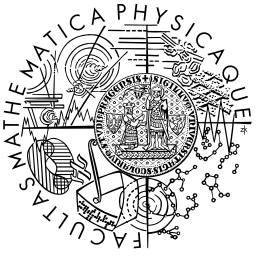

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

Mathematical Statistics 2

NMSA 332

Course notes

Stanislav Nagy

Last updated: April 30, 2024

Contents

The essential reference is the textbook [1, Chapters 7 and 8]. In fact, much of what follows draws from [1], and I strongly advise to consult that reference. When doing so, however, one has to read with care. In [1], some quite different (and perhaps occasionally confusing) notations are used. For example, by Ω (capital omega) we standardly denote the probability space on which all random variables are defined. In [1], $\Omega \subseteq \mathbb{R}^m$ is used also for the parameter space, which we here denote by $\boldsymbol{\Theta} \subseteq \mathbb{R}^m$. Further, the prime as in $\boldsymbol{x}'$ or $\boldsymbol{h}'$ is used for both the derivative of a function and the transposition of a matrix in [1]. Here we distinguish this and f' stands for the derivative of f, while $\boldsymbol{x}^\mathsf{T}$ is the transposition of a vector $\boldsymbol{x}$. The quantile functions of common distributions, such as the standard normal distribution or the χ^2-distribution, are also denoted and used slightly differently. For example, for $\alpha \in (0,1)$, by $u(\alpha)$ is in [1, p. 72] not denoted the α-quantile of the standard normal distribution $\mathsf{N}(0,1)$, but rather the critical value of $\mathsf{N}(0,1)$, which is defined as the $(1-\alpha)$-quantile of $\mathsf{N}(0,1)$. That is, in [1] the quantity $u(\alpha)$ is defined as $\Phi^{-1}(1-\alpha)$ for Φ the distribution function of $\mathsf{N}(0,1)$. The true value of a parameter $\boldsymbol{\theta}$ is here denoted by $\boldsymbol{\theta}_X$; in [1] the notation $\boldsymbol{\theta}_0$ is used, which is the same as for the value of the parameter that corresponds to a simple null hypothesis in a testing problem.

1 Theory of point estimation

1.1 The task of point estimation

All probability distributions throughout these notes are Borel, defined on the Borel σ-algebra of an appropriate topological space (typically the real line $\mathbb{R}$ or the Euclidean space $\mathbb{R}^d$ with $d \in \mathbb{N}$). All random variables are defined on a common probability space $(\Omega, \mathcal{A}, \mathsf{P})$.

Definition 1. A *statistical model* is a collection of probability distributions $\mathcal{F} = \{P_{\boldsymbol{\theta}} \colon \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ indexed by a parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. The set $\boldsymbol{\Theta}$ is called the *parameter space* of the model $\mathcal{F}$. Depending on the dimension of $\boldsymbol{\Theta}$, the model can be

- *parametric* if $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$ for $p \in \mathbb{N}$; or
- *nonparametric* if $\boldsymbol{\Theta}$ is infinite-dimensional, i.e. if $\mathcal{F}$ cannot be expressed as a parametric model.

Example 1.1. Several parametric models:

$$\mathcal{F}_1 = \{\mathsf{Exp}(\lambda) \colon \lambda > 0\},$$
$$\mathcal{F}_2 = \left\{\mathsf{N}(\mu, \sigma^2) \colon \mu \in \mathbb{R} \text{ and } \sigma > 0\right\},$$
$$\mathcal{F}_3 = \left\{\mathsf{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \colon \boldsymbol{\mu} \in \mathbb{R}^d \text{ and } \boldsymbol{\Sigma} \in \mathbb{R}^{d\times d} \text{ symmetric and positive definite}\right\},$$
$$\mathcal{F}_4 = \{\text{distributions } P \text{ in } \mathbb{R} \text{ such that } P(\{1, 2, \ldots, K\}) = 1\}.$$

In $\mathcal{F}_1$, $\mathcal{F}_2$, and $\mathcal{F}_3$ we have $p = 1$, 2, and $d + d(d+1)/2$, respectively. In $\mathcal{F}_4$ the number K is fixed in advance. What is the dimension p of the parameter space in model $\mathcal{F}_4$?

Examples of nonparametric models are

$$\mathcal{F}_5 = \left\{\text{all distributions in } \mathbb{R}^d\right\},$$
$$\mathcal{F}_6 = \left\{\text{all distributions in } \mathbb{R}^d \text{ with a continuous distribution function}\right\},$$
$$\mathcal{F}_7 = \left\{\text{all distributions in } \mathbb{R}^d \text{ with a density w.r.t.}^{1} \text{ the Lebesgue measure on } \mathbb{R}^d\right\},$$
$$\mathcal{F}_8 = \left\{\text{all distributions in } \mathbb{R}^d \text{ with a bounded density w.r.t. the Lebesgue measure on } \mathbb{R}^d\right\},$$
$$\mathcal{F}_9 = \left\{\text{all distributions in } \mathbb{R} \text{ with a symmetric density}\right\}.$$

We see that $\mathcal{F}_5 \supset \mathcal{F}_6 \supset \mathcal{F}_7 \supset \mathcal{F}_8$, but in all cases, $\boldsymbol{\Theta}$ is a space of distributions (or a function space) of infinite dimension. △

Special classes of nonparametric models are the *semiparametric* ones, where the parameter space is naturally a product of a finite-dimensional and a functional component, e.g.

$$\mathcal{F} = \left\{\text{densities of the form } f(x) = g(x-\delta) \text{ for all } x \in \mathbb{R}, \text{ with } \delta \in \mathbb{R} \text{ and a symmetric density } g\right\}.$$

This model has a parametric part ($\delta \in \mathbb{R}$) and a nonparametric part (the function g), but the dimension of $\boldsymbol{\Theta}$ is infinite.

Typically, all distributions in a parametric model $\mathcal{F}$ are supposed to be absolutely continuous with respect to a given σ-finite measure μ (typically the Lebesgue measure on $\mathbb{R}^d$, or an appropriate counting measure). We also say that the system $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is *dominated* by μ. For statistical models $\mathcal{F} = \{P_{\boldsymbol{\theta}_n} : n \in \mathbb{N}\}$ with countably many elements this is always possible, as one can take for μ a properly scaled sum of all the measures $\mu = \sum_{n=1}^{\infty} P_{\boldsymbol{\theta}_n}/(n^2)$ defined by $\mu(B) = \sum_{n=1}^{\infty} P_{\boldsymbol{\theta}_n}(B)/(n^2)$ for each B Borel, which is by definition σ-finite. For uncountable systems the situation is more complicated. Can you think of a system of measures that is not dominated by any σ-finite measure?

A reasonable model also must be *identifiable*, meaning that if $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$, then necessarily $P_{\boldsymbol{\theta}_1} \neq P_{\boldsymbol{\theta}_2}$. This could be written also as the mapping $\boldsymbol{\theta} \mapsto P_{\boldsymbol{\theta}}$ being invertible, meaning that from the knowledge of $P_{\boldsymbol{\theta}}$ it is possible to identify a unique parameter $\boldsymbol{\theta}$.

Example 1.2. In the following model of analysis of variance

$$\mathcal{F}_1 = \left\{ \mathsf{N}_2\left(\begin{pmatrix} \mu + \lambda_1 \\ \mu + \lambda_2 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) : \mu, \lambda_1, \lambda_2 \in \mathbb{R} \right\}$$

the parameter $\boldsymbol{\theta} = (\mu, \lambda_1, \lambda_2)^{\mathsf{T}} \in \mathbb{R}^3$ is not identifiable, as there are many combinations of parameters $\boldsymbol{\theta}$ that can lead to the same bivariate normal distribution in $\mathcal{F}_1$. An identifiable

[1] with respect to

reparametrization of this model is e.g.

$$\mathcal{F}_2 = \left\{ \mathsf{N}_2 \left(\begin{pmatrix} \mu \\ \mu + \lambda \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) : \mu, \lambda \in \mathbb{R} \right\},$$

which shows that the model has, in fact, only a two-dimensional parameter. △

A basic problem of statistical inference is the estimation of the unknown parameter $\boldsymbol{\theta}$ of a model $\mathcal{F} = \{P_{\boldsymbol{\theta}} \colon \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ from a random vector $\boldsymbol{X}$. We suppose that $\boldsymbol{X}$ is distributed according to some $P_{\boldsymbol{\theta}_X}$ with $\boldsymbol{\theta}_X \in \boldsymbol{\Theta}$. Alternatively, $\boldsymbol{X}$ can be a random sample[2] from $P_{\boldsymbol{\theta}_X}$.

The situation when $\boldsymbol{X}$ is a random sample is a special case of the first scenario. If $\boldsymbol{X}$ is composed of a random sample of size $n \in \mathbb{N}$, we could simply expand the statistical model $\mathcal{F}$ and consider the model with product measures $\mathcal{F}_n = \{\bigotimes_{i=1}^n P_{\boldsymbol{\theta}} \colon \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ instead. Of course, a single random vector $\boldsymbol{X}$ from $\mathcal{F}_n$ is equivalent with a random sample of size n from $\mathcal{F}$. It will be always clear from the context which of the situations we consider.

Remark 1 (Boldface notation)**.** In our notation, each individual random variable X_i can be one-dimensional, or also a random vector in $\mathbb{R}^d$ of dimension $d \in \mathbb{N}$. When considering the random sample of n random variables (or vectors) X_i as a whole, we write also $\boldsymbol{X} = \left(X_1^\mathsf{T}, \ldots, X_n^\mathsf{T}\right)^\mathsf{T}$, but for simplicity we will omit the transpositions inside that vector and write only $\boldsymbol{X} = (X_1, \ldots, X_n)^\mathsf{T}$. If $d = 1$, we have $\boldsymbol{X} \in \mathbb{R}^n$; for general $d \in \mathbb{N}$ we have to consider $\boldsymbol{X}$ as a $(d\,n)$-dimensional vector in $\mathbb{R}^{dn}$.

Observe that in our notation, we distinguish multi-dimensional parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ in a boldface font in contrast to the one-dimensional parameters $\theta \in \Theta \subseteq \mathbb{R}$. But, we do not make this distinction for random variables or random vectors X_i as elements of $\boldsymbol{X}$, for $i = 1, \ldots, n$. This slight abuse of notation will be useful in what follows, as most of the presented theory differs when considering the dimensionality $p \in \mathbb{N}$ of the parameter, but remains the same w.r.t. the dimension of the random sample $d \in \mathbb{N}$.

We suppose that the model $\mathcal{F}$ is known, but the *true value of the parameter* $\boldsymbol{\theta}_X \in \boldsymbol{\Theta}$ from which $\boldsymbol{X} = (X_1, \ldots, X_n)^\mathsf{T}$ is drawn is not. Our intention is to find a *point estimator* of $\boldsymbol{\theta}_X$ of the form $\widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n(\boldsymbol{X})$, a function of the random sample that does not depend on the unknown parameter $\boldsymbol{\theta}$, that estimates (approximates) the true value of the parameter $\boldsymbol{\theta}_X$ "well". Because in our problem the true value of the parameter $\boldsymbol{\theta}_X$ is not known, we call $\widehat{\boldsymbol{\theta}}_n$ an estimator of $\boldsymbol{\theta}$. The quality of the estimator $\widehat{\boldsymbol{\theta}}_n$ is assessed by its properties. The estimator itself is a random variable, as it depends on the random variable $\boldsymbol{X}$. We can thus talk about its basic characteristics.

[2] That is, we are given $\boldsymbol{X} = (X_1, \ldots, X_n)^\mathsf{T}$ whose elements $X_1, \ldots, X_n$ form a random sample (a sequence of independent, identically distributed random variables) from a distribution $P_{\boldsymbol{\theta}_X} \in \mathcal{F}$.

Definition 2. Let $\widehat{\boldsymbol{\theta}}_n$ be an estimator of a parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ in model $\mathcal{F}$. If the expectation of $\widehat{\boldsymbol{\theta}}_n$ exists, we can write

$$\mathsf{E}_{\boldsymbol{\theta}}\, \widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta} + \boldsymbol{b}_n(\boldsymbol{\theta}) \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

for a deterministic (that is, non-random) function $\boldsymbol{b}_n \colon \boldsymbol{\Theta} \to \mathbb{R}^p$. The function $\boldsymbol{b}_n$ is called the *bias* of the estimator $\widehat{\boldsymbol{\theta}}_n$. If the bias is a constant zero function, we call the estimator $\widehat{\boldsymbol{\theta}}_n$ *unbiased*. If $\boldsymbol{b}_n(\boldsymbol{\theta}) \xrightarrow[n\to\infty]{} \mathbf{0}$ for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we call $\widehat{\boldsymbol{\theta}}_n$ *asymptotically unbiased*.

Note that some authors [8] say that $\widehat{\boldsymbol{\theta}}_n$ is asymptotically unbiased for $\boldsymbol{\theta}$ if the expectation of the asymptotic distribution of $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)$ is $\mathbf{0}$ for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. This is a condition different from our definition of asymptotic unbiasedness.

A second order characteristic of a point estimator $\widehat{\theta}_n$ for $p = 1$ is the *mean squared error* defined as

$$\mathsf{MSE}_\theta\left(\widehat{\theta}_n\right) = \mathsf{E}_\theta\left(\widehat{\theta}_n - \theta\right)^2 \text{ for } \theta \in \Theta,$$

whenever the integral on the right hand side exists. The mean squared error takes into account both the bias and the variance of the estimator. The following lemma was proved already in the course NMSA331 [6, Section 3.1].

Lemma 1. *We can decompose the mean squared error into a squared bias term, and a variance term*

$$\mathsf{MSE}_\theta\left(\widehat{\theta}_n\right) = (b_n(\theta))^2 + \mathsf{var}_\theta\, \widehat{\theta}_n,$$

where $b_n(\theta)$ is the bias of $\widehat{\theta}_n$.

Proof. Write

$$\begin{aligned}\mathsf{MSE}_\theta\left(\widehat{\theta}_n\right) &= \mathsf{E}_\theta\left(\widehat{\theta}_n - \mathsf{E}_\theta\,\widehat{\theta}_n + \mathsf{E}_\theta\,\widehat{\theta}_n - \theta\right)^2 \\ &= \mathsf{E}_\theta\left(\widehat{\theta}_n - \mathsf{E}_\theta\,\widehat{\theta}_n\right)^2 + \mathsf{E}_\theta\left(\mathsf{E}_\theta\,\widehat{\theta}_n - \theta\right)^2 + 2\,\mathsf{E}_\theta\left(\widehat{\theta}_n - \mathsf{E}_\theta\,\widehat{\theta}_n\right)\left(\mathsf{E}_\theta\,\widehat{\theta}_n - \theta\right).\end{aligned}$$

The first summand on the right hand side is $\mathsf{var}_\theta\,\widehat{\theta}_n$. In the second summand the quantity $\left(\mathsf{E}_\theta\,\widehat{\theta}_n - \theta\right)^2 = (b_n(\theta))^2$ is deterministic, and equals the square of the bias of $\widehat{\theta}_n$. Likewise, in the third summand the factor $\left(\mathsf{E}_\theta\,\widehat{\theta}_n - \theta\right)$ is not random, and can be pulled out of the expectation. What remains is $\mathsf{E}_\theta\left(\widehat{\theta}_n - \mathsf{E}_\theta\,\widehat{\theta}_n\right) = 0$ for each $\theta \in \Theta$. □

With the mean squared error we usually consider only the case $p = 1$ of one-dimensional parameters. In the higher-dimensional case we can analyse the estimator component-wise.

A good estimator $\widehat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ must be consistent, meaning that

$$\widehat{\boldsymbol{\theta}}_n \xrightarrow[n\to\infty]{\mathsf{P}} \boldsymbol{\theta} \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

In this expression, $n \to \infty$ formally means that the size of the random vector $\boldsymbol{X}$ grows to infinity; usually this is intended for models $\mathcal{F}_n = \{\bigotimes_{i=1}^n P_{\boldsymbol{\theta}} \colon \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ corresponding to $\boldsymbol{X} = \boldsymbol{X}_n$ being random samples $X_1, \ldots, X_n$ from some $P_{\boldsymbol{\theta}}$. The consistency of a one-dimensional asymptotically unbiased estimator $\widehat{\theta}_n$ follows, for example, from the analysis of its variance, or equivalently its mean squared error. We know this result already from the course NMSA331 [6, Theorem 3.1].

Lemma 2. *Let $p = 1$, and let $\widehat{\theta}_n$ be an asymptotically unbiased estimator of $\theta \in \Theta \subseteq \mathbb{R}$ whose variance $\mathsf{var}_\theta \widehat{\theta}_n$ is finite for each $\theta \in \Theta$ and $n \in \mathbb{N}$. Suppose that $\mathsf{var}_\theta \widehat{\theta}_n \xrightarrow[n\to\infty]{} 0$ for each $\theta \in \Theta$. Then $\widehat{\theta}_n$ is consistent.*

Proof. Take $\varepsilon > 0$ and write by the Chebyshev inequality and Lemma 1

$$\mathsf{P}_\theta\left(\left|\widehat{\theta}_n - \theta\right| > \varepsilon\right) \leq \frac{\mathsf{E}_\theta\left(\widehat{\theta}_n - \theta\right)^2}{\varepsilon^2} = \frac{\mathsf{MSE}_\theta\left(\widehat{\theta}_n\right)}{\varepsilon^2} = \frac{(b_n(\theta))^2 + \mathsf{var}_\theta\, \widehat{\theta}_n}{\varepsilon^2},$$

where by our assumptions both $b_n(\theta)$ and $\mathsf{var}_\theta \widehat{\theta}_n$ decay to zero as $n \to \infty$ for all $\theta \in \Theta$. We have verified the convergence in probability as needed. □

The moment-based analysis of the estimator $\widehat{\boldsymbol{\theta}}_n$ performed using its expectation (that is, bias) and variance (or mean squared error) is quite useful, and will be followed throughout most of the course. It is, however, not the only possible approach. More generally, a quantitative measure of the quality of an estimator is based on the loss function and the risk.

Definition 3. A *loss function* is any measurable map

$$Q \colon \boldsymbol{\Theta} \times \boldsymbol{\Theta} \to [0, \infty).$$

The *risk*, or the *expected loss*, of an estimator $\widehat{\boldsymbol{\theta}}_n$ of parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, is defined as

$$R(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}) = \mathsf{E}_{\boldsymbol{\theta}}\, Q(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}) \text{ for } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

The loss $Q(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta})$ of an estimator $\widehat{\boldsymbol{\theta}}_n$ at $\boldsymbol{\theta}$ is a random variable, and the risk is its numerical summary. They both assess the degree of "discrepancy" of the estimator from the true parameter. It should therefore be $Q(\boldsymbol{\theta}, \boldsymbol{\theta}) = 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and the loss should increase as its arguments depart from each other. The use of the loss is consistent with our analysis based on the moments from above. For $p = 1$ we can choose $Q(x, y) = (x - y)^2$, in which case the risk equals the mean squared error. More generally, typically chosen loss functions are the *quadratic loss* for $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$

$$Q(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^2 \tag{1}$$

or the *absolute loss* $p = 1$ and $x, y \in \Theta \subseteq \mathbb{R}$

$$Q(x, y) = |x - y| .$$

If Θ is a bounded interval or, e.g. $\Theta = (0, \infty)$ as in the case of a variance parameter, other loss functions may be more appropriate.

For a given estimator $\widehat{\boldsymbol{\theta}}_n$ and a loss Q, the risk is a deterministic function $\boldsymbol{\Theta} \to [0, \infty] \colon \boldsymbol{\theta} \mapsto R(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta})$. Our main task is to search for estimators $\widehat{\boldsymbol{\theta}}_n$ minimizing the risk, if possible uniformly (that is, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$) as a function on $\boldsymbol{\Theta}$. We shall be mostly concerned with the quadratic loss (1) and $p = 1$, as this theory is standard, well explored, and relatively simple to handle.[3] Our main task is to find reasonable estimators uniformly minimizing the risk w.r.t. the quadratic loss function (1). As we see in the following example, in this task we need to impose additional restrictions — without them, our problem is not well posed.

Example 1.3. Let $\mathcal{F} = \{\mathsf{N}(\theta, 1) \colon \theta \in \mathbb{R}\}$, that is $\Theta = \mathbb{R}$ and $p = 1$. Take the estimator $\widetilde{\theta}_n = 3$ (almost surely). This estimator does not depend on the data, and is clearly not reasonable. Its (quadratic) risk is

$$R(\widetilde{\theta}_n, \theta) = \mathsf{MSE}_\theta \left(\widetilde{\theta}_n\right) = \mathsf{E}_\theta \, (3 - \theta)^2 = (3 - \theta)^2 .$$

As a function of $\theta \in \mathbb{R}$, the risk takes the value 0 at $\theta = 3$. That is, if the true value of the parameter equals $\theta = 3$, we get a perfect estimator, and any truly random estimator of θ must have a higher risk at $\theta = 3$. Thus, the risk function is impossible to be minimized for a single estimator uniformly over $\Theta = \mathbb{R}$.

For comparison, suppose that we observe only a single realisation ($n = 1$) of a random variable $X \sim \mathsf{N}(\theta, 1)$, and take $\widehat{\theta}_1 = X$. The risk of this estimator is

$$R(\widehat{\theta}_1, \theta) = \mathsf{MSE}_\theta \left(\widehat{\theta}_1\right) = \mathsf{E}_\theta \, (X - \mathsf{E}_\theta \, X)^2 = \mathsf{var}_\theta \, X = 1.$$

Similarly, for a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)^\mathsf{T}$ from $\mathsf{N}(\theta, 1)$ we could consider $\widehat{\theta}_n(\boldsymbol{X}) = \widehat{\theta}_n = \sum_{i=1}^n X_i / n = \bar{X}_n$ the sample mean, and obtain

$$R(\widehat{\theta}_n, \theta) = \mathsf{MSE}_\theta \left(\widehat{\theta}_n\right) = \mathsf{E}_\theta \, \left(\bar{X}_n - \mathsf{E}_\theta \, X_1\right)^2 = \mathsf{var}_\theta \, \bar{X}_n = 1/n.$$

None of these estimators is better than the trivial $\widetilde{\theta}_n$ if the true parameter θ_X is close enough to 3. However, in contrast to $\widetilde{\theta}_n$, the sample means $\widehat{\theta}_n$ are unbiased for θ. △

[3] It is however important to note that choosing a different loss function, other estimators would be preferred, and a different theory of estimation arises. The choice of the absolute loss, for example, results in the so-called theory of *robust estimation*, giving estimators that are less sensitive to measurement errors and misspecified statistical models. That theory is covered in more advanced courses.

For this reason, in the task of minimizing the risk, we cannot consider just any estimator, but need to restrict to a smaller sensible class of estimators. This is typically the class of the unbiased estimators of $\boldsymbol{\theta}$. An unbiased estimator that minimizes the quadratic risk for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ among all unbiased estimators of $\boldsymbol{\theta}$ is called the *best unbiased estimator (BUE)* of a parameter $\boldsymbol{\theta}$. In the first part of this course, we will search for best unbiased estimators. As we will show later, the sample mean $\bar{X}_n$ in Example 1.3 is BUE in that setting.

1.2 Fisher information and Rao-Cramér theorem

1.2.1 One-dimensional parameter

We are given a random vector composed of $X_1, \ldots, X_n$, denoted also by $\boldsymbol{X} = (X_1, \ldots, X_n)^\mathsf{T}$. The distribution of $\boldsymbol{X}$ depends only on a parameter $\theta \in \Theta$. For the beginning, we suppose that we deal with a one-dimensional parameter $\Theta \subseteq \mathbb{R}$, and a known measurable function $g\colon \Theta \to \mathbb{R}$. To deal with our estimation problem more generally, we want to estimate the transformed parameter $g(\theta) \in \mathbb{R}$. We call $g(\theta)$ a *parametric function* of θ. We intend to estimate θ or some $g(\theta)$ based on $\boldsymbol{X}$. Not to confuse θ and its estimators with the estimators of its parametric functions $g(\theta)$, we shall write also $T = T(\boldsymbol{X})$, or equivalently also $T_n = T_n(\boldsymbol{X})$ if the length n of the vector $\boldsymbol{X}$ plays a role, for the estimators of $g(\theta)$ based on $\boldsymbol{X} = (X_1, \ldots, X_n)^\mathsf{T}$. We will use $\widehat{\theta}_n$ only for the estimator of θ. The common situation is with g an identity function, in which case $g(\theta) = \theta$; the general function g allows to consider also possible reparametrizations of our problem.

Often, the elements $X_1, \ldots, X_n$ of $\boldsymbol{X}$ form a random sample from a distribution parametrized by θ. Nevertheless, for the following theory this is not necessary, and the random variables $X_1, \ldots, X_n$ are allowed to be dependent, or may fail to be identically distributed. The only requirement is that the distribution of the whole vector $\boldsymbol{X}$ depends only on θ. Recall that according to our Remark 1, our setup also applies to the situation when $p = 1$, but we observe n multivariate random vectors, each of dimension $d \in \mathbb{N}$. In that case, the length of the vector $\boldsymbol{X}$ is in fact $d\,n$. This is not a problem as we did not assume anything about the joint distribution of $\boldsymbol{X}$. Instead, to search for the BUE of $g(\theta)$, we need the following regularity conditions.

