big) \right\} dy}_{\int_{\mathbb{R}} f_\theta(y) dy = 1} \right) \\
&= e^{-\frac{b(\theta)}{\tau^2}} \tau^2 \frac{\partial}{\partial \theta} \left(e^{\frac{b(\theta)}{\tau^2}} \right) = e^{-\frac{b(\theta)}{\tau^2}} \cdot \tau^2 \frac{b'(\theta)}{\tau^2} e^{\frac{b(\theta)}{\tau^2}} = b'(\theta).
\end{aligned}$$

2. Show (analogously to 1) that

Exercise 5.1.5.

$$\mathrm{Var}Y = \tau^2 b''(\theta).$$

□

5.2 Link functions

The goal variables Y_i, $i = 1, \ldots, n$ are i.i.d. with a distribution which is a member of the exponential family and a probability (density or) mass function as in (5.3) resp. (5.2). Assume that $b : \Theta \to \mathbb{R}$ is two times continuously differentiable with $b''(\theta) > 0$ for all $\theta \in \Theta$. Additionally assume a generalized linear model as in (5.1).

Definition 5.2.1. (Natural link functions) The link function $g : G \to \mathbb{R}$ is called *natural*, if $g = (b')^{-1}$, $G = \{b'(\theta) : \theta \in \Theta\}$ and g is two times continuously differentiable with $g'(x) \neq 0$ for all $x \in G$.

The question why the link function is called "natural" is answered in the following Lemma.

Lemma 5.2.2. If the generalized linear model (5.1) has the natural link function, then $(\theta_1, \ldots, \theta_n)^\top = X\beta$

Proof Since $b''(\theta) > 0$, it holds that $b'(\theta)$ is monotonically increasing and thus invertible. Define

$$\mu_i = \mathbb{E}Y_i, \quad \eta_i = x_i^\top \beta, \quad x_i = (x_{i1}, \ldots, x_{im})^\top, \quad i = 1, \ldots, n.$$

Since g is invertible it holds that

$$\mu_i = g^{-1}(x_i^\top \beta) = g^{-1}(\eta_i), \quad i = 1, \ldots, n.$$

On the other hand, it holds that $\mu_i = b'(\theta_i)$ by Lemma 5.1.4, so

$$b'(\theta_i) = g^{-1}(\eta_i) \overset{\text{Def. } 5.2.1}{=} b'(\eta_i), \quad i = 1, \ldots, n.$$

Because of the monotonicity of b' the assertion $\theta_i = \eta_i$, $i = 1, \ldots, n$ holds. □

Example 5.2.3. In the following the natural link function for the distributions of Example 5.1.3.

1. **Normal distribution**: Since $b(\mu) = \frac{\mu^2}{2}$, it holds that

$$b'(x) = \frac{2x}{2} = x \text{ and thus } g(x) = (b')^{-1}(x) = x$$

The natural link function is given by $g(x) = x$, thus it holds that

$$(\mu_1, \ldots, \mu_n)^\top = (\mathbb{E}Y_1, \ldots, \mathbb{E}Y_n)^\top = X\beta$$

This is exactly the case of linear regression.

2. **Bernoulli distribution**: Since $b(\theta) = \log(1 + e^\theta)$, it holds that

$$\begin{aligned} & b'(x) = \frac{1}{1+e^x} \cdot e^x = y \\ \Leftrightarrow & \frac{1}{e^{-x}+1} = y \\ \Leftrightarrow & \frac{1}{y} - 1 = e^{-x} \\ \Leftrightarrow & x = -\log \frac{1-y}{y} = \log \frac{y}{1-y} \\ \Rightarrow & g(x) = (b')^{-1}(x) = \log \frac{x}{1-x} \end{aligned}$$

The generalized linear regression model in the case of the Bernoulli distribution is called *binary (categorical) regression*. If it is used with the natural link function, it is called *logistic regression*. In this case it holds that

$$\begin{aligned} & (p_1, \ldots, p_n)^\top = (\mathbb{E}Y_1, \ldots, \mathbb{E}Y_n)^\top, \\ & \theta_i = \log \frac{p_i}{1-p_i} = x_i^\top \beta, \quad i = 1, \ldots, n \\ \Leftrightarrow & e^{\theta_i} = \frac{p_i}{1-p_i} \\ \Leftrightarrow & p_i = \frac{e^{\theta_i}}{1+e^{\theta_i}} \\ \Leftrightarrow & p_i = \frac{e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}}, \quad i = 1, \ldots, n. \end{aligned}$$

The ratio

$$\frac{p_i}{1-p_i} = \frac{P(Y_i = 1)}{P(Y_i = 0)}, \quad i = 1, \ldots, n$$

is called *Odds*. The logarithm of the Odds is called *Logit*:

$$\log \frac{p_i}{1-p_i}, \quad i = 1, \ldots, n.$$

Logits are thus "new goal variables", which are estimated by using the linear combinations $x_i^\top \beta$.
An alternative link function which is often used is defined by $g(x) = \Phi^{-1}(x)$, which is the *Quantile function of the normal distribution* . It is however, not a natural link function. By using them, the so called *Probit model*:

$$p_i = \Phi(x_i^\top \beta), \quad i = 1, \ldots, n$$

can be obtained.

3. **Poisson distribution**: Since $b(\theta) = e^\theta$, in this case

$$g(x) = (b')^{-1}(x) = \log x, \quad x > 0$$

is the natural link function. Thus the generalized linear model with natural link function has the representation

$$(\log \lambda_1, \ldots, \log \lambda_n)^\top = X\beta \quad \text{or} \quad \lambda_i = e^{x_i^\top \beta},\ i = 1, \ldots, n.$$

5.3 Maximum likelihood estimator for β

Since the probability mass (or density) function of Y_i is given by

$$\exp\left\{\tfrac{1}{\tau^2}(y\theta_i + a(y,\tau) - b(\theta_i))\right\}$$

and the Y_i are independent, the log-likelihood function of the random sample $Y = (Y_1, \ldots, Y_n)$ can be written as follows

$$\log L(Y,\theta) = \log \prod_{i=1}^n f_{\theta_i}(Y_i) = \frac{1}{\tau^2} \sum_{i=1}^n \Big(Y_i\theta_i + a(Y_i,\tau) - b(\theta_i)\Big). \tag{5.4}$$

The proof of Lemma 5.2.2 implies that

$$\theta_i = (b')^{-1}(g^{-1}(x_i^\top \beta)), \quad i = 1, \ldots, n\,, \tag{5.5}$$

which implies that $\log L(Y,\theta)$ is a function of the parameter β. From now on the notation $\log L(Y,\beta)$ is used to emphasize this fact.

The maximum likelihood estimator $\hat\beta$ for β is desired:

$$\hat\beta = \operatorname*{argmax}_\beta \log L(Y,\beta)$$

The necessary condition for an extrema

$$\frac{\partial \log L(Y,\beta)}{\partial \beta_i} = 0, \quad i = 1, \ldots, m,$$

needs to be checked. Introduce the following notation

$$\begin{aligned} U_i(\beta) &= \frac{\partial \log L(Y,\beta)}{\partial \beta_i}, \quad i = 1, \ldots, m, \\ U(\beta) &= (U_1(\beta), \ldots, U_m(\beta))^\top, \\ I_{ij}(\beta) &= \mathbb{E}[U_i(\beta)U_j(\beta)], \quad i,j = 1, \ldots, m, \end{aligned}$$

Definition 5.3.1.

1. The matrix $I(\beta) = (I_{ij}(\beta))_{i,j=1}^m$ is called *Fisher information matrix.*

2. Introduce the *Hesse matrix* $W(\beta)$ as a random matrix

$$W(\beta) = (W_{ij}(\beta))_{i,j=1}^m \quad \text{with} \quad W_{ij}(\beta) = \frac{\partial^2}{\partial\beta_i \partial\beta_j} \log L(Y, \beta).$$

This $(m \times m)$ matrix contains the 2nd order derivative of the log-likelihood function, which will be relevant for solving the maximisation problem

$$\log L(Y, \beta) \to \max_\beta .$$

Theorem 5.3.2. It can be shown that $U(\beta)$ and $I(\beta)$ have the following explicit form.

1.

$$U_j(\beta) = \sum_{i=1}^n x_{ij} \left(Y_i - \mu_i(\beta)\right) \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \frac{1}{\sigma_i^2(\beta)}, \quad j = 1, \ldots, m$$

2.

$$I_{jk}(\beta) = \sum_{i=1}^n x_{ij} x_{ik} \left(\frac{\partial g^{-1}(\eta_i)}{\partial \eta_i}\right)^2 \frac{1}{\sigma_i^2(\beta)}, \quad j, k = 1, \ldots, m \ ,$$

where $\eta_i = x_i^\top \beta$, $\mu_i(\beta) = g^{-1}(x_i^\top \beta)$ is the expectation of Y_i and

$$\sigma_i^2(\beta) \overset{\text{Lemma 5.1.4}}{=} \tau^2 b''(\theta_i) \overset{(5.5)}{=} \tau^2 b''((b')^{-1}(g^{-1}(x_i^\top \beta))), \quad i = 1, \ldots, n,$$

is the variance of Y_i.

Proof

1. Introduce the notation

$$l_i(\beta) = \frac{1}{\tau^2} \left(Y_i \theta_i + a\left(Y_i, \tau\right) - b(\theta_i)\right), \quad i = 1, \ldots, n \, .$$

Then,

$$U_j(\beta) = \sum_{i=1}^n \frac{\partial l_i(\beta)}{\partial \beta_j}, \quad j = 1, \ldots, m.$$

Applying the chain rule several times yields

$$\frac{\partial l_i(\beta)}{\partial \beta_j} = \frac{\partial l_i(\beta)}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j}, \quad i = 1, \ldots, n, \ j = 1, \ldots, m.$$

Since

$$\frac{\partial l_i(\beta)}{\partial \theta_i} = \frac{1}{\tau^2}\Big(Y_i - b'(\theta_i)\Big) \overset{\text{Lemma } 5.1.4}{=} \frac{1}{\tau^2}\Big(Y_i - \mu_i(\beta)\Big),$$

$$\frac{\partial \theta_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \theta_i}\right)^{-1} = \Big(\big(b'(\theta_i)\big)'\Big)^{-1} = \big(b''(\theta_i)\big)^{-1} \overset{\text{Lemma } 5.1.4}{=} \left(\frac{\sigma_i^2(\beta)}{\tau^2}\right)^{-1}$$

$$= \frac{\tau^2}{\sigma_i^2(\beta)}\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i}$$

because $\mu_i = \mathrm{E}Y_i = g^{-1}(\eta_i)$,

$$\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial (x_i^\top \beta)}{\partial \beta_j} = x_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m,$$

we have

$$U_j(\beta) = \frac{1}{\tau^2}\sum_{i=1}^{n} x_{ij}\,(Y_i - \mu_i(\beta)) \cdot \frac{\tau^2}{\sigma_i^2(\beta)} \cdot \frac{\partial g^{-1}(\eta_i)}{\partial_{\eta_i}}$$

$$= \sum_{i=1}^{n} x_{ij}(Y_i - \mu_i(\beta))\frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \cdot \frac{1}{\sigma_i^2(\beta)}, \quad j = 1, \ldots, m.$$

2. For all $i, j = 1, \ldots, m$ it holds that

$$I_{ij}(\beta) = \mathrm{E}(U_i(\beta)U_j(\beta))$$

$$= \sum_{k,l=1}^{n} x_{ki}x_{lj} \underbrace{\mathrm{Cov}\,(Y_k, Y_l)}_{\delta_{k_l}\sigma_k^2(\beta)} \cdot \frac{\partial g^{-1}(\eta_k)}{\partial \eta_k}\frac{\partial g^{-1}(\eta_l)}{\partial \eta_l}\frac{1}{\sigma_k^2(\beta)\sigma_l^2(\beta)}$$

$$= \sum_{k=1}^{n} x_{ki}x_{kj}\left(\frac{\partial g^{-1}(\eta_k)}{\partial \eta_k}\right)^2 \frac{1}{\sigma_k^2(\beta)}.$$

□

Remark 5.3.3. In case of the natural link function, simplify the equation above so that the log-likelihood function is given by

$$\log L(Y, \beta) = \frac{1}{\tau^2}\sum_{i=1}^{n}\Big(Y_i x_i^\top \beta + a(Y_i, \tau) - b(x_i^\top \beta)\Big).$$

Since in this case $g^{-1}(\eta_i) = b'(\eta_i)$, $\eta_i = x_i^\top \beta = \theta_i$ holds,

$$\frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} = b''(\theta_i) \overset{\text{Lemma } 5.1.4}{=} \frac{1}{\tau^2}\sigma_i^2(\beta),$$

and thus

$$U_j(\beta) = \frac{1}{\tau^2}\sum_{i=1}^{n} x_{ij}\,(Y_i - \mu_i(\beta)), \quad j = 1,\dots,m,$$
$$I_{jk}(\beta) = \frac{1}{\tau^4}\sum_{i=1}^{n} x_{ij}x_{ik}\sigma_i^2(\beta), \quad j,k = 1,\dots,m.$$

Theorem 5.3.4.

$$W_{jk}(\beta) = \sum_{i=1}^{n} x_{ij}x_{ik}\left(\Big(\,Y_i - \mu_i(\beta)\Big)\nu_i - \frac{u_i^2}{\sigma_i^2(\beta)}\right), \quad j,k = 1,\dots,m$$

where

$$u_i = \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i}$$
$$\nu_i = \frac{1}{\tau^2}\cdot\frac{\partial^2((b')^{-1}\circ g^{-1}(\eta_i))}{\partial \eta_i^2}$$
$$\mu_i(\beta) = \mathrm{E}Y_i,\ \sigma_i^2(\beta) = \mathrm{Var}Y_i,$$
$$\eta_i = x_i^\top\beta$$

for $i = 1,\dots,n$.