Definition 4 (Regular system of densities — one-dimensional parameter)**.** Let the distribution of the random vector $\boldsymbol{X}$ depend only on the parameter $\theta \in \Theta$. Suppose that $\boldsymbol{X}$ has a density $f(\boldsymbol{x}; \theta)$ w.r.t. a given σ-finite measure μ. The system of densities $\{f(\boldsymbol{x}; \theta)\colon \theta \in \Theta\}$ is called a *regular system of densities* if the following conditions hold:

(R_1) The parameter space $\Theta \subseteq \mathbb{R}$ is a non-empty open set.

(R$_2$) The set $M = \{\boldsymbol{x}\colon f(\boldsymbol{x};\theta) > 0\}$ does not depend on θ.

(R$_3$) For μ-almost all $\boldsymbol{x} \in M$ there exists a finite partial derivative

$$f'(\boldsymbol{x};\theta) = \frac{\partial f(\boldsymbol{x};\theta)}{\partial\theta}.$$

(R$_4$) For all $\theta \in \Theta$ we can write $\int_M f'(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) = 0$.

(R$_5$) The integral

$$J_n(\theta) = \int_M \left(\frac{f'(\boldsymbol{x};\theta)}{f(\boldsymbol{x};\theta)}\right)^2 f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) \tag{2}$$

is finite and non-zero for every $\theta \in \Theta$.

For a regular system of densities, the function $J_n\colon \Theta \to (0,\infty)$ from (2) is called the *Fisher information* of θ contained in $\boldsymbol{X}$.

Conditions (R$_1$)–(R$_5$) are quite natural, and all lead to the definition of the Fisher information in (2). In (R$_1$) we require that we can take a derivative w.r.t. θ. Condition (R$_2$) deals with the support[4] of the distribution of $\boldsymbol{X}$, and requires that this support does not depend on the unknown parameter. Note that, since densities are defined uniquely only μ-almost everywhere, (R$_2$) in fact states that there exists a version of the density f with this property. Condition (R$_3$) is needed for the definition of the Fisher information (2), and in (R$_4$) we can recognize a swap of a limit and an integral. Indeed, since all $f(\cdot;\theta)$ are densities, we know that

$$\int_M f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) = \int_{\mathbb{R}^n} f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) = 1 \text{ for all } \theta \in \Theta,$$

and taking the derivative of both sides of this equation w.r.t. θ we obtain

$$\frac{\partial \int_M f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x})}{\partial\theta} = 0 \text{ for all } \theta \in \Theta.$$

Condition (R$_4$) states that the derivative and the integral on the left hand side can be interchanged, which is for common densities usually true. Finally, the Fisher information (2) in (R$_5$) can be also seen as an integral over $\mathbb{R}^n$ (or $\mathbb{R}^{dn}$) instead of over M, because for $\boldsymbol{x} \notin M$ we have by (R$_2$) that $f(\boldsymbol{x};\theta) = 0$ for all $\theta \in \Theta$. Thus, the integral can be rewritten as an expectation in the forms

$$J_n(\theta) = \mathsf{E}_\theta\left(\frac{f'(\boldsymbol{X};\theta)}{f(\boldsymbol{X};\theta)}\right)^2 = \mathsf{E}_\theta\left(\frac{\partial \log f(\boldsymbol{X};\theta)}{\partial\theta}\right)^2.$$

[4]The *support* of (the distribution of) a random vector $\boldsymbol{X} \in \mathbb{R}^n$ is the smallest closed set $\mathcal{S} \subseteq \mathbb{R}^n$ such that $\mathsf{P}(\boldsymbol{X} \in \mathcal{S}) = 1$. We will use this term more freely, and also the set $M \subseteq \mathbb{R}^n$ of all points where the density of $\boldsymbol{X}$ is positive will be called the support of $\boldsymbol{X}$. Note the small difference between these two terms; take, for instance, $n = 1$ and the density $f(x) = 1$ for $x \in (0,1)$, $f(x) = 0$ elsewhere.

Similarly, also Condition (R_4) can be rewritten as an expectation

$$0 = \int_M f'(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) = \mathsf{E}_\theta\left(\frac{f'(\boldsymbol{X};\theta)}{f(\boldsymbol{X};\theta)}\right) = \mathsf{E}_\theta\left(\frac{\partial \log f(\boldsymbol{X};\theta)}{\partial\theta}\right) \text{ for all } \theta\in\Theta.$$

This formula will appear frequently in our proofs.

Example 1.4 (Normal distribution and μ)**.** Let $\boldsymbol{X} = (X_1,\dots,X_n)^{\mathsf{T}}$ be a random sample from $\mathsf{N}(\mu,1)$ for $\mu\in\mathbb{R}$. The corresponding system of densities is easily seen to be regular. For instance, $\Theta=\mathbb{R}$ in (R_1) and $M=\mathbb{R}^n$ in (R_2). We compute the Fisher information of μ contained in $\boldsymbol{X}$, first by its definition in (R_5). We have

$$f(\boldsymbol{x};\mu) = \left(\frac{1}{\sqrt{2\,\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i-\mu)^2\right) \text{ for } \boldsymbol{x}=(x_1,\dots,x_n)^{\mathsf{T}}\in\mathbb{R}^n.$$

This gives

$$\log f(\boldsymbol{x};\mu) = -\frac{n}{2}\log(2\,\pi) - \frac{1}{2}\sum_{i=1}^n (x_i-\mu)^2,$$

and

$$\frac{\partial \log f(\boldsymbol{x};\mu)}{\partial\mu} = \sum_{i=1}^n (x_i-\mu).$$

For the Fisher information of μ we get

$$\begin{aligned} J_n(\mu) &= \mathsf{E}_\mu\left(\frac{\partial \log f(\boldsymbol{X};\mu)}{\partial\mu}\right)^2 = \mathsf{E}_\mu\left(\sum_{i=1}^n (X_i-\mu)\right)^2 \\ &= \mathsf{var}_\mu\left(\sum_{i=1}^n (X_i-\mu)\right) + \left(\mathsf{E}_\mu\sum_{i=1}^n (X_i-\mu)\right)^2 = n. \end{aligned}$$

△

For the next example, let $\boldsymbol{X} = (X_1,\dots,X_n)^{\mathsf{T}}$ be a random sample from $\mathsf{N}(0,\sigma^2)$ with $\sigma^2>0$. Now, there are two natural parametrizations of this system: either (i) $\theta=\sigma^2$, or (ii) $\theta=\sigma$. In both situations, the corresponding systems of densities are regular, and we could compute both $J_n(\sigma^2)$ in the first case, or also $J_n(\sigma)$ in the second. These two Fisher informations are not the same. Let us first explore what happens with $\theta=\sigma^2$; the situation with $\theta=\sigma$ will be treated below in Example 1.9.

Example 1.5 (Normal distribution and σ^2)**.** For $\theta=\sigma^2$ we get

$$f(\boldsymbol{x};\theta) = \left(\frac{1}{\sqrt{2\,\pi\,\theta}}\right)^n \exp\left(-\frac{1}{2\,\theta}\sum_{i=1}^n x_i^2\right) \text{ for } \boldsymbol{x}=(x_1,\dots,x_n)^{\mathsf{T}}\in\mathbb{R}^n,$$

which gives

$$\log f(\boldsymbol{x};\theta) = -\frac{n}{2}\log(2\,\pi) - \frac{n}{2}\log\theta - \frac{1}{2\,\theta}\sum_{i=1}^n x_i^2.$$

We take a derivative w.r.t. θ to get

$$\frac{\partial \log f(\boldsymbol{x};\theta)}{\partial \theta} = -\frac{n}{2\,\theta} + \frac{1}{2\,\theta^2}\sum_{i=1}^{n} x_i^2.$$

For the Fisher information of $\theta = \sigma^2$ we have

$$\begin{aligned} J_n(\theta) &= \mathsf{E}_\theta \left(\frac{\partial \log f(\boldsymbol{X};\theta)}{\partial \theta}\right)^2 = \mathsf{E}_\theta \left(-\frac{n}{2\,\theta} + \frac{1}{2\,\theta^2}\sum_{i=1}^{n} X_i^2\right)^2 \\ &= \frac{n^2}{4\,\theta^2} + \mathsf{E}_\theta \left(\frac{1}{2\,\theta^2}\sum_{i=1}^{n} X_i^2\right)^2 - 2\,\mathsf{E}_\theta \frac{n}{2\,\theta}\frac{1}{2\,\theta^2}\sum_{i=1}^{n} X_i^2 \\ &= \frac{n^2}{4\,\theta^2} + \frac{1}{4\,\theta^4}\mathsf{E}_\theta \left(\sum_{i=1}^{n} X_i^2\right)^2 - \frac{n}{2\,\theta^3}\mathsf{E}_\theta \sum_{i=1}^{n} X_i^2. \end{aligned}$$

Now, since the variables X_i are independent and each distributed as $\mathsf{N}(0,\theta)$, we get that $Y_i = X_i/\sigma$ are independent and identically $\mathsf{N}(0,1)$-distributed, with $\sum_{i=1}^n X_i^2 = \sum_{i=1}^n (\sigma Y_i)^2 = \sigma^2 T_n$ for T_n having the chi-squared distribution χ^2_n. We know that $\mathsf{E}\,T_n = n$ and $\mathsf{var}\,T_n = 2\,n$, which gives

$$\begin{aligned} J_n(\theta) &= \frac{n^2}{4\,\theta^2} + \frac{1}{4\,\theta^4}\mathsf{E}_\theta \left(\sigma^2 T_n\right)^2 - \frac{n}{2\,\theta^3}\mathsf{E}_\theta\, \sigma^2 T_n. \\ &= \frac{n^2}{4\,\theta^2} + \frac{1}{4\,\theta^2}\left(\mathsf{var}\,T_n + (\mathsf{E}\,T_n)^2\right) - \frac{n}{2\,\theta^2}\mathsf{E}\,T_n \\ &= \frac{n^2}{4\,\theta^2} + \frac{1}{4\,\theta^2}\left(2\,n + n^2\right) - \frac{2\,n}{4\,\theta^2}n = \frac{n}{2\,\theta^2} = \frac{n}{2\,\sigma^4}. \end{aligned}$$

△

Very roughly speaking, the Fisher information quantifies how much "information", or "knowledge" about θ can be extracted from the random vector $\boldsymbol{X}$ — the higher the Fisher information is, the better estimators of θ can be constructed. It has interesting properties.

The end of lecture 1 (20.2.2024)

Theorem 1. *Let $\{f(\boldsymbol{x};\theta)\colon \theta \in \Theta\}$ be a regular system of densities. Suppose further that the second derivative*

$$f''(\boldsymbol{x};\theta) = \frac{\partial^2 f(\boldsymbol{x};\theta)}{\partial \theta^2}$$

exists for μ-almost all $\boldsymbol{x} \in M$, and a condition analogous to (R_4)

$$\int_M f''(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) = 0 \tag{3}$$

is valid for all $\theta \in \Theta$. Then we can write

$$J_n(\theta) = -\int_M \frac{\partial^2 \log f(\boldsymbol{x};\theta)}{\partial \theta^2} f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) = -\mathsf{E}_\theta \frac{\partial^2 \log f(\boldsymbol{X};\theta)}{\partial \theta^2}.$$

Proof. A direct computation gives

$$\frac{\partial^2 \log f(\boldsymbol{x};\theta)}{\partial\theta^2} = \frac{\partial f'(\boldsymbol{x};\theta)/f(\boldsymbol{x};\theta)}{\partial\theta} = \frac{f''(\boldsymbol{x};\theta)\, f(\boldsymbol{x};\theta) - \left(f'(\boldsymbol{x};\theta)\right)^2}{\left(f(\boldsymbol{x};\theta)\right)^2} = \frac{f''(\boldsymbol{x};\theta)}{f(\boldsymbol{x};\theta)} - \left(\frac{f'(\boldsymbol{x};\theta)}{f(\boldsymbol{x};\theta)}\right)^2,$$

which results in

$$-\int_M \frac{\partial^2 \log f(\boldsymbol{x};\theta)}{\partial\theta^2} f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) = -\int_M f''(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) + \int_M \left(\frac{f'(\boldsymbol{x};\theta)}{f(\boldsymbol{x};\theta)}\right)^2 f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) = J_n(\theta),$$

since the first summand on the right hand side vanishes by our assumption (3). □

Theorem 1 simplifies the computation of the Fisher information substantially. Take, for example, the normal distribution as above.

Example 1.6. For $\boldsymbol{X} = (X_1, \ldots, X_n)^{\mathsf{T}}$ a random sample from $\mathsf{N}(\mu, 1)$ as in Example 1.4 we get

$$\frac{\partial^2 \log f(\boldsymbol{x};\mu)}{\partial\mu^2} = \frac{\partial\left(\sum_{i=1}^n (x_i - \mu)\right)}{\partial\mu} = -n,$$

and directly

$$J_n(\mu) = -\mathsf{E}_\mu \frac{\partial^2 \log f(\boldsymbol{X};\mu)}{\partial\mu^2} = n.$$

Similarly, in the setup of Example 1.5, where we assumed a random sample from $\mathsf{N}\left(0, \sigma^2\right)$ and $\theta = \sigma^2$, we have

$$\frac{\partial^2 \log f(\boldsymbol{x};\mu)}{\partial\theta^2} = \frac{\partial\left(-\frac{n}{2\,\theta} + \frac{1}{2\,\theta^2}\sum_{i=1}^n x_i^2\right)}{\partial\theta} = \frac{n}{2\,\theta^2} - \frac{1}{\theta^3}\sum_{i=1}^n x_i^2,$$

and for the Fisher information of θ we can write

$$J_n(\theta) = -\mathsf{E}_\theta\left(\frac{n}{2\,\theta^2} - \frac{1}{\theta^3}\sum_{i=1}^n X_i^2\right) = -\frac{n}{2\,\theta^2} + \frac{n}{\theta^2} = \frac{n}{2\,\theta^2}.$$

△

If the random variables $X_1, \ldots, X_n$ from $\boldsymbol{X} = (X_1, \ldots, X_n)^{\mathsf{T}}$ are independent, the Fisher information of $\boldsymbol{X}$ is the sum of the Fisher informations of its components.

Theorem 2. *Let the distributions of the random vectors $\boldsymbol{Y} = (Y_1, \ldots, Y_{m_1})^{\mathsf{T}}$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_{m_2})^{\mathsf{T}}$ both depend only on the parameter $\theta \in \Theta$, and suppose that $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are independent of each other. Let both $\boldsymbol{Y}$ and $\boldsymbol{Z}$ correspond to regular systems of densities w.r.t. the σ-finite measures μ_1 and μ_2, respectively. Denote the Fisher information of $\boldsymbol{Y}$ by $J_{\boldsymbol{Y}}$ and the Fisher information of $\boldsymbol{Z}$ by $J_{\boldsymbol{Z}}$. Then the joint random vector $\boldsymbol{X} = \left(\boldsymbol{Y}^{\mathsf{T}}, \boldsymbol{Z}^{\mathsf{T}}\right)^{\mathsf{T}}$ has a regular system of densities, and its Fisher information $J_{\boldsymbol{X}}$ equals $J_{\boldsymbol{Y}} + J_{\boldsymbol{Z}}$.*

Proof. Write $f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)$ for the density of $\boldsymbol{Y}$ w.r.t. μ_1, and $f_{\boldsymbol{Z}}(\boldsymbol{z};\theta)$ for the density of $\boldsymbol{Z}$ w.r.t. μ_2. Because $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are independent, the joint density $f(\boldsymbol{x};\theta)$ of the vector $\boldsymbol{X}$ takes the form of the product

$$f(\boldsymbol{x};\theta) = f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)\, f_{\boldsymbol{Z}}(\boldsymbol{z};\theta) \text{ for } \boldsymbol{x} = \left(\boldsymbol{y}^{\mathsf{T}}, \boldsymbol{z}^{\mathsf{T}}\right)^{\mathsf{T}} \in \mathbb{R}^{m_1+m_2}.$$

This density is taken w.r.t. the product measure $\mu = \mu_1 \times \mu_2$ on $\mathbb{R}^{m_1+m_2}$. Verification of (R_1)–(R_5) for this system of densities is straightforward. For instance, if $M_1 \subseteq \mathbb{R}^{m_1}$ and $M_2 \subseteq \mathbb{R}^{m_2}$ are the sets from (R_2) for $\boldsymbol{Y}$ and $\boldsymbol{Z}$, respectively, then the support of $\boldsymbol{X}$ takes the form $M = M_1 \times M_2$ and also does not depend on θ. The derivative of f in (R_3) is

$$f'(\boldsymbol{x};\theta) = f'_{\boldsymbol{Y}}(\boldsymbol{y};\theta)\, f_{\boldsymbol{Z}}(\boldsymbol{z};\theta) + f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)\, f'_{\boldsymbol{Z}}(\boldsymbol{z};\theta).$$

We compute the Fisher information of $\boldsymbol{X}$. We have for $\theta \in \Theta$

$$\begin{aligned}
J_{\boldsymbol{X}}(\theta) &= \int_M \left(\frac{f'(\boldsymbol{x};\theta)}{f(\boldsymbol{x};\theta)}\right)^2 f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) \\
&= \int_{M_1}\int_{M_2} \left(\frac{f'_{\boldsymbol{Y}}(\boldsymbol{y};\theta)\, f_{\boldsymbol{Z}}(\boldsymbol{z};\theta) + f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)\, f'_{\boldsymbol{Z}}(\boldsymbol{z};\theta)}{f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)\, f_{\boldsymbol{Z}}(\boldsymbol{z};\theta)}\right)^2 f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)\, f_{\boldsymbol{Z}}(\boldsymbol{z};\theta)\,\mathrm{d}\,\mu_2(\boldsymbol{z})\,\mathrm{d}\,\mu_1(\boldsymbol{y}) \\
&= \int_{M_1}\int_{M_2} \left(\frac{f'_{\boldsymbol{Y}}(\boldsymbol{y};\theta)}{f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)} + \frac{f'_{\boldsymbol{Z}}(\boldsymbol{z};\theta)}{f_{\boldsymbol{Z}}(\boldsymbol{z};\theta)}\right)^2 f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)\, f_{\boldsymbol{Z}}(\boldsymbol{z};\theta)\,\mathrm{d}\,\mu_2(\boldsymbol{z})\,\mathrm{d}\,\mu_1(\boldsymbol{y}) \\
&= \int_{M_1}\int_{M_2} \left(\frac{f'_{\boldsymbol{Y}}(\boldsymbol{y};\theta)}{f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)}\right)^2 f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)\, f_{\boldsymbol{Z}}(\boldsymbol{z};\theta)\,\mathrm{d}\,\mu_2(\boldsymbol{z})\,\mathrm{d}\,\mu_1(\boldsymbol{y}) \\
&\quad + \int_{M_1}\int_{M_2} \left(\frac{f'_{\boldsymbol{Z}}(\boldsymbol{z};\theta)}{f_{\boldsymbol{Z}}(\boldsymbol{z};\theta)}\right)^2 f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)\, f_{\boldsymbol{Z}}(\boldsymbol{z};\theta)\,\mathrm{d}\,\mu_2(\boldsymbol{z})\,\mathrm{d}\,\mu_1(\boldsymbol{y}) \\
&\quad + 2\int_{M_1}\int_{M_2} \frac{f'_{\boldsymbol{Y}}(\boldsymbol{y};\theta)}{f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)}\frac{f'_{\boldsymbol{Z}}(\boldsymbol{z};\theta)}{f_{\boldsymbol{Z}}(\boldsymbol{z};\theta)} f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)\, f_{\boldsymbol{Z}}(\boldsymbol{z};\theta)\,\mathrm{d}\,\mu_2(\boldsymbol{z})\,\mathrm{d}\,\mu_1(\boldsymbol{y}) \\
&= \int_{M_1} \left(\frac{f'_{\boldsymbol{Y}}(\boldsymbol{y};\theta)}{f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)}\right)^2 f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)\,\mathrm{d}\,\mu_1(\boldsymbol{y}) \int_{M_2} f_{\boldsymbol{Z}}(\boldsymbol{z};\theta)\,\mathrm{d}\,\mu_2(\boldsymbol{z}) \\
&\quad + \int_{M_1} f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)\,\mathrm{d}\,\mu_1(\boldsymbol{y}) \int_{M_2} \left(\frac{f'_{\boldsymbol{Z}}(\boldsymbol{z};\theta)}{f_{\boldsymbol{Z}}(\boldsymbol{z};\theta)}\right)^2 f_{\boldsymbol{Z}}(\boldsymbol{z};\theta)\,\mathrm{d}\,\mu_2(\boldsymbol{z}) \\
&\quad + 2\int_{M_1} f'_{\boldsymbol{Y}}(\boldsymbol{y};\theta)\,\mathrm{d}\,\mu_1(\boldsymbol{y}) \int_{M_2} f'_{\boldsymbol{Z}}(\boldsymbol{z};\theta)\,\mathrm{d}\,\mu_2(\boldsymbol{z}) \\
&= J_{\boldsymbol{Y}}(\theta) + J_{\boldsymbol{Z}}(\theta),
\end{aligned}$$

where in the final equality we used that both $f_{\boldsymbol{Y}}$ and $f_{\boldsymbol{Z}}$ are densities and thus integrate to one, the definition of the Fisher information (2), and Condition (R_4) for both $\boldsymbol{Y}$ and $\boldsymbol{Z}$. □

An important consequence of Theorem 2 comes for a vector $\boldsymbol{X}$ whose elements form a random sample $X_1, \ldots, X_n$. Applying Theorem 2 several times, we obtain that in this case $J_n(\theta) = n\, J_1(\theta)$, meaning that the Fisher information contained in a random sample is n-times larger than the Fisher information contained in each individual random variable. We

already saw this is Examples 1.4, 1.5, and 1.6. Using Theorem 2, we actually did not need to work with the whole density of $\boldsymbol{X}$ in any of those examples, but a density of a single variable X_1 would be enough to obtain that $J_1(\mu) = 1$ in Example 1.4, and $J_1(\theta) = 1/(2\,\theta^2)$ in Example 1.5.

Our first application of the Fisher information is a lower bound on the variance of an unbiased estimator.

Theorem 3 (Rao-Cramér)**.** *Let $T_n = T_n(\boldsymbol{X})$ be an unbiased estimator of a parametric function $g(\theta)$ that satisfies $\mathsf{var}_\theta\, T_n < \infty$ for all $\theta \in \Theta$. Let the following conditions be satisfied:*

(RC_1) the system of densities $\{f(\boldsymbol{x};\theta)\colon \theta \in \Theta\}$ of $\boldsymbol{X}$ is regular;

(RC_2) the derivative $g'(\theta)$ of g exists for every $\theta \in \Theta$;

(RC_3) the following interchange of a derivative and an integral is valid for all $\theta \in \Theta$

$$\frac{\partial}{\partial\theta}\int_M T_n(\boldsymbol{x}) f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) = \int_M T_n(\boldsymbol{x}) f'(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}).$$

Then it holds true that

$$\mathsf{var}_\theta\, T_n = \mathsf{E}_\theta\,(T_n - g(\theta))^2 \geq \frac{(g'(\theta))^2}{J_n(\theta)} \quad \text{for all } \theta \in \Theta. \tag{4}$$

Proof. Because T_n is an unbiased estimator of $g(\theta)$, we know that for each $\theta \in \Theta$ we have

$$\mathsf{E}_\theta\, T_n = \int_M T_n(\boldsymbol{x}) f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) = g(\theta).$$

Take a derivative of the previous equality. Using (RC_2) and (RC_3) we obtain

$$g'(\theta) = \int_M T_n(\boldsymbol{x}) f'(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) = \int_M T_n(\boldsymbol{x}) \frac{f'(\boldsymbol{x};\theta)}{f(\boldsymbol{x};\theta)} f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}). \tag{5}$$

At the same time (R_4) gives for all $\theta \in \Theta$

$$0 = \int_M f'(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) = \int_M g(\theta) \frac{f'(\boldsymbol{x};\theta)}{f(\boldsymbol{x};\theta)} f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}),$$

which together with (5) allows us to write

$$g'(\theta) = \int_M (T_n(\boldsymbol{x}) - g(\theta)) \frac{f'(\boldsymbol{x};\theta)}{f(\boldsymbol{x};\theta)} f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}).$$

We apply the Cauchy-Schwarz inequality to the last integral. The integral is taken w.r.t. the probability measure given by $f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x})$, that is, the measure on $\mathbb{R}^n$ (or $\mathbb{R}^{dn}$, or M) with density $f(\cdot;\theta)$ w.r.t. μ. We obtain

$$\begin{aligned} g'(\theta) &\leq \sqrt{\int_M (T_n(\boldsymbol{x}) - g(\theta))^2 f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) \int_M \left(\frac{f'(\boldsymbol{x};\theta)}{f(\boldsymbol{x};\theta)}\right)^2 f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x})} \\ &= \sqrt{J_n(\theta)\,\mathsf{var}_\theta\, T_n}, \end{aligned} \tag{6}$$

as needed. □

The proof above is quite straightforward but its idea is not completely clear; the following alternative approach may be more insightful.