Proof For arbitrary $j,k = 1,\dots,m$ it holds that

$$\begin{aligned}
W_{jk}(\beta) = \frac{\partial}{\partial\beta_k}U_j(\beta) &\overset{\text{Theorem 5.3.2}}{=} \frac{\partial}{\partial\beta_k}\sum_{i=1}^{n} x_{ij}\,(Y_i - \mu_i(\beta))\frac{\partial g^{-1}(\eta_i)}{\partial\eta_i}\frac{1}{\sigma_i^2(\beta)}\\
&= \sum_{i=1}^{n} x_{ij}\left((Y_i - \mu_i(\beta))\frac{\partial}{\partial\beta_k}\left(\frac{\partial g^{-1}(\eta_i)}{\partial\eta_i}\frac{1}{\sigma_i^2(\beta)}\right) - \right.\\
&\qquad\left. - \frac{\partial g^{-1}(\eta_i)}{\partial\eta_i}\frac{1}{\sigma_i^2(\beta)}\frac{\partial\mu_i(\beta)}{\partial\beta_k}\right)\\
&= \sum_{i=1}^{n}\left(x_{ij}(Y_i - \mu_i(\beta))\frac{\partial}{\partial\beta_k}\left(\frac{1}{\tau^2}\frac{\partial\theta_i}{\partial\eta_i}\right)\right.\\
&\qquad\left. - \left(\frac{\partial g^{-1}(\eta_i)}{\partial\eta_i}\right)^2\frac{1}{\sigma_i^2(\beta)}x_{ik}\right)\\
&= \sum_{i=1}^{n} x_{ij}x_{ik}\left((Y_i - \mu_i(\beta))\nu_i - u_i^2\frac{1}{\sigma_i^2(\beta)}\right),
\end{aligned}$$

where

$$\frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \cdot \frac{1}{\sigma_i^2(\beta)} \overset{\text{Lemma 5.1.4 and Theorem 5.3.2}}{=} \frac{\partial b'(\theta_i)}{\partial \eta_i} \cdot \frac{1}{\tau^2} \cdot \frac{1}{b''(\theta_i)}$$
$$= \frac{\partial b'(\theta_i)}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \eta_i} \frac{1}{\tau^2} \frac{1}{b''(\theta_i)} = \frac{1}{\tau^2} \frac{\partial \theta_i}{\partial \eta_i}$$

and

$$\frac{\partial}{\partial \beta_k} \left(\frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \cdot \frac{1}{\sigma_i^2(\beta)} \right) = \frac{1}{\tau^2} \frac{\partial^2 \theta_i}{\partial \eta_i^2} \cdot \frac{\partial \eta_i}{\partial \beta_k} \overset{\eta_i = x_i^\top \beta}{=} \frac{1}{\tau^2} \frac{\partial^2 \theta_i}{\partial \eta_i^2} \cdot x_{ik},$$

with

$$\frac{\partial \overbrace{g^{-1}(\eta_i)}^{\mu_i(\beta)}}{\partial \beta_k} = \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_k} = \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \cdot x_{ik}$$

and $\theta_i = (b')^{-1} \circ g^{-1}(\eta_i)$, $i = 1, \ldots, n$. □

Moreover for generalized linear models with natural link function

$$W(\beta) = -I(\beta) = \left(-\frac{1}{\tau^4} \sum_{i=1}^{n} x_{ij} x_{ik} \sigma_i^2(\beta) \right)_{j,k=1}^{m} \tag{5.6}$$

holds, since in this case $\nu_i = 0$ for all $i = 1, \ldots, n$. $W(\beta)$ is therefore deterministic. Indeed, by Lemma 5.2.2 $\theta_i = x_i^\top \beta = \eta_i$ and thus $\frac{\partial^2 \theta_i}{\partial \eta_i^2} = 0$, $i = 1, \ldots, n$.
Remark 5.3.3 implies $u_i^2 = \frac{1}{\tau^4} \sigma_i^4(\beta)$.

Example 5.3.5. What do $U(\beta), I(\beta)$ and $W(\beta)$ look like for the models introduced in Example 5.2.3 (natural link function)?

1. **Normal distribution**: this case corresponds to the usual multivariate linear regression with normally distributed error terms. In this case it holds that $\mu = X\beta$, $\tau^2 = \sigma^2$.

   Remark 5.3.3 implies

$$U(\beta) = \frac{1}{\sigma^2} X^\top (Y - X\beta),$$
$$I(\beta) = \left(\mathbb{E}\left(U_i(\beta) \cdot U_j(\beta) \right) \right)_{i,j=1,\ldots,m} = \frac{1}{\sigma^2} X^\top X,$$
$$W(\beta) = -I(\beta).$$

2. **Logistic regression**: It holds that $\tau^2 = 1,\ \mu_i = p_i,\ \sigma_i^2 = p_i(1-p_i),$ $i = 1, \ldots, n,\ p_i \in (0,1)$ and thus

$$\begin{aligned} U(\beta) &= X^\top (Y - p) \\ I(\beta) &= X^\top diag(p_i(1-p_i))X \\ W(\beta) &= -I(\beta) \end{aligned}$$

   where $p = (p_1, \ldots, p_n)^\top$.

3. **Poisson regression**: It holds that $\tau^2 = 1,\ \mu_i = \lambda_i = \sigma_i^2,\ i = 1, \ldots, n$ and thus

$$\begin{aligned} U(\beta) &= X^\top (Y - \lambda), \\ I(\beta) &= X^\top diag(\lambda_i)X, \\ W(\beta) &= -I(\beta), \end{aligned}$$

   where $\lambda = (\lambda_1, \ldots, \lambda_n)^\top$

When is the solution to the equation $U(\beta) = 0$ maximizing the function $\log L(Y, \beta)$?

In other words: When does an unique MLE $\hat{\beta}$ of β exist?

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \log L(Y, \beta).$$

The sufficient condition of a maximum implies that the Hesse matrix $W(\beta)$ is negative definite.
Consider the case of the natural link function. Then Remark 5.3.3 implies that

- The system of equations $U(\beta) = 0$ can be rewritten as $U(\beta) = \frac{1}{\tau^2} X^\top (Y - \mu(\beta)) = 0$
- The matrix $W(\beta) = -\frac{1}{\tau^4} X^\top diag(\sigma_i^2(\beta))X$ is negative definite, has $\operatorname{rank}(X) = m$ and $0 < \sigma_i^2(\beta) < \infty$ for all $i = 1, \ldots, n$. Under those conditions there exists an unique MLE $\hat{\beta}$ for β.

In the following, two numerical algorithms for solving the system of (in general nonlinear) equations $U(\beta) = 0$ are introduced. These approaches are iterative, i.e. they approximate the MLE $\hat{\beta}$ incrementally.

1. **Newton's method**

   Choose a suitable starting value $\hat{\beta}_0 \in \mathbb{R}^m$.

   In step $k+1$, calculate $\hat{\beta}_{k+1}$ from $\hat{\beta}_k,\ k = 0, 1, \ldots$ as follows:

   - Take the first order Taylor expansion of $U(\beta)$ at $\hat{\beta}_k$: $U(\beta) \approx U(\hat{\beta}_k) + W(\hat{\beta}_k)(\beta - \hat{\beta}_k)$.

- Solving for 0: $U(\hat{\beta}_k) + W(\hat{\beta}_k)(\beta - \hat{\beta}_k) = 0$
- The solution for this system of equations is $\hat{\beta}_{k+1}$:

$$\hat{\beta}_{k+1} = \hat{\beta}_k - W^{-1}(\hat{\beta}_k) \cdot U(\hat{\beta}_k), \quad k = 0, 1, 2, \ldots$$

assuming that $W(\hat{\beta}_k)$ is invertible.

Stop the iteration process once $|\hat{\beta}_{k+1} - \hat{\beta}_k| < \delta$ for a predetermined boundary $\delta > 0$.

The convergence of this method heavily depends on the choice of $\hat{\beta}_0$, since $\hat{\beta}_0$ has to be close enough to $\hat{\beta}$. Another disadvantage of this method is that the random matrix $W(\beta)$ might not be invertible. That is why a modification of the Newton method is presented in which $W(\beta)$ is replaced by the expectation

$$\mathbb{E}W(\beta) = -I(\beta). \tag{5.7}$$

It can be shown that the identity (5.7) holds by using Theorem 5.3.4 and the fact that $\mathbb{E}Y_i = \mu_i,\ i = 1, \ldots, n$. If it is assumed that $\operatorname{rank}(X) = m$ and $u_i \neq 0,\ i = 1, \ldots, n$, then by Theorem 5.3.2, $I(\beta)$ is invertible. This method is called *Fisher's scoring method*.
The only difference of Newton's method compared to Fisher's Scoring is that in the second step the iterative equation

$$\hat{\beta}_{k+1} = \hat{\beta}_k + I^{-1}(\hat{\beta}_k)U(\hat{\beta}_k),\ k = 0, 1, \ldots$$

is used.
In the case of the natural link function (cf. Remark 5.3.3)

$$\begin{aligned}\hat{\beta}_{k+1} &= \hat{\beta}_k + \tau^4\Big(X^\top diag(\sigma_i^2(\hat{\beta}_k))X\Big)^{-1}\frac{1}{\tau^2}\Big(X^\top(Y - \mu(\hat{\beta}_k))\Big) \\ &= \hat{\beta}_k + \tau^2\Big(X^\top diag(\sigma_i^2(\hat{\beta}_k))X\Big)^{-1}\Big(X^\top(Y - \mu(\hat{\beta}_k))\Big).\end{aligned}$$

5.4 Asymptotic tests for β

The goal of this section is to construct a test for the hypotheses

$$\begin{aligned}&H_0 : \beta = \beta_0 \text{ vs.} \\ &H_1 : \beta \neq \beta_0\end{aligned}$$

with $\beta = (\beta_1, \ldots, \beta_m)^\top$ and $\beta_0 = (\beta_{01}, \ldots, \beta_{0m})^\top$. In particular, the hypotheses $H_0 : \beta = 0$ resp. $H_0 : \beta_j = 0$ are of interest, because they imply

that the test variables $Y = (Y_1, \dots, Y_n)^\top$ does not depend on several output variables (e.g. $(x_{1j}, \dots, x_{nj})^\top$ in case of the hypothesis $\beta_j = 0$).
In order to test these kinds of hypotheses, the test statistics T_n are used, which have an asymptotic (for $n \to \infty$) reference distribution (e.g. multivariate normal or χ^2 distribution). Some groundwork has to be done beforehand though. Let

$$g(\mathbb{E}Y_i) = X_i\beta, \quad i = 1, \dots, n$$

be a generalized linear model with natural link function g. Let $L(Y, \beta)$ be the likelihood function, $U(\beta)$ the partial derivatives of $\log L(Y, \beta)$ and $I(\beta)$ the Fisher information matrix in this model.
$\hat{\beta}_n = \hat{\beta}(Y_1, \dots, Y_n, X)$ denotes a sequence of maximum likelihood estimators for β.
Assume that

1. $\exists$ compact subspace $K \subset \mathbb{R}^m$, such that all rows X_i, $i = 1, \dots, n$, $n \in \mathbb{N}$, of X are in K. Here $\theta = x^\top\beta \in \Theta$ for all $\beta \in \mathbb{R}^m$ and $x \in K$.

2. There exists a sequence $\{\Gamma_n\}_{n\in\mathbb{N}}$ of diagonal $(m \times m)$ matrices $\Gamma_n = \Gamma_n(\beta)$ with the properties

   (a) $\gamma^n_{i,i} > 0$, $i \in \{1, \dots, m\}$

   (b) $\lim\limits_{n\to\infty} \Gamma_n = 0$,

   (c) $\lim\limits_{n\to\infty} \Gamma_n^\top I_n(\beta)\Gamma_n = K^{-1}(\beta)$, where $K(\beta)$ is a symmetric positive definite $(m \times m)$ matrix for all $\beta \in \mathbb{R}^m$.

Theorem 5.4.1. Under the conditions above, there exists a Γ_n consistent sequence of MLE $\{\hat{\beta}_n\}$ for β,
(i.e. $P\left(\Gamma_n^{-1}|\hat{\beta}_n - \beta| \le \varepsilon, U(\hat{\beta}_n) = 0\right) \to 1$ for $n \to \infty$), such that

1. $T_n^* = \Gamma_n^{-1}(\hat{\beta}_n - \beta) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, K(\beta))$,

2. $T_n = 2(\log L(Y, \hat{\beta}_n) - \log L(Y, \beta)) \xrightarrow[n\to\infty]{d} \chi^2_m$,

where $m = \dim \beta$.

Remark 5.4.2. (cf. [27], p.288-292)

1. Usually $\Gamma_n = diag\left(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}\right)$ is used.

2. Until now we always assumed that the dispersion term τ^2 is known. If that is not the case, then τ^2 can be estimated by

$$\hat{\tau}^2 = \frac{1}{n-m}\sum_{i=1}^{n} \frac{\left(Y_i - \mu_i(\hat{\beta}_n)\right)^2}{b''(\hat{\theta}_{ni})}$$

where $\hat{\theta}_{ni} = (b')^{-1}(\mu_i(\hat{\beta}_n))$, $i = 1, \ldots, n$. This estimator is an empirical analogue to the equation $\tau^2 = \frac{\mathrm{Var} Y_i}{b''(\theta_i)}$ of Lemma 5.1.4.

3. The second assertion of Theorem 5.4.1 also holds, if the unknown parameter τ^2 can be replaced by a consistent estimator τ_n^2.

How can Theorem 5.4.1 be used to test the hypothesis

$$\begin{aligned} H_0 &: \beta = \beta_0 \quad \text{vs.} \\ H_1 &: \beta \neq \beta_0 \end{aligned}$$

or component wise

$$\begin{aligned} H_0 &: \beta_j = \beta_{j0} \ , \ j = 1, \ldots, m \quad \text{vs.} \\ H_1 &: \exists j_1 : \beta_{j_1} \neq \beta_{j_1 0} \quad ? \end{aligned}$$

Let

$$g(\mathbb{E} Y_i) = \sum_{j=1}^{m} x_{ij} \beta_j, \quad i = 1, \ldots, n$$

be a generalized linear model with natural link function g.
Using Remark 5.3.3, it holds that

$$\log L(Y, \beta) = \frac{1}{\tau^2} \sum_{i=1}^{n} \Big(Y_i x_i^\top \beta + a(Y_i, \tau) - b(x_i^\top \beta) \Big)$$

where $Y = (Y_1, \ldots, Y_n)^\top$ and $x_i = (x_{i1}, \ldots, x_{im})^\top$. Thus is holds that

$$T_n = \frac{2}{\tau^2} \sum_{i=1}^{n} \Big(Y_i x_i^\top (\hat{\beta}_n - \beta_0) - b(x_i^\top \hat{\beta}_n) + b(x_i^\top \beta_0) \Big)$$

By specifying an exponential model (τ, b are known), with respect to the random sample of the goal variable Y and the design matrix X, H_0 is rejected if $T_n > \chi^2_{m,1-\alpha}$, where m is the number of parameters in the model, $\chi^2_{m,1-\alpha}$ the $(1-\alpha)$ quantile of the χ^2_m- distribution and $\alpha \in (0,1)$ is the significance level of the asymptotic test. This test can only be applied for relatively large n. Type I errors have the (for $n \to \infty$) asymptotic probability α. If a simple hypothesis

$$\begin{aligned} H_0 &: \beta_j = 0 \ \text{vs.} \\ H_1 &: \beta_j \neq 0 \end{aligned}$$

is tested, the test statistic T_n^1 is used which can be derived from T_n^*:
H_0 is rejected, if

$$|T_n^1| = \frac{|\hat{\beta}_{nj}|}{(\Gamma_n(\hat{\beta}_n))_{jj}} > z_{1-\frac{\alpha}{2}}$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$ quantile of the $N(0,1)$ distribution. Here $\{\Gamma_n\}$ is chosen in a way that $K(\beta) = Id$, for all $\beta \in \mathbb{R}^m$. This is an asymptotic test with confidence level α, since

$$P_{H_0}(|T_n^1| > z_{1-\frac{\alpha}{2}}) = 1 - P_{H_0}(|T_n^1| \leq z_{1-\frac{\alpha}{2}}) \xrightarrow[n\to\infty]{} 1 - \Phi(z_{1-\frac{\alpha}{2}}) + \underbrace{\Phi(-z_{1-\frac{\alpha}{2}})}_{1-\Phi(z_{1-\frac{\alpha}{2}})}$$

$$= 1 - \left(1 - \frac{\alpha}{2}\right) + 1 - \left(1 - \frac{\alpha}{2}\right) = \alpha,$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

is the cumulative distribution function of the $N(0,1)$ distribution.