Alternative proof of Theorem 3. Fix $\theta \in \Theta$ and denote

$$S_n(\boldsymbol{x};\theta) = \frac{f'(\boldsymbol{x};\theta)}{f(\boldsymbol{x};\theta)} \quad \text{and} \quad S_n = S_n(\boldsymbol{X};\theta),$$

where S_n is random, and the random vector $\boldsymbol{X}$ plugged into S_n has a distribution with density $f(\boldsymbol{x};\theta)$ for our particular value of $\theta \in \Theta$. By (R_4) we know that

$$\mathsf{E}_\theta\, S_n = \int_M S_n(\boldsymbol{x};\theta) f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) = 0 \quad \text{for all } \theta \in \Theta,$$

and by (R_5)

$$\mathsf{var}_\theta\, S_n = \int_M \left(S_n(\boldsymbol{x};\theta)\right)^2 f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) = J_n(\theta).$$

For the covariance of S_n and T_n we can write using (RC_3) and (R_4)

$$\begin{aligned}
\mathsf{cov}\,(T_n, S_n) &= \int_M (T_n(\boldsymbol{x}) - g(\theta)) S_n(\boldsymbol{x};\theta) f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) \\
&= \int_M (T_n(\boldsymbol{x}) - g(\theta)) f'(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) \\
&= \int_M T_n(\boldsymbol{x}) f'(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) - g(\theta) \int_M f'(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) \\
&= \frac{\partial}{\partial\theta} \int_M T_n(\boldsymbol{x}) f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) - 0 = g'(\theta).
\end{aligned}$$

The covariance matrix of the vector $(T_n, S_n)^\mathsf{T}$ thus takes the form

$$\boldsymbol{\Sigma}_n = \begin{pmatrix} \mathsf{var}_\theta\, T_n & g'(\theta) \\ g'(\theta) & J_n(\theta) \end{pmatrix}.$$

As a covariance matrix, $\boldsymbol{\Sigma}_n$ must be positive semi-definite, and in particular its determinant must be non-negative for any $\theta \in \Theta$. Thus,

$$0 \le \det \boldsymbol{\Sigma}_n = J_n(\theta)\, \mathsf{var}_\theta\, T_n - \left(g'(\theta)\right)^2.$$

□

In Theorem 3 we found a lower bound on the variance of an unbiased estimator. Let us explore what this theorem gives for the normal distribution.

Example 1.7. In the scenario of Examples 1.3 and 1.4 of $X_1, \ldots, X_n$ a random sample from $\mathsf{N}(\mu, 1)$ we have $J_n(\mu) = n$, and Theorem 3 gives that for any (regular enough) unbiased estimator $T_n = T_n(\boldsymbol{X})$ of $g(\mu) = \mu$ we have

$$\mathsf{var}_\mu\, T_n \ge \frac{\left(g'(\mu)\right)^2}{J_n(\mu)} = \frac{1}{n} \text{ for all } \mu \in \mathbb{R}.$$

On the other hand, for $T_n = \bar{X}_n = \sum_{i=1}^n X_i/n$ we know that

$$\mathsf{var}_\mu T_n = \frac{1}{n} \text{ for all } \mu \in \mathbb{R}.$$

Thus, the Rao-Cramér bound gives that the sample average is the best unbiased estimator of μ in the model $\mathsf{N}(\mu, 1)$ (at least among all estimators that satisfy condition (RC_3)).

In the situation of Example 1.5 where $X_1, \ldots, X_n$ is a random sample from $\mathsf{N}(0, \theta)$ we have from Theorem 3 that for any (regular enough) unbiased estimator $T_n = T_n(\boldsymbol{X})$ of $\theta = \sigma^2$ we have

$$\mathsf{var}_\theta T_n \geq \frac{1}{J_n(\theta)} = \frac{2\,\theta^2}{n} = \frac{2\,\sigma^4}{n} \text{ for all } \theta = \sigma^2 > 0.$$

For the estimator $V_n = \sum_{i=1}^n X_i^2/n$ it is easy to see that we attain the Rao-Cramér bound. Thus, V_n is the best unbiased estimator of σ^2 in the model $\mathsf{N}(0, \sigma^2)$ (again, among all estimators that satisfy (RC_3)).

Later on we will see that in the more common situation when both parameters μ and σ^2 are unknown, the situation with best unbiased estimators is slightly more involved. △

Combined with Theorem 2, the Rao-Cramér theorem gives that under appropriate regularity conditions, no unbiased estimator of a parameter θ based on a random sample of observations of size n can have the variance of order smaller than $\mathcal{O}(n^{-1})$. It is important to note that if the regularity conditions are not met, there can still exist better estimators, as we show in the following example.

Example 1.8. Let $X_1, \ldots, X_n$ be a random sample from the uniform distribution on the interval $[0, \theta]$ for $\theta > 0$. The corresponding system of densities is not regular, as Condition (R_2) is clearly violated. Consider the estimator $T_n = (n+1)\max_{i=1,\ldots,n} X_i/n$. It is easy to show that T_n is unbiased for θ, and at the same time

$$\mathsf{var}_\theta T_n = \frac{\theta^2}{n(n+2)}.$$

We see that the rate of convergence of the variance of T_n to zero is $\mathcal{O}(n^{-2})$, and T_n does not obey the Rao-Cramér bound. In fact, because (R_2) is violated, the Fisher information of θ contained in $X_1, \ldots, X_n$ is not even well defined. △

It is interesting to observe that for a general parametric function $g(\theta)$, the term $g'(\theta)$ in the Rao-Cramér bound corresponds to the Fisher information of the transformed parameter $g(\theta)$. To see this, we first return to Example 1.5 with the unknown parameter being the variance of a normal distribution.

Example 1.9 (Normal distribution with σ and σ^2)**.** In the situation of $\boldsymbol{X} = (X_1, \ldots, X_n)^{\mathsf{T}}$ being a random sample from $\mathsf{N}(0, \sigma^2)$, take $\theta = \sigma^2$, and $g(\theta) = \sigma = \sqrt{\theta}$. The Rao-Cramér bound of Theorem 3 then gives that for any regular unbiased estimator T_n of σ we have

$$\mathsf{var}_\theta\, T_n \geq \frac{(g'(\theta))^2}{J_n(\theta)} = \frac{1}{4\,\theta}\frac{2\,\theta^2}{n} = \frac{\theta}{2\,n} = \frac{\sigma^2}{2\,n} \quad \text{for all } \theta = \sigma^2 > 0. \tag{7}$$

On the other hand, take now $\theta = \sigma$ with $\Theta = (0, \infty)$, and compute the Fisher information of σ contained in $\boldsymbol{X}$. For a single observation X_1 we have

$$\log f(x;\theta) = -\frac{1}{2}\log(2\,\pi) - \log\theta - \frac{x^2}{2\,\theta^2} \quad \text{for } x \in \mathbb{R},$$

and

$$\frac{\partial^2 \log f(x;\theta)}{\partial\theta^2} = \frac{\partial\left(-\frac{1}{\theta} + \frac{x^2}{\theta^3}\right)}{\partial\theta} = \frac{1}{\theta^2} - \frac{3\,x^2}{\theta^4}.$$

The Fisher information of $\theta = \sigma$ contained in $\boldsymbol{X}$ is therefore

$$J_n(\theta) = -n\,\mathsf{E}_\theta\,\frac{\partial^2 \log f(X_1;\theta)}{\partial\theta^2} = n\left(-\frac{1}{\theta^2} + \frac{3\,\mathsf{E}\,X_1^2}{\theta^4}\right) = \frac{2\,n}{\theta^2} = \frac{2\,n}{\sigma^2} \text{ for } \theta = \sigma > 0.$$

Applying Theorem 3 with $\theta = \sigma$ and $g(\theta) = \theta$, we thus also obtain that for any unbiased estimator T_n of σ we get (7). △

The fact that both Rao-Cramér bounds in Example 1.9 are equal is, of course, not a coincidence. Let σ be the original parameter, and let $g(\sigma) = \sigma^2$. In both cases, to compute the Fisher information we need the density $f(\boldsymbol{x};\sigma)$ of $\boldsymbol{X}$ that can be parametrized either by σ, or by $g(\sigma) = \sigma^2$. For the Fisher information of σ we then take

$$J_n(\sigma) = \mathsf{E}_\sigma\left(\frac{\partial \log f(\boldsymbol{X};\sigma)}{\partial\sigma}\right)^2, \tag{8}$$

while for the Fisher information of $\eta = g(\sigma) = \sigma^2$ we need to find

$$\widetilde{J}_n(\eta) = \mathsf{E}_\sigma\left(\frac{\partial \log f(\boldsymbol{X};\sigma)}{\partial g(\sigma)}\right)^2.$$

We have added a tilde above the Fisher information of $\eta = g(\sigma)$ to distinguish this notation from the Fisher information of the original parameter σ. To relate the two expressions J_n and $\widetilde{J}_n$, we use the chain rule for derivatives, and express the integrand in (8) as

$$\frac{\partial \log f(\boldsymbol{x};\sigma)}{\partial\sigma} = \frac{\partial \log f(\boldsymbol{x};\sigma)}{\partial g(\sigma)}\,\frac{\partial g(\sigma)}{\partial\sigma} \text{ for all } \boldsymbol{x} \in \mathbb{R}^n.$$

Plugging this into (8) we get

$$J_n(\sigma) = \mathsf{E}_\sigma\left(\frac{\partial \log f(\boldsymbol{X};\sigma)}{\partial \sigma}\right)^2 = \mathsf{E}_\sigma\left(\frac{\partial \log f(\boldsymbol{X};\sigma)}{\partial g(\sigma)}\frac{\partial g(\sigma)}{\partial \sigma}\right)^2$$
$$= \left(\frac{\partial g(\sigma)}{\partial \sigma}\right)^2 \mathsf{E}_\sigma\left(\frac{\partial \log f(\boldsymbol{X};\sigma)}{\partial g(\sigma)}\right)^2 = \left(g'(\sigma)\right)^2 \widetilde{J}_n(\eta) = \left(g'(\sigma)\right)^2 \widetilde{J}_n(g(\sigma)),$$

where on the left hand side we have the Fisher information of σ, and on the right hand side the Fisher information of $g(\sigma) = \eta$. Observe that in our argument, we used only the chain rule for derivatives and the fact that the function g is differentiable. As in Condition (R_5) we require that Fisher informations are positive, we also need to impose $g'(\sigma) \neq 0$. This gives the following more general result.

Theorem 4. *Let $J_n(\theta)$ be the Fisher information of $\theta \in \Theta$ contained in a random vector $\boldsymbol{X}$, and let $g\colon \Theta \to \mathbb{R}$ be a differentiable function such that $g'(\theta) \neq 0$ for all $\theta \in \Theta$. Then the Fisher information $\widetilde{J}_n(g(\theta))$ of the parametric function $g(\theta)$ contained in $\boldsymbol{X}$ is*

$$\widetilde{J}_n(g(\theta)) = \frac{J_n(\theta)}{(g'(\theta))^2} \quad \textit{for all } \theta \in \Theta.$$

We now explore when, under the given regularity conditions, can the Rao-Cramér bound of Theorem 3 be attained. Looking at the first proof of Theorem 3, the bound followed by a direct application of the Cauchy-Schwarz inequality in (6). We know that the Cauchy-Schwarz inequality is strict unless the two functions multiplied in the integrand are linearly dependent. Suppose therefore that there is equality in the Rao-Cramér inequality (4) for each $\theta \in \Theta$. Necessarily, for every $\theta \in \Theta$ we then have a constant $c(\theta) \in \mathbb{R}$ such that

$$c(\theta)\,(T_n(\boldsymbol{x}) - g(\theta)) = \frac{f'(\boldsymbol{x};\theta)}{f(\boldsymbol{x};\theta)} = \frac{\partial \log f(\boldsymbol{x};\theta)}{\partial \theta} \quad \text{for } \mu\text{-almost all } \boldsymbol{x} \in M.$$

We integrate the previous equality w.r.t. θ. Denoting by $C(\theta)$ the primitive function of $c(\theta)$ and by $G(\theta)$ the primitive function of $c(\theta)\,g(\theta)$, we can express

$$C(\theta)T_n(\boldsymbol{x}) - G(\theta) + H(\boldsymbol{x}) = \log f(\boldsymbol{x};\theta) \quad \text{for } \mu\text{-almost all } \boldsymbol{x} \in M,$$

where $H(\boldsymbol{x})$ is the constant term obtained by integration which does not depend on θ. Take an exponential of both sides of the last formula, giving

$$f(\boldsymbol{x};\theta) = \exp\left(C(\theta)T_n(\boldsymbol{x})\right) u(\boldsymbol{x})v(\theta) \quad \text{for } \mu\text{-almost all } \boldsymbol{x} \in M, \tag{9}$$

where $u(\boldsymbol{x}) = \exp\left(H(\boldsymbol{x})\right)$ and $v(\theta) = \exp\left(G(\theta)\right)$. In addition to (9) we know by (R_2) that $f(\boldsymbol{x};\theta) = 0$ for $\boldsymbol{x} \notin M$, meaning that if the Rao-Cramér bound can be attained, the density of $\boldsymbol{X}$ must possess a very special form where the contribution of $\boldsymbol{x}$ and θ factorizes, except

for the single exponential term $\exp\left(C(\theta)T_n(\boldsymbol{x})\right)$. This is a special case of a density called a *density of an exponential type*. The class of exponential-type densities is very important and will be dealt with in a greater detail later in the course. Many common densities are of an exponential type.

Given an unbiased estimator T_n of $g(\theta)$, Theorem 3 allows us to check whether T_n is BUE. If $\mathsf{var}_\theta\, T_n$ attains the Rao-Cramér bound, by Lemma 1 we know that the mean squared error of T_n is the smallest among all unbiased estimators. This leads to the following definition.

Definition 5. We call an estimator T a *regular estimator of* θ if it satisfies all the assumptions of Theorem 3 with $g(\theta) = \theta$. For a regular estimator of θ, the *efficiency* of T is defined as the ratio

$$e(\theta) = \frac{1}{J_n(\theta)\,\mathsf{var}_\theta\, T} \quad \text{for } \theta \in \Theta.$$

A regular estimator of θ is called *efficient* if $e(\theta) = 1$ for all $\theta \in \Theta$.

The end of lecture 2 (27.2.2024)

By Theorem 3, the efficiency of a regular estimator of θ is bounded in the interval $(0, 1]$. An efficient estimator is surely BUE, but we saw that efficient estimators exist only for special types of systems of densities. For many other models, efficient estimators do not exist. Thus, it is of interest to refine the Rao-Cramér bound. One such statement is found in a theorem of Bhattacharya that will follow. Before stating its result, we provide a useful lemma about determinants and inverses of matrices divided into blocks. This lemma will be used in the next, and several other proofs in the sequel.

Lemma 3. *Let*

$$\boldsymbol{J} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}$$

be a square matrix whose blocks $\boldsymbol{A}$ *and* $\boldsymbol{D}$ *are also square. Let* $\boldsymbol{D}$ *be non-singular. Then the determinant of* $\boldsymbol{J}$ *can be written in the form*

$$\det(\boldsymbol{J}) = \det(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}) \det(\boldsymbol{D}).$$

If, in addition, also $\boldsymbol{J}$ *and* $\boldsymbol{A}$ *are non-singular, then we can express the inverse of* $\boldsymbol{J}$ *as*

$$\begin{aligned} \boldsymbol{J}^{-1} &= \begin{pmatrix} \left(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1} \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{I} & -\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\boldsymbol{C}\boldsymbol{A}^{-1} & \boldsymbol{I} \end{pmatrix} \\ &= \begin{pmatrix} \left(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1} & -\left(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} & \left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1} \end{pmatrix}. \end{aligned}$$

Proof. For the statement about the determinant, notice that we can write with $\boldsymbol{I}$ the identity matrix

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{D}^{-1}\boldsymbol{C} & \boldsymbol{D}^{-1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C} & \boldsymbol{B}\boldsymbol{D}^{-1} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix}.$$

A determinant of a block triangular matrix equals the product of determinants of the blocks on the diagonal of a matrix, as follows from the Laplace expansion of the determinant. Also, the determinant commutes with matrix multiplication, meaning that we can apply determinant to both sides of the previous formula to get

$$\det(\boldsymbol{J})\det(\boldsymbol{I}) = \det(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})\det(\boldsymbol{D}),$$

as we wanted to show.

For the expression for the inverse matrix, write

$$\begin{aligned}\boldsymbol{M}_1 &= \left(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1},\\ \boldsymbol{M}_2 &= \left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1}\end{aligned}$$

By direct multiplication, starting from the right

$$\begin{aligned}&\begin{pmatrix}\boldsymbol{M}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{M}_2\end{pmatrix}\cdot\begin{pmatrix}\boldsymbol{I} & -\boldsymbol{B}\boldsymbol{D}^{-1}\\ -\boldsymbol{C}\boldsymbol{A}^{-1} & \boldsymbol{I}\end{pmatrix}\cdot\begin{pmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{pmatrix}\\ &\qquad = \begin{pmatrix}\boldsymbol{M}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{M}_2\end{pmatrix}\cdot\begin{pmatrix}\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\end{pmatrix}\\ &\qquad = \begin{pmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}\end{pmatrix}\end{aligned}$$

for $\boldsymbol{I}$ the unit square matrix of an appropriate dimension. □

We are now ready to prove Bhattacharya's extension of the Rao-Cramér theorem.

Theorem 5 (Bhattacharya)**.** *Let $T = T(\boldsymbol{X})$ be an unbiased estimator of the parametric function $g(\theta)$ that satisfies $\mathsf{var}_\theta\, T < \infty$ for all $\theta \in \Theta$. Suppose further that for some $k \in \mathbb{N}$ the following is true:*

(B_1) the system of densities $\{f(\boldsymbol{x};\theta)\colon \theta \in \Theta\}$ is regular;

(B_2) the k-th derivative $g^{(k)}(\theta)$ exists for all $\theta \in \Theta$;

(B_3) for each $j = 1, \ldots, k$ there exists the j-th derivative $f^{(j)}(\boldsymbol{x};\theta) = \frac{\partial^j f(\boldsymbol{x};\theta)}{\partial\theta^j}$ and

$$\int_M f^{(j)}(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) = 0;$$

(B_4) for each $j = 1, \ldots, k$ the following interchange of a derivative and an integral is valid for all $\theta \in \Theta$

$$\frac{\partial^j}{\partial\theta^j}\int_M T(\boldsymbol{x})f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) = \int_M T(\boldsymbol{x})f^{(j)}(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}).$$

(B_5) for each $j, \ell = 1, \dots, k$ and $\theta \in \Theta$ we have

$$\int_M \left| \frac{f^{(j)}(\boldsymbol{x};\theta)}{f(\boldsymbol{x};\theta)} \frac{f^{(\ell)}(\boldsymbol{x};\theta)}{f(\boldsymbol{x};\theta)} \right| f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) < \infty;$$

(B_6) for each $j, \ell = 1, \dots, k$ denote

$$V_{j,\ell}(\theta) = \int_M \frac{f^{(j)}(\boldsymbol{x};\theta)}{f(\boldsymbol{x};\theta)} \frac{f^{(\ell)}(\boldsymbol{x};\theta)}{f(\boldsymbol{x};\theta)} f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) = \mathsf{E}_\theta \frac{f^{(j)}(\boldsymbol{X};\theta)}{f(\boldsymbol{X};\theta)} \frac{f^{(\ell)}(\boldsymbol{X};\theta)}{f(\boldsymbol{X};\theta)}.$$

Let the matrix $\boldsymbol{V}(\theta) = (V_{j,\ell}(\theta))_{j,\ell=1}^k$ be non-singular for all $\theta \in \Theta$.

Denote $\boldsymbol{h}(\theta) = \big(g'(\theta), \dots, g^{(k)}(\theta)\big)^\mathsf{T}$. Then we can bound

$$\mathsf{var}_\theta\, T = \mathsf{E}\ (T - g(\theta))^2 \geq \boldsymbol{h}(\theta)^\mathsf{T} \boldsymbol{V}(\theta)^{-1} \boldsymbol{h}(\theta). \tag{10}$$

Proof. Let for $j = 1, \dots, k$

$$S_j(\boldsymbol{x};\theta) = \frac{f^{(j)}(\boldsymbol{x};\theta)}{f(\boldsymbol{x};\theta)} \quad \text{and} \quad S_j = S_j(\boldsymbol{X};\theta),$$

where S_j is random, and the random vector $\boldsymbol{X}$ plugged into S_j has a distribution corresponding to θ. By (B_3) we know that

$$\mathsf{E}_\theta\, S_j = \int_M S_j(\boldsymbol{x};\theta) f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) = 0 \text{ for each } j = 1, \dots, k,$$

and thus for the elements of the covariance matrix of T and all S_j we can write using (B_4)

$$\begin{aligned}
\mathsf{cov}\,(T, S_j) &= \int_M (T(\boldsymbol{x}) - g(\theta)) S_j(\boldsymbol{x};\theta) f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) \\
&= \int_M (T(\boldsymbol{x}) - g(\theta)) f^{(j)}(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) = \int_M T(\boldsymbol{x}) f^{(j)}(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) \\
&= \frac{\partial^j}{\partial \theta^j} \int_M T(\boldsymbol{x}) f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) = g^{(j)}(\theta), \\
\mathsf{cov}\,(S_j, S_\ell) &= \int_M S_j(\boldsymbol{x};\theta) S_\ell(\boldsymbol{x};\theta) f(\boldsymbol{x};\theta)\,\mathrm{d}\,\mu(\boldsymbol{x}) = V_{j,\ell}(\theta).
\end{aligned}$$

The covariance matrix of the vector $(T, S_1, \dots, S_k)^\mathsf{T}$ thus takes the form

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathsf{var}_\theta\, T & \boldsymbol{h}(\theta)^\mathsf{T} \\ \boldsymbol{h}(\theta) & \boldsymbol{V}(\theta) \end{pmatrix}.$$

As a covariance matrix, $\boldsymbol{\Sigma}$ must be positive semi-definite, and in particular its determinant must be non-negative. Using the rule for the determinant of a block matrix from Lemma 3 we arrive at

$$\det(\boldsymbol{\Sigma}) = \det(\boldsymbol{V}) \det(\mathsf{var}_\theta\, T - \boldsymbol{h}(\theta)^\mathsf{T} \boldsymbol{V}(\theta)^{-1} \boldsymbol{h}(\theta)) = \det(\boldsymbol{V}) \left(\mathsf{var}_\theta\, T - \boldsymbol{h}(\theta)^\mathsf{T} \boldsymbol{V}(\theta)^{-1} \boldsymbol{h}(\theta)\right) \geq 0.$$

Because also $\boldsymbol{V}$ is a covariance matrix, its determinant is non-negative, meaning that necessarily $\mathsf{var}_\theta\, T - \boldsymbol{h}(\theta)^\mathsf{T} \boldsymbol{V}(\theta)^{-1} \boldsymbol{h}(\theta) \geq 0$ as we wanted to show. □

Certainly, $V_{1,1}(\theta) = J_n(\theta)$ is the Fisher information of $\boldsymbol{X}$, and Theorem 5 applied with $k = 1$ gives the Rao-Cramér bound of Theorem 3. Using a little matrix algebra [10, p. 347], it is also possible to show that the sequence of Bhattacharya's bounds is, in fact, non-decreasing in $k \in \mathbb{N}$. In particular, the Bhattacharya bound is always an improvement over the Rao-Cramér bound. This is easiest to see if one considers $g(\theta) = \theta$ the identity function; for the general case of a parametric function $g(\theta)$, reparametrization relations such as those from Theorem 4 can be applied.

Theorem 6. *Suppose that $g(\theta) = \theta$ and that the assumptions of Theorem 5 are satisfied for both $k \in \mathbb{N}$ and $k+1$. Denote by $B_k(\theta)$ the constant on the right hand side of the Bhattacharya bound (10). Then*

$$B_{k+1}(\theta) \geq B_k(\theta) \text{ for all } \theta \in \Theta.$$

Proof. In the case of $g(\theta) = \theta$, Theorem 5 gives with $\boldsymbol{h}(\theta) = (1, 0, \ldots, 0)^{\mathsf{T}}$ the bound

$$\mathsf{var}_\theta\, T \geq B_k(\theta) = \boldsymbol{V}_k^{1,1}(\theta), \tag{11}$$

where $\boldsymbol{V}_k^{1,1}(\theta)$ stands for the first diagonal element of the matrix $\boldsymbol{V}_k(\theta)^{-1}$, where we have emphasized by a subscript k that the matrix $\boldsymbol{V}(\theta) = \boldsymbol{V}_k(\theta)$ is considered with k derivatives taken in Theorem 5. Our task is to show that with k replaced by $k+1$ in (11) we obtain a stronger bound, meaning that

$$B_k(\theta) = \boldsymbol{V}_k^{1,1}(\theta) \leq \boldsymbol{V}_{k+1}^{1,1}(\theta) = B_{k+1}(\theta) \text{ for all } \theta \in \Theta. \tag{12}$$

We need the following matrix identity.