Example 5.4.3. (Credit risk assessment)[1]
The following data is provided by a southern German bank from the 90's: Results from credit risk assessment for $n = 1000$ credit applications (ca. 700 good credits and 300 bad credits) analysed:

Goal variable $Y_i = \begin{cases} 0, & \text{if the credit of customer } i \text{ has been paid} \\ 1, & \text{if the credit of customer } i \text{ has not been paid} \end{cases}$.

The design matrix X contains the following additional information about the customer:

x_{i1} - Account management with the bank $= \begin{cases} 1, & \text{no account} \\ 0, & \text{else} \end{cases}$

x_{i2} - Assessment of account management $= \begin{cases} 1, & \text{good account} \\ 0, & \text{no- or bad account} \end{cases}$

x_{i3} - Term of credit in months

x_{i4} - Value of Credit in DM

x_{i5} - Payment history of customer $= \begin{cases} 1, & \text{good} \\ 0, & \text{else} \end{cases}$

x_{i6} - Reference $= \begin{cases} 1, & \text{private} \\ 0, & \text{business} \end{cases}$

Question: How should $\hat{\beta}$ be estimated?

As a model, the logit model is used with $p_i = P(Y_i = 1),\ i = 1, \ldots, n$:

$$\log \frac{p_i}{1-p_i} = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + x_{i4}\beta_4 + x_{i5}\beta_5 + x_{i6}\beta_6$$

for $i = 1, \ldots, n$, where $\beta = (\beta_0, \ldots, \beta_6)^\top, m = 7$.

[1] *cf. Fahrmeir, L., Kneib, T., Lang, S. - Regression, p.208*

		$Y=1$	$Y=0$
x_1	no account	45.0	20.0
x_2	good	15.3	49.8
	bad	39.7	30.2
x_4	Credit value	$Y=1$	$Y=0$
	$0 < \ldots \leq 500$	1.00	2.14
	$500 < \ldots \leq 1000$	11.33	9.14
	$1000 < \ldots \leq 1500$	17.00	19.86
	$1500 < \ldots \leq 2500$	19.67	24.57
	$2500 < \ldots \leq 5000$	25.00	28.57
	$5000 < \ldots \leq 7500$	11.33	9.71
	$7500 < \ldots \leq 10000$	6.67	3.71
	$10000 < \ldots \leq 15000$	7.00	2.00
	$15000 < \ldots \leq 20000$	1.00	0.29
x_5	Credit history	$Y=1$	$Y=0$
	good	82.33	94.95
	bad	17.66	5.15
x_6	Reference	$Y=1$	$Y=0$
	private	57.53	69.29
	business	42.47	30.71

Table 5.1: Abstract of the data

Goal: Estimate $\beta_0, \ldots, \beta_6$ and check, which factors are important for future credit risk assessment.

$H_0 : \beta_i = 0$ (feature x_i does not affect the credit risk assessment) is rejected, if the p-value $\leq \alpha$. It can also be noticed that β_4 is not relevant for credit risk assessment, contrary to belief. A refinement of the model is necessary:

New model:

$$g(\mathbb{E}Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3^1 x_{i3} + \beta_3^2 x_{i3}^2 + \beta_4^1 x_{i4} + \beta_4^2 x_{i4}^2 + \beta_5 x_{i5} + \beta_6 x_{i6}$$

Question: Which model is better?

$\overline{x}_1$	$\overline{x}_2$	$\overline{x}_3$	$\overline{x}_4$	$\overline{x}_5$	$\overline{x}_6$
0.274	0.393	20.903	3271	0.911	0.657

Table 5.2: Means $\overline{x}_j$ of x_{ij} in the data

	Value	$\sqrt{(I_n^{-1}(\hat{\beta}))_{ii}}$	T_n^1	p-value
β_0	0.281	0.303	-0.94	0.347
β_1	0.618	0.175	3.53	< 0.001
β_2	-1.338	0.201	-6.65	< 0.001
β_3	0.033	0.008	4.29	< 0.001
β_4	0.023	0.033	0.72	0.474
β_5	-0.986	0.251	-3.93	< 0.001
β_6	-0.426	0.266	-2.69	0.007

Table 5.3: Results of the ML estimation by using the Fisher Scoring method, where $\sqrt{(I_n^{-1}(\hat{\beta}))_{ii}}$ is used as an asymptotic standard deviation of $\hat{\beta}_i$. Significance level : $\alpha = 0.001$.

In other words, the following hypotheses are tested:

$H_0 : \beta_3^2 = 0$ (linear model) vs. $H_1 : \beta_3^2 \neq 0$ (quadratic model) resp.
$H_0 : \beta_4^2 = 0$ (linear model) vs. $H_1 : \beta_4^2 \neq 0$ (quadratic model)

Here the type of statistical hypothesis is generalized as follows:

$$H_0 : C\beta = d \text{ vs. } H_1 : C\beta \neq d$$

is tested, where C is a $(r \times m)$ - matrix with rank $C = r \leq m$ and $d \in \mathbb{R}^r$. For comparison, the hypothesis

$$H_0 : \beta = \beta_0 \text{ vs. } H_1 : \beta \neq \beta_0, \quad \beta, \beta_0 \in \mathbb{R}^m$$

was tested before. Obviously $\beta = \beta_0$ is a special case of $C\beta = d$ with $C =$ Id, $d = \beta_0$. The new hypotheses include assertions about the linear combinations of the parameters. How should H_0 vs. H_1 be tested?

Let $\widetilde{\beta}_n$ be the MLE of β under H_0, i.e. $\widetilde{\beta}_n = \underset{\beta \in \mathbb{R}^m : C\beta = d}{\operatorname{argmax}} \log L(Y, \beta)$
Let $\hat{\beta}_n$ be the MLE of β unrestricted, i.e. $\hat{\beta}_n = \underset{\beta \in \mathbb{R}^m}{\operatorname{argmax}} \log L(Y, \beta)$.
The idea behind the following tests is to compare $\widetilde{\beta}_n$ with $\hat{\beta}_n$.

	Value	$\sqrt{(I_n^{-1}(\hat{\beta}))_{ii}}$	T_n^1	p-Wert
β_0	-0.488	0.390	-1.25	0.211
β_1	0.618	0.176	3.51	< 0.001
β_2	-1.337	0.202	-6.61	< 0.001
β_3^1	0.092	0.025	3.64	< 0.001
β_3^2	-0.001	< 0.001	-2.20	0.028
β_4^1	-0.264	0.099	-2.68	0.007
β_4^1	0.023	0.007	3.07	0.002
β_5	-0.995	0.255	-3.90	< 0.001
β_6	-0.404	0.160	-2.52	0.012

Table 5.4: p-values for the regression coefficients of the new model

If the deviation $\hat{\beta}_n - \widetilde{\beta}_n$ is big, H_0 should be rejected.

Theorem 5.4.4. Let $\log L(Y, \beta)$ be the log-likelihood function of the random sample of the goal variable $Y = (Y_1, \ldots, Y_n)^\top$, $I_n(\beta)$ be the fisher information matrix, $U(\beta)$ be the score function of the generalized linear model with natural link function g:

$$g(\mathbb{E}Y_i) = X_i\beta, \quad i = 1, \ldots, n.$$

Consider the following test statistics

1. **likelihood-ratio test statistic**:

$$\widetilde{T}_n = 2(\log L(Y, \hat{\beta}_n) - \log L(Y, \widetilde{\beta}_n))$$

2. **Wald statistic**:

$$\widetilde{T}_n^* = (C\hat{\beta}_n - d)^\top (CI_n^{-1}(\hat{\beta}_n)C^\top)^{-1}(C\hat{\beta}_n - d)$$

3. **Score statistic**:

$$\overline{T}_n^* = U(\widetilde{\beta}_n)^\top I_n^{-1}(\widetilde{\beta}_n)U(\widetilde{\beta}_n)$$

Under certain conditions for the estimators $\hat{\beta}$ and $\widetilde{\beta}$ (cf. Theorem 5.4.1) the test statistics 1 - 3 are asymptotically χ_m^2 distributed, e.g. for the likelihood-ratio-test statistic it holds that

$$\widetilde{T}_n \xrightarrow[n\to\infty]{d} \chi_m^2.$$

Corollary 5.4.5. Theorem 5.4.4 provides the following decision rule: H_0 is rejected, if

$$\widetilde{T}_n(\widetilde{T}_n^*, \overline{T}_n) > \chi^2_{m,1-\alpha}.$$

This is an asymptotic test with confidence level α.

Example 5.4.6 (Continued)**.** The following values for the test statistics are obtained:

- $\widetilde{T}_n = 12.44$ p-value: 0.0020
- $\widetilde{T}_n^* = 11.47$ p-value: 0.0032

for $\alpha = 0.005$ it holds that the p-value $\leq \alpha$, thus $H_0 : \beta_4^2 = 0$ is rejected $\Rightarrow$ the quadratic generalized linear model is preferred.

5.5 Criteria for model selection or model adjustment

It is known that the goodness of fit of a parametric model to the data rises, if the number of parameters increases. A goal of a statistician is to find a well fitted model with as little variables as possible in order to avoid overfitting. The Akaike information criterion can help achieving this goal by comparing models with (possibly) different parameter estimators.
Akaike information coefficient:

$$\text{AIC} = -2 \log L(Y, \hat{\beta}) + 2m$$

where $Y = (Y_1, \ldots, Y_n)$ is the random sample of the goal variable in the generalized linear model and $\hat{\beta}$ the corresponding MLE. The value of the AIC takes the required maximality of the log-likelihood function $\log L(Y, \hat{\beta})$ into account and punishes models with a large number of parameters m. The models with the smallest AIC is considered the better model. Sometimes instead of the AIC, the standardized AIC given by $\frac{\text{AIC}}{n}$ is used.

Example 5.5.1 (Continued)**.** Calculate the AIC for the linear and quadratic Logit model in the example of credit risk assessment:

$$\text{Linear model} : \text{AIC} = 1043.815$$
$$\text{Quadratic model} : \text{AIC} = 1035.371$$

By considering the AIC, it can be noticed that the quadratic model seems to be better.

A disadvantage of making a decision based on the AIC alone is, that the final decision is up to the statistician. Thus it is desirable to construct a

statistical test which can asses the goodness of fit for the model. The χ^2-test aims to solve this issue.
Let

$$g(\mathbb{E}Y_i) = X_i\beta, \quad i = 1, \ldots, n$$

be a generalized linear model with link function g and parameter vector $\beta = (\beta_1, \ldots, \beta_m)^\top$. Split the goal variables $Y_1, \ldots, Y_n$ in k groups, such that they are as homogeneous as possible with respect to the parameters that need to be estimated. A said partition can be achieved by splitting the domain of the goal variable Y_i „skillfully“ in $k > m$[2] intervals $(a_l, b_l]$:

$$-\infty \leq a_1 < b_1 = a_2 < b_2 = a_3 < \ldots < b_{k-1} = a_k < b_k \leq +\infty$$

Group l contains all observations Y_i, which are in $(a_l, b_l]$. Here $(a_l, b_l]$ need to be chosen in a way, such that $\hat{\mu}_j = g^{-1}(X_j\hat{\beta})$ are constant within the respective groups: $\hat{\mu}_j \equiv \hat{\mu}_l \ \forall\ j$ of group l.[3] Let

- $n_l = \#\ \{Y_j : Y_j \ \in \ (a_l, b_l]\}$ be class size of class l.
- $\overline{Y}_l = \frac{1}{n_l}\sum Y_j$ be the arithmetic mean of class l.
- $\hat{\beta}$ be the MLE of β, which was obtained by using Y.
- $l_l(\beta) = \sum \log f_\theta(Y_j)$ be the log-likelihood function of the goal variables Y_i within the group l.
- $\hat{\mu}_l = g^{-1}(X_l\hat{\beta})$ and $v(\hat{\mu}_l)$ be the expectation estimator resp. variance estimator $\mu_l = \mathbb{E}Y_l$, which are obtained by using the MLE $\hat{\beta}$.

Here $v(\hat{\mu}_l) = \tau^2 b''(b'^{-1}(\hat{\mu}_l))$, where $b(\cdot)$ is the corresponding coefficient in the probability density function f_θ in the exponential family. The following test statistic is obtained:

$$\chi^2 = \sum_{l=1}^{k} \frac{(\overline{Y}_l - \hat{\mu}_l)^2}{v(\hat{\mu}_l)/n_l},$$

$$D = -2\tau^2 \sum_{l=1}^{k} \Big(l_l(\hat{\mu}_l) - l_l(\overline{Y}_l) \Big).$$

Theorem 5.5.2. If $n \to \infty$ and $n_l \to \infty\ \forall\ l$, then under certain conditions it holds that

$$\chi^2 \xrightarrow[n\to\infty]{d} \chi^2_{k-m-1}$$

$$D \xrightarrow[n\to\infty]{d} \chi^2_{k-m-1}$$

[2] $k \leq m \Rightarrow D \xrightarrow[n\to\infty]{d} \chi^2_{\underbrace{k-m-1}_{<0}}$

[3] This is an informal description of the methodology, in which for each Y_i, n_i independent copies of Y_i are created, which compose the i-th class.

Corollary 5.5.3. The hypotheses

$$H_0 : Y = (Y_1, \ldots, Y_n) \text{ is from the model } g(\mathbb{E}Y_i) = X_i$$

vs.

$$H_1 : Y = (Y_1, \ldots, Y_n) \text{ is not from the model } g(\mathbb{E}Y_i) = X_i\beta$$

for $i = 1, \ldots, n$ can be tested as follows:
H_0 is rejected (for large n) at significance level α, if

$$\chi^2 > \chi^2_{k-m-1,1-\alpha} \quad \text{resp.} \quad D > \chi^2_{k-m-1,1-\alpha}.$$

Those tests should not be used if the class sizes n_l are small.

Example 5.5.4. What do the tests described above look like in the Logit- resp. Poisson regression?

1. **logit model**: $y_i \sim \text{Bernoulli}(p_i),\ i = 1, \ldots, n$

   $$\Rightarrow \text{ generalized linear model} \quad \log \frac{p_i}{1-p_i} = x_i\beta$$

   for $i = 1, \ldots, n$. Divide $y_1, \ldots, y_n$ in k classes, such that the probability of occurring 1 in each class is estimated as good as possible by $\overline{y}_l = \frac{1}{n_l} \sum y_i$. Thus it holds that

   - $\hat{\mu}_l = \hat{p}_l = g^{-1}(X_l\hat{\beta}) = \frac{e^{X_l^\top \hat{\beta}}}{1+e^{X_l^\top \hat{\beta}}}$,
   - $v(\hat{p}_l) = \hat{p}_l(1-\hat{p}_l)$,
   - $\chi^2 = \sum_{l=1}^k \frac{(\overline{Y}_l - \hat{p}_l)^2}{\hat{p}_l(1-\hat{p}_l)/n_l}$.

2. **Poisson model**: $Y_i \sim \text{Poisson}(\lambda)$,

   $$\Rightarrow \text{ generalized linear model} \quad \log \lambda_i = X_i\beta$$

   for $i = 1, \ldots, n$. Thus it holds that $\hat{\mu}_l = \hat{\lambda}_l = e^{X_l\hat{\beta}},\ v(\hat{\lambda}_l) = \hat{\lambda}_l$ and

   $$\chi^2 = \sum_{l=1}^k \frac{(\overline{Y}_l - \hat{\lambda}_l)^2}{\hat{\lambda}_l/n_l}.$$

Chapter 6

Principal Component Analysis

In this chapter, methods for reducing the complexity of big statistical data is presented in form of the principal component analysis (PCA). PCA aims to reduce a high dimensional datasets $X = (X_1, \dots, X_n)^\top \in \mathbb{R}^n$ to very few but important components $\varphi = AX \in \mathbb{R}^d$ with $d \ll n$. Those components should then explain most of the variability of the original dataset X. Here, A is a $(d \times n)$ matrix which can be found if some restrictions (given in (6.1)) are fulfilled. Another applications of PCA is the visualization of complex datasets, outlier detection, cluster analysis and so on. For an overview see [18].

6.1 Introduction

In order to motivate the following problem, consider the example of text mining in automotive:

Example 6.1.1. A car manufacturer is interested in minimizing its losses which are due to fraud and incompetence in warranty repairs in one of the subsidiaries. That's why he wants to introduce a conspicuousness analysis of repair visits in said warranty subsidiary, which is supposed to find suspicious reports with the help of a computer that can be manually checked afterwards. Another motivation for the automatized early detection system is the comprehensive examination of few subsidiaries in irregular time intervals (due to high costs) which could be marginalized. A typical repair log consists of a maximum of 300.000 technical terms. That's why the logs should be saved as vectors $x = (x_1, \dots, x_n)^\top$ of length $n = 300.000$, where

$$x_i = \begin{cases} 1, & \text{the word } i \text{ is in text } x \\ 0, & \text{else} \end{cases}$$

These vectors x are normed such that they are on the sphere S^{n-1}. Within one year a huge dataset of those vectors is created with several million entries. The task of a statistician is the drastic reduction of the dimension $n-1$ of the data such that a visualization is possible. A possible solution is the usage of PCA. The groundwork for PCA has been done by Beltran (1873) and Jordan (1874) who introduced the singular value decomposition (SVD). In a more or less modern form (cf. (6.1)) it is presented in the work of K. Pearson (1901) and H. Hotelling (1933). Also, the name PCA was introduced by Hotelling. A more developed version of the method was introduced by Girshick (1939), Anderson (1963), Rao (1964) and some others. Without a computer the calculation of principal components for $n > 4$ turns out to be rather difficult, thus this methodology has found its practical applications after their invention.
Since the 1980's there is a rapid increase in applications of PCA in the whole knowledge domain (especially in in engineering), where multivariate datasets are analyzed.

6.2 PCA on model level

This section aims to introduce the main concept of PCA for random samples $X = (X_1, \dots, X_n)^\top$ with known covariance structure. Let $X = (X_1, \dots, X_n)^\top$ be a random sample of random variables X_i with known covariance matrix Σ and $\mathrm{Var} X_i \in (0, \infty)$, $i = 1, \dots, n$. Let $\lambda_1 > \lambda_2 > \ldots > \lambda_n > 0$ be the eigenvalues of Σ, which are sorted in descending order and all different from each other. The goal is to find linear combinations $\alpha^\top X$ of X_i which have the biggest variance, whereas the vectors α are normed respectively, e.g. such that $\alpha \in S^{n-1}$ with the euclidean norm.

Definition 6.2.1. The linear combination $\alpha_i^\top X$, $i = 1, \dots, n$, is called i-th principal component of X, if it has the biggest variance under the condition that $\alpha_i \in S^{n-1}$ and $\alpha_1^\top X, \alpha_2^\top X, \dots, \alpha_{i-1}^\top X$ and $\alpha_i^\top X$ are uncorrelated:

$$\begin{cases} \mathrm{Var}\ \alpha^\top X \to \max\limits_{\alpha}, \\ |\alpha| = 1, \\ \mathrm{Cov}\ (\alpha^\top X, \alpha_j^\top X) = 0, \quad j = 1, \dots, i-1. \end{cases} \tag{6.1}$$

Here α_i is the coefficient vector of the i-th principal component $\alpha_i^\top X$.

Theorem 6.2.2. The i-th principal component of X is given by

$$Y_i = \alpha_i^\top X,$$

where α_i is the eigenvector of Σ with eigenvalue λ_i. Here

$$\mathrm{Var}(Y_i) = \lambda_i, \quad i = 1, \dots, n.$$

Proof Show that the assertion holds for $i = 1, 2$. For $i > 2$ the proof works in the same spirit.
For $i = 1$ there is a constraint $|\alpha| = 1$ in (6.1), which is taken into the Lagrange function

$$f(\alpha) = \mathrm{Var}(\alpha^\top X) + \lambda(|\alpha|^2 - 1).$$

Furthermore

$$\begin{aligned}\mathrm{Var}(\alpha^\top X) &= \mathbb{E}\Big(\alpha^\top X - \mathbb{E}\alpha^\top X\Big)^2 = \mathbb{E}\Big(\alpha^\top (X - \mathbb{E}X)\Big)^2 \\ &= \mathbb{E}\alpha^\top (X - \mathbb{E}X)(X - \mathbb{E}X)^\top \alpha = \alpha^\top \mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)^\top \alpha \\ &= \alpha^\top \Sigma \alpha,\end{aligned}$$

$|\alpha|^2 = \alpha^\top \cdot \alpha$, and $f(\alpha) = \alpha^\top \Sigma \alpha + \lambda(\alpha^\top \alpha - 1)$.
The necessary conditions for a maximum is given by

$$\frac{\partial f}{\partial \alpha} = 0, \quad \frac{\partial f}{\partial \lambda} = 0,$$

where the second equation represents the constraint $|\alpha| = 1$.
$\frac{\partial f}{\partial \alpha} = \left(\frac{\partial f}{\partial \alpha^1}, \ldots, \frac{\partial f}{\partial \alpha^n}\right)$, where $\alpha = (\alpha^1, \ldots, \alpha^n)^\top$ and $\frac{\partial f}{\partial \alpha} = 0$ can be rewritten as $\Sigma\alpha - \lambda\alpha = 0$ vectorial or $\Sigma\alpha = \lambda\alpha$, which means, that α is an eigenvector of Σ with eigenvalue λ. Since $\mathrm{Var}(\alpha^\top X) = \alpha^\top \Sigma \alpha$ is supposed to be maximized, it holds that

$$\mathrm{Var}(\alpha^\top X) = \alpha^\top \lambda \alpha = \lambda \underbrace{\alpha^\top \alpha}_{1} = \lambda$$

and $\lambda = \lambda_1 > \lambda_2 > \ldots > \lambda_n \Rightarrow \lambda = \lambda_1$ and $\alpha = \alpha_1$.
For $i = 2$, the maximization task

$$\begin{cases} \alpha^\top \Sigma \alpha \to \max\limits_{\alpha}, \\ \alpha^\top \cdot \alpha = 1, \\ \mathrm{Cov}\ (\alpha_1^\top X, \alpha^\top X) = 0 \end{cases}$$

needs to be solved with respect to α, where

$$\mathrm{Cov}\ (\alpha_1 X, \alpha^\top X) = \alpha_1^\top \Sigma \alpha = \alpha^\top \Sigma \alpha_1 = \alpha^\top \lambda_1 \alpha_1 = \lambda_1 \alpha^\top \alpha_1.$$

That means, the following function needs to be maximized:

$$f(\alpha) = \alpha^\top \Sigma \alpha + \lambda(\alpha^\top \alpha - 1) + \delta \alpha^\top \alpha_1.$$

Similarly as above it holds that

$$\frac{\partial f}{\partial \alpha} = \Sigma\alpha + \lambda\alpha + \delta\alpha_1 = 0.$$

The constraint $\alpha_1^\top \Sigma\alpha = 0$ and $\alpha_1^\top \alpha = 0$ (see above) imply

$$\alpha_1^\top \frac{\partial f}{\partial \alpha} = \delta \underbrace{\alpha_1^\top \alpha_1}_{1} = \delta = 0,$$

which means, that $\Sigma\alpha = \lambda\alpha$ and α is again, an eigenvector of Σ with eigenvalue λ. Since α is supposed to be orthogonal to α_1 and $\mathrm{Var}(\alpha^\top X) = \lambda$ is supposed to be maximized, it holds that

$$\alpha = \alpha_2 \text{ and } \lambda = \lambda_2 \Rightarrow Y_2 = \alpha_2^\top X.$$

□

Exercise 6.2.3. Work out the proof for $i > 2$!

Let now $A = (\alpha_1, \ldots, \alpha_n)$. This is a orthogonal $(n \times n)$ matrix, for which it holds that (by Theorem 6.2.2)

$$\Sigma A = A\Lambda, \quad \Lambda = \mathrm{diag}\ (\lambda_1, \ldots, \lambda_n),$$

or equivalently

$$A^\top \Sigma A = \Lambda, \quad \Sigma = A\Lambda A^\top. \tag{6.2}$$

Theorem 6.2.4. For a $(n \times m)$ matrix B, with orthogonal columns b_i, $i = 1, \ldots, m$, $m \leq n$, let $Y = B^\top X$ and $\Sigma_Y = \mathrm{Cov}\ (Y) = B^\top \Sigma B$ be the covariance matrix of Y. Then

$$A_m = \underset{B}{\mathrm{argmax}}\, \mathrm{trace}(\Sigma_Y),$$

where $A_m = (\alpha_1, \ldots, \alpha_m)$.

Proof Since $\alpha_1, \ldots, \alpha_n$ is a basis of $\mathbb{R}^n$, it holds that

$$b_k = \sum_{i=1}^{n} c_{ik}\alpha_i, \quad k = 1, \ldots, m,$$

where $B = (b_1, \ldots, b_m)$, or matrix wise, $B = AC$, with $C = (c_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, m$. Thus it holds that

$$\Sigma_Y = B^\top \Sigma B = C^\top \underbrace{A^\top \Sigma A}_{\Lambda} C = C^\top \Lambda C = \sum_{j=1}^{n} \lambda_j c_j c_j^\top,$$

where $c_j^\top$ is the j-th row of C. Thus it holds that

$$\mathrm{trace}(\Sigma_Y) = \sum_{j=1}^{n} \lambda_j \mathrm{trace}(c_j c_j^\top) = \sum_{j=1}^{n} \lambda_j \mathrm{trace}(c_j^\top c_j) = \sum_{j=1}^{n} \lambda_j |c_j|^2.$$

Since $C = A^{-1}B = A^\top B$, it holds that

$$C^\top C = B^\top \underbrace{AA^\top}_{I_n} B = \underbrace{B^\top B}_{I_m} = I_m,$$

where

$$I_k = \text{diag } (\underbrace{1, \ldots, 1}_{k}).$$

Thus

$$\sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij}^2 = m,$$

and the columns of C are orthonormal. Thus C can be seen as a part (the first m columns) of an orthonormal $(n \times n)$ matrix D. Since the rows of D are also orthonormal vectors and $c_i^\top$ are the first m entries of the rows of D, it holds that

$$c_i^\top c_i = \sum_{j=1}^{m} c_{ij}^2 \leq 1, \quad i = 1, \ldots, n.$$

Since

$$\text{trace}(\Sigma_Y) = \sum_{i=1}^{n} \lambda_i \underbrace{\sum_{j=1}^{m} c_{ij}^2}_{\beta_i} = \sum_{i=1}^{n} \beta_i \lambda_i,$$

where $\beta_i \leq 1$, $i = 1, \ldots, n$, $\sum_{i=1}^{n} \beta_i = m$ and

$$\lambda_1 > \lambda_2 > \ldots > \lambda_n, \quad \sum_{i=1}^{n} \beta_i \lambda_i \to \max$$

for $\beta_1 = \ldots = \beta_m = 1$, $\beta_{m+1} = \ldots = \beta_n = 0$. If $B = A_m$, then

$$c_{ij} = \begin{cases} 1 & , 1 \leq i = j \leq m \\ 0 & , \text{else} \end{cases},$$

which implies $\beta_1 = \ldots = \beta_m = 1$, $\beta_{m+1} = \ldots = \beta_n = 0$. Thus A_m is the solution of $\text{trace}(\Sigma_Y) \to \max_B$. □

The assertion of Theorem 6.2.4 implies that

$$\text{Var} \left(\sum_{i=1}^{m} Y_i \right) = \text{Var} \left(\sum_{i=1}^{m} \alpha_i^\top X \right)$$

is maximized for all $m = 1, \ldots, n$, if Y_i are principal components of X.

Corollary 6.2.5. *Spectral representation of* Σ.
It holds that

$$\Sigma = \sum_{i=1}^{n} \lambda_i \cdot \alpha_i \cdot \alpha_i^\top. \tag{6.3}$$

Proof The representation is obtained by using (6.2), since

$$\Sigma = (\alpha_1, \ldots, \alpha_n) \cdot \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \cdot (\alpha_1, \ldots, \alpha_n)^\top.$$

□

Remark 6.2.6.

1. Since $\lambda_1 > \lambda_2 > \ldots > \lambda_n$ with $|\alpha_i| = 1$, $\forall i$, the representation (6.3) implies, that the first principal components do not only explain the biggest ratio of the variance of X_i, but also to the covariance. This value decreases with an increasing $i = 1, \ldots, n$.

2. If rank $(\Sigma) = r < n$, then (6.3) implies, that Σ can be completely determined by considering the first r principal components and coefficient vectors.

Lemma 6.2.7. Let Σ be a positive definite and symmetric $(n \times n)$ matrix with eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_n > 0$ and corresponding eigenvectors $\alpha_1, \ldots, \alpha_n$, $|\alpha_i| = 1$, $i = 1, \ldots, n$. Then

$$\lambda_k = \sup_{\alpha \in S_k, \alpha \neq 0} \frac{\alpha^\top \Sigma \alpha}{|\alpha|^2},$$

where $S_k = \langle \alpha_1, \ldots, \alpha_{k-1} \rangle^\perp$ for arbitrary $k = 1, \ldots, n$.[1]

Proof Let

$$c = \sup_{\alpha \in S_k} \frac{\alpha^\top \Sigma \alpha}{|\alpha|^2}.$$

Show that $\lambda_k \leq c \leq \lambda_k$.