Lemma 4 (Woodbury matrix identity)**.** *For any $\boldsymbol{A}$ and $\boldsymbol{D}$ non-singular square matrices and any matrices $\boldsymbol{B}$, $\boldsymbol{C}$ with conformable sizes we have*

$$\left(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1} = \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}.$$

Proof. We can proceed by a direct multiplication. Writing $\boldsymbol{X} = \left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1}$ we have

$$\begin{aligned}
\left(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)\left(\boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{X}\boldsymbol{C}\boldsymbol{A}^{-1}\right) &= \boldsymbol{I} + \boldsymbol{B}\boldsymbol{X}\boldsymbol{C}\boldsymbol{A}^{-1} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{X}\boldsymbol{C}\boldsymbol{A}^{-1} \\
&= \left(\boldsymbol{I} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}\right) + \left(\boldsymbol{B}\boldsymbol{X}\boldsymbol{C}\boldsymbol{A}^{-1} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{X}\boldsymbol{C}\boldsymbol{A}^{-1}\right) \\
&= \left(\boldsymbol{I} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}\right) + \boldsymbol{B}\left(\boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)\left(\boldsymbol{X}\boldsymbol{C}\boldsymbol{A}^{-1}\right) \\
&= \left(\boldsymbol{I} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}\right) + \boldsymbol{B}\boldsymbol{D}^{-1}\left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)\left(\boldsymbol{X}\boldsymbol{C}\boldsymbol{A}^{-1}\right) \\
&= \left(\boldsymbol{I} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}\right) + \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{X}^{-1}\boldsymbol{X}\boldsymbol{C}\boldsymbol{A}^{-1} \\
&= \left(\boldsymbol{I} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}\right) + \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} \\
&= \boldsymbol{I}.
\end{aligned}$$

□

We have the matrix $\boldsymbol{V}_{k+1}(\theta)$ and intend to express the first diagonal element of its inverse. To do this, we partition $\boldsymbol{V}_{k+1}(\theta)$ into a block matrix

$$\boldsymbol{V}_{k+1}(\theta) = \begin{pmatrix} \boldsymbol{V}_k(\theta) & \boldsymbol{v}_{k+1}(\theta) \\ \boldsymbol{v}_{k+1}(\theta)^\mathsf{T} & V_{k+1,k+1}(\theta) \end{pmatrix} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}, \tag{13}$$

where $\boldsymbol{v}_{k+1}(\theta) = (V_{1,k+1}(\theta), \ldots, V_{k,k+1}(\theta))^\mathsf{T}$, and $V_{j,k+1}(\theta)$ is defined analogously as in (B_6) but with k replaced by $k+1$. To find the first diagonal element of $\boldsymbol{V}_{k+1}(\theta)$ we use Lemma 3 to get that the top left $k \times k$ block of the inverse $\boldsymbol{V}_{k+1}(\theta)$ takes the form

$$\left(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1} = \left(\boldsymbol{V}_k(\theta) - \boldsymbol{v}_{k+1}(\theta) V_{k+1,k+1}(\theta)^{-1} \boldsymbol{v}_{k+1}(\theta)^\mathsf{T}\right)^{-1}$$

which by the Woodbury identity from Lemma 4 equals

$$\begin{aligned} &\boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} \\ &\quad = \boldsymbol{V}_k(\theta)^{-1} + \boldsymbol{V}_k(\theta)^{-1}\boldsymbol{v}_{k+1}(\theta)\left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{v}_{k+1}(\theta)^\mathsf{T}\boldsymbol{V}_k(\theta)^{-1} \\ &\quad = \boldsymbol{V}_k(\theta)^{-1} + \frac{\boldsymbol{b}\cdot\boldsymbol{b}^\mathsf{T}}{V_{k+1,k+1}(\theta) - \boldsymbol{v}_{k+1}(\theta)^\mathsf{T}\boldsymbol{V}_k(\theta)^{-1}\boldsymbol{v}_{k+1}(\theta)} \end{aligned} \tag{14}$$

with $\boldsymbol{b} = \boldsymbol{V}_k(\theta)^{-1}\boldsymbol{v}_{k+1}(\theta)$. For the denominator on the right hand side we have, in the notation from (13), that

$$\frac{1}{V_{k+1,k+1}(\theta) - \boldsymbol{v}_{k+1}(\theta)^\mathsf{T}\boldsymbol{V}_k(\theta)^{-1}\boldsymbol{v}_{k+1}(\theta)} = \left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1}.$$

This is by Lemma 3 the last diagonal element of the matrix $\boldsymbol{V}_{k+1}(\theta)^{-1}$. Because $\boldsymbol{V}_{k+1}(\theta)$ is assumed to be positive definite for all $\theta \in \Theta$ in (B_6), this term must be positive as well.

We thus obtained in equation (14) that the first diagonal term of $\boldsymbol{V}_{k+1}(\theta)^{-1}$ is equal to the first diagonal term of $\boldsymbol{V}_k(\theta)^{-1}$ plus a non-negative term, meaning that (12) is true as we wanted to show. □

Example 1.10. Take $\boldsymbol{X} = (X_1, \ldots, X_n)^\mathsf{T}$ a random sample from $\mathsf{N}(\mu, 1)$, and suppose that we want to estimate $g(\mu) = \mu^2$. We start from $U_n = U_n(\boldsymbol{X}) = (\sum_{i=1}^n X_i/n)^2 = (\bar{X}_n)^2$, which is a consistent estimator of $g(\mu)$, and observe that since $\bar{X}_n \sim \mathsf{N}(\mu, 1/n)$, we have that U_n has the same distribution as $Z/\sqrt{n} + \mu$ for $Z \sim \mathsf{N}(0,1)$. Thus we can write

$$\mathsf{E}_\mu U_n = \mathsf{E}_\mu \left(Z/\sqrt{n} + \mu\right)^2 = \mathsf{E}_\mu Z^2/n + \mu^2 + \frac{2\mu}{\sqrt{n}}\mathsf{E}_\mu Z = g(\mu) + \frac{1}{n},$$

which gives an unbiased estimator of $g(\mu)$

$$T_n = (\bar{X}_n)^2 - \frac{1}{n}.$$

For the variance of T_n we have

$$\begin{aligned}\mathsf{var}_\mu\, T_n &= \mathsf{var}_\mu \left(\left(\bar{X}_n\right)^2 - \frac{1}{n} \right) = \mathsf{var}_\mu \left(Z/\sqrt{n} + \mu \right)^2 \\ &= \frac{1}{n^2} \mathsf{var}_\mu\, Z^2 + \left(\frac{2\,\mu}{\sqrt{n}} \right)^2 \mathsf{var}_\mu\, Z = \frac{4\,\mu^2}{n} + \frac{2}{n^2},\end{aligned} \tag{15}$$

where we used that Z and Z^2 are uncorrelated, due to the symmetry of Z.

The standard Rao-Cramér bound from Theorem 3 gives

$$\mathsf{var}_\mu\, T_n \geq \frac{\left(g'(\mu)\right)^2}{J_n(\mu)} = \frac{4\,\mu^2}{n},$$

which is only the first term in (15), and T_n does not attain this bound. For a second order bound we use Bhattacharya's Theorem 5 with $k = 2$. First we compute the matrix $\boldsymbol{V}(\mu)$, for which we need the first two derivatives of the density $f(\boldsymbol{x};\mu)$ w.r.t. μ. We have for $\boldsymbol{x} = (x_1, \ldots, x_n)^\mathsf{T} \in \mathbb{R}^n$

$$\begin{aligned}\frac{f^{(1)}(\boldsymbol{x};\mu)}{f(\boldsymbol{x};\mu)} &= \frac{f(\boldsymbol{x};\mu) \sum_{i=1}^n (x_i - \mu)}{f(\boldsymbol{x};\mu)} = \sum_{i=1}^n (x_i - \mu), \\ \frac{f^{(2)}(\boldsymbol{x};\mu)}{f(\boldsymbol{x};\mu)} &= \frac{f(\boldsymbol{x};\mu) \left(\sum_{i=1}^n (x_i - \mu)\right)^2 - n\, f(\boldsymbol{x};\mu)}{f(\boldsymbol{x};\mu)} = \left(\sum_{i=1}^n (x_i - \mu) \right)^2 - n,\end{aligned}$$

which gives

$$\begin{aligned}V_{1,1}(\mu) &= J_n(\mu) = n, \\ V_{1,2}(\mu) &= \mathsf{E}_\mu \left(\frac{f^{(1)}(\boldsymbol{X};\mu)}{f(\boldsymbol{X};\mu)} \frac{f^{(2)}(\boldsymbol{X};\mu)}{f(\boldsymbol{X};\mu)} \right) \\ &= \mathsf{E}_\mu \left(\sum_{i=1}^n (X_i - \mu) \left(\left(\sum_{i=1}^n (X_i - \mu) \right)^2 - n \right) \right) = 0, \\ V_{2,2}(\mu) &= \mathsf{E}_\mu \left(\frac{f^{(2)}(\boldsymbol{X};\mu)}{f(\boldsymbol{X};\mu)} \right)^2 = \mathsf{E}_\mu \left(\left(\sum_{i=1}^n (X_i - \mu) \right)^2 - n \right)^2 = 2\,n^2.\end{aligned}$$

The matrix $\boldsymbol{V}(\mu)$ takes the form

$$\boldsymbol{V}(\mu) = \begin{pmatrix} n & 0 \\ 0 & 2\,n^2 \end{pmatrix},$$

and the Bhattacharya bound from Theorem 5 is

$$\mathsf{var}_\mu\, T_n \geq \left(g'(\mu), g^{(2)}(\mu) \right) \boldsymbol{V}(\mu)^{-1} \left(g'(\mu), g^{(2)}(\mu) \right)^\mathsf{T} = \frac{4\,\mu^2}{n} + \frac{2}{n^2},$$

which matches the variance of T_n. Therefore, we have found that T_n is the best unbiased estimator of $g(\mu) = \mu^2$, even though it does not attain the Rao-Cramér bound. △

1.2.2 Multi-dimensional parameter

Having dealt with the one-dimensional parameter θ, we now turn to the multi-dimensional situation. Suppose that we have a model parametrized by $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ with $p \in \mathbb{N}$. Our interest is in the estimation of either a one-dimensional parametric function $g(\boldsymbol{\theta}) \in \mathbb{R}$ with $g\colon \boldsymbol{\Theta} \to \mathbb{R}$ measurable, or a simultaneous estimation of a vector of parameters $\boldsymbol{g}(\boldsymbol{\theta}) \in \mathbb{R}^k$ for a measurable map $\boldsymbol{g}\colon \boldsymbol{\Theta} \to \mathbb{R}^k$.

We now extend the theory from the situation $p = 1$ to the general setup. We begin with the generalized Fisher information.

Definition 6 (Regular system of densities — multi-dimensional parameter)**.** Let the distribution of the random vector $\boldsymbol{X}$ depend only on parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\mathsf{T} \in \boldsymbol{\Theta}$. Suppose that $\boldsymbol{X}$ has a density $f(\boldsymbol{x}; \boldsymbol{\theta})$ w.r.t. a given σ-finite measure μ. The system of densities $\{f(\boldsymbol{x}; \boldsymbol{\theta})\colon \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is called a *regular system of densities* if the following conditions hold:

(R_1) The parameter space $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$ is a non-empty open set.

(R_2) The set $M = \{\boldsymbol{x}\colon f(\boldsymbol{x}; \boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta}$.

(R_3) For μ-almost all $\boldsymbol{x} \in M$ and all $j = 1, \ldots, p$ there exist finite partial derivatives

$$f'_j(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{\partial f(\boldsymbol{x}; \boldsymbol{\theta})}{\partial \theta_j}.$$

(R_4) For all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and all $j = 1, \ldots, p$ we can write $\int_M f'_j(\boldsymbol{x}; \boldsymbol{\theta}) \,\mathrm{d}\,\mu(\boldsymbol{x}) = 0$.

(R_5) For all $j, k = 1, \ldots, p$ the integral

$$J_{j,k,n}(\boldsymbol{\theta}) = \int_M \frac{f'_j(\boldsymbol{x}; \boldsymbol{\theta})}{f(\boldsymbol{x}; \boldsymbol{\theta})} \frac{f'_k(\boldsymbol{x}; \boldsymbol{\theta})}{f(\boldsymbol{x}; \boldsymbol{\theta})} f(\boldsymbol{x}; \boldsymbol{\theta}) \,\mathrm{d}\,\mu(\boldsymbol{x})$$

is finite for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

(R_6) The matrix $\boldsymbol{J}_n(\boldsymbol{\theta}) = (J_{j,k,n}(\boldsymbol{\theta}))_{j,k=1}^p$ is positive definite for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

For a regular system of densities, the matrix-valued function $\boldsymbol{J}_n\colon \boldsymbol{\Theta} \to \mathbb{R}^{p \times p}$ from (2) that for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ takes values in the space of positive definite matrices is called the *Fisher information matrix* of $\boldsymbol{\theta}$ contained in $\boldsymbol{X}$.

For the Fisher information matrix, results analogous to Theorems 1 and 2 hold true. We provide explicitly only the first one, but also the statement on the Fisher information matrix of independent random vectors from Theorem 2 generalizes readily, using an analogous proof. In particular, for a single random vector $\boldsymbol{X}$ whose n elements constitute a random sample we have $\boldsymbol{J}_n(\boldsymbol{\theta}) = n\,\boldsymbol{J}_1(\boldsymbol{\theta})$.

Theorem 7. *Let $\{f(\boldsymbol{x};\boldsymbol{\theta})\colon \boldsymbol{\theta}\in\boldsymbol{\Theta}\}$ be a regular system of densities. Suppose further that the second derivatives*

$$f''_{j,k}(\boldsymbol{x};\boldsymbol{\theta}) = \frac{\partial^2 f(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}$$

exist for μ-almost all $\boldsymbol{x}\in M$ and all $j,k=1,\dots,p$. Also, let a condition analogous to (R_4)

$$\int_M f''_{j,k}(\boldsymbol{x};\boldsymbol{\theta})\,\mathrm{d}\,\mu(\boldsymbol{x}) = 0$$

be valid for all $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ and $j,k=1,\dots,p$. Then we can write

$$J_{j,k,n}(\boldsymbol{\theta}) = -\int_M \frac{\partial^2 \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k} f(\boldsymbol{x};\boldsymbol{\theta})\,\mathrm{d}\,\mu(\boldsymbol{x}) = -\mathsf{E}_{\boldsymbol{\theta}}\,\frac{\partial^2 \log f(\boldsymbol{X};\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}.$$

Proof. Completely analogous to that of Theorem 1. □

Example 1.11. Let $\boldsymbol{X} = (X_1,\dots,X_n)^{\mathsf{T}}$ be a random sample from distribution $\mathsf{N}\left(\mu,\sigma^2\right)$, and take $\boldsymbol{\theta} = \left(\mu,\sigma^2\right)^{\mathsf{T}} \in \boldsymbol{\Theta} = \mathbb{R}\times(0,\infty)$. To compute the Fisher information matrix of $\boldsymbol{\theta}$ we take the the density of X_1 and use Theorem 7. We get

$$\begin{aligned} f(x;\boldsymbol{\theta}) &= \frac{1}{\sqrt{2\,\pi\,\sigma^2}}\exp\left(-\frac{1}{2\,\sigma^2}\,(x-\mu)^2\right),\\ \log f(x;\boldsymbol{\theta}) &= -\frac{1}{2}\log\left(2\,\pi\right) - \frac{1}{2}\log\left(\sigma^2\right) - \frac{1}{2\,\sigma^2}\,(x-\mu)^2, \end{aligned}$$

from which we get

$$\begin{aligned} \frac{\partial \log f(x;\boldsymbol{\theta})}{\partial\mu} &= \frac{x-\mu}{\sigma^2},\\ \frac{\partial \log f(x;\boldsymbol{\theta})}{\partial\sigma^2} &= -\frac{1}{2\,\sigma^2} + \frac{(x-\mu)^2}{2\,(\sigma^2)^2}, \end{aligned}$$

and

$$\begin{aligned} \frac{\partial^2 \log f(x;\boldsymbol{\theta})}{\partial\mu^2} &= -\frac{1}{\sigma^2},\\ \frac{\partial^2 \log f(x;\boldsymbol{\theta})}{\partial\mu\,\partial\sigma^2} &= -\frac{x-\mu}{(\sigma^2)^2},\\ \frac{\partial^2 \log f(x;\boldsymbol{\theta})}{\partial\,(\sigma^2)^2} &= \frac{1}{2\,(\sigma^2)^2} - \frac{(x-\mu)^2}{(\sigma^2)^3}. \end{aligned}$$

Now we replace x by the random variable X_1 and take the negative expectation of the previous formulae to get

$$\begin{aligned} J_{1,1,1}(\boldsymbol{\theta}) &= -\mathsf{E}_{\boldsymbol{\theta}}\,\frac{\partial^2 \log f(X_1;\boldsymbol{\theta})}{\partial\mu^2} = \frac{1}{\sigma^2},\\ J_{1,2,1}(\boldsymbol{\theta}) &= -\mathsf{E}_{\boldsymbol{\theta}}\,\frac{\partial^2 \log f(X_1;\boldsymbol{\theta})}{\partial\mu\,\partial\sigma^2} = 0,\\ J_{2,2,1}(\boldsymbol{\theta}) &= -\mathsf{E}_{\boldsymbol{\theta}}\,\frac{\partial^2 \log f(X_1;\boldsymbol{\theta})}{\partial\,(\sigma^2)^2} = \frac{1}{2\,\sigma^4}. \end{aligned}$$

Together, we computed that

$$\boldsymbol{J}_1(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\,\sigma^4} \end{pmatrix},$$

and $\boldsymbol{J}_n(\boldsymbol{\theta}) = n\,\boldsymbol{J}_1(\boldsymbol{\theta})$. Of course, the diagonal terms of the Fisher information matrix agree with the Fisher informations of the individual elements μ and σ^2 from Example 1.6. △

Having introduced the Fisher information matrix, we provide two versions of the theorem of Rao and Cramér for multi-dimensional parameters. The first theorem deals with a scalar-valued parametric function; the latter with a vector-valued estimated parameter.

Theorem 8 (Rao-Cramér, multi-dimensional I)**.** *Let $T_n = T_n(\boldsymbol{X})$ be an unbiased estimator of a (scalar) parametric function $g(\boldsymbol{\theta})$ that satisfies $\mathsf{var}_{\boldsymbol{\theta}}\, T_n < \infty$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$. Let the following conditions be satisfied:*

(RC_1) the system of densities $\{f(\boldsymbol{x};\boldsymbol{\theta})\colon \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ of $\boldsymbol{X}$ is regular;

(RC_2) the partial derivatives $g'_j(\boldsymbol{\theta}) = \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_j}$ of g exist for every $j = 1, \dots, p$ in every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$;

(RC_3) the following interchange of a derivative and an integral is valid for all $j = 1, \dots, p$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

$$\frac{\partial}{\partial \theta_j} \int_M T_n(\boldsymbol{x}) f(\boldsymbol{x};\boldsymbol{\theta})\,\mathrm{d}\,\mu(\boldsymbol{x}) = \int_M T_n(\boldsymbol{x}) \frac{\partial f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_j}\,\mathrm{d}\,\mu(\boldsymbol{x}).$$

Denote $\boldsymbol{h}(\boldsymbol{\theta}) = \big(g'_1(\boldsymbol{\theta}), \dots, g'_p(\boldsymbol{\theta})\big)^{\mathsf{T}}$. Then we can bound

$$\mathsf{var}_{\boldsymbol{\theta}}\, T_n = \mathsf{E}_{\boldsymbol{\theta}}\, (T_n - g(\boldsymbol{\theta}))^2 \geq \boldsymbol{h}(\boldsymbol{\theta})^{\mathsf{T}}\, (\boldsymbol{J}_n(\boldsymbol{\theta}))^{-1}\, \boldsymbol{h}(\boldsymbol{\theta}) \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

Proof. The proof method is quite similar to that of the usual Rao-Cramér theorem (Theorem 3) and the Bhattacharya theorem (Theorem 5). We use the score statistics

$$S_j = \frac{\partial \log f(\boldsymbol{X};\boldsymbol{\theta})}{\partial \theta_j} \text{ for } j = 1, \dots, p.$$

By (R_4) we know that $\mathsf{E}_{\boldsymbol{\theta}}\, S_j = 0$ for each $j = 1, \dots, p$. Computing the covariance matrix of the random vector $(T_n, S_1, \dots, S_p)^{\mathsf{T}}$ we obtain, using (RC_3), the fact that T_n is an unbiased estimator of $g(\boldsymbol{\theta})$, and (R_4), that

$$\begin{aligned}
\mathsf{cov}_{\boldsymbol{\theta}}\,(T_n, S_j) &= \int_M (T_n(\boldsymbol{x}) - g(\boldsymbol{\theta})) f'_j(\boldsymbol{x};\boldsymbol{\theta})\,\mathrm{d}\,\mu(\boldsymbol{x}) \\
&= \frac{\partial}{\partial \theta_j} \int_M T_n(\boldsymbol{x}) f(\boldsymbol{x};\boldsymbol{\theta})\,\mathrm{d}\,\mu(\boldsymbol{x}) - g(\boldsymbol{\theta}) \int_M f'_j(\boldsymbol{x};\boldsymbol{\theta})\,\mathrm{d}\,\mu(\boldsymbol{x}) \\
&= g'_j(\boldsymbol{\theta}), \\
\mathsf{cov}_{\boldsymbol{\theta}}\,(S_j, S_k) &= \int_M \frac{f'_j(\boldsymbol{x};\boldsymbol{\theta})}{f(\boldsymbol{x};\boldsymbol{\theta})} \frac{f'_k(\boldsymbol{x};\boldsymbol{\theta})}{f(\boldsymbol{x};\boldsymbol{\theta})} f(\boldsymbol{x};\boldsymbol{\theta})\,\mathrm{d}\,\mu(\boldsymbol{x}) = J_{j,k,n}(\boldsymbol{\theta})
\end{aligned}$$

for each $j, k = 1, \ldots, p$. The covariance matrix thus takes the form

$$\mathsf{cov}_{\boldsymbol{\theta}} \left((T_n, S_1, \ldots, S_p)^{\mathsf{T}} \right) = \begin{pmatrix} \mathsf{var}_{\boldsymbol{\theta}}\, T_n & \boldsymbol{h}(\boldsymbol{\theta})^{\mathsf{T}} \\ \boldsymbol{h}(\boldsymbol{\theta}) & \boldsymbol{J}_n(\boldsymbol{\theta}) \end{pmatrix}. \tag{16}$$

Exactly in the same way as argued in the proof of Theorem 5, the positive semi-definiteness of this matrix gives the desired bound. □

In formula (16) in the proof of Theorem 8 we see that the Fisher information matrix is, in fact, the covariance matrix of the p-dimensional random vector

$$\boldsymbol{S} = \frac{\partial \log f(\boldsymbol{X}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{17}$$

As such, it must be always positive semi-definite, and the additional requirement (R_6) only states that it should not be singular for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. The random vector (17) is sometimes called the *score*, or the *score statistic*. Note, however, that the latter term is not appropriate as the score in general still depends on the unknown value of the parameter $\boldsymbol{\theta}$, and thus is formally speaking not a statistic. The term score statistic is therefore more fitting in the situation when the unknown $\boldsymbol{\theta}$ is in (17) substituted by a fixed, known value $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, as we will do in the second part of this course. For $\boldsymbol{x}$ fixed, when considered as a function of the parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the function

$$\boldsymbol{\Theta} \to \mathbb{R} \colon \boldsymbol{\theta} \mapsto \frac{\partial \log f(\boldsymbol{X}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

is also called the *score function*. The scores and the score functions will be of great importance in the maximum likelihood estimation.

Example 1.12. For a random sample $X_1, \ldots, X_n$ from $\mathsf{N}(\mu, \sigma^2)$ and $\boldsymbol{\theta} = \left(\mu, \sigma^2\right)^{\mathsf{T}}$ as in Example 1.11 we have

$$\boldsymbol{J}_n(\boldsymbol{\theta}) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\,\sigma^4} \end{pmatrix}.$$

Consider a parametric function $g(\boldsymbol{\theta}) = \mu + c\,\sigma$ with $c > 0$ fixed and given. In the notation from Theorem 8 we therefore get

$$\boldsymbol{h}(\boldsymbol{\theta})^{\mathsf{T}} = \left(g_1'(\boldsymbol{\theta}), g_2'(\boldsymbol{\theta})\right)^{\mathsf{T}} = \left(1, \frac{c}{2\,\sigma}\right)^{\mathsf{T}},$$

and by Theorem 8 we have for any unbiased estimator T_n of $g(\boldsymbol{\theta})$ the Rao-Cramér bound

$$\mathsf{var}_{\boldsymbol{\theta}}\, T_n \geq \boldsymbol{h}(\boldsymbol{\theta})^{\mathsf{T}} \left(\boldsymbol{J}_n(\boldsymbol{\theta})\right)^{-1} \boldsymbol{h}(\boldsymbol{\theta}) = \frac{\sigma^2}{n} \left(1 + \frac{c^2}{2}\right) \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

△

We conclude this section by another extension of the Rao-Cramér theorem, this time to the setup of vector-valued parametric functions. For that result, we first provide an auxiliary lemma about the characterization of positive semi-definiteness of block matrices.