1. $c \geq \lambda_k$: For $\alpha = \alpha_k$ prove that

$$c \geq \frac{\alpha_k^\top \Sigma \alpha_k}{\alpha_k^\top \alpha_k} = \frac{\lambda_k \alpha_k^\top \alpha_k}{\alpha_k^\top \alpha_k} = \lambda_k.$$

2. $c \leq \lambda_k$: It needs to be shown that

$$\alpha^\top \Sigma \alpha \leq \lambda_k |\alpha|^2, \quad \forall \alpha \in S_k, \quad \alpha \neq 0, \quad \forall \alpha \in \mathbb{R}^n \quad \alpha = \sum_{i=1}^{n} c_i \alpha_i,$$

[1] $\langle \ldots \rangle$ denotes the span

since $\{\alpha_i\}_{i=1}^n$ form an orthonormal basis.

$$\alpha \in S_k \quad \Rightarrow \quad c_1 = \ldots = c_{k-1} = 0.$$

That means

$$\alpha = \sum_{i=k}^{n} c_i \alpha_i, \quad \Sigma\alpha = \sum_{i=1}^{n} c_i \Sigma\alpha_i = \sum_{i=1}^{n} c_i \lambda_i \alpha_i,$$

$$\alpha^\top \Sigma\alpha = \left(\sum_{i=1}^{n} c_i \alpha_i\right)^\top \left(\sum_{i=1}^{n} \lambda_i c_i \alpha_i\right)$$

$$= \sum_{i,j=1}^{n} c_i c_j \lambda_i \underbrace{\alpha_j^\top \alpha_i}_{\delta_{ij}} = \sum_{i=1}^{n} c_i^2 \lambda_i, \quad |\alpha|^2 = \sum_{i=1}^{n} c_i^2.$$

Thus it holds that $\alpha \in S_k$

$$\alpha^\top \Sigma\alpha = \sum_{i=k}^{n} c_i^2 \lambda_i \leq \sum_{i=k}^{n} \lambda_k c_i^2 = \lambda_k \sum_{i=k}^{n} c_i^2 = \lambda_k |\alpha|^2,$$

and $c \leq \lambda_k$ since $\lambda_k > \lambda_j$, $j > k$.

□

Theorem 6.2.8. Let B, Y and Σ_Y such as in Theorem 6.2.4. Then

$$A_m = \operatorname*{argmax}_B \det(\Sigma_Y),$$

where $A_m = (\alpha_1, \ldots, \alpha_m)$.

Proof Let $k \in \{1, \ldots, m\}$ be fix. Introduce $S_k = \langle \alpha_1, \ldots, \alpha_{k-1} \rangle^\perp \subset \mathbb{R}^k$ (as in Lemma 6.2.7). Let $\mu_1 > \mu_2 > \ldots > \mu_m$ be the eigenvalues of $\Sigma_Y = B^\top \Sigma B$ with corresponding eigenvectors $\gamma_1, \ldots, \gamma_m$, which are orthonormal. Let $T_k = \langle \gamma_{k+1}, \ldots, \gamma_m \rangle \subset \mathbb{R}^m$. It obviously holds, that

$$\text{Dim } (S_k) = n - k + 1, \quad \text{Dim } T_k = k.$$

As in Lemma 6.2.7, it can be shown, that $\forall \gamma \neq 0$, $\gamma \in T_k$ it holds that

$$\frac{\gamma^\top \Sigma\gamma}{|\gamma|^2} \geq \mu_k.$$

Consider $\widetilde{S}_k = B(T_k) \subset \mathbb{R}^n$. Since B is an orthonormal transformation, it is thus unique Dim $(\widetilde{S}_k) =$ Dim $(T_k) = k$. The formula

$$\text{Dim } (S_k \cup \widetilde{S}_k) + \text{Dim } (S_k \cap \widetilde{S}_k) = \text{Dim } S_k + \text{Dim } \widetilde{S}_k$$

implies

$$\text{Dim } (S_k \cap \widetilde{S}_k) = \underbrace{\text{Dim } S_k}_{n-k+1} + \underbrace{\text{Dim } \widetilde{S}_k}_{k} - \underbrace{\text{Dim } (S_k \cup \widetilde{S}_k)}_{\leq n} \geq n-k+1+k-n = 1$$

that means, $\exists \alpha \in S_k \cap \widetilde{S}_k$, $\alpha \neq 0$. For this α it holds that $\alpha = B\gamma$, $\gamma \in T_k$ and thus

$$\mu_k \leq \frac{\gamma^\top \Sigma \gamma^2}{|\gamma|^2} = \frac{\gamma^\top B^\top \Sigma B \gamma}{\underbrace{\gamma^\top \gamma}_{\gamma^\top B^\top B \gamma}} = \frac{\alpha^\top \Sigma \alpha}{\alpha^\top \alpha} \leq \lambda_k$$

since $|\gamma| = |B\gamma|$, because B is preserving distances. That's why $\mu_k \leq \lambda_k$ for all $k = 1, \ldots, m$ and

$$\det(\Sigma_Y) = \prod_{i=1}^{m} \mu_k \leq \prod_{k=1}^{m} \lambda_k \quad \Rightarrow \quad \max_B \det(\Sigma_Y) \leq \prod_{k=1}^{m} \lambda_k.$$

However, since $B = A_m$, $\mu_k = \lambda_k$, $k = 1, \ldots, m$, it holds that

$$A_m = \operatorname*{argmax}_B \det(\Sigma_Y).$$

□

Now geometric properties of principal components are considered.

Proposition 6.2.9. The principal component coefficients $\alpha_1, \ldots, \alpha_n$ are the principle axis of the ellipsoids $x^\top \Sigma^{-1} x = c$, with semi-axis length $\sqrt{c\lambda_i}$, $i = 1, \ldots, n$.

Proof The principal components of X are given by $Z = A^\top X$, where $A = (\alpha_1, \ldots, \alpha_n)$ is an orthonormal transformation and thus $A^\top = A^{-1}$, $X = AZ$. Therefore for the ellipsoid it holds that

$$x^\top \Sigma^{-1} x \underbrace{=}_{Subst. x = Az} z^\top A^\top \Sigma^{-1} A z = z^\top \Lambda^{-1} z = c,$$

where

$$A^\top \Sigma^{-1} A = \Lambda^{-1} = \text{diag } \left(\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_n} \right), \quad \Lambda = \text{diag } (\lambda_1, \ldots, \lambda_n),$$

since Σ^{-1} has the same eigenvectors with eigenvalues $\frac{1}{\lambda_i}$. That's why the ellipsoid $z^\top \Lambda^{-1} z = c$ can be represented in its normed form as

$$\sum_{k=1}^{n} \frac{z_k^2}{c\lambda_k} = 1.$$

That implies that the α_i point towards the principal axis and that the half-axis are given by $\sqrt{c\lambda_i}$. □

Remark 6.2.10. (*Multivariate normal distribution*). If $X \sim N(0, \Sigma)$ then $x^\top \Sigma^{-1} x = c$ is an ellipsoid of constant probability for X, since the probability density function of X

$$f_X(x) = \frac{1}{\sqrt{\det \Sigma}} \exp\left\{-\frac{1}{2} x^\top \Sigma^{-1} x\right\} \cdot \frac{1}{(2\pi)^{\frac{n}{2}}}, \quad x \in \mathbb{R}^n,$$

is constant on this ellipsoid. Else $x^\top \Sigma^{-1} x = c$ defines contours of the constant probability for X. Here the vector α_1 points towards the largest variance of $\alpha^\top X$ (it is the biggest principal axis with length $\sqrt{c\lambda_1}$ of the ellipsoid); α_2 points towards the second largest variance (half-axis with length $\sqrt{c\lambda_2}$), and so on (cf. Condition (6.1)).

Remark 6.2.11. Another form of PCA is possible, if instead of $X = (X_1, \ldots, X_n)^\top$ the normed random sample $X_\omega = (X_1/\omega_1, \ldots, X_n/\omega_n)^\top$ is used, where the weights $\omega = (\omega_1, \ldots, \omega_n)^\top$ contain advance information which represent a certain preference in the analysis. A usual choice is

$$\omega_i = \sqrt{\sigma_{ii}} = \sqrt{\mathrm{Var} X_i},$$

which leads to a PCA of $X^* = (X_1^*, \ldots, X_n^*)$, $X_i^* = \frac{X_i}{\sqrt{\mathrm{Var} X_i}}$, $i = 1, \ldots, n$ by using the correlation matrix $\Sigma^* = (\mathrm{Corr}\ (X_j, X_i))_{i,j=1}$ with

$$\mathrm{Corr}\ (X_i, X_j) = \frac{\mathrm{Cov}\ (X_i, X_j)}{\sqrt{\mathrm{Var} X_i \mathrm{Var} X_j}} = \mathrm{Cov}\ (X_i^*, X_j^*), \quad i, j = 1, \ldots, n.$$

By doing so, other principal components $\alpha_i^{*T} X^*$ can be obtained for which $\alpha_i^* \neq \alpha_i$ holds for $i = 1, \ldots, n$.

What are the advantages and disadvantages of PCA based on (X, Σ) and (X^*, Σ^*)?

Disadvantages of (X, Σ)-PCA:

1. PCA based on (X^*, Σ^*) does not depend on the unit measurements of X. Thus comparisons between results of PCA for several samples of different origin are possible.

2. If the variances X_i are varying a lot, the variables X_i with the largest variance are determined to be the first principal component. The PCA based on (X^*, Σ^*) does not have this disadvantage. The (X, Σ)-PCA is not significant in such a case, because it sorts the variables X_i (in a slightly different form) in a variance wise descending order.

   Example 6.2.12. Let $X = (X_1, X_2)$, where X_1 is the length and X_2 the weight. X_1 can be measured in *cm* or *m*, X_2 only in *kg*. In those two cases, the covariance matrices X are given by

$$\Sigma_1 = \begin{pmatrix} 80 & 44 \\ 44 & 80 \end{pmatrix} \quad \text{resp.} \quad \Sigma_2 = \begin{pmatrix} 8000 & 4400 \\ 4400 & 8800 \end{pmatrix}.$$

Calculating the first principal components in both cases yields

$$\alpha_1^\top X = 0,707X_1 + 0,707X_2 \text{ for } \Sigma_1 \text{ resp.}$$
$$\alpha_1^\top X = 0,998X_1 + 0,055X_2 \text{ for } \Sigma_2.$$

Note that, in the first case, both X_1 and X_2 have the same absolute value with respect to the 1. principal component, whereas in the 2. case X_1 is a dominating factor. Moreover it holds that $\frac{\lambda_1}{\lambda_1+\lambda_2} \cdot 100\% = 77,5\%$ in the first case and $\frac{\lambda_1}{\lambda_1+\lambda_2} \cdot 100\% = 99,3\%$ in the 2. case (it is the ratio of the variation of the first principal component to the whole variation).

3. If random variables X_i in X have a differing origin (as in the example above), then the interpretation of the ratio of the variation is rather problematic, since the sum $\lambda_1 + \ldots + \lambda_n$ contains m^2, kg^2 and so on. The PCA based on (X^*, Σ^*) only considers values without unit, such that the sum $\lambda_1 + \ldots + \lambda_n$ can be interpreted.

Advantages of (X, Σ)-PCA:

1. If instead of Σ resp. Σ^* the empirical analogues $\hat{\Sigma}$ resp. $\hat{\Sigma}^*$ are used (if $\Sigma(\Sigma^*)$ are not known, they have to be estimated by using the available data), then $(X, \hat{\Sigma})$-PCA has some advantages, since the statistical methods are easier to use in this case compared to using them in $(X^*, \hat{\Sigma}^*)$-PCA.

2. If all X_i in X have the same unit, then the (X, Σ)-PCA is sometimes preferable, since during the standardisation of (X, Σ) to (X^*, Σ^*) the relation to the units of X are lost.

Remark 6.2.13. Sometimes instead of $|\alpha| = 1$ the standardization $|\alpha_k| = \sqrt{\lambda_k}$, $k = 1, \ldots, n$ in Definition 6.2.1 is used (cf. optimisation problem (6.1)). This is in particular the case in the correlation based PCA.

Remark 6.2.14. (*Equal eigenvalues* λ_i). If some eigenvalues of Σ are equal, e.g. $\lambda_1 = \lambda_2 = \ldots = \lambda_k > \lambda_{k+1} > \ldots > \lambda_m$ implies that there exists a linear subspace of dimension k, in which an arbitrary basis represents the first k eigenvectors. This means, with respect to PCA that the first k eigenvectors can not be defined uniquely. Geometrical interpretation: The first k half-axis of $x^\top \Sigma^{-1} x = c$ are equal, i.e., the ellipsoid $x^\top \Sigma^{-1} x = c$ has a spherical k-dimensional cross section through the origin, where the directions of the half-axis can be chosen (orthogonal to each other) arbitrarily.

Remark 6.2.15 ($\lambda_i = 0$)**.** If $\lambda_1 > \ldots > \lambda_{n-k} > \lambda_{n-k+1} = \ldots = \lambda_n = 0$, then there are only $n-k$ linear independent random vectors X_i in the random sample X. That's why only those $n-k$ variables should be good for the analysis.

6.3 PCA on data level

In this section it is not assumed that the covariance matrix Σ is known. That's why it is replaced by empirical covariance matrix $\hat{\Sigma}$.
Let $X^1, X^2, \ldots, X^m$ be independent realizations of a n-dimensional random vector $X = (X_1, \ldots, X_n)^\top$, $X^i = (X_1^i, \ldots, X_n^i)^\top$, $i = 1, \ldots, m$. X^i is interpreted as a sample of X.

Definition 6.3.1. Define the n-dimensional random vector a_k by

$$a_k = \underset{a \in \mathbb{R}^n}{\operatorname{argmax}} \frac{1}{m-1} \sum_{i=1}^{m} (Y_i - \overline{Y})^2$$

with constraint $|a| = 1$, a uncorrelated to $a_1, \ldots, a_{k-1}$ for all $k = 1, \ldots, n$, where

$$Y_i = a^\top X^i, \quad i = 1, \ldots, m, \quad \overline{Y} = \frac{1}{m} \sum_{i=1}^{m} Y_i.$$

Thus $a_k^\top X$ defines the k-th principal component of X with coefficient vector a_k. $Y_{ik} = a_k^\top X^i$ is the evaluation of the k-ten principal component of the i-th observation X^i of X_i, $i = 1, \ldots, m$, $k = 1, \ldots, n$.