Lemma 5 (Schur's complement lemma)**.** *Let $\boldsymbol{M}$ be any symmetric block matrix*

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^\mathsf{T} & \boldsymbol{D} \end{pmatrix},$$

such that $\boldsymbol{D}$ is positive definite. Then $\boldsymbol{M}$ is positive semi-definite if and only if $\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{B}^\mathsf{T}$ is positive semi-definite.

Proof. It is easy to verify that we can write

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^\mathsf{T} & \boldsymbol{D} \end{pmatrix} = \underbrace{\begin{pmatrix} \boldsymbol{I} & \boldsymbol{B}\boldsymbol{D}^{-1} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix}}_{=\boldsymbol{N}^\mathsf{T}} \underbrace{\begin{pmatrix} \boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{B}^\mathsf{T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} \end{pmatrix}}_{=\boldsymbol{T}} \underbrace{\begin{pmatrix} \boldsymbol{I} & \boldsymbol{B}\boldsymbol{D}^{-1} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix}^\mathsf{T}}_{=\boldsymbol{N}},$$

where

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{B}^\mathsf{T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} \end{pmatrix}$$

is a symmetric square matrix. Since obviously

$$\begin{pmatrix} \boldsymbol{I} & \boldsymbol{B}\boldsymbol{D}^{-1} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{I} & -\boldsymbol{B}\boldsymbol{D}^{-1} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix},$$

the matrix

$$\boldsymbol{N} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{B}\boldsymbol{D}^{-1} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix}^\mathsf{T}$$

is invertible. We found that we can write

$$\boldsymbol{M} = \boldsymbol{N}^\mathsf{T}\boldsymbol{T}\boldsymbol{N}.$$

For any compatible vector $\boldsymbol{x} \neq \boldsymbol{0}$ we thus have

$$\boldsymbol{x}^\mathsf{T}\boldsymbol{M}\boldsymbol{x} = \boldsymbol{x}^\mathsf{T}\boldsymbol{N}^\mathsf{T}\boldsymbol{T}\boldsymbol{N}\boldsymbol{x} = (\boldsymbol{N}\boldsymbol{x})^\mathsf{T}\boldsymbol{T}(\boldsymbol{N}\boldsymbol{x})^\mathsf{T},$$

meaning that $\boldsymbol{M}$ is positive semi-definite if and only if $\boldsymbol{T}$ is. Because $\boldsymbol{D}$ is assumed to be positive definite, we get that $\boldsymbol{M}$ is positive semi-definite if and only if the first block on the diagonal of $\boldsymbol{T}$, which is $\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{B}^\mathsf{T}$, is positive semi-definite. □

Theorem 9 (Rao-Cramér, multi-dimensional II)**.** *Let $\boldsymbol{T}_n = \boldsymbol{T} = (T_1, \ldots, T_k)^\mathsf{T} = \boldsymbol{T}(\boldsymbol{X})$ be an unbiased estimator of a vector-valued parametric function $\boldsymbol{g}(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), \ldots, g_k(\boldsymbol{\theta}))^\mathsf{T}$, where $\boldsymbol{g}\colon \boldsymbol{\Theta} \to \mathbb{R}^k$ for $k \in \mathbb{N}$. Suppose that for each $j = 1, \ldots, k$ this estimator satisfies $\mathsf{var}_{\boldsymbol{\theta}}\, T_j < \infty$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$. Let the following conditions be satisfied:*

(RC$_1$) the system of densities $\{f(\boldsymbol{x};\boldsymbol{\theta})\colon \boldsymbol{\theta}\in\boldsymbol{\Theta}\}$ *of* $\boldsymbol{X}$ *is regular;*

(RC$_2$) all partial derivatives

$$\frac{\partial g_j(\boldsymbol{\theta})}{\partial\theta_\ell} \text{ for } j=1,\dots,k \text{ and } \ell=1,\dots,p$$

of $\boldsymbol{g}$ *exist for every* $\boldsymbol{\theta}\in\boldsymbol{\Theta}$*;*

(RC$_3$) the following interchange of a derivative and an integral is valid for all $j=1,\dots,k$, $\ell=1,\dots,p$ *and* $\boldsymbol{\theta}\in\boldsymbol{\Theta}$

$$\frac{\partial}{\partial\theta_\ell}\int_M T_j(\boldsymbol{x})f(\boldsymbol{x};\boldsymbol{\theta})\,\mathrm{d}\,\mu(\boldsymbol{x}) = \int_M T_j(\boldsymbol{x})\frac{\partial f(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_\ell}\,\mathrm{d}\,\mu(\boldsymbol{x}).$$

Denote

$$\boldsymbol{H}(\boldsymbol{\theta}) = \left(\frac{\partial g_j(\boldsymbol{\theta})}{\partial\theta_\ell}\right)_{j=1,\ell=1}^{k,p}.$$

Then we can bound for $\mathsf{var}_{\boldsymbol{\theta}}\,\boldsymbol{T}_n = \mathsf{E}_{\boldsymbol{\theta}}\,(\boldsymbol{T}_n-\boldsymbol{g}(\boldsymbol{\theta}))\,(\boldsymbol{T}_n-\boldsymbol{g}(\boldsymbol{\theta}))^{\mathsf{T}}$

$$\mathsf{var}_{\boldsymbol{\theta}}\,\boldsymbol{T}_n - \boldsymbol{H}(\boldsymbol{\theta})\,(\boldsymbol{J}_n(\boldsymbol{\theta}))^{-1}\,\boldsymbol{H}(\boldsymbol{\theta})^{\mathsf{T}} \geq 0 \text{ for all } \boldsymbol{\theta}\in\boldsymbol{\Theta}.$$

The last inequality means that the resulting $k\times k$*-matrix on the left hand side is positive semi-definite.*

Proof. The proof is completely analogous to that of Theorem 8. We compute the covariance matrix of the vector $(T_1,\dots,T_k,S_1,\dots,S_p)^{\mathsf{T}}$ and establish that it takes the form

$$\begin{pmatrix}\mathsf{var}_{\boldsymbol{\theta}}\,\boldsymbol{T}_n & \boldsymbol{H}(\boldsymbol{\theta})\\ \boldsymbol{H}(\boldsymbol{\theta})^{\mathsf{T}} & \boldsymbol{J}_n(\boldsymbol{\theta})\end{pmatrix}.$$

To conclude, we apply the Schur complement lemma stated as Lemma 5 for this positive semi-definite matrix. □

1.3 Sufficiency and its role in estimation

We now turn to the problem of finding best unbiased estimators of parametric functions. In this respect, an eminent role is played by the concept of sufficiency. Intuitively, a sufficient statistic extracts all the (Fisher) information about a parameter $\boldsymbol{\theta}$ from the random vector $\boldsymbol{X}$ whose distribution depends on $\boldsymbol{\theta}$. Formally, it is defined as follows.

Definition 7. A statistic $\boldsymbol{S}=\boldsymbol{S}(\boldsymbol{X})$ is called *sufficient* for parameter $\boldsymbol{\theta}$ if the conditional distribution of the random vector $\boldsymbol{X}$ given $\boldsymbol{S}$ does not depend on $\boldsymbol{\theta}$.

Sufficient statistics are of great importance in the estimation theory. They allow us to reduce the problem of finding the best unbiased estimators to the task of finding unbiased functions of certain sufficient statistics. The simplest sufficient statistic is the vector $\boldsymbol{X}$ itself, as obviously the distribution of $\boldsymbol{X}$ given $\boldsymbol{X} = \boldsymbol{x}$ is the trivial Dirac measure at $\boldsymbol{x}$, which does not involve $\boldsymbol{\theta}$ at all.

Example 1.13. Let $\boldsymbol{X} = (X_1, X_2)^\mathsf{T}$ be a vector composed of $X_1, X_2 \sim \mathsf{Bernoulli}(p)$ with $p \in (0,1)$, independent of each other. The vector $\boldsymbol{X}$ is distributed according to the following frequency table

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	$(1-p)^2$	$p(1-p)$
$X_1 = 1$	$p(1-p)$	p^2

Take $T = X_1 + X_2$, and find the conditional distribution of $\boldsymbol{X}$ given $T = t$. The random variable T takes one of the values $t = 0, 1, 2$, almost surely. Trivially, for $t = 0$ we get that the conditional distribution $(\boldsymbol{X} \mid T = 0)$ corresponds to the constant $(0,0)^\mathsf{T}$ almost surely, and analogously the distribution of $(\boldsymbol{X} \mid T = 2)$ is the Dirac measure at $(1,1)^\mathsf{T}$. In the remaining situation $t = 1$ we have

$$\mathsf{P}_p\left(\boldsymbol{X} = (0,1)^\mathsf{T} \mid T = 1\right) = \frac{\mathsf{P}_p\left(\boldsymbol{X} = (0,1)^\mathsf{T}\right)}{\mathsf{P}_p\left(T = 1\right)} = \frac{p(1-p)}{2\,p(1-p)} = 1/2,$$
$$\mathsf{P}_p\left(\boldsymbol{X} = (1,0)^\mathsf{T} \mid T = 1\right) = \frac{\mathsf{P}_p\left(\boldsymbol{X} = (1,0)^\mathsf{T}\right)}{\mathsf{P}_p\left(T = 1\right)} = \frac{p(1-p)}{2\,p(1-p)} = 1/2.$$

Thus, the distribution of $(\boldsymbol{X} \mid T = 1)$ is uniform in the two points $(0,1)^\mathsf{T}$ and $(1,0)^\mathsf{T}$. Overall, the distribution of $\boldsymbol{X}$ given T does not depend on the parameter p, and T is a sufficient statistic for p. △

Observe that since any statistic $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$ is a function of $\boldsymbol{X}$, it necessarily reduces (or, more precisely, cannot increase) the information available in the random vector $\boldsymbol{X}$. Indeed, knowing only the observed value $\boldsymbol{t}$ of $\boldsymbol{T}$ we are usually unable to recover the original observed value $\boldsymbol{x}$ of the random vector $\boldsymbol{X}$. In Example 1.13, for instance, if $t = 0$ or $t = 2$, we identify the observed value $\boldsymbol{x}$ of $\boldsymbol{X}$ uniquely — it is $\boldsymbol{x} = (0,0)^\mathsf{T}$ if $t = 0$, or $\boldsymbol{x} = (1,1)^\mathsf{T}$ if $t = 2$. In the situation when $t = 1$, we are however not able to determine whether $\boldsymbol{x} = (0,1)^\mathsf{T}$ or $\boldsymbol{x} = (1,0)^\mathsf{T}$ was observed. Naturally, any statistic $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$ therefore introduces a certain "partitioning" of the sample space into collections of sets

$$\boldsymbol{T}^{-1}(\boldsymbol{t}) = \{\boldsymbol{x} \colon \boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{t}\} \text{ for all possible values } \boldsymbol{t} \text{ of } \boldsymbol{T}. \tag{18}$$

Roughly speaking, the larger these sets are, the greater reduction of information occurs when working with $\boldsymbol{T}$ instead of with $\boldsymbol{X}$.

From this perspective, the intuitive meaning of sufficiency becomes more clear. The distribution of $\boldsymbol{X}$ depends on our parameter of interest $\boldsymbol{\theta}$. We choose a statistic $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$. Directly from the definition of the conditional distribution, we know that knowing (i) the distribution of $\boldsymbol{X}$ is equivalent with the knowledge of (ii) the marginal distribution of $\boldsymbol{T}$, and the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{T}$. If the statistic $\boldsymbol{T}$ is sufficient, the conditional distribution $(\boldsymbol{X} \mid \boldsymbol{T})$ does not depend on $\boldsymbol{\theta}$ at all. Therefore, everything we can obtain from $\boldsymbol{X}$ about $\boldsymbol{\theta}$ is contained in the distribution of $\boldsymbol{T}$ alone.

Take again Example 1.13, and suppose that we know only the distribution of T, which is of course the binomial distribution $T \sim \mathsf{Bi}(2, p)$. Because T is sufficient for p and the distribution of $\boldsymbol{X}$ given T does not depend on p, we could use the (completely known) conditional distribution of $(\boldsymbol{X} \mid T)$ and reconstruct the whole distribution of the random vector $\boldsymbol{X}$. But, why would we do this? In the process of reconstruction of $\boldsymbol{X}$ we would only add randomness that does not involve p to our problem. Therefore, it appears that using sufficient statistics allows us to simplify the problem of inference about $\boldsymbol{\theta}$ substantially; instead of the whole observed vector $\boldsymbol{X}$, we can only extract a (typically much "smaller") statistic $\boldsymbol{T}$, without losing anything interesting about $\boldsymbol{\theta}$.

The end of lecture 3 (5.3.2024)

The easiest way to determine whether a statistic $\boldsymbol{S}$ is sufficient is given by the following criterion. In full generality, its proof is rather technical and is given in e.g. [9, Corollary 2.6.1].

Theorem 10 (Neyman's factorization criterion)**.** *Let $\boldsymbol{X}$ be a random vector with density $f(\boldsymbol{x}; \boldsymbol{\theta})$ w.r.t. a σ-finite measure μ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Then the statistic $\boldsymbol{S} = \boldsymbol{S}(\boldsymbol{X})$ is sufficient for $\boldsymbol{\theta}$ if and only if there exist measurable functions g and h such that*

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = g\left(\boldsymbol{S}(\boldsymbol{x}), \boldsymbol{\theta}\right) h(\boldsymbol{x}) \text{ for } \mu\text{-almost all } \boldsymbol{x}.$$

Proof. We prove the theorem only for $\boldsymbol{X}$ discrete. In that case, also $\boldsymbol{S}$ is necessarily discrete, and we can take μ to be the counting measure on the union of the supports of $\boldsymbol{X}$ and $\boldsymbol{S}$, which must be an at most countable set.

Let $\boldsymbol{S}$ be sufficient, and take $\boldsymbol{x}$ such that $\mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{X} = \boldsymbol{x}\right) > 0$. Then $\mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{S}(\boldsymbol{X}) = \boldsymbol{S}(\boldsymbol{x})\right) \geq \mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{X} = \boldsymbol{x}\right) > 0$, and the density rewrites into

$$\mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{X} = \boldsymbol{x}\right) = \mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{S}(\boldsymbol{X}) = \boldsymbol{S}(\boldsymbol{x})\right) = \mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{S}(\boldsymbol{X}) = \boldsymbol{S}(\boldsymbol{x})\right) \mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{S}(\boldsymbol{X}) = \boldsymbol{S}(\boldsymbol{x})\right).$$

Take $g(\boldsymbol{S}(\boldsymbol{x}), \boldsymbol{\theta}) = \mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{S}(\boldsymbol{X}) = \boldsymbol{S}(\boldsymbol{x})\right)$ and note that indeed, g depends on $\boldsymbol{x}$ only through $\boldsymbol{S}(\boldsymbol{x})$. Further, with $h(\boldsymbol{x}) = \mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{S}(\boldsymbol{X}) = \boldsymbol{S}(\boldsymbol{x})\right)$ we know that, by the assumption of sufficiency of $\boldsymbol{S}$, the term on the right hand side does not depend on $\boldsymbol{\theta}$, and h is therefore only a function of $\boldsymbol{x}$. We have factorized the density as required.

For the other implication, suppose that the density f factorizes as in the statement of the theorem. For any $\boldsymbol{s}$ we have

$$\mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{S}(\boldsymbol{X})=\boldsymbol{s}\right)=\sum_{\{\boldsymbol{x}\colon \boldsymbol{S}(\boldsymbol{x})=\boldsymbol{s}\}}\mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{X}=\boldsymbol{x}\right)=g(\boldsymbol{s},\boldsymbol{\theta})\sum_{\{\boldsymbol{x}\colon \boldsymbol{S}(\boldsymbol{x})=\boldsymbol{s}\}}h(\boldsymbol{x}). \tag{19}$$

We compute the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{S}$ directly. If $\mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{S}(\boldsymbol{X})=\boldsymbol{s}\right)>0$ we have

$$\mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{X}=\boldsymbol{x}\mid\boldsymbol{S}(\boldsymbol{X})=\boldsymbol{s}\right)=\frac{\mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{X}=\boldsymbol{x},\boldsymbol{S}(\boldsymbol{X})=\boldsymbol{s}\right)}{\mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{S}(\boldsymbol{X})=\boldsymbol{s}\right)}=\begin{cases}0 & \text{if } \boldsymbol{S}(\boldsymbol{x})\neq\boldsymbol{s},\\ \frac{g(\boldsymbol{s},\boldsymbol{\theta})\,h(\boldsymbol{x})}{g(\boldsymbol{s},\boldsymbol{\theta})\sum_{\{\boldsymbol{y}\colon \boldsymbol{S}(\boldsymbol{y})=\boldsymbol{s}\}}h(\boldsymbol{y})} & \text{if } \boldsymbol{S}(\boldsymbol{x})=\boldsymbol{s}.\end{cases}$$

Since $\mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{S}(\boldsymbol{X})=\boldsymbol{s}\right)>0$, the factor $g(\boldsymbol{s},\boldsymbol{\theta})$ cannot be zero if $\boldsymbol{S}(\boldsymbol{x})=\boldsymbol{s}$, and thus it cancels out. We have computed the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{S}(\boldsymbol{X})$, and see that it does not depend on $\boldsymbol{\theta}$. Therefore, $\boldsymbol{S}$ is sufficient. □

Remark 2 (Density of a sufficient statistic)**.** It is important to observe that in the proof of Theorem 10 we obtained the form of the density $f_{\boldsymbol{S}}(\boldsymbol{s};\boldsymbol{\theta})$ of the sufficient statistic $\boldsymbol{S}$. For $\boldsymbol{S}$ with a discrete distribution, we derived in (19) that this density takes the form

$$f_{\boldsymbol{S}}(\boldsymbol{s};\boldsymbol{\theta})=g(\boldsymbol{s},\boldsymbol{\theta})H(\boldsymbol{s}) \text{ for all } \boldsymbol{s}, \tag{20}$$

with g the function from the factorization of the density of $\boldsymbol{X}$, and H a function that does not depend on $\boldsymbol{\theta}$. An analogous result can be shown also more generally. For instance, for distributions with densities w.r.t. the Lebesgue measure in $\mathbb{R}^n$, this can be seen using the standard transformation of densities [4, Theorem 1 on p. 318]. Indeed, let $f(\boldsymbol{x};\boldsymbol{\theta})$ be the density of $\boldsymbol{X}$ and suppose for simplicity that $\boldsymbol{x}\mapsto\boldsymbol{S}(\boldsymbol{x})$ is invertible with an inverse function $\boldsymbol{S}^{-1}$. The density of $\boldsymbol{S}(\boldsymbol{X})$ at $\boldsymbol{s}$ is then given by

$$\begin{aligned}f_{\boldsymbol{S}}(\boldsymbol{s};\boldsymbol{\theta})&=f(\boldsymbol{S}^{-1}(\boldsymbol{s});\boldsymbol{\theta})\mathcal{J}_{\boldsymbol{S}^{-1}}(\boldsymbol{s})=g(\boldsymbol{S}(\boldsymbol{S}^{-1}(\boldsymbol{s}));\boldsymbol{\theta})\mathcal{J}_{\boldsymbol{S}^{-1}}(\boldsymbol{s})h(\boldsymbol{S}^{-1}(\boldsymbol{s}))\\&=g(\boldsymbol{s};\boldsymbol{\theta})\mathcal{J}_{\boldsymbol{S}^{-1}}(\boldsymbol{s})h(\boldsymbol{S}^{-1}(\boldsymbol{s})),\end{aligned}$$

for $\mathcal{J}_{\boldsymbol{S}^{-1}}$ the Jacobian determinant of the inverse function $\boldsymbol{S}^{-1}$. We see that (20) is true with $H(\boldsymbol{s})=\mathcal{J}_{\boldsymbol{S}^{-1}}(\boldsymbol{s})h(\boldsymbol{S}^{-1}(\boldsymbol{s}))$, which does not depend on $\boldsymbol{\theta}$.

Directly from the definition of the sufficient statistic we also see that any one-to-one mapping of a sufficient statistic is itself a sufficient statistic.

Theorem 11. *Let $\boldsymbol{S}=\boldsymbol{S}(\boldsymbol{X})$ be a sufficient statistic, and let $\boldsymbol{t}$ be any measurable mapping that does not depend on $\boldsymbol{\theta}$, which has an inverse $\boldsymbol{\tau}$. Then the statistic $\boldsymbol{T}=\boldsymbol{t}(\boldsymbol{S})=\boldsymbol{t}(\boldsymbol{S}(\boldsymbol{X}))$ is also sufficient.*

Proof. Follows directly from the definition of sufficiency, and the definition of the conditional distribution. Alternatively, one could use Theorem 10 and see that the density of $\boldsymbol{X}$ can be rewritten also in the form

$$f(\boldsymbol{x};\boldsymbol{\theta}) = g\left(\boldsymbol{S}(\boldsymbol{x}),\boldsymbol{\theta}\right) h(\boldsymbol{x}) = g\left(\boldsymbol{\tau}\left(\boldsymbol{T}(\boldsymbol{x})\right),\boldsymbol{\theta}\right) h(\boldsymbol{x}) \text{ for } \mu\text{-almost all } \boldsymbol{x}.$$

Set $g_1(\boldsymbol{T}(\boldsymbol{x}),\boldsymbol{\theta}) = g(\boldsymbol{\tau}\left(\boldsymbol{T}(\boldsymbol{x})\right),\boldsymbol{\theta})$ and observe that we found a factorization of the density for statistic $\boldsymbol{T}$ as in Theorem 10. Thus, also $\boldsymbol{T}$ is a sufficient statistic. □

Example 1.14. Take $\boldsymbol{X} = (X_1,\dots,X_n)^\mathsf{T}$ whose elements are independent random variables with distribution $\mathsf{N}\left(\mu,\sigma^2\right)$, for $\boldsymbol{\theta} = \left(\mu,\sigma^2\right)^\mathsf{T} \in \mathbb{R}\times(0,\infty)$. The joint density of $\boldsymbol{X}$ w.r.t. the Lebesgue measure in $\mathbb{R}^n$ is

$$\begin{aligned} f(\boldsymbol{x};\boldsymbol{\theta}) &= \prod_{i=1}^n \left(\frac{1}{\sqrt{2\,\pi\,\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\,\sigma^2}\right)\right) \\ &= \frac{1}{(2\,\pi\,\sigma^2)^{n/2}}\exp\left(-\frac{1}{2\,\sigma^2}\left(\sum_{i=1}^n x_i^2 - 2\,\mu\sum_{i=1}^n x_i + n\,\mu^2\right)\right) \end{aligned}$$

for $\boldsymbol{x} = (x_1,\dots,x_n)^\mathsf{T} \in \mathbb{R}^n$. Theorem 10 gives that a sufficient statistic for $\boldsymbol{\theta}$ is $\boldsymbol{S} = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)^\mathsf{T}$, with g the whole function f and $h(\boldsymbol{x}) \equiv 1$. By Theorem 11, another sufficient statistic for $\boldsymbol{\theta}$ is $\boldsymbol{T} = \left(\bar{X}_n, S_n^2\right)^\mathsf{T}$, where $\bar{X}_n$ is the sample average and $S_n^2 = \sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2/(n-1)$ is the sample variance of $\boldsymbol{X}$. By Remark 2 we are also able to determine the first factor $g(\boldsymbol{t},\boldsymbol{\theta})$ of the density of $\boldsymbol{T}$. This is not difficult to be verified to indeed correspond to the exact distribution of $\boldsymbol{T}$, which we know to be the product of $\mathsf{N}\left(\mu,\sigma^2/n\right)$ and a multiple of the χ^2_{n-1} distribution by the factor $\sigma^2/(n-1)$, see [6, Theorem 2.8]. △

A sufficient statistic $\boldsymbol{S}$ possesses the same Fisher information about $\boldsymbol{\theta}$ as the original random vector $\boldsymbol{X}$. In this sense, we see that also formally, no reduction of information occurs when working with sufficient statistics.