Lemma 6.3.2. It holds that

$$\frac{1}{m-1} \sum_{i=1}^{m} (Y_{ik} - \overline{Y}_k)^2 = l_k, \quad k = 1, \ldots, n,$$

where

$$\overline{Y}_k = \frac{1}{m} \sum_{i=1}^{m} Y_{ik}, \ \overline{X}_k = \frac{1}{m} \sum_{i=1}^{m} X_k^i, \quad k = 1, \ldots, n$$

and l_k is the eigenvalue of the empirical covariance matrix $\hat{\Sigma} = (\hat{\sigma}_{ij})_{i,j=1}^n$,

$$\hat{\sigma}_{ij} = \frac{1}{m-1} \sum_{t=1}^{m} (X_i^t - \overline{X}_i)(X_j^t - \overline{X}_j), \quad i, j = 1, \ldots, n, \quad l_1 > l_2 > \ldots > l_n.$$

a_k is the eigenvector of $\hat{\Sigma}$ with eigenvalue l_k, $k = 1, \ldots, n$.

Proof

Exercise 6.3.3. cf. proof of Theorem 6.2.2.

□

In the following, replace X^i with $X^i - \overline{X}$ but keep the notation X^i, $i = 1, \ldots, n$.

Remark 6.3.4. The properties of PCA as formulated in Theorem 6.2.4, Corollary 6.2.5 and Proposition 6.2.9 are preserved in the statistical version (cf. Definition 6.3.1) by using the following modification: Σ is replaces by $\hat{\Sigma}$, $A = (\alpha_1, \ldots, \alpha_n)$ by $A = (a_1, \ldots, a_n)$, $A_m = (\alpha_1, \ldots, \alpha_m)$ by $A_m = (a_1, \ldots, a_m)$ and Σ_Y by the empirical covariance matrix $\hat{\Sigma}_Y$ of Y. Thus use the spectral representation of $\hat{\Sigma}$:

$$\hat{\Sigma} = \sum_{i=1}^{n} l_i a_i a_i^\top. \tag{6.4}$$

Exercise 6.3.5. Prove that!

In the following another property of the empirical PCA, which can also be seen as an equivalent definition is presented:

Theorem 6.3.6. Let B be a $(n \times p)$ matrix, $p \leq n$, with orthogonal columns. Let $Z_i = B^\top X^i$, $i = 1, \ldots, m$ be a projection of X^i, $i = 1, \ldots, m$, on a p-dimensional subspace L_B. Define

$$G(B) = \sum_{i=1}^{m} \left| X^i - Z_i \right|^2.$$

Then

$$A_p = (a_1, \ldots, a_p) = \operatorname*{argmin}_B G(B).$$

Proof By the Pythagorean Theorem it holds that $\left|X^i\right|^2 = |Z_i|^2 + |X^i - Z_i|^2$, that's why

$$G(B) = \sum_{i=1}^{m} \left| X^i \right|^2 - \sum_{i=1}^{m} |Z_i|^2 \to \min$$

if

$$\widetilde{G}(B) = \sum_{i=1}^{m} |Z_i|^2 = \sum_{i=1}^{m} Z_i^\top Z_i = \sum_{i=1}^{m} X^{iT} B B^\top X^i \to \max_B.$$

It holds that

$$\begin{aligned}
\widetilde{G}(B) &= \operatorname{trace}\left(\sum_{i=1}^{m} \left(X^{iT} B B^\top X^i \right) \right) = \sum_{i=1}^{m} \operatorname{trace}\left(X^{iT} B B^\top X^i \right) \\
&= \sum_{i=1}^{m} \operatorname{trace}\left(B^\top X^i X^{iT} B \right) = \operatorname{trace}\Bigg(B^\top \underbrace{\left(\sum_{i=1}^{m} X^i X^{iT} \right)}_{^2 (m-1)\hat{\Sigma}} B \Bigg) \\
&= (m-1)\operatorname{trace}(B^\top \hat{\Sigma} B).
\end{aligned}$$

In summary:

$$\widetilde{G}(B) = (m-1)\text{trace}\left(B^\top \hat{\Sigma} B\right),$$

since it is maximized by Remark 6.3.4 and Theorem 6.2.4, if $B = A_p$. □

Remark 6.3.7. How can Theorem 6.3.6 be used as an equivalent definition of the empirical PCA? a_i are defined as orthogonal vectors, which is the span of a linear subspace $L_p = \langle a_1, \ldots, a_p \rangle$, $p = 1, \ldots, n-1$, with the property, that the sum of the quadratic orthogonal distances of X^i to L_p are minimized. Thus for $p = 1$, L_1 would be the best line for approximating the data set $X^1, \ldots, X^m$. For $p = n-1$, L_{n-1} would be the best hyperplane with the same property (cf. linear regression).

The following theorem provides a new interpretation for PCA as well as more efficient method for the calculation.

Theorem 6.3.8. (Singular value decomposition.)
Let $\widetilde{X} = \left(X^1 - \overline{X}, X^2 - \overline{X}, \ldots, X^m - \overline{X}\right)^\top$ be a $(m \times n)$ matrix, with centered observations X^i of X. Let rank $(\widetilde{X}) = r \leq n, m$. The following decomposition holds:

$$\widetilde{X} = ULA_r^\top, \tag{6.5}$$

where U is a $(m \times r)$ matrix with orthonormal columns.

$$L = \text{diag}\,(\widetilde{l}_1, \ldots, \widetilde{l}_r) \quad \text{where} \quad \widetilde{l}_i = \sqrt{(m-1)l_i}$$

is the square root of the i-th (non trivial) eigenvalue of $\widetilde{X}^\top \widetilde{X} = (m-1)\hat{\Sigma}$, $i = 1, \ldots, r$. $A_r = (a_1, \ldots, a_r)$ is the $(n \times r)$ matrix with columns a_i.

Proof Define $U = (u_1, \ldots, u_r)$ with columns $u_i = \widetilde{X} a_i / \widetilde{l}_i$, $i = 1, \ldots, r$. In the following it is shown, that the representation (6.5) holds. Using the spectral representation 6.4 it holds that

$$(m-1)\hat{\Sigma} = \widetilde{X}^\top \widetilde{X} = \sum_{i=1}^{r} \widetilde{l}_i^2 a_i a_i^\top, \quad \text{since} \quad l_i = 0, i = r+1, \ldots, n.$$

Thus

$$ULA_r^\top = U \begin{pmatrix} \widetilde{l}_1 a_1^\top \\ \vdots \\ \widetilde{l}_r a_r^\top \end{pmatrix} = \sum_{i=1}^{r} \widetilde{X} \frac{a_i}{\widetilde{l}_i} \widetilde{l}_i a_i^\top = \sum_{i=1}^{r} \widetilde{X} a_i a_i^\top \overset{l_i=0, i>r}{=} \sum_{i=1}^{n} \widetilde{X} a_i a_i^\top.$$

[2] Since X^i was replaced by $X^i - \overline{X}$.

It holds that $\widetilde{X}a_i = 0$, $i = r+1, \ldots, n$, since $\operatorname{rank}(\widetilde{X}) = r$ and because of the centered columns of $\widetilde{X}$ by $\overline{X}$. By the orthogonality of the a_i it holds that

$$ULA_r^\top = \widetilde{X}\sum_{i=1}^{n} a_i a_i^\top = \widetilde{X}I = \widetilde{X}.$$

□

Remark 6.3.9. The matrix U provides the following versions of evaluations

$$Y_{ik} = a_k^\top X^i = X^{iT} a_k, \quad Y_{ik} = u_{ik}\widetilde{l_k}, \quad i = 1, \ldots, m,\ k = 1, \ldots, n.$$

It holds that

$$\operatorname{Var}(u_{ik}) = \frac{\operatorname{Var}(Y_{ik})}{\widetilde{l}_k^2} = \frac{l_k}{(m-1)l_k} = \frac{1}{m-1}, \quad \forall i, k.$$

6.4 Asymptotic distributions of principal components for normal distributed random samples

Let now $X \sim N(\mu, \Sigma)$, Σ have the eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_n > 0$ and corresponding eigenvectors α_k, $k = 1, \ldots, n$. Calculate

$$\begin{aligned}
&\lambda = (\lambda_1, \ldots, \lambda_n)^\top, \quad l = (l_1, \ldots, l_n)^\top, \\
&\alpha_k = (\alpha_{k1}, \ldots, \alpha_{kn})^\top, \quad a_k = (a_{k1}, \ldots, a_{kn})^\top, \\
&k = 1, \ldots, n
\end{aligned}$$

Theorem 6.4.1.

1. l is asymptotically (for $m \to \infty$) independent of a_k, $k = 1, \ldots, n$.
2. l and a_k, $k = 1, \ldots, n$ are asymptotic $m \to \infty$ multivariate normal distributed, with asymptotic expectation

$$\lim_{m\to\infty} \mathbb{E}(l) = \lambda \quad \text{and} \quad \lim_{m\to\infty} \mathbb{E}(a_k) = \alpha_k, \quad k = 1, \ldots, n.$$

3. It holds that

$$\operatorname{Cov}(l_k, l_{k'}) \sim \begin{cases} \frac{2\lambda_k^2}{m-1}, & k = k' \\ 0, & k \neq k' \end{cases} \quad \text{for } m \to \infty,$$

$$\operatorname{Cov}(a_{kj}, a_{k'j'}) \sim \begin{cases} \frac{\lambda_k}{m-1} \sum_{l=1, l\neq k}^{n} \frac{\lambda_l \alpha_{lj} \alpha_{lj'}}{(\lambda_l - l_k)^2}, & k = k' \\ -\frac{\lambda_k \lambda_{k'} \alpha_{kj} \alpha_{k'j'}}{(m-1)(\lambda_k - \lambda_{k'})^2}, & k \neq k' \end{cases} \quad \text{for } m \to \infty.$$

Without proof!
The assertion of Theorem 6.4.1 can be used for constructing MLE and confidence intervals for λ and α_k.

Exercise 6.4.2.

1. Show that an MLE of Σ is given by $\frac{m-1}{m}\hat{\Sigma}$.
2. Show, that the MLE

$$\begin{cases} \text{for } \lambda \text{ is given by} & \hat{\lambda} = \frac{m-1}{m} l. \\ \text{for } \alpha_k \text{ is given by} & \hat{\alpha}_k = a_k, k = 1, \ldots, n. \end{cases}$$

3. Show that the MLE in 2. coincide with the moment estimators λ and α_k, which can be obtained from Theorem 6.4.1.

Corollary 6.4.3 (Confidence intervals for λ_k)**.** An asymptotic confidence interval for λ_k $(m \to \infty)$ with confidence level $1-\alpha$ is given by

$$\left[l_k \left(1 - \sqrt{\frac{2}{m-1}} z_{\frac{\alpha}{2}} \right)^{-1}, l_k \left(1 + \sqrt{\frac{2}{m-1}} z_{\frac{\alpha}{2}} \right)^{-1} \right],$$

where m is large enough such that $-\sqrt{\frac{2}{m-1}} z_{\frac{\alpha}{2}} < 1$.

Proof Since $l_k \sim N\left(\lambda_k, \frac{2\lambda_k^2}{m-1}\right)$ for $m \to \infty$ by Theorem 6.4.1, 2. and 3., it holds that

$$\frac{l_k - \lambda_k}{\sqrt{\frac{2}{m-1}}\lambda_k} \sim N(0,1) \quad \text{for} \quad m \to \infty.$$

This implies, that

$$\lim_{m\to\infty} P\left(z_{\frac{\alpha}{2}} \leq \frac{l_k - \lambda_k}{\lambda_k} \sqrt{\frac{m-1}{2}} \leq z_{1-\frac{\alpha}{2}} \right) = 1 - \alpha,$$

or for $m \to \infty$

$$\sqrt{\frac{2}{m-1}} z_{\frac{\alpha}{2}} \leq \frac{l_k}{\lambda_k} - 1 \leq \sqrt{\frac{2}{m-1}} \underbrace{z_{1-\frac{\alpha}{2}}}_{=-z_{\frac{\alpha}{2}}},$$

$$\frac{l_k}{1 - \sqrt{\frac{2}{m-1}} z_{\frac{\alpha}{2}}} \leq \lambda_k \leq \frac{l_k}{1 + \sqrt{\frac{2}{m-1}} z_{\frac{\alpha}{2}}}$$

with probability $1-\alpha$. □

Since all l_k, $k = 1, \ldots, n$ are asymptotically $(m \to \infty)$ independent, a simultaneous confidence interval for l can be denoted by a cartesian product of the confidence intervals for l_k as in Corollary 6.4.3.

Lemma 6.4.4. It holds that

$$(m-1)\alpha_k^\top \left(l_k \hat{\Sigma}^{-1} + l_k^{-1} \hat{\Sigma} - 2I_n \right) \alpha_k \xrightarrow[m\to\infty]{d} \chi^2_{n-1}.$$

Without proof!
As a consequence of the lemma above, the (asymptotic) confidence ellipsoid for α_k with confidence level $1 - \beta$

$$\left\{ y \in \mathbb{R}^n : (m-1) y^\top \left(l_k \hat{\Sigma}^{-1} + l_k^{-1} \hat{\Sigma} - 2I_n \right) y \leq \chi^2_{n-1,\beta} \right\}$$

is obtained.

Remark 6.4.5. Corollary 6.4.3 resp. Lemma 6.4.4 can be used to construct statistical tests forλ_k resp. α_k as follows:

1. Test $H_0 : \lambda_k = \lambda_{k_o}$ v.s. $H_1 : \lambda_k \neq \lambda_{k_0}$
   H_0 can be rejected, if

   $$\left| \frac{l_k - \lambda_{k_0}}{\sqrt{\frac{2}{m-1}} \lambda_{k_0}} > z_{\frac{\alpha}{2}} \right|.$$

   This is an asymptotic test $(m \to \infty)$ with confidence level α.

2. Test $H_0 : \alpha_k = \alpha_{k_0}$ v.s. $H_1 : \alpha_k \neq \alpha_{k_0}$
   H_0 can be rejected, if

   $$(m-1)\alpha_{k_0}^\top \left(l_k \hat{\Sigma}^{-1} + l_k^{-1} \hat{\Sigma} - 2I_n \right) \alpha_{k_0} \geq \chi^2_{n-1,\alpha}.$$

   This is an asymptotic test $(m \to \infty)$ with confidence level α.

6.5 Outlier detection

In this section it is assumed that the random sample $X^1, X^2, \ldots, X^m$ can contain some outliers. How can an outlier be defined? In statistical literature there is no coherent definition. Generally speaking, an observation X^i is an outlier if it attains an unusual value (with respect to the distribution of X). For example, an unusual value of some coordinates X_i could be significantly bigger or smaller than the others. An outlier could also occur in form of an unusual combination of the coordinate values of some coordinates X^i. A reason for those anomalies could lie in the data, or simply occur because of measurement errors.

Example 6.5.1. Let $X = (X_1, X_2)$, where X_1 = "height" (in cm) and X_2 = "weight" (in kg) of children between the age of 5 and 15. The feature X is obtained n times in a medical survey. Here the features $X^i = (250, 80)$ and $X^j = (175, 25)$ are considered as outliers, because $X^i = 250$cm is an abnormal height and for X^j, $X_1^j = 175$ and $X_2^j = 25$ as a combination are highly unlikely.

How can outliers be detected? One way to identify outliers of X^i is to plot the dataset $X^1, \ldots, X^m$ and spot values which are outside of a larger agglomeration of values. If the dimension n of X is high, it is rather difficult to visualize the data. It can thus be helpful to generate a data point of the first 2-3 principal components of $(X^1, \ldots, X^m)$. By looking at them, outliers of X_k^i can also be identified quickly. In order to detect unusual relationships between coordinate values X_k^i, the last few principal components should be considered. Let $a_1, \ldots, a_n$ be the coefficient vectors of the principal components of $(X^1, \ldots, X^m)$, $Y_{ik} = a_k^\top X^i$, $i = 1, \ldots, m$, $k = 1, \ldots, n$ be realizations of the principal components of the observation X^i and l_k, $k = 1, \ldots, n$ be the eigenvalues of the empirical covariance matrix $\hat{\Sigma}$ of $(X^1, \ldots, X^m)$. For $1 \leq n_0 \leq n$, define the statistic

$$d_i^{(1)}(n_0) = \sum_{k=n-n_0+1}^{n} Y_{ik}^2, \quad d_i^{(2)}(n_0) = \sum_{k=n-n_0+1}^{n} \frac{Y_{ik}^2}{l_k},$$
$$d_i^{(3)}(n_0) = \sum_{k=n-n_0+1}^{n} l_k Y_{ik}^2, \quad d_i^{(4)}(n_0) = \max_{n-n_0+1 \leq k \leq n} \frac{|Y_{ik}|}{\sqrt{l_k}},$$

for $i = 1, \ldots, m$.

Lemma 6.5.2. It holds that

$$d_j^{(2)}(n) = \left(X^i - \overline{X}\right)^\top \hat{\Sigma}^{-1} \left(X^i - \overline{X}\right), \quad i = 1, \ldots, m,$$

where Y_{ik} are centered, i.e. Y_{ik} is replaced by $Y_{ik} - \overline{Y_k}$, $k = 1, \ldots, n$, $i = 1, \ldots, m$.

Proof It holds that

$$\hat{\Sigma} = ALA^\top, \quad \text{where} \quad L = \text{diag}\,(l_1, \ldots, l_n) \quad \text{and} \quad A = (a_1, \ldots, a_n).$$

Thus

$$\hat{\Sigma}^{-1} = AL^{-1}A^\top \quad \text{with} \quad L^{-1} = \text{diag}\,(l_1^{-1}, \ldots, l_n^{-1}).$$

Since additionally $Y_i = A^\top X^i$ for $Y_i = (Y_{i1}, \ldots, Y_{in})^\top$, $i = 1, \ldots, n$, it holds that

$$X^i = A^{\top -1} Y_i = AY_i, \quad X^{i\top} = Y_i^\top A^\top, \quad i = 1, \ldots, n$$

and thus

$$\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X^i = A\overline{Y}, \quad \overline{Y} = \frac{1}{m}\sum_{i=1}^{m} Y^i, \quad \overline{X}^\top = \overline{Y}^\top A^\top.$$

This implies that

$$\begin{aligned}\left(X^i - \overline{X}\right)^\top \hat{\Sigma}^{-1} \left(X^i - \overline{X}\right) &= \left(Y_i - \overline{Y}\right)^\top \underbrace{A^\top A}_{I} L^{-1} \underbrace{A^\top A}_{I} \left(Y_i - \overline{Y}\right) \\ &= \left(Y_i - \overline{Y}\right)^\top L^{-1} \left(Y_i - \overline{Y}\right) = \sum_{k=1}^{n} \frac{Y_{ik}^2}{l_k} = d_i^{(2)}(n).\end{aligned}$$

□

In order to identify outliers in $(X^1, \ldots, X^m)$, the values $d_i^{(j)}(n)$, $i = 1, \ldots, m$, $j = 1, \ldots, n$ for $n = 1, 2, 3$ are calculated. Observations X^i with the largest value $d_i^{(j)}(n)$ are classified as possible outliers. Additionally the plot of the point cloud, defined by

$$D = \left\{ \left(d_i^{(2)}(n) - d_i^{(2)}(n_0), d_i^{(2)}(n_0) \right), i = 1, \ldots, m \right\}$$

can be helpful. X^i is considered an outlier, if

$$\left(d_i^{(2)}(n) - d_i^{(2)}(n_0), d_i^{(2)}(n_o) \right)$$

is isolated from the remaining point cloud D.

Remark 6.5.3. If $X \sim N(\mu, \Sigma)$ with known μ and Σ and PCA is conducted on model level, the distributions of $d_i^{(j)}(n_0)$ can be explicitly stated. They are (except for $d_i^{(4)}$) gamma distributed with known parameters e.g. $d_i^{(2)}(n_0) \sim \chi^2_{n_0}$, $i = 1, \ldots, m$. The distribution function of $d_j^{(4)}(n_0)$ is given by $\Phi^{n_0}(x)$, where $\Phi(x)$ is the distribution function of a $N(0, 1)$ distribution. Confidence intervals for $d_i^{(j)}(n_0)$ can provide a decision rule, whether X^i is an outlier. Even though this approach is based on a strict mathematical basis, it is rather uncommon in practice, since normally distributed data (with known μ and Σ!) are relatively rare.

Remark 6.5.4. The statistics $d_i^{(2)}, d_i^{(4)}$ emphasize the last statistics more than $d_i^{(1)}$ (because of the corresponding standardization). That's why they are sufficient for the detection of unusual correlations in the data (cf. Example 6.5.1, observation $X^j = (175, 25)$. The statistic $d_j^{(3)}$ emphasizes the first principal component. Thus it can be used to detect unusual large (small) values of the coordinates X_k^i (cf. Example 6.5.1 X_1^i =250).

6.6 PCA and regression

Consider the multivariate regression model: $Y = X\beta + \varepsilon$, where $Y = (Y_1, \ldots, Y_n)^\top$ is the vector of goal variables,

$$X = (X_{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,m}}$$

the $(n \times m)$ matrix of output variables, rank $(X) = m$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ the vector of error terms, where ε_i are independent of $\mathbb{E}\varepsilon_i = 0$, $\mathrm{Var}\varepsilon_i = \sigma^2$, $i = 1, \ldots, n$. W.l.o.g. assume, that X (as in Theorem 6.3.8) is centered, i.e., the empirical mean of X is zero, or in detail, X_{ij} is replaced by $X_{ij} - \overline{X_j}$, where

$$\overline{X_j} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}, \quad j = 1, \ldots, m.$$

Assuming that some of the variables X_{ij} in X are almost linearly dependent, i.e. $\det(X^\top X) \approx 0$, causes the estimators $\hat{\beta}$ of β to be affected in form of an instability for the calculation, since $\mathrm{Cov}\,(\hat{\beta}) = \sigma^2(X^\top X)^{-1}$ (cf. Theorem 6.3.8) only contains little variance of $\hat{\beta}_j$. A solution to this problem is the usage of generalizations as in Section 6.3. Another application of PCA is the detection of linear dependencies in X by looking at the last principal components and eliminating variables β_j based on those. This application will be discussed in more detail below.
Let $a_1, \ldots, a_m$ be the coefficient vectors of the principal components (i.e. the eigenvectors) of $X^\top X$. Let $Z_{ik} = a_k^\top X^i$ be the realization of the k-th principal component of the i-th row X^i of X, $i = 1, \ldots, n$, $k = 1, \ldots, m$. With $Z = (Z_{ik})$ it holds that $Z = XA$, where $A = (a_1, \ldots, a_m)$ is an orthogonal $(m \times m)$ matrix. The regression equation $Y = X\beta + \mathcal{E}$ is given by:

$$Y = X\underbrace{AA^\top}_{I}\beta + \mathcal{E} = \underbrace{XA}_{Z}\underbrace{A^\top\beta}_{\gamma} + \mathcal{E} = Z\gamma + \mathcal{E}, \text{ where } \gamma = A^\top\beta. \tag{6.6}$$

By doing so, the old output variables β are replaced by the transformation $\gamma = A^\top\beta$. The estimation of γ is obtained with Theorem 4.2.1:

$$\hat{\gamma} = \left(Z^\top Z\right)^{-1} Z^\top Y = L^{-1}Z^\top Y, \tag{6.7}$$

where $L = \mathrm{diag}\,(l_1, \ldots, l_m)$ contains the eigenvalues l_i of $X^\top X$. This holds, since Z has orthogonal columns. Thus

$$\hat{\beta} = A\hat{\gamma} = AL^{-1}Z^\top Y = \underbrace{AL^{-1}A^\top}_{(X^\top X)^{-1}} X^\top Y = \sum_{k=1}^{m} l_k^{-1} a_k a_k^\top X^\top Y,$$

where in the last part of the equation the terms (6.6), (6.7) and the spectral representation (Corollary 6.2.5) of $(X^\top X)^{-1}$ have been used. Theorem 6.2.4 implies furthermore, that

$$\mathrm{Var}(\hat{\beta}) = \sigma^2 \sum_{k=1}^{m} l_k^{-1} a_k a_k^\top.$$

Thus the following assertion is proved:

Lemma 6.6.1. The solution of the OLS equation $Y = X\beta + \mathcal{E}$ is given by

$$\hat{\beta} = \sum_{k=1}^{m} l_k^{-1} a_k a_k^\top X^\top Y.$$

Here it holds that

$$\mathrm{Cov}\ (\hat{\beta}) = \sigma^2 \sum_{k=1}^{m} l_k^{-1} a_k a_k^\top.$$

Remark 6.6.2. What are the advantages of the in (6.6)-(6.7) introduced methodology?

1. After calculating the principal components of $X^\top X$, the calculation of $\hat{\gamma} = L^{-1} Z^\top Y$ is fast and easy, since (6.7) does not include any inverted matrices ($L^{-1} = \mathrm{diag}\ (l_1^{-1}, \ldots, l_m^{-1})$ is explicitly known).

2. If some of the l_k are close to zero or rank $(X) < m$, some of the last few principal components (with small or even zero variance) of $X^\top X$ can simply be excluded from the regression. This can be realized with the new estimator given by

$$\widetilde{\beta} = \sum_{k=1}^{p} l_k^{-1} a_k a_k^\top X^\top Y$$

$p < m$.

Lemma 6.6.3. Let rank $(X) = m$:

1. The estimator $\widetilde{\beta}$ is biased:

$$\mathbb{E}\widetilde{\beta} = \left(I - \sum_{k=p+1}^{m} a_k a_k^\top \right) \beta.$$

2. It holds that:

$$\mathrm{Cov}\ (\widetilde{\beta}) = \sigma^2 \sum_{k=1}^{p} l_k^{-1} a_k a_k^\top$$

.

Proof

1. Since
$$\widetilde{\beta} = \hat{\beta} - \sum_{k=p+1}^{m} l_k^{-1} a_k a_k^\top X^\top Y$$
and $\hat{\beta}$ is biased, it holds that
$$\begin{aligned}\mathbb{E}\widetilde{\beta} &= \mathbb{E}\hat{\beta} - \sum_{k=p+1}^{m} l_k^{-1} a_k a_k^\top X^\top \mathbb{E}Y \\ &= \beta - \sum_{k=p+1}^{m} l_k^{-1} a_k \underbrace{a_k^\top X^\top X}_{l_k a_k^\top} \beta = \beta - \sum_{k=p+1}^{m} a_k a_k^\top \beta \\ &= \left(I - \sum_{k==p+1}^{m} a_k a_k^\top \right) \beta\end{aligned}$$
2. **Exercise 6.6.4.**

□

Another equivalent formulation for regression with PCA is given in the following. Instead of using $\gamma = A^\top \beta$, use singular value decomposition (cf. Theorem 6.3.8) for X:
$$X = U L^{\frac{1}{2}} A^\top,$$
where U is a $(n \times m)$ matrix with orthonormal columns and L a diagonal matrix with $L^{\frac{1}{2}} = \text{diag}\,(\sqrt{l_1}, \ldots, \sqrt{l_m})$. Define
$$\delta = L^{\frac{1}{2}} A^\top \beta, \tag{6.8}$$
then
$$Y = X\beta + \mathcal{E} = U \underbrace{L^{\frac{1}{2}} A^\top \beta}_{\delta} + \mathcal{E} = U\delta + \mathcal{E}.$$
The MLE for δ is given by
$$\hat{\delta} = \underbrace{(U^\top U)^{-1}}_{I} U^\top Y = U^\top Y,$$
since U has orthonormal columns. (6.8) implies $\beta = AL^{-\frac{1}{2}}\delta$ and thus
$$\hat{\beta} = AL^{-\frac{1}{2}}\hat{\delta} = AL^{-\frac{1}{2}}U^\top Y.$$
Here the relationship between γ and δ is given by:
$$\gamma = A^\top \beta = A^\top \left(AL^{-\frac{1}{2}}\delta \right) = \underbrace{A^\top A}_{I} L^{-\frac{1}{2}}\delta = L^{-\frac{1}{2}}\delta.$$
Thus the following Lemma has been proven.

Lemma 6.6.5. The principal components $Y = U\delta + \mathcal{E}$ of the regression $Y = X\beta + \mathcal{E}$ has the MLE solution $\hat{\delta} = U^\top Y$ resp.

$$\hat{\beta} = AL^{-\frac{1}{2}}U^\top Y. \tag{6.9}$$

Here the parameter vector δ is simply a standardized version of γ: $\delta = L^{\frac{1}{2}}\gamma$.

Remark 6.6.6.

1. Since there are efficient algorithms for calculating a singular value decomposition, the term (6.9) can be calculated more efficiently compared to $\hat{\beta} = (X^\top X)^{-1}X^\top Y$, since $X^\top X$ has to be inverted in the latter.

2. Instead of removing the last $m - p$ principal components of $X^\top X$ from the regression (cf. Remark 6.6.2, 2.), it is generally possible to calculate $\widetilde{\beta}$ on a subset M of $\{1, \ldots, m\}$:

$$\widetilde{\beta}_M = \sum_{k \in M} l_k^{-1} a_k a_k^\top X^\top Y.$$

   Here, only principal components l_k, $k \in M$, are used for the regression. Then it also holds that

$$\text{Cov}\,(\widetilde{\beta}_M) = \sigma^2 \sum_{k \in M} l_k^{-1} a_k a_k^\top,$$

   cf. Exercise 6.6.4. This approach uses the elimination of components γ_k, $k \notin M$ of $\gamma = (\gamma_1, \ldots, \gamma_m)^\top$ of the ML estimation. Equivalently it can be thought of the exclusion of the components δ_k, $k \notin M$ of $\delta = (\delta_1, \ldots, \delta_m)^\top$, since $\delta = L^{\frac{1}{2}}$, with $\delta_k = \sqrt{l_k}\gamma_k$ for all k.

What are possible strategies for determining M?