Theorem 12. *Let $\boldsymbol{X}$ correspond to a regular system of densities with Fisher information matrix $\boldsymbol{J}_n(\boldsymbol{\theta})$, and let $\boldsymbol{S} = \boldsymbol{S}(\boldsymbol{X})$ be sufficient for $\boldsymbol{\theta}$. Suppose that also $\boldsymbol{S}$ has a regular system of densities, and denote its Fisher information matrix by $\widetilde{\boldsymbol{J}}_n(\boldsymbol{\theta})$. Then $\widetilde{\boldsymbol{J}}_n(\boldsymbol{\theta}) = \boldsymbol{J}_n(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}\in\boldsymbol{\Theta}$.*

Proof. We give only an outline of the formal proof, relying on the statement of Remark 2 that we proved only partially.

Any regular system of densities is dominated by some σ-finite measure μ. By Theorem 10 we therefore get that the density of $\boldsymbol{X}$ factorizes into the product $f(\boldsymbol{x};\boldsymbol{\theta}) = g\left(\boldsymbol{S}(\boldsymbol{x}),\boldsymbol{\theta}\right)\, h(\boldsymbol{x})$,

where h does not depend on $\boldsymbol{\theta}$. The partial derivative f'_j of f w.r.t. θ_j is therefore for each $j, k = 1, \ldots, p$

$$f'_j(\boldsymbol{x}; \boldsymbol{\theta}) = g'_j\left(\boldsymbol{S}(\boldsymbol{x}), \boldsymbol{\theta}\right) h(\boldsymbol{x})$$

where g'_j is, of course, the partial derivative of g w.r.t. θ_j. For the Fisher information matrix of $\boldsymbol{X}$ we therefore get

$$J_{j,k,n}(\boldsymbol{\theta}) = \mathsf{E}_{\boldsymbol{\theta}}\left(\frac{f'_j(\boldsymbol{X}; \boldsymbol{\theta})}{f(\boldsymbol{X}; \boldsymbol{\theta})} \frac{f'_k(\boldsymbol{X}; \boldsymbol{\theta})}{f(\boldsymbol{X}; \boldsymbol{\theta})}\right) = \mathsf{E}_{\boldsymbol{\theta}}\left(\frac{g'_j(\boldsymbol{S}(\boldsymbol{X}), \boldsymbol{\theta})}{g(\boldsymbol{S}(\boldsymbol{X}), \boldsymbol{\theta})} \frac{g'_k(\boldsymbol{S}(\boldsymbol{X}), \boldsymbol{\theta})}{g(\boldsymbol{S}(\boldsymbol{X}), \boldsymbol{\theta})}\right).$$

It remains to realise that by our Remark 2, the density of the sufficient statistic $\boldsymbol{S}$ takes the form $g(\boldsymbol{s}, \boldsymbol{\theta})H(\boldsymbol{s})$, and H does not depend on $\boldsymbol{\theta}$. The last expression for $J_{j,k,n}(\boldsymbol{\theta})$ above therefore equals the element $\widetilde{J}_{j,k,n}(\boldsymbol{\theta})$ of the Fisher information matrix of $\boldsymbol{S} = \boldsymbol{S}(\boldsymbol{X})$, and the two matrices are the same. □

From what we saw, there are clearly many sufficient statistics. For example, any one-to-one map of the random vector $\boldsymbol{X}$ is sufficient. It will be interesting to find a sufficient statistic which, in a sense, is the smallest possible. It turns out that this notion of a "small" sufficient statistic is not captured well by the dimensionality of the statistic. Consider the following example, and compare it with Example 1.16 below.

Example 1.15. In the setup of Example 1.13 we found two sufficient statistics — the trivial $\boldsymbol{X} = (X_1, X_2)^\mathsf{T}$, and $T = X_1 + X_2$. We also saw that T reduces the information provided in $\boldsymbol{X}$, because from the knowledge of $T = 1$ we cannot recover the original value of $\boldsymbol{X}$. Therefore, T is in this sense "smaller" than $\boldsymbol{X}$. Take, however, another sufficient statistic $\boldsymbol{S} = \left(X_1 + X_2, (X_1 + X_2)^2, (X_1 + X_2)^3\right)^\mathsf{T}$. Because $X_1 + X_2 \in \{0, 1, 2\}$ almost surely, any element of $\boldsymbol{S}$ carries the same information about $\boldsymbol{X}$ as T does. Indeed, $\boldsymbol{S}$ takes only one of the values $(0, 0, 0)^\mathsf{T}$, $(1, 1, 1)^\mathsf{T}$, $(2, 4, 8)^\mathsf{T}$ almost surely, and using the preimage map (18) we get

$$\begin{aligned}
\boldsymbol{S}^{-1}\left((0,0,0)^\mathsf{T}\right) &= \{(0,0)\} = T^{-1}(0), \\
\boldsymbol{S}^{-1}\left((1,1,1)^\mathsf{T}\right) &= \{(0,1),(1,0)\} = T^{-1}(1), \\
\boldsymbol{S}^{-1}\left((2,4,8)^\mathsf{T}\right) &= \{(1,1)\}.
\end{aligned}$$

Therefore, the partitioning of the sample space of $\boldsymbol{X}$ induced by $\boldsymbol{S}$ is exactly the same as for the statistic T. In fact, knowing T and knowing $\boldsymbol{S}$ is equivalent; from the value of one of them we can obtain the other. Thus, even though the dimension of $\boldsymbol{S}$ is 3 and the dimension of T is 1, the information reduction obtained by $\boldsymbol{S}$ and T are the same. △

We follow Theorem 11 and define the smallest sufficient statistic to be one that can be written as a function of any other sufficient statistic.

Definition 8. A sufficient statistic $\boldsymbol{S}$ is called *minimal sufficient* if it can be written as a measurable function of any other sufficient statistic.

The following general lemma is useful for dealing with functions of statistics.

Lemma 6. *Let X, Y, Z be sets and let $g\colon X \to Y$ and $f\colon X \to Z$ be any functions. Then there exists a function $h\colon Y \to Z$ such that $f(x) = (h \circ g)(x) = h(g(x))$ for all $x \in X$ if and only if the following implication holds true for all $x, y \in X$:*

$$g(x) = g(y) \quad \textit{implies} \quad f(x) = f(y). \tag{21}$$

Proof. One direction is trivial; if the implication (21) is not true, then $h(g(x)) = h(g(y))$ would have to take both values $f(x)$ and $f(y)$, which is impossible. For the other direction, one defines $h(z) = f(g^{-1}(z))$ and verifies that if (21) holds true, then $h\colon Y \to Z$ is well-defined and satisfies $f = h \circ g$ as needed. □

Using the idea of preimages of statistics from (18), the definition of a minimal sufficient statistics is a natural one. The minimal sufficient statistic generates the "coarsest" set of preimages in the sample space of $\boldsymbol{X}$ that still corresponds to a sufficient statistic. In this sense, it reduces the information present in $\boldsymbol{X}$ to the maximum possible extent, without losing anything of interest about $\boldsymbol{\theta}$.

Example 1.16. Take $X \sim \mathsf{N}\left(0, \sigma^2\right)$ and let $\theta = \sigma^2$. Since we have only a single random variable X, one could guess that X is the minimal sufficient statistic. This is, however, not true, as $T = X^2$ is also sufficient for θ by Theorem 10, but X cannot be written as a function of T. △

Clearly, not every function of a sufficient statistic is sufficient. On the other hand, if $\boldsymbol{T}$ is sufficient, and $\boldsymbol{T}$ is a function of $\boldsymbol{S}$, then also $\boldsymbol{S}$ must be sufficient. In the following theorem we give a simple criterion on how to verify that a statistic is minimal sufficient.

Theorem 13 (Lehmann-Scheffé on minimal sufficient statistics)**.** *Let $\boldsymbol{X}$ be a random vector with density $f(\boldsymbol{x}; \boldsymbol{\theta})$ with respect to a σ-finite measure μ on $\mathbb{R}^n$, and let $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Suppose that the support $M = \{\boldsymbol{x} \in \mathbb{R}^n\colon f(\boldsymbol{x}; \boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Let $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$ be sufficient for $\boldsymbol{\theta}$, and denote*

$$h(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) = \frac{f(\boldsymbol{x}; \boldsymbol{\theta})}{f(\boldsymbol{y}; \boldsymbol{\theta})} \quad \textit{for } \boldsymbol{x}, \boldsymbol{y} \in M \textit{ and } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

Suppose that for every $\boldsymbol{x}, \boldsymbol{y} \in M$ the fraction $h(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$ implies that $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{T}(\boldsymbol{y})$. Then $\boldsymbol{T}(\boldsymbol{X})$ is a minimal sufficient statistic.

Proof. First note that for $\boldsymbol{y} \in M$ is the density $f(\boldsymbol{y};\boldsymbol{\theta})$ positive for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Thus, the fraction $h(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})$ is well defined.

We need to show that $\boldsymbol{T}$ is a function of any other sufficient statistic $\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{X})$ for $\boldsymbol{\theta}$. Thanks to the Neyman factorization criterion from Theorem 10 applied to $\boldsymbol{U}$ there exist functions g_0 and h_0 such that

$$f(\boldsymbol{x};\boldsymbol{\theta}) = g_0\left(\boldsymbol{U}(\boldsymbol{x}),\boldsymbol{\theta}\right)\, h_0(\boldsymbol{x}).$$

Suppose that $\boldsymbol{x},\boldsymbol{y} \in M$ are such that $\boldsymbol{U}(\boldsymbol{x}) = \boldsymbol{U}(\boldsymbol{y})$. Then necessarily $g_0\left(\boldsymbol{U}(\boldsymbol{x}),\boldsymbol{\theta}\right) = g_0\left(\boldsymbol{U}(\boldsymbol{y}),\boldsymbol{\theta}\right)$, and in the ratio of densities we have

$$h(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta}) = \frac{f(\boldsymbol{x};\boldsymbol{\theta})}{f(\boldsymbol{y};\boldsymbol{\theta})} = \frac{g_0\left(\boldsymbol{U}(\boldsymbol{x}),\boldsymbol{\theta}\right)\, h_0(\boldsymbol{x})}{g_0\left(\boldsymbol{U}(\boldsymbol{x}),\boldsymbol{\theta}\right)\, h_0(\boldsymbol{y})} = \frac{h_0(\boldsymbol{x})}{h_0(\boldsymbol{y})}.$$

The last inequality is valid because for $\boldsymbol{y} \in M$ we know that $f(\boldsymbol{y};\boldsymbol{\theta}) > 0$. We see that the fraction $h(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$, and by our assumptions it follows that $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{T}(\boldsymbol{y})$.

We have verified that for any $\boldsymbol{x}$ and $\boldsymbol{y}$ the equality $\boldsymbol{U}(\boldsymbol{x}) = \boldsymbol{U}(\boldsymbol{y})$ implies $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{T}(\boldsymbol{y})$. This means that by Lemma 6, $\boldsymbol{T}$ can be written as a function of $\boldsymbol{U}$ for $\boldsymbol{x} \in M$. Because $\mathsf{P}\left(\boldsymbol{X} \in M\right) = \int_M f(\boldsymbol{x};\boldsymbol{\theta})\,\mathrm{d}\,\mu(\boldsymbol{x}) = \int_{\mathbb{R}^n} f(\boldsymbol{x};\boldsymbol{\theta})\,\mathrm{d}\,\mu(\boldsymbol{x}) = 1$, it follows that $\boldsymbol{T}(\boldsymbol{X})$ is a function of $\boldsymbol{U}(\boldsymbol{X})$ almost surely, and we have proved that $\boldsymbol{T}(\boldsymbol{X})$ is a minimal sufficient statistic. □

The criterion from Theorem 13 gives only one implication. It does not allow us to conclude that a sufficient statistic is not minimal.

Example 1.17. In the setup of normal distributions $\mathsf{N}(\mu,\sigma^2)$ with $\boldsymbol{\theta} = \left(\mu,\sigma^2\right)^\mathsf{T}$ from Example 1.14 we have $M = \mathbb{R}^n$ and

$$h(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta}) = \frac{f(\boldsymbol{x};\boldsymbol{\theta})}{f(\boldsymbol{y};\boldsymbol{\theta})} = \exp\left(-\frac{1}{2\,\sigma^2}\left(\sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i^2\right) + \frac{\mu}{\sigma^2}\left(\sum_{i=1}^n x_i - \sum_{i=1}^n y_i\right)\right).$$

The last expression does not depend on $\boldsymbol{\theta}$ if and only if $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$ and at the same time $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2$. Thus, Theorem 13 gives that the statistic $\boldsymbol{S} = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)^\mathsf{T}$ is minimal sufficient. Is the other sufficient statistic $\boldsymbol{T} = \left(\bar{X}_n, S_n^2\right)^\mathsf{T}$ from Example 1.14 minimal sufficient? △

The final important notion is that of a complete statistic.

Definition 9. A statistic $\boldsymbol{S}$ is said to be *complete* if for every measurable real-valued function w we have that

$$\mathsf{E}_{\boldsymbol{\theta}}\, w(\boldsymbol{S}) = 0 \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta} \text{ implies } w(\boldsymbol{S}) = 0 \text{ almost surely for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

The end of lecture 4 (12.3.2024)

If a minimal sufficient statistic exists, any complete sufficient statistic must be already minimal. The assumption of the existence of a minimal sufficient statistic is a weak one. Under mild regularity conditions it can be shown that a minimal sufficient statistic of a system of measures dominated by a σ-finite measure always exists.

Theorem 14. *Suppose that a minimal sufficient statistic exists. Let $\boldsymbol{S} = \boldsymbol{S}(\boldsymbol{X})$ be a complete sufficient statistic such that $\mathsf{E}_{\boldsymbol{\theta}}\, \boldsymbol{S}$ exists for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Then $\boldsymbol{S}$ is a minimal sufficient statistic.*

Proof. Let $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$ be any minimal sufficient statistic. Then $\boldsymbol{T}$ must be a function of the (complete) sufficient statistic $\boldsymbol{S}$, say $\boldsymbol{T} = \boldsymbol{h}(\boldsymbol{S})$. But then the difference $\boldsymbol{S} - \mathsf{E}\,(\boldsymbol{S} \mid \boldsymbol{T}) = \boldsymbol{S} - \mathsf{E}\,(\boldsymbol{S} \mid \boldsymbol{h}(\boldsymbol{S}))$ is also a function of $\boldsymbol{S}$ only, and at the same time $\mathsf{E}_{\boldsymbol{\theta}}\,(\boldsymbol{S} - \mathsf{E}\,(\boldsymbol{S} \mid \boldsymbol{h}(\boldsymbol{S}))) = \mathsf{E}_{\boldsymbol{\theta}}\, \boldsymbol{S} - \mathsf{E}_{\boldsymbol{\theta}}\, \boldsymbol{S} = \boldsymbol{0}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Because $\boldsymbol{S}$ is complete, necessarily $\boldsymbol{S} = \mathsf{E}\,(\boldsymbol{S} \mid \boldsymbol{T})$ almost surely, and therefore we are able to write $\boldsymbol{S}$ as a function of $\boldsymbol{T}$. We have found a one-to-one mapping between the statistics $\boldsymbol{S}$ and $\boldsymbol{T}$. Because $\boldsymbol{T}$ is minimal sufficient, it is a function of any other sufficient statistic. But, since $\boldsymbol{S}$ is a function of $\boldsymbol{T}$, it is also a function of any other sufficient statistic. Necessarily, also $\boldsymbol{S}$ is minimal sufficient. □

Being complete sufficient is a better property than being minimal sufficient. There are statistics that are minimal sufficient, but not complete. In general, it is not easy to determine whether a statistic is complete. Sometimes, this is possible to be verified directly from the definition. More often, the following theorem is quite useful.

Theorem 15. *Let $X_1, \ldots, X_n$ be a random sample from a distribution whose density $f(x; \boldsymbol{\theta})$ w.r.t. a σ-finite measure μ is of exponential type, meaning that*

$$f(x; \boldsymbol{\theta}) = q(\boldsymbol{\theta})\, h(x) \exp\left(\sum_{j=1}^{p} b_j(\boldsymbol{\theta}) R_j(x) \right) \quad \text{for all } x \text{ and } \boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathsf{T}},$$

for some functions q, h, b_j and R_j, $j = 1, \ldots, p$. Suppose that

- *the set $\left\{ (b_1(\boldsymbol{\theta}), \ldots, b_p(\boldsymbol{\theta}))^{\mathsf{T}} : \boldsymbol{\theta} \in \boldsymbol{\Theta} \right\} \subseteq \mathbb{R}^p$ has non-empty interior; and*
- *the set of functions $\{b_j\}_{j=1}^{p}$ is affinely independent, meaning that*

$$\sum_{j=1}^{p} \lambda_j\, b_j(\boldsymbol{\theta}) \equiv \lambda_0 \text{ for some } \lambda_j \in \mathbb{R},\ j = 0, \ldots, p, \text{ implies that } \lambda_j = 0 \text{ for all } j = 0, \ldots, p. \tag{22}$$

Denote

$$T_j = \sum_{i=1}^{n} R_j(X_i) \text{ and } \boldsymbol{T} = (T_1, \ldots, T_p)^{\mathsf{T}}.$$

Then $\boldsymbol{T}$ is a complete minimal sufficient statistic for $\boldsymbol{\theta}$.

Proof. The statistic $\boldsymbol{T}$ is sufficient directly by the factorization criterion of Theorem 10. Its minimality follows from Theorem 13. First observe that $M = \{x\colon f(x;\boldsymbol{\theta}) > 0\} = \{x\colon h(x) > 0\}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, meaning that the support of X_1 does not depend on $\boldsymbol{\theta}$. We have[5] for any $\boldsymbol{x} = (x_1, \dots, x_n)^\mathsf{T}$ and $\boldsymbol{y} = (y_1, \dots, y_n)^\mathsf{T}$ in the support M^n of $\boldsymbol{X} = (X_1, \dots, X_n)^\mathsf{T}$ that

$$\begin{aligned}\frac{\prod_{i=1}^n f(x_i;\boldsymbol{\theta})}{\prod_{i=1}^n f(y_i;\boldsymbol{\theta})} &= \frac{q(\boldsymbol{\theta})^n \prod_{i=1}^n h(x_i)}{q(\boldsymbol{\theta})^n \prod_{i=1}^n h(y_i)} \exp\left(\sum_{i=1}^n \sum_{j=1}^p b_j(\boldsymbol{\theta}) R_j(x_i) - \sum_{i=1}^n \sum_{j=1}^p b_j(\boldsymbol{\theta}) R_j(y_i)\right) \\ &= \frac{\prod_{i=1}^n h(x_i)}{\prod_{i=1}^n h(y_i)} \exp\left(\sum_{j=1}^p b_j(\boldsymbol{\theta}) \sum_{i=1}^n \left(R_j(x_i) - R_j(y_i)\right)\right) \\ &= \frac{\prod_{i=1}^n h(x_i)}{\prod_{i=1}^n h(y_i)} \exp\left(\sum_{j=1}^p b_j(\boldsymbol{\theta}) a_j\right)\end{aligned}$$

where we denoted $a_j = \sum_{i=1}^n \left(R_j(x_i) - R_j(y_i)\right) \in \mathbb{R}$, $j = 1, \dots, p$. By the assumption of affine independence of $\{b_j\}_{j=1}^p$ in (22) we get that the ratio of densities above does not depend on $\boldsymbol{\theta}$ implies that all $a_j = 0$, meaning that

$$\sum_{i=1}^n R_j(x_i) = \sum_{i=1}^n R_j(y_i) \text{ for all } j = 1, \dots, p.$$

We have verified the condition from Theorem 13, and $\boldsymbol{T}$ is minimal sufficient.

The proof of completeness of $\boldsymbol{T}$ is technical, and involves the theory of analytic functions in the complex plane. It will not be given here in full; a complete proof can be found in e.g. [5, Lemma 2.13]. For the main argument, we use the observation from Remark 2 that the density of the sufficient statistic $\boldsymbol{T}$ takes the form

$$f_{\boldsymbol{T}}(\boldsymbol{t};\boldsymbol{\theta}) = g(\boldsymbol{t},\boldsymbol{\theta}) H(\boldsymbol{t}) = q(\boldsymbol{\theta})^n \exp\left(\sum_{j=1}^p b_j(\boldsymbol{\theta})\, t_j\right) H(\boldsymbol{t}) \text{ for all } \boldsymbol{t} = (t_1, \dots, t_p)^\mathsf{T},$$

where g is the function from the factorization of the density of $\boldsymbol{X}$, and H does not depend on $\boldsymbol{\theta}$. This density is taken w.r.t. a σ-finite measure ν on $\mathbb{R}^p$ (not necessarily equal to μ). The assumption $\mathsf{E}_{\boldsymbol{\theta}}\, w(\boldsymbol{T}) = 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ from the definition of completeness rewrites into

$$0 = \int_{\mathbb{R}^p} w(\boldsymbol{t}) f_{\boldsymbol{T}}(\boldsymbol{t};\boldsymbol{\theta})\, \mathrm{d}\,\nu(\boldsymbol{t}) = q(\boldsymbol{\theta})^n \int_{\mathbb{R}^p} w(\boldsymbol{t})\, H(\boldsymbol{t}) \exp\left(\sum_{j=1}^p b_j(\boldsymbol{\theta})\, t_j\right) \mathrm{d}\,\nu(\boldsymbol{t}) \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

To simplify the expression, divide both sides by the positive term $q(\boldsymbol{\theta})^n$ and write $\eta_j = b_j(\boldsymbol{\theta})$ for $j = 1, \dots, p$. We get an integral equation

$$0 = \int_{\mathbb{R}^p} w(\boldsymbol{t})\, H(\boldsymbol{t}) \exp\left(\sum_{j=1}^p \eta_j\, t_j\right) \mathrm{d}\,\nu(\boldsymbol{t}) \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}, \tag{23}$$

[5] Here we follow the notational convention from Remark 1 and even with $d > 1$ we write $\boldsymbol{x} = (x_1, \dots, x_n)^\mathsf{T} \in \mathbb{R}^{dn}$ instead of $\boldsymbol{x} = \left(x_1^\mathsf{T}, \dots, x_n^\mathsf{T}\right)^\mathsf{T} \in \mathbb{R}^{dn}$ for simplicity.

and by our assumption of non-empty interior of the transformed parameter space, this must be also true for all $(\eta_1, \ldots, \eta_p)^\mathsf{T}$ in an open subset of $\mathbb{R}^p$. The right hand side of (23) is an integral transform of the function $w(\boldsymbol{t})\, H(\boldsymbol{t})$ that can be interpreted as a special Laplace transform. This transform can be shown to be injective, meaning that (23) already implies $w(\boldsymbol{t})\, H(\boldsymbol{t}) = 0$ for ν-almost all $\boldsymbol{t}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, or equivalently $w(\boldsymbol{T}) = 0$ almost surely for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ as we wanted to show. □

Note that the random variables X_i in Theorem 15 can be one-dimensional (that is, $d = 1$), or also d-dimensional random vectors with $d > 1$. The assumption of affine independence in Theorem 15 is needed to guarantee that the density f cannot be reparametrized using a smaller number of parameters.

Example 1.18. For the distribution $\mathsf{N}(\mu, \sigma^2)$ with $p = 2$ and $\boldsymbol{\theta} = \left(\mu, \sigma^2\right)^\mathsf{T}$ we have

$$f(x; \boldsymbol{\theta}) = \frac{\exp\left(-\mu^2/(2\,\sigma^2)\right)}{\sqrt{2\,\pi\,\sigma^2}} \exp\left(\frac{\mu}{\sigma^2} x - \frac{1}{2\,\sigma^2} x^2\right) \quad \text{for } x \in \mathbb{R}.$$

Take

$$b_1(\boldsymbol{\theta}) = \frac{\mu}{\sigma^2}, \qquad b_2(\boldsymbol{\theta}) = -\frac{1}{2\,\sigma^2},$$

and

$$R_1(x) = x, \qquad R_2(x) = x^2.$$

The pair of functions $\{b_1(\boldsymbol{\theta}), b_2(\boldsymbol{\theta})\}$ is affinely independent. The transformed parameter space

$$\left\{(b_1(\boldsymbol{\theta}), b_2(\boldsymbol{\theta}))^\mathsf{T} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\right\} = \left\{\left(\frac{\mu}{\sigma^2}, -\frac{1}{2\,\sigma^2}\right)^\mathsf{T} : \left(\mu, \sigma^2\right)^\mathsf{T} \in \mathbb{R} \times (0, \infty)\right\} = \mathbb{R} \times (-\infty, 0)$$

has non-empty interior in $\mathbb{R}^2$. Thus, Theorem 15 applies and we get that

$$\boldsymbol{S} = \left(\sum_{i=1}^n R_1(X_i), \sum_{i=1}^n R_2(X_i)\right)^\mathsf{T} = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)^\mathsf{T}$$

is a complete minimal sufficient statistic for $\boldsymbol{\theta}$ based on $X_1, \ldots, X_n$ sampled independently from $\mathsf{N}(\mu, \sigma^2)$. △

The main relevance of complete and minimal sufficient statistics rests in the following three crucial theorems. They substantially simplify the search for the best unbiased estimators.

Theorem 16 (Rao-Blackwell)**.** *Let $\boldsymbol{S}$ be a sufficient statistic and let $a(\boldsymbol{\theta})$ be a parametric function of $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$. Let $T = T(\boldsymbol{X})$ be any statistic that satisfies $\mathsf{var}_{\boldsymbol{\theta}}\, T < \infty$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Denote $U = \mathsf{E}\,(T \mid \boldsymbol{S})$. Then $U = U(\boldsymbol{S})$ is also a statistic which satisfies*

$$\mathsf{E}_{\boldsymbol{\theta}}\, U = \mathsf{E}_{\boldsymbol{\theta}}\, T \text{ and } \mathsf{E}_{\boldsymbol{\theta}}\, (T - a(\boldsymbol{\theta}))^2 \geq \mathsf{E}_{\boldsymbol{\theta}}\, (U - a(\boldsymbol{\theta}))^2 \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

The last inequality turns to an equality if and only if $U = T$ almost surely.

Proof. Because $\boldsymbol{S}$ is sufficient, the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{S}$ does not depend on $\boldsymbol{\theta}$. Hence also the conditional distribution of $T = T(\boldsymbol{X})$ given $\boldsymbol{S}$ does not depend on $\boldsymbol{\theta}$, and so also the expression for the random variable $U = \mathsf{E}\ (T \mid \boldsymbol{S})$ does not depend on $\boldsymbol{\theta}$. We have verified that U is a statistic. For U we have

$$\mathsf{E}_{\boldsymbol{\theta}}\, U = \mathsf{E}_{\boldsymbol{\theta}}\, \mathsf{E}\ (T \mid \boldsymbol{S}) = \mathsf{E}_{\boldsymbol{\theta}}\, T \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta},$$

and

$$\begin{aligned}
\mathsf{E}_{\boldsymbol{\theta}}\ (T - a(\boldsymbol{\theta}))^2 &= \mathsf{E}_{\boldsymbol{\theta}}\ (T - U + U - a(\boldsymbol{\theta}))^2 \\
&= \mathsf{E}_{\boldsymbol{\theta}}\ (T-U)^2 + \mathsf{E}_{\boldsymbol{\theta}}\ (U - a(\boldsymbol{\theta}))^2 + 2\,\mathsf{E}_{\boldsymbol{\theta}}\ (T-U)\,(U - a(\boldsymbol{\theta})) \\
&= \mathsf{E}_{\boldsymbol{\theta}}\ (T-U)^2 + \mathsf{E}_{\boldsymbol{\theta}}\ (U - a(\boldsymbol{\theta}))^2 + 2\,\mathsf{E}_{\boldsymbol{\theta}}\, \mathsf{E}\ ((T - U(\boldsymbol{S}))\,(U(\boldsymbol{S}) - a(\boldsymbol{\theta})) \mid \boldsymbol{S}) \\
&= \mathsf{E}_{\boldsymbol{\theta}}\ (T-U)^2 + \mathsf{E}_{\boldsymbol{\theta}}\ (U - a(\boldsymbol{\theta}))^2 + 2\,\mathsf{E}_{\boldsymbol{\theta}}\ (U(\boldsymbol{S}) - a(\boldsymbol{\theta}))\, \mathsf{E}\ ((T - \mathsf{E}\,(T \mid \boldsymbol{S})) \mid \boldsymbol{S}) \\
&= \mathsf{E}_{\boldsymbol{\theta}}\ (T-U)^2 + \mathsf{E}_{\boldsymbol{\theta}}\ (U - a(\boldsymbol{\theta}))^2 + 2\,\mathsf{E}_{\boldsymbol{\theta}}\ (U(\boldsymbol{S}) - a(\boldsymbol{\theta}))\,(\mathsf{E}\,(T \mid \boldsymbol{S}) - \mathsf{E}\,(T \mid \boldsymbol{S})) \\
&= \mathsf{E}_{\boldsymbol{\theta}}\ (T-U)^2 + \mathsf{E}_{\boldsymbol{\theta}}\ (U - a(\boldsymbol{\theta}))^2 \geq \mathsf{E}_{\boldsymbol{\theta}}\ (U - a(\boldsymbol{\theta}))^2 .
\end{aligned}$$

The inequality turns to equality if and only if $\mathsf{E}_{\boldsymbol{\theta}}\,(T-U)^2 = 0$, in which case $T = U$ almost surely. □

In words, the Rao-Blackwell theorem says that by conditioning on a sufficient statistic, any initial statistic T can be changed to a statistic with the same expectation, but a smaller mean squared error. If T is unbiased for $a(\boldsymbol{\theta})$, then also U must be unbiased for $a(\boldsymbol{\theta})$, and Theorem 16 guarantees that the variance of U cannot be larger than that of T. In the particular situation with T unbiased, Theorem 16 takes a quite familiar form. Indeed, using the law of total variance we know that

$$\mathsf{var}_{\boldsymbol{\theta}}\, T = \mathsf{var}_{\boldsymbol{\theta}}\, \mathsf{E}\ (T \mid \boldsymbol{S}) + \mathsf{E}_{\boldsymbol{\theta}}\, \mathsf{var}\ (T \mid \boldsymbol{S}) = \mathsf{var}_{\boldsymbol{\theta}}\, U + \mathsf{E}_{\boldsymbol{\theta}}\, \mathsf{var}\ (T \mid \boldsymbol{S}) \geq \mathsf{var}_{\boldsymbol{\theta}}\, U.$$

We have used that $\mathsf{var}\ (T \mid \boldsymbol{S}) \geq 0$ almost surely, with equality if and only if $T = \mathsf{E}\ (T \mid \boldsymbol{S}) = U$ almost surely. This is precisely the Rao-Blackwell theorem for T unbiased for $a(\boldsymbol{\theta})$.

Example 1.19. Let $X_1, \ldots, X_n$ be independent from a distribution with a density $f(\cdot; \boldsymbol{\theta})$ w.r.t. a σ-finite measure μ on $\mathbb{R}$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$. Denote by $X_{(1)} \leq \cdots \leq X_{(n)}$ the ordered random sample, and let $\boldsymbol{S} = \left(X_{(1)}, \ldots, X_{(n)}\right)^{\mathsf{T}}$. The density of $\boldsymbol{X} = (X_1, \ldots, X_n)^{\mathsf{T}}$ w.r.t. the product measure of n copies of μ is

$$\prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_{(i)}; \boldsymbol{\theta}) \text{ for all } x_1, \ldots, x_n \in \mathbb{R},$$

where by $x_{(1)} \leq \cdots \leq x_{(n)}$ we denoted the ordered points $x_1, \ldots, x_n$. By the Neyman factorization criterion from Theorem 10 we therefore see that the ordered random sample $\boldsymbol{S}$ is always a sufficient statistic for $\boldsymbol{\Theta}$.

Take now a parametric function $a(\boldsymbol{\theta})$, and suppose that $T = h(X_1)$ is unbiased for $a(\boldsymbol{\theta})$, i.e. $\mathsf{E}_{\boldsymbol{\theta}}\, h(X_1) = a(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. By the Rao-Blackwell theorem, we can improve T by conditioning on the ordered random sample $\boldsymbol{S}$. We obtain

$$\begin{aligned} \mathsf{E}\ (T \mid \boldsymbol{S}) &= \mathsf{E}\ \big(h(X_1) \mid X_{(1)}, \ldots, X_{(n)}\big) \\ &= \mathsf{E}\ \big(h(X_i) \mid X_{(1)}, \ldots, X_{(n)}\big) \ \text{ for all } i = 1, \ldots, n, \end{aligned} \tag{24}$$

because by the assumption of independence and identical distribution, the random vector $\boldsymbol{X}$ is exchangeable, meaning that each permutation of its elements has the same distribution. By summing all the equations in (24) we obtain

$$\begin{aligned} \mathsf{E}\ (T \mid \boldsymbol{S}) &= \frac{1}{n} \sum_{i=1}^{n} \mathsf{E}\ \big(h(X_i) \mid X_{(1)}, \ldots, X_{(n)}\big) = \frac{1}{n} \mathsf{E}\ \left(\sum_{i=1}^{n} h(X_{(i)}) \mid X_{(1)}, \ldots, X_{(n)} \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} h(X_{(i)}) = \frac{1}{n} \sum_{i=1}^{n} h(X_i). \end{aligned}$$

The second equality follows because $\sum_{i=1}^{n} h(X_i) = \sum_{i=1}^{n} h(X_{(i)})$ almost surely, the third one because $\sum_{i=1}^{n} h(X_{(i)})$ is a function of $\boldsymbol{S}$. The Rao-Blackwell theorem gives that whenever $T = h(X_1)$ is unbiased for $a(\boldsymbol{\theta})$, its symmetrized version $U_n = \sum_{i=1}^{n} h(X_i)/n$ is (of course) also unbiased for $a(\boldsymbol{\theta})$, and the variance of U_n is smaller than that of T.

This observation can be generalized substantially. Suppose for example that $h(X_1, X_2)$ is unbiased for $a(\boldsymbol{\theta})$. Using an argument similar to (24) we obtain that

$$U_n = \frac{1}{\binom{n}{2}} \sum_{1 \le i_1 < i_2 \le n} h(X_{i_1}, X_{i_2})$$

is unbiased for $a(\boldsymbol{\theta})$ too, and its variance cannot exceed that of $h(X_1, X_2)$. Analogous conclusions hold true for statistics of the type $h(X_1, \ldots, X_m)$ with $m \le n$. Symmetric functions of the data that can be written as U_n above are called *U-statistics*. They play an important role in the estimation theory. Overall, the Rao-Blackwell theorem says that whenever the random vector $\boldsymbol{X}$ has an exchangeable distribution, symmetrization w.r.t. the data always improves the estimators. △

We saw that the Rao-Blackwell theorem gives a way to improve unbiased estimators by means of conditioning. This brings a natural question: under what conditions can we guarantee that the improved estimator U is already the best unbiased? An answer is given in the following two theorems.

Theorem 17 (first Lehmann-Scheffé theorem)**.** *Let T be an unbiased estimator of a parametric function $a(\boldsymbol{\theta})$ such that $\mathsf{var}_{\boldsymbol{\theta}}\, T < \infty$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Let $\boldsymbol{S}$ be a complete sufficient statistic, and define $U = \mathsf{E}\ (T \mid \boldsymbol{S})$. Then U is the unique best unbiased estimator of $a(\boldsymbol{\theta})$.*

Proof. By Theorem 16 we know that $U = U(\boldsymbol{S})$ is an unbiased estimator of $a(\boldsymbol{\theta})$ with the property $\mathsf{var}_{\boldsymbol{\theta}}\, U \leq \mathsf{var}_{\boldsymbol{\theta}}\, T$. For any other unbiased estimator Z of $a(\boldsymbol{\theta})$ with finite variance we can use Theorem 16 again and find $V = \mathsf{E}\,(Z \mid \boldsymbol{S})$. But then both U and V are functions of $\boldsymbol{S}$ only, and we have

$$\mathsf{E}_{\boldsymbol{\theta}}\,(U(\boldsymbol{S}) - V(\boldsymbol{S})) = a(\boldsymbol{\theta}) - a(\boldsymbol{\theta}) = 0 \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

Because $\boldsymbol{S}$ is complete, necessarily $U = V$ almost surely, and thus also $\mathsf{var}_{\boldsymbol{\theta}}\, U = \mathsf{var}_{\boldsymbol{\theta}}\, V$ for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. □

Theorem 18 (second Lehmann-Scheffé theorem)**.** *Let $\boldsymbol{S}$ be a complete sufficient statistic, and let $W = g(\boldsymbol{S})$ be an unbiased estimator of a parametric function $a(\boldsymbol{\theta})$ such that $\mathsf{var}_{\boldsymbol{\theta}}\, W < \infty$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Then W is the unique best unbiased estimator of $a(\boldsymbol{\theta})$.*

Proof. Since W is a function of $\boldsymbol{S}$, we have $\mathsf{E}\,(W \mid \boldsymbol{S}) = W$. The rest follows directly from Theorem 17. □

Lehmann-Scheffé's theorems guide the construction of best unbiased estimators. Given that a complete sufficient statistic $\boldsymbol{S}$ is known, to find a unique BUE it is enough to either condition any unbiased estimator T on $\boldsymbol{S}$, or to find any unbiased function of $\boldsymbol{S}$.

Example 1.20. For $X_1, \ldots, X_n$ independent from $\mathsf{N}\left(\mu, \sigma^2\right)$ with $\left(\mu, \sigma^2\right)^{\mathsf{T}} = \boldsymbol{\theta}$ we know by Example 1.18 that $\boldsymbol{S} = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)^{\mathsf{T}}$ is a complete minimal sufficient statistic for $\boldsymbol{\theta}$. We also know that for $a_1(\boldsymbol{\theta}) = \mu$, the sample mean $\bar{X}_n$ is an unbiased estimator of $a_1(\boldsymbol{\theta})$ that can be written as a function of $\boldsymbol{S}$. By Theorem 18 we see that $\bar{X}_n$ is the BUE of μ in our model. Analogously, for $a_2(\boldsymbol{\theta}) = \sigma^2$ we know that the sample variance S_n^2 is a an unbiased estimator of $a_2(\boldsymbol{\theta})$ that is also a function of $\boldsymbol{S}$. Thus, it is the BUE of σ^2 for a random sample from a normal distribution.

Comparing our results with Example 1.12 and the Rao-Cramér bound from Theorem 8 we see that $\bar{X}_n$ attains the Rao-Cramér bound and thus it must be the BUE for μ. But, for S_n^2 we have by [6, Theorem 2.8] that

$$\mathsf{var}_{\boldsymbol{\theta}}\, S_n^2 = \frac{2\,\sigma^4}{n-1} > \frac{2\,\sigma^4}{n},$$

where on the right hand side we have the corresponding Rao-Cramér bound for σ^2. We see that even though the Rao-Cramér bound is not attained, there is no better unbiased estimator of σ^2 in our model than the sample variance. △

1.4 Ancillary statistics

In connection with the sufficient statistics, sometimes also the term ancillarity appears. Ancillary statistics are in a sense the opposite to sufficient statistics. While we saw that sufficient statistics exhaust the complete information about a parameter $\boldsymbol{\theta}$, ancillary statistics do not carry any information about $\boldsymbol{\theta}$.

Definition 10. Let the distribution of a random vector $\boldsymbol{X}$ depend only on parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. A statistic $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$ is called *ancillary* for parameter $\boldsymbol{\theta}$ if the distribution of $\boldsymbol{T}$ does not depend on $\boldsymbol{\theta}$.

An interesting result about ancillary statistics is the following theorem.

Theorem 19 (Basu)**.** *Let $\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{X})$ be a complete sufficient statistic, and let $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{X})$ be an ancillary statistic for parameter $\boldsymbol{\theta}$. Then $\boldsymbol{U}$ is independent of $\boldsymbol{V}$ for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.*

Proof. The proof is given only for the situation of discrete distribution of $\boldsymbol{X}$; the general case is analogous, yet requires to work with conditional distributions.

Because by our assumptions the distribution of $\boldsymbol{V}$ does not depend on $\boldsymbol{\theta}$, we can denote

$$\mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{V} = \boldsymbol{v}\right) = h(\boldsymbol{v})$$

for some function h that does not depend on $\boldsymbol{\theta}$. Since $\boldsymbol{U}$ is a sufficient statistic, the conditional distribution $\mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{V} = \boldsymbol{v} \mid \boldsymbol{U} = \boldsymbol{u}\right)$ also does not depend on $\boldsymbol{\theta}$. Thus, we can also denote

$$\mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{V} = \boldsymbol{v} \mid \boldsymbol{U} = \boldsymbol{u}\right) = g(\boldsymbol{u}, \boldsymbol{v}) \tag{25}$$

for a function g that also does not depend on $\boldsymbol{\theta}$. Combining both these results and noting that $\boldsymbol{U}$ is a statistic, we found that also the function $\boldsymbol{x} \mapsto g(\boldsymbol{U}(\boldsymbol{x}), \boldsymbol{v}) - h(\boldsymbol{v})$ does not depend on $\boldsymbol{\theta}$, and therefore the random variable $g(\boldsymbol{U}, \boldsymbol{v}) - h(\boldsymbol{v}) = g(\boldsymbol{U}(\boldsymbol{X}), \boldsymbol{v}) - h(\boldsymbol{v})$ is a statistic. For its expectation, we have

$$\begin{aligned}\mathsf{E}_{\boldsymbol{\theta}}\, g(\boldsymbol{U}, \boldsymbol{v}) &= \sum_{\boldsymbol{u}} g(\boldsymbol{u}, \boldsymbol{v}) \mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{U} = \boldsymbol{u}\right) = \sum_{\boldsymbol{u}} \mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{V} = \boldsymbol{v} \mid \boldsymbol{U} = \boldsymbol{u}\right) \mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{U} = \boldsymbol{u}\right) \\ &= \sum_{\boldsymbol{u}} \mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{V} = \boldsymbol{v}, \boldsymbol{U} = \boldsymbol{u}\right) = \mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{V} = \boldsymbol{v}\right) = h(\boldsymbol{v}),\end{aligned}$$

where the sum is taken over those $\boldsymbol{u}$ such that $\mathsf{P}_{\boldsymbol{\theta}}(\boldsymbol{U} = \boldsymbol{u}) > 0$. Hence,

$$\mathsf{E}_{\boldsymbol{\theta}}\ \left(g(\boldsymbol{U}, \boldsymbol{v}) - h(\boldsymbol{v})\right) = 0 \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

We assumed that $\boldsymbol{U}$ is complete, meaning that necessarily $g(\boldsymbol{U}, \boldsymbol{v}) = h(\boldsymbol{v})$ almost surely for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Thus, $g(\boldsymbol{U}, \boldsymbol{v})$ is almost surely a constant $h(\boldsymbol{v})$ that depends only on $\boldsymbol{v}$. Plugging this into (25) we get that also for any $\boldsymbol{u}$ such that $\mathsf{P}_{\boldsymbol{\theta}}(\boldsymbol{U} = \boldsymbol{u}) > 0$ we have that

$$\mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{V} = \boldsymbol{v} \mid \boldsymbol{U} = \boldsymbol{u}\right) = g(\boldsymbol{u}, \boldsymbol{v}) = h(\boldsymbol{v}) = \mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{V} = \boldsymbol{v}\right)$$

does not depend on $\boldsymbol{u}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. That is equivalent with the independence of $\boldsymbol{U}$ and $\boldsymbol{V}$ for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. □

Example 1.21. Let $X_1, \ldots, X_n$ be independent from $\mathsf{N}(\mu, \sigma^2)$ with $\mu = \theta \in \mathbb{R}$ an unknown parameter and $\sigma^2 > 0$ fixed and known. The sample mean $\bar{X}_n$ is a complete sufficient statistic for θ by Theorem 15. The sample variance

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left((X_i - \mu) - \frac{1}{n} \sum_{j=1}^{n} (X_j - \mu) \right)^2$$

can be written as a function the random sample $X_1 - \mu, \ldots, X_n - \mu$ from distribution $\mathsf{N}(0, \sigma^2)$ that does not depend on θ. Thus, S_n^2 is ancillary for θ. By Theorem 19 the sample mean and the sample variance must be independent of each other. This result was shown, in a completely different way, already in [6, Theorem 2.8]. △

The end of lecture 5 (19.3.2024)

2 Maximum likelihood estimation

2.1 The maximum likelihood method

In the second large section of this course, we study the principles of maximum likelihood-based procedures. The main idea of maximum likelihood is relatively simple. Our task is to estimate the unknown parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, based on a realisation of a random vector $\boldsymbol{X}$ whose distribution depends on $\boldsymbol{\theta}$. Typically, the random vector $\boldsymbol{X}$ has elements forming a random sample of d-dimensional vectors with $d \in \mathbb{N}$, but this is not always needed. Denote the observed value of $\boldsymbol{X}$ by $\boldsymbol{x} \in \mathbb{R}^{dn}$. Since all we know about the parameter $\boldsymbol{\theta}$ is contained in $\boldsymbol{x}$, we can ask how much "likely" it was that, for a given value of a parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the vector $\boldsymbol{x}$ was observed. If the distribution of $\boldsymbol{X}$ is discrete, we can evaluate the probability of observing $\boldsymbol{x}$ directly, and consider $\mathsf{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x})$. The higher value this function of $\boldsymbol{\theta}$ takes, the more "probable" it was that the true distribution of $\boldsymbol{X}$ corresponded to the parameter $\boldsymbol{\theta}$. The maximizer of the function $\boldsymbol{\Theta} \to [0, \infty) \colon \boldsymbol{\theta} \mapsto \mathsf{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x})$ is called the maximum likelihood estimator of $\boldsymbol{\theta}$.

Take now the situation when the distribution of $\boldsymbol{X}$ is absolutely continuous w.r.t. the Lebesgue measure ν on $\mathbb{R}^{dn}$, for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Now, the evaluation of $\mathsf{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x})$ directly does not help, as the probability of observing any fixed $\boldsymbol{x} \in \mathbb{R}^{dn}$ is zero, for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Instead of considering the probability directly, we therefore consider the density $f(\boldsymbol{x}; \boldsymbol{\theta})$ as a surrogate for the probability. The intuition behind this comes from the Lebesgue differentiation theorem,

roughly stating that for ν-almost all $\boldsymbol{x} \in \mathbb{R}^{dn}$ we have that

$$\lim_{\varepsilon \to 0+} \frac{\int_{B(\boldsymbol{x},\varepsilon)} f(\boldsymbol{y};\boldsymbol{\theta}) \,\mathrm{d}\, \nu(\boldsymbol{y})}{\int_{B(\boldsymbol{x},\varepsilon)} 1 \,\mathrm{d}\, \nu(\boldsymbol{y})} = f(\boldsymbol{x};\boldsymbol{\theta}). \tag{26}$$

Here, $B(\boldsymbol{x},\varepsilon)$ is the closed ball in $\mathbb{R}^{dn}$ centred at $\boldsymbol{x}$ with radius $\varepsilon > 0$. The numerator on the left hand side in (26) is the probability $\mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{X} \in B(\boldsymbol{x},\varepsilon)\right)$; the denominator is the volume of the ball $B(\boldsymbol{x},\varepsilon)$. We see that taking the density of $\boldsymbol{X}$ instead of the exact probability $\mathsf{P}_{\boldsymbol{\theta}}\left(\boldsymbol{X} = \boldsymbol{x}\right)$, we in fact approximate the probability of $\boldsymbol{X}$ lying "near" $\boldsymbol{x}$, and therefore basing our inference on the value of the density is still well interpretable. Again, any maximizer of the density of $\boldsymbol{X}$ as a function of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is called the maximum likelihood estimator of $\boldsymbol{\theta}$.

Our motivation leads us to the following important definition.

Definition 11. Let $\boldsymbol{X} = (X_1, \ldots, X_n)^{\mathsf{T}}$ be a random vector whose distribution depends on the parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Suppose that the density $f(\boldsymbol{x};\boldsymbol{\theta})$ of $\boldsymbol{X}$ exists w.r.t. some σ-finite measure ν, for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Then, for a fixed value of $\boldsymbol{x}$, is the function

$$\boldsymbol{\Theta} \to [0,\infty) \colon \boldsymbol{\theta} \mapsto f(\boldsymbol{x};\boldsymbol{\theta})$$

called the *likelihood function* of $\boldsymbol{\theta}$ at point $\boldsymbol{x} \in \mathbb{R}^{dn}$. Any value $\widehat{\boldsymbol{\theta}}_n$ that maximizes the likelihood function of $\boldsymbol{\theta}$ is called the *maximum likelihood estimator (MLE)* of $\boldsymbol{\theta}$.

The maximum likelihood estimator is a function of $\boldsymbol{x}$ that does not depend on $\boldsymbol{\theta}$. We consider the maximum likelihood estimator from two different perspectives. From a practical point of view, once the observed value $\boldsymbol{x}$ of $\boldsymbol{X}$ is known, the MLE is a function of $\boldsymbol{x}$, and is thus a fixed value in $\boldsymbol{\Theta}$. From the theoretical perspective, it is important to understand how MLE behaves when the distribution of $\boldsymbol{X}$ is involved. In this case, we do not take any particular value of $\boldsymbol{x}$ but instead we consider $\boldsymbol{X}$ to be random, and explore the distribution of the MLE being a function of $\boldsymbol{X}$. Thus, the MLE is still a function of $\boldsymbol{x}$, but this time we do not restrict to the particular observed value $\boldsymbol{x}$, but let $\boldsymbol{x}$ to be random and distributed as $\boldsymbol{X}$. In this way, the MLE is random and therefore, it is a statistic.

Very often, the random vector $\boldsymbol{X}$ has elements forming a random sample from a distribution with a density $f(x;\boldsymbol{\theta})$ with $x \in \mathbb{R}^d$. In that case, the joint density of $\boldsymbol{X}$ takes the form of a product $f(\boldsymbol{x};\boldsymbol{\theta}) = \prod_{i=1}^n f(x_i;\boldsymbol{\theta})$ with $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathsf{T}} \in \mathbb{R}^{dn}$. This function is cumbersome to maximize, as taking a derivative w.r.t. $\boldsymbol{\theta}$ of a multiple product is tedious. Therefore, a simple trick is used. Because the logarithm log is a strictly increasing function on $(0,\infty)$, we first take the logarithm of $f(\boldsymbol{x};\boldsymbol{\theta})$, and only then optimize for $\boldsymbol{\theta}$. This transformation does not change the arguments of maxima of the likelihood function, and is much easier to work with as

$$\log \prod_{i=1}^n f(x_i;\boldsymbol{\theta}) = \sum_{i=1}^n \log f(x_i;\boldsymbol{\theta}).$$

The logarithm of a density is already well known to us. In the form of scores and the score function we already encountered it in (17) in the theory of point estimation.