1. $M = \{k : l_k > l^*\}$ for a predetermined threshold $l^* > 0$. If

$$\bar{l} = \frac{1}{m}\sum_{i=1}^{m} l_i$$

   are close to 1, $l^* \in (0.01, 0.1)$. The disadvantage of this methodology is that some of the (possibly important for the forecast of Y) principal components, might have a small variance and are thus eliminated from the model.

2. Let σ_{ii}^2 be the i-th diagonal element of $(X^\top X)^{-1}$. It holds that $\sigma_{ii}^2 = \frac{\text{Var}\hat{\beta}_i}{\sigma^2}$ (cf. Theorem 6.2.4), $i = 1, \ldots, m$. Then $M = \{k : \sigma_{kk}^2 > \sigma^*\}$ can be chosen for a sufficient threshold σ^*. For the choice of σ^* see [18], p. 174. This methodology has the same disadvantages as 1.

3. Define $M = \{1, \dots, p\}$, where p is the biggest number $\leq m$, for which one of the following criteria is met:

   a) It holds that:

   $$\sum_{i=1}^{m} \mathbb{E}(\widetilde{\beta}_{M_i} - \beta_i)^2 \leq \sum_{i=1}^{m} \mathbb{E}(\hat{\beta}_i - \beta_i)^2, \tag{6.10}$$

   for all $\beta = (\beta_1, \dots, \beta_m)^\top \in \mathbb{R}^m$.

   b) It holds that:

   $$\mathbb{E}(c^\top \widetilde{\beta}_M - c^\top \beta)^2 \leq \mathbb{E}(c^\top \hat{\beta} - c^\top \beta)^2 \quad \forall \beta \in \mathbb{R}^m, c \in \mathbb{R}^m$$

   c) It holds that:

   $$\mathbb{E}\left|X\widetilde{\beta}_M - X\beta\right|^2 \leq \mathbb{E}\left|X\hat{\beta} - X\beta\right|^2$$

   Here the criteria a) is similar to the task of estimating β as precise as possible. Criteria b) and c) on the other hand deliver the best possible estimation of of $\mathbb{E}Y = X\beta$ with $X\hat{\beta}_M$ resp. $X\hat{\beta}$. All terms in a)-c) are mean squared errors, which contain both the bias and the variance of $\widetilde{\beta}_M$.

Many more strategies are described in statistical literature, which provide a better estimator $\widetilde{\beta}_M$ compared to $\hat{\beta}$ depending on the given situation. The question on how to choose M is still unanswered.

An alternative approach of eliminating principal components in the regression is given by the following estimator $\widetilde{\beta}_R$:

$$\widetilde{\beta}_R = \sum_{k=1}^{m} (l_k + K_k)^{-1} a_k a_k^\top X^\top Y,$$

where $K_1, \dots, K_m > 0$ are weights, which represent additional influencing factors with respect to the regression. By using those weights it can be achieved, that $l_k \approx 0$ does not have a destabilizing influence on the estimation.

Exercise 6.6.7. Show that

- $$\text{Cov}\,(\widetilde{\beta}_R) = \sigma^2 \sum_{k=1}^{m} \frac{l_k}{(l_k + K_k)^2} a_k a_k^\top$$
- $\widetilde{\beta}_R$ is a biased estimator of β. Find the bias of $\widetilde{\beta}_R$!

$\widetilde{\beta}_R$ is called *Ridge Regression*. Here the question arises on how to choose K_k, $k = 1, \ldots, m$. In practice, $K_k = K$, $k = 1, \ldots, m$ is usually used, where K is to be chosen small.

Another application of PCA in regression is the so called *latent root regression*. This form of regression aims to only remove principal components, if they have a small variance l_k and do not add any additional value to the estimation of $\mathbb{E}Y$ with $X\beta$. Here the PCA is applied to the $(m+1) \times (m+1)$ matrix $\widetilde{X}^\top \widetilde{X}$ with $\widetilde{X} = (Y, X)$. Let $\tilde{a}_k$, $k = 0, \ldots, m$ be the coefficients of the PCA of $\widetilde{X}^\top \widetilde{X}$, with corresponding eigenvalues $\widetilde{l}_k$, $k = 0, \ldots, m$. Let $\tilde{a_k} = (a_{k0}, \ldots, a_{km})^\top$, $k = 0, \ldots, m$.
Define the index set of the principal components that are to be eliminated as $M_L = \{k = 0, \ldots, m : \widetilde{l}_k \leq l^*, |a_{k0}| \leq a^*\}$. This is the index set of those principal components, that have small variance and do not influence the estimation of Y a lot. Let $M = \{0, \ldots, m\} \backslash M_L$. Define $\hat{\beta}_L = \sum_{k \in M} \widetilde{c}_k \widetilde{a}_k$, where $\{\widetilde{c}_k, k \in M\} = \operatorname{argmin}_\beta |Y - X\beta|^2$ with $\beta = \sum_{k \in M} c_k \widetilde{a}_k$.

Theorem 6.6.8. It holds that

$$\widetilde{c_k} = -\frac{a_{k0} \sqrt{\sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2}}{\widetilde{l}_k \sum_{i \in M} \frac{a_{i0}^2}{\widetilde{l}_i}}, \quad k \in M.$$

Without proof!
Thresholds l^* and a^* are still to be chosen empirically.

6.7 Numeric calculation of principal components

In order to understand how statistical software packages calculates principal components, it is important to know the algorithms. By knowing the algorithms one can gain awareness about why some results might be bad (e.g. with eigenvalues that are almost equal) or what kind of restrictions there are with respect to size of the datasets (e.g. storage wise or runtime wise). In the following a short overview for those methods is given. Since the PCA is mainly based on calculating eigenvalues λ_i and eigenvectors α_i of a positive semi-definite $(m \times m)$ matrix Σ, the focus will mainly be on this calculation.
Let Σ thus be a $(m \times m)$ matrix with eigenvectors $\alpha_1, \ldots, \alpha_m$ and eigenvalues $\lambda_1, \ldots, \lambda_m$, which is positive semi-definite. In statistical literature there are at least four approaches for calculating α_i and λ_i:

1. Power iteration,

2. QR decomposition,

3. Singular value decomposition,

4. Neural networks.

Here only the essence of power iteration is presented: it represents an iterative algorithm for determining λ_1 and α_1, if $\lambda_1 >> \lambda_2 > \ldots > \lambda$. Let $u_0 \in \mathbb{R}^m$ be a starting vector. Define $u_r = \Sigma u_{r-1} = \Sigma^r u_0$ for all $r \in \mathbb{N}$. If

$$u_0 = \sum_{i=1}^{m} c_i \alpha_i,$$

where $\alpha_1, \ldots, \alpha_m$ are the orthonormal basis vectors and $c_1, \ldots, c_m$ are coordinates then

$$u_r = \Sigma^r u_0 = \sum_{i=1}^{m} c_i \Sigma^r \alpha_i = \sum_{i=1}^{m} c_i \lambda_i^r \alpha_i, \quad r \in \mathbb{N}.$$

Let $u_r = (u_{r1}, \ldots, u_{rm})^\top$, $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{im})^\top$.

Lemma 6.7.1. It holds that

$$\frac{u_{ri}}{u_{r-1,i}} \xrightarrow[r\to\infty]{} \lambda_1$$

for $i = 1, \ldots, m$ and

$$\frac{u_r}{c_i \lambda_1^r} \xrightarrow[r\to\infty]{} \alpha_1.$$

Proof For $j = 1, \ldots, m$ it holds that

$$u_{rj} = \sum_{i=1}^{m} c_i \lambda_i^r \alpha_{ij}$$

and thus

$$\begin{aligned} \frac{u_{rj}}{u_{r-1,j}} &= \frac{\sum_{i=1}^m c_i \lambda_i^r \frac{\alpha_{ij}}{\lambda_1^{r-1}}}{\sum_{i=1}^m c_i \lambda_i^{r-1} \frac{\alpha ij}{\lambda_1^{r-1}}} \\ &= \frac{c_1 \alpha_{1j} \lambda_1 + \sum_{i=2}^m c_i \left(\frac{\lambda_i}{\lambda_1}\right)^{r-1} \lambda_i \alpha_{ij}}{c_1 \alpha_{1j} + \sum_{i=2}^m c_i \left(\frac{\lambda_i}{\lambda_1}\right)^{r-1} \alpha_{ij}} \xrightarrow[r\to\infty]{} \frac{c_1 \alpha_{1j}}{c_1 \alpha_{1j}} \lambda_1 = \lambda_1, \end{aligned}$$

since $\frac{\lambda_i}{\lambda_1} < 1, i = 2, \ldots$. Furthermore,

$$\frac{u_r}{u \lambda_1^r} = \alpha_1 + \sum_{i=2}^{m} \frac{c_i}{c_1} \left(\frac{\lambda_i}{\lambda_1}\right)^r \alpha_i \xrightarrow[r\to\infty]{} \alpha_1.$$

□

The fact, that c_1 is unknown, is nothing to be worried about, since $\frac{u_r}{\lambda_1^r}$ can be standardized. The proof of Lemma 6.6.5 implies, that the rate of convergence of $\frac{u_{ri}}{u_{r-1,i}}$ to λ_1 and $\frac{u_r}{c_1\lambda_1^r}$ to α_1 is getting worse, if and only if $\lambda_1 \approx \lambda_2$, or in this case $\frac{\lambda_2}{\lambda_1} \approx 1$.

What should be done in the case that $\lambda_1 \approx \lambda_2$, in order to increase the rate of convergence? Instead of using Σ, $\Sigma - \rho I$ can be used for the iteration, in order to decrease the ratio $\frac{\lambda_2 - \rho}{\lambda_1 - \rho}$. Furthermore Σ can be replaced with $(\Sigma - \rho I)^{-1}$, which leads to solving the system of equations $(\Sigma - \rho I)\, u_r = u_{r-1}$ for every $r \in \mathbb{N}$. Thus for a suitable choice of ρ a convergence to α_k, $k = 1, \ldots, m$ is possible (in the second case).

Exercise 6.7.2. Construct those vectors and proof the convergence!

An increase in the rate of convergence can also be achieved, if one considers the sequence $\{u_{2r}\}$ instead of $\{u_r\}$ where $u_{2r} = T^{2^r} u_0$, $r \in \mathbb{N}$. Further methodologies for improving the algorithm of power iteration can be found in [18], p. 410-411.

Bibliography

[1] H. Albrecher, S. A. Ladoucette, and J. L. Teugels. Asymptotics of the sample coefficient of variation and the sample dispersion. *J. Statist. Plann. Inference*, 140(2):358–368, 2010.

[2] H. Dehling, B. Haupt. *Einf"uhrung in die Wahrscheinlichkeitstheorie und Statistik*. Springer, Berlin, 2003.

[3] P. Bickel and K. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*. Prentice Hall, London, 2001. 2nd ed., Vol. l.

[4] A. A. Borovkov. *Mathematical Statistics*. Gordon & Breach, 1998.

[5] M. Burkschat, E. Cramer, and U. Kamps. *Beschreibende Statistik, Grundlegende Methoden*. Springer, Berlin, 2004.

[6] G. Casella and R. L. Berger. *Statistical Inference*. Pacific Grove (CA), Duxbury, 2002.

[7] E. Cramer, K. Cramer, U. Kamps, and Zuckschwerdt. *Beschreibende Statistik, Interaktive Grafiken*. Springer, Berlin, 2004.

[8] E. Cramer and U. Kamps. *Grundlagen der Wahrscheinlichkeitsrechnung und Statistik*. Springer, Berlin, 2007.

[9] P. Dalgaard. *Introductory Statistics with R*. Springer, Berlin, 2002.

[10] A.J Dobson. *An Introduction to Generalizes Linear Models*. Chapmen & Hall, Boca Raton, 2002.

[11] Joseph L Doob. The limiting distributions of certain statistics. *The Annals of Mathematical Statistics*, 6(3):160–169, 1935.

[12] L. Fahrmeir, T. Kneib, and S. Lang. *Regression. Modelle, Methoden und Anwendungen*. Springer, Berlin, 2007.

[13] L. Fahrmeir, R. Künstler, I. Pigeot, and G. Tutz. *Statistik. Der Weg zur Datenanalyse*. Springer, Berlin, 2001.

[14] H. O. Georgii. *Stochastik*. de Gruyter, Berlin, 2002.

[15] J. Hartung, B. Elpert, and K. H. Klösener. *Statistik*. R. Oldenbourg Verlag, München, 1993. 9. Auflage.

[16] C. C. Heyde and E. Seneta. *Statisticians of the Centuries*. Springer, Berlin, 2001.

[17] A. Irle. *Wahrscheinlichkeitstheorie und Statistik, Grundlagen, Resultate, Anwendungen*. Teubner, 2001.

[18] I. T. Jolliffe. *Principal component analysis*. Springer, 2nd edition edition, 2002.

[19] L. J. Kazmir. *Wirtschaftsstatistik*. McGraw-Hill, 1996.

[20] K. R. Koch. *Parameter Estimation and Hypothesis Testing in Linear Models*. Springer, Berlin, 1999.

[21] A. Krause and M. Olson. *The Basics of S-PLUS*. Springer, Berlin, 2002. Third Ed.

[22] E. L. Lehmann. *Elements of Large-Sample Theory*. Springer, New York, 1999.

[23] J. Lehn and H. Wegmann. *Einführung in die Statistik*. Teubner, Stuttgart, 2000. 3. Auflage.

[24] J. Maindonald and J. Braun. *Data Analysis and Graphics Using R*. Cambridge University Press, 2003.

[25] M. Overbeck-Larisch and W. Dolejsky. *Stochastik mit Mathematica*. Vieweg, Braunschweig, 1998.

[26] H. Pruscha. *Angewandte Methoden der Mathematischen Statistik*. Teubner, Stuttgart, 2000.

[27] H. Pruscha. *Vorlesungen über Mathematische Statistik*. Teubner, Stuttgart, 2000.

[28] L. Sachs. *Angewandte Statistik*. Springer, 2004.

[29] L. Sachs and J. Hedderich. *Angewandte Statistik, Methodensammlung mit R*. Springer, Berlin, 2006.

[30] Robert J Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

[31] M. R. Spiegel and L. J. Stephens. *Statistik*. McGraw-Hill, 1999.

[32] E. Spodarev. *Wahrscheinlichkeitstheorie und stochastische Prozesse*. Ulm, 2020.

[33] E. Spodarev. *Elementare Wahrscheinlichkeitsrechnung und Statistik*. Ulm, 2022.

[34] V. Spokoiny and T. Dickhaus. *Basics of modern mathematical statistics*. Springer, 2015.

[35] W. A. Stahel. *Statistische Datenanalyse*. Vieweg, 1999.

[36] W. Venables and D. Ripley. *Modern applied statistics with S-PLUS*. Springer, 1999. 3rd ed.

[37] L. Wasserman. *All of Statistics. A Concise Course in Statistical Inference*. Springer, 2004.

Index