Definition 12. The logarithm of the likelihood function $f(\boldsymbol{x};\boldsymbol{\theta})$ is called the *log-likelihood function* of $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$. For $\boldsymbol{x} \in \mathbb{R}^{dn}$ fixed it is denoted by

$$L_n(\boldsymbol{\theta};\boldsymbol{x}) = \log f(\boldsymbol{x};\theta).$$

The map L_n is formally a function of both $\boldsymbol{\theta}$ and $\boldsymbol{x}$. Since $\boldsymbol{x}$ is, however, often held fixed, we will also frequently denote

$$L_n(\boldsymbol{\theta}) = L_n(\boldsymbol{\theta};\boldsymbol{x})$$

where no confusion can arise. If the function L_n is differentiable (as a function of $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_p)^\mathsf{T}$), the vector of its partial derivatives

$$\frac{\partial L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left(\frac{\partial L_n(\boldsymbol{\theta})}{\partial \theta_1},\ldots,\frac{\partial L_n(\boldsymbol{\theta})}{\partial \theta_p}\right)^\mathsf{T}$$

is called the *score function* of $\boldsymbol{\theta}$. The system of p equations

$$\frac{\partial L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0} \tag{27}$$

is called the *system of likelihood equations*.

If we can write the system of likelihood equations, any MLE $\widehat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ must be a root of this system. We therefore usually search for MLE by solving (27).

Example 2.1. Let $X_1,\ldots,X_n$ be independent from $\mathsf{N}\left(\mu,\sigma^2\right)$ and let $\boldsymbol{\theta} = \left(\mu,\sigma^2\right)^\mathsf{T} \in \mathbb{R}\times(0,\infty)$. The density of $\boldsymbol{X} = (X_1,\ldots,X_n)^\mathsf{T}$ w.r.t. the n-dimensional Lebesgue measure is for $\boldsymbol{x} = (x_1,\ldots,x_n)^\mathsf{T} \in \mathbb{R}^n$ given by

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

The log-likelihood function of $\boldsymbol{\theta}$ is

$$L_n(\boldsymbol{\theta}) = \log f(\boldsymbol{x};\boldsymbol{\theta}) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2.$$

This function is differentiable in $\boldsymbol{\theta}$, and we can express the likelihood equations as

$$\begin{aligned}
\frac{\partial L_n(\boldsymbol{\theta})}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^n (x_i-\mu), \\
\frac{\partial L_n(\boldsymbol{\theta})}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (x_i-\mu)^2.
\end{aligned}$$

Solving the system (27) we obtain a root

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \widehat{\sigma_n^2} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \widehat{\mu}_n)^2 . \tag{28}$$

To verify that $\widehat{\boldsymbol{\theta}}_n = \left(\widehat{\mu}_n, \widehat{\sigma_n^2}\right)^{\mathsf{T}}$ is indeed the maximizer of the likelihood, consider the Hessian matrix of L_n. This is the matrix of all second partial derivatives of L_n w.r.t. the elements of $\boldsymbol{\theta}$. We obtain

$$\frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} = \begin{pmatrix} \frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial \mu^2} & \frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial \mu \partial (\sigma^2)} \\ \frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial (\sigma^2) \partial \mu} & \frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial (\sigma^2)^2} \end{pmatrix} = \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu) \\ -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2 \end{pmatrix}.$$

Plugging in the solution to the likelihood equations $\widehat{\boldsymbol{\theta}}_n$ we get

$$\frac{\partial^2 L_n(\widehat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} = \begin{pmatrix} -\frac{n}{\widehat{\sigma_n^2}} & 0 \\ 0 & -\frac{n}{2\left(\widehat{\sigma_n^2}\right)^2} \end{pmatrix}. \tag{29}$$

This matrix is negative definite, meaning that $\widehat{\boldsymbol{\theta}}_n$ is a local maximum of L_n over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Since it is the only local extreme, it is easy to see that $\widehat{\boldsymbol{\theta}}_n$ is indeed the maximum likelihood estimator of $\boldsymbol{\theta}$. Our estimator in (28) was considered for fixed observed values $\boldsymbol{x} \in \mathbb{R}^n$; to obtain the true maximum likelihood estimator of $\boldsymbol{\theta}$, we need to replace the fixed observed values $\boldsymbol{x}$ in (28) by the random vector $\boldsymbol{X}$ again, and conclude that our maximum likelihood estimator of $\boldsymbol{\theta}$ takes the form

$$\widehat{\boldsymbol{\theta}}_n = \left(\bar{X}_n, \frac{n-1}{n} S_n^2\right)^{\mathsf{T}},$$

where as usual, $\bar{X}_n$ and S_n^2 stand for the sample mean and the sample variance of $\boldsymbol{X}$. Note that maximum likelihood estimation does not guarantee unbiased estimators of $\boldsymbol{\theta}$; our estimator $\widehat{\sigma_n^2}$ is a biased one. △

Observe that in formula (29) it is not the first time that we see the Hessian of the log-density of $\boldsymbol{X}$; we already encountered it in Example 1.11 when we computed the Fisher information matrix of $\mathsf{N}(\mu, \sigma^2)$. More generally, from Theorem 7 we see that the negative expected Hessian is, in fact, the Fisher information matrix of $\boldsymbol{\theta}$. This gives another interpretation to the Fisher information matrix — the "more negative definite" the Hessian is at $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the "more positive definite" the Fisher information matrix $\boldsymbol{J}_n(\boldsymbol{\theta})$ is. That is, high Fisher information corresponds to "more peaked" density $f(\boldsymbol{x}; \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$. High peakedness of f in turn, makes it easier to estimate $\boldsymbol{\theta}$ from the data using the maximum likelihood principle. These arguments will recur several times in the theory that follows.

Our first observation about the maximum likelihood method concerns the behaviour of MLE and reparametrizations. That is, we first transform the parameter $\boldsymbol{\theta}$ using a function

$\boldsymbol{u}\colon \boldsymbol{\Theta} \to \mathbb{R}^k$ to a different parameter $\boldsymbol{\tau} = \boldsymbol{u}(\boldsymbol{\theta})$, then express the likelihood as a function of $\boldsymbol{\tau}$, and finally find the MLE of $\boldsymbol{\tau}$ by maximizing the likelihood w.r.t. $\boldsymbol{\tau}$. If the transformation $\boldsymbol{u}$ is one-to-one, the previous procedure is simple and naturally

$$L_n(\boldsymbol{\theta}) = L_n(\boldsymbol{u}^{-1}(\boldsymbol{\tau})) \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta},$$

meaning that L_n is maximized in $\boldsymbol{\theta}$ at $\widehat{\boldsymbol{\theta}}_n$ if and only if it is maximized in $\boldsymbol{\tau}$ in $\widehat{\boldsymbol{\tau}}_n = \boldsymbol{u}(\widehat{\boldsymbol{\theta}}_n)$. We shall, however, prove a more general invariance principle, which covers also the case when $\boldsymbol{u}$ is not injective. In that case, we need to specify what exactly do we mean by the likelihood w.r.t. $\boldsymbol{\tau}$, as several values of $\boldsymbol{\theta}$ can correspond to a single value $\boldsymbol{\tau}$. Therefore, define for $\boldsymbol{\tau} \in \mathbb{R}^k$ its pre-image via $\boldsymbol{u}$ as $\boldsymbol{u}^{-1}(\boldsymbol{\tau}) = \{\boldsymbol{\theta} \in \boldsymbol{\Theta}\colon \boldsymbol{u}(\boldsymbol{\theta}) = \boldsymbol{\tau}\}$, see also (18). Note that $\boldsymbol{u}^{-1}(\boldsymbol{\tau})$ may be an empty set. The *induced log-likelihood function* of $\boldsymbol{\tau}$ is defined as

$$L_n^*(\boldsymbol{\tau}) = \sup_{\boldsymbol{\theta} \in \boldsymbol{u}^{-1}(\boldsymbol{\tau})} L_n(\boldsymbol{\theta}) \text{ for } \boldsymbol{\tau} \in \mathbb{R}^k. \tag{30}$$

In words, we set the likelihood of $\boldsymbol{\tau}$ to be the maximum likelihood of those $\boldsymbol{\theta}$ that map to $\boldsymbol{\tau}$. Recall that the supremum of an empty set is $-\infty$, meaning that in the case when no $\boldsymbol{\theta}$ maps to $\boldsymbol{\tau}$, $L_n^*(\boldsymbol{\tau}) = -\infty$. Any maximizer of the induced log-likelihood is called the *maximum likelihood estimator of the parameter* $\boldsymbol{\tau}$.

Theorem 20 (Zehna's invariance principle)**.** *Let a random vector $\boldsymbol{X}$ have a density $f(\boldsymbol{x};\boldsymbol{\theta})$ w.r.t. some σ-finite measure, with $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$. Let $\boldsymbol{u}\colon \boldsymbol{\Theta} \to \mathbb{R}^k$. If $\widehat{\boldsymbol{\theta}}_n$ is the maximum likelihood estimator of $\boldsymbol{\theta}$, then the maximum likelihood estimator of $\boldsymbol{\tau} = \boldsymbol{u}(\boldsymbol{\theta})$ is $\widehat{\boldsymbol{\tau}}_n = \boldsymbol{u}(\widehat{\boldsymbol{\theta}}_n)$.*

Proof. The collection of sets $\left\{\boldsymbol{u}^{-1}(\boldsymbol{\tau})\colon \boldsymbol{\tau} \in \mathbb{R}^k\right\}$ forms a partitioning of the parameter space $\boldsymbol{\Theta}$. Since $\widehat{\boldsymbol{\theta}}_n \in \boldsymbol{\Theta}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$, it maximizes the log-likelihood over $\boldsymbol{\Theta}$. Then for $\widehat{\boldsymbol{\tau}}_n = \boldsymbol{u}(\widehat{\boldsymbol{\theta}}_n)$ we have that $\widehat{\boldsymbol{\theta}}_n \in \boldsymbol{u}^{-1}(\widehat{\boldsymbol{\tau}}_n)$, and we can write

$$\begin{aligned} L_n(\widehat{\boldsymbol{\theta}}_n) &\leq \sup_{\boldsymbol{\theta} \in \boldsymbol{u}^{-1}(\widehat{\boldsymbol{\tau}}_n)} L_n(\boldsymbol{\theta}) = L_n^*(\widehat{\boldsymbol{\tau}}_n) \leq \sup_{\boldsymbol{\tau} \in \mathbb{R}^k} L_n^*(\boldsymbol{\tau}) \\ &= \sup_{\boldsymbol{\tau} \in \mathbb{R}^k} \sup_{\boldsymbol{\theta} \in \boldsymbol{u}^{-1}(\boldsymbol{\tau})} L_n(\boldsymbol{\theta}) = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L_n(\boldsymbol{\theta}) = L_n(\widehat{\boldsymbol{\theta}}_n). \end{aligned}$$

In the first inequality we used that $\widehat{\boldsymbol{\theta}}_n \in \boldsymbol{u}^{-1}(\widehat{\boldsymbol{\tau}}_n)$. In the second last equality we used that the sets $\boldsymbol{u}^{-1}(\boldsymbol{\tau})$ cover the whole parameter space $\boldsymbol{\Theta}$. In the last equality we use that $\widehat{\boldsymbol{\theta}}_n$ is MLE for $\boldsymbol{\theta}$. It follows that $L_n^*(\widehat{\boldsymbol{\tau}}_n) = \sup_{\boldsymbol{\tau} \in \mathbb{R}^k} L_n^*(\boldsymbol{\tau})$, which means that $\widehat{\boldsymbol{\tau}}_n$ is the maximum likelihood estimator of $\boldsymbol{\tau}$. Because $\widehat{\boldsymbol{\theta}}_n \in \boldsymbol{u}^{-1}(\widehat{\boldsymbol{\tau}}_n)$, we see that $\boldsymbol{u}(\widehat{\boldsymbol{\theta}}_n) = \widehat{\boldsymbol{\tau}}_n$ as we wanted to show. □

Theorem 20 is valuable not only because it allows us to determine the maximum likelihood estimator of a one-to-one transformation of $\boldsymbol{\theta}$. It is important especially because the map $\boldsymbol{u}$ does not have to be one-to-one. Consider the following example.

Example 2.2. Let $\boldsymbol{X} = (X_1, \ldots, X_n)^\mathsf{T}$ be a random sample from $\mathsf{N}(\mu, \sigma^2)$, where $\boldsymbol{\theta} = (\mu, \sigma^2)^\mathsf{T} \in \boldsymbol{\Theta} = \mathbb{R} \times (0, \infty)$. By Example 2.1, for $\boldsymbol{x} = (x_1, \ldots, x_n)^\mathsf{T} \in \mathbb{R}^n$ the observed values of $\boldsymbol{X}$, the log-likelihood of $\boldsymbol{\theta}$ takes the form

$$L_n(\boldsymbol{\theta}) = -\frac{n}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 .$$

Say that we are interested only in parameter μ, and σ^2 is a nuisance parameter. We thus choose a transform $\boldsymbol{u}\colon \boldsymbol{\Theta} \to \mathbb{R}\colon (\mu, \sigma^2)^\mathsf{T} \mapsto \mu$. Theorem 20 allows us to simplify the log-likelihood. Because $\boldsymbol{u}^{-1}(\mu)$ corresponds to the interval $\{\mu\} \times (0, \infty) \subset \boldsymbol{\Theta}$ for each $\mu \in \mathbb{R}$, we can write the induced log-likelihood of μ as

$$\begin{aligned} L_n^*(\mu) &= \sup_{\sigma^2 > 0} L_n\left((\mu, \sigma^2)^\mathsf{T}\right) = L_n\left(\left(\mu, \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}\right)^\mathsf{T}\right) \\ &= -\frac{n}{2} \log\left(\frac{2\pi \sum_{i=1}^n (x_i - \mu)^2}{n}\right) - \frac{n}{2}. \end{aligned} \tag{31}$$

By Theorem 20 we know that maximizing this likelihood function over $\mu \in \mathbb{R}$, we obtain the maximizer $\sum_{i=1}^n x_i/n$, which leads to the maximum likelihood estimator of μ in the form $\widehat{\mu}_n = \bar{X}_n$ as we obtained in Example 2.1. △

Of course, the log-likelihood in Example 2.2 is easy to maximize also simultaneously in both μ and σ^2 as we did in Example 2.1. What is interesting about the approach that we took in Example 2.2 is that this application of Zehna's principle allows us to reduce also fairly complex maximum likelihood problems into simpler (lower-dimensional) optimization tasks. The one-dimensional likelihood function in (31) is easier to visualise or understand than the full log-likelihood $L_n(\boldsymbol{\theta})$. We will return to this idea with the notion of the *profile likelihood* in Section 2.4.

An important question is that of consistency of the maximum likelihood estimators. As a first step, we have the following result.

Theorem 21. *For each $n \in \mathbb{N}$, let $\boldsymbol{X} = (X_1, \ldots, X_n)^\mathsf{T}$ be a random vector such that $X_1, \ldots, X_n$ form a random sample of d-variate random vectors from a distribution which depends on $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Let $f(x_1; \boldsymbol{\theta})$ for $x_1 \in \mathbb{R}^d$ be the density of X_1 w.r.t. a σ-finite measure μ on $\mathbb{R}^d$. Suppose that the support $M = \{x_1 \in \mathbb{R}^d\colon f(x_1; \boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and assume that $f(x_1; \boldsymbol{\theta}_1) = f(x_1; \boldsymbol{\theta}_2)$ for μ-almost all $x_1 \in \mathbb{R}^d$ if and only if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$. Denote by $\boldsymbol{\theta}_X$ the true value of the parameter from which $X_1, \ldots, X_n$ are sampled. Then for any fixed $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that $\boldsymbol{\theta} \neq \boldsymbol{\theta}_X$ we have that*

$$\mathsf{P}_{\boldsymbol{\theta}_X}\left(\prod_{i=1}^n f(X_i; \boldsymbol{\theta}_X) > \prod_{i=1}^n f(X_i; \boldsymbol{\theta})\right) \xrightarrow[n\to\infty]{} 1.$$

Proof. The inequality $\prod_{i=1}^n f(X_i;\boldsymbol{\theta}_X) > \prod_{i=1}^n f(X_i;\boldsymbol{\theta})$ is equivalent with

$$\frac{1}{n}\sum_{i=1}^n \log \frac{f(X_i;\boldsymbol{\theta})}{f(X_i;\boldsymbol{\theta}_X)} < 0. \tag{32}$$

This is an average of independent and identically distributed random variables with expectation

$$\begin{aligned} \mathsf{E}_{\boldsymbol{\theta}_X} \log \frac{f(X_i;\boldsymbol{\theta})}{f(X_i;\boldsymbol{\theta}_X)} < \log \mathsf{E}_{\boldsymbol{\theta}_X} \frac{f(X_i;\boldsymbol{\theta})}{f(X_i;\boldsymbol{\theta}_X)} &= \log \int_M \frac{f(x_i;\boldsymbol{\theta})}{f(x_i;\boldsymbol{\theta}_X)} f(x_i;\boldsymbol{\theta}_X)\,\mathrm{d}\,\mu(x_i) \\ &= \log \int_{\mathbb{R}^d} f(x_i;\boldsymbol{\theta})\,\mathrm{d}\,\mu(x_i) = \log 1 = 0. \end{aligned} \tag{33}$$

The inequality in (33) follows from Jensen's inequality applied to the strictly concave function $\log(x)$ for $x \in (0,\infty)$. The inequality is strict because of our identifiability assumption, which guarantees that $f(X_i;\boldsymbol{\theta})/f(X_i;\boldsymbol{\theta}_X)$ is not a constant almost surely. Indeed, because $\boldsymbol{\theta}_X \neq \boldsymbol{\theta}$, we know that $f(x_1;\boldsymbol{\theta})/f(x_1;\boldsymbol{\theta}_X) \neq 1$ in a set of $x_1 \in \mathbb{R}^d$ of positive μ-measure. The ratio $f(x_1;\boldsymbol{\theta})/f(x_1;\boldsymbol{\theta}_X)$ also cannot take any other constant value $\lambda \neq 1$ because in that case

$$1 = \int_{\mathbb{R}^d} f(x_1;\boldsymbol{\theta})\,\mathrm{d}\,\mu(x_1) = \lambda \int_{\mathbb{R}^d} f(x_1;\boldsymbol{\theta}_X)\,\mathrm{d}\,\mu(x_1) = \lambda,$$

a contradiction. The equality on the second line of (33) follows from our assumption about the common support of all densities in the system. From (33) we see that in (32) we have an average of independent identically distributed random variables with a negative expectation. Thus, the law of large numbers applies, and almost surely as $n \to \infty$ the left hand side of (32) converges to a negative quantity. Note that this argument applies also to the case when the expectation is $-\infty$, in which case it can be shown that the sum on the left hand side of (32) converges almost surely to $-\infty$. The proof of the last claim follows via a truncation argument, i.e. by applying the standard law of large numbers to the sequence

$$\frac{1}{n}\sum_{i=1}^n \max\left\{-M, \log \frac{f(X_i;\boldsymbol{\theta})}{f(X_i;\boldsymbol{\theta}_X)}\right\} \qquad \text{for } M > 0,$$

and then taking $M \to \infty$; for details see [3, Theorem 2.4.5]. □

Theorem 21 is very general, but does not immediately guarantee the consistency of the maximum likelihood estimator, since the limit expression is valid only for a single parameter value $\boldsymbol{\theta} \neq \boldsymbol{\theta}_X$ fixed. It therefore guarantees consistency only if the set $\boldsymbol{\Theta}$ has finitely many elements, which is a rare situation. To prove better results about the consistency of MLE, more assumptions need to be introduced, and the results become more elaborate. We will see some of such results in our separate treatment of one-dimensional, and multi-dimensional MLE, respectively.

We now turn to the additional properties of the maximum likelihood estimators and derived quantities. We begin with the simpler case of a parameter of dimension $p = 1$.

The end of lecture 6 (26.3.2024)

2.2 Properties of MLE — one-dimensional parameter

Throughout this section, we need the following assumptions.

(P_1) The parametric space $\Theta \subseteq \mathbb{R}$ contains the true value of the parameter θ_X in its interior. In other words, there exists an open interval $\Theta_0 \subseteq \Theta$ such that $\theta_X \in \Theta_0$.

(P_2) The random vector $\boldsymbol{X} = (X_1, \ldots, X_n)^{\mathsf{T}}$ corresponds to a random sample $X_1, \ldots, X_n$, where the random variable X_i has a density $f(x;\theta)$ w.r.t. a σ-finite measure μ on $\mathbb{R}^d$.

(P_3) The support $M = \{x \in \mathbb{R}^d \colon f(x;\theta) > 0\}$ does not depend on $\theta \in \Theta$.

(P_4) For any $\theta_1, \theta_2 \in \Theta$ we have that $f(x;\theta_1) = f(x;\theta_2)$ for μ-almost all $x \in \mathbb{R}^d$ if and only if $\theta_1 = \theta_2$.

Conditions (P_1)–(P_4) are again quite natural. (P_1) is here to be able to differentiate the likelihood. Condition (P_2) ensures that the joint density of $\boldsymbol{X}$ takes the form $f(\boldsymbol{x};\theta) = \prod_{i=1}^n f(x_i;\theta)$ when considered w.r.t the product measure $\nu = \bigotimes_{i=1}^n \mu$. The stated form of ($P_2$) is still somewhat strict, and weaker versions of our results can be found in the literature even if the data are not forming a random sample. Assumption (P_3) is standard, and allows us to avoid pathologies such as those arising in Example 1.8, where estimators outperforming the Rao-Cramér bound exist. Finally, (P_4) is a simple identifiability criterion that guarantees that no two different parameter values can correspond to the same density. Note that for $p = 1$, these conditions are exactly analogous to the assumptions of Theorem 21. For $p = 1$, we can however prove more about the consistency of the maximum likelihood estimator.

Theorem 22. *Let the conditions (P_1)–(P_4) be satisfied, and in addition suppose that for every $\theta \in \Theta_0$ the derivative $f'(x;\theta) = \frac{\partial f(x;\theta)}{\partial \theta}$ exists for μ-almost all $x \in M$. Then as $n \to \infty$ with probability converging to one does the likelihood equation*

$$\frac{\partial L_n(\theta;\boldsymbol{x})}{\partial \theta} = 0 \tag{34}$$

have a root $\widehat{\theta}_n(\boldsymbol{x})$ such that the estimator $\widehat{\theta}_n(\boldsymbol{X})$ converges to the true value θ_X in probability.

Proof. Let $\varepsilon > 0$ be small enough so that $[\theta_X - \varepsilon, \theta_X + \varepsilon] \subset \Theta_0$. Set

$$S_n = \left\{\boldsymbol{x} \in \mathbb{R}^{dn} \colon L_n(\theta_X;\boldsymbol{x}) > L_n(\theta_X - \varepsilon;\boldsymbol{x}) \text{ and } L_n(\theta_X;\boldsymbol{x}) > L_n(\theta_X + \varepsilon;\boldsymbol{x})\right\}.$$

Here, the points $\theta_X - \varepsilon$ and $\theta_X + \varepsilon$ are both elements of $\boldsymbol{\Theta}$ and are fixed. For $A^{\mathsf{C}} = \Omega \setminus A$, we use the simple inequality $\mathsf{P}(A \cap B) = 1 - \mathsf{P}(A^{\mathsf{C}} \cup B^{\mathsf{C}}) \geq 1 - \mathsf{P}(A^{\mathsf{C}}) - \mathsf{P}(B^{\mathsf{C}})$ to get that

$$\begin{aligned}\mathsf{P}_{\theta_X}(\boldsymbol{X} \in S_n) &= \mathsf{P}_{\theta_X}\left((L_n(\theta_X;\boldsymbol{X}) > L_n(\theta_X - \varepsilon;\boldsymbol{X})) \cap (L_n(\theta_X;\boldsymbol{X}) > L_n(\theta_X + \varepsilon;\boldsymbol{X}))\right) \\ &\geq 1 - \mathsf{P}_{\theta_X}(L_n(\theta_X;\boldsymbol{X}) \leq L_n(\theta_X - \varepsilon;\boldsymbol{X})) - \mathsf{P}_{\theta_X}(L_n(\theta_X;\boldsymbol{X}) \leq L_n(\theta_X + \varepsilon;\boldsymbol{X}))\end{aligned}$$

where the right hand side converges to 1 because of Theorem 21. We obtain that

$$\mathsf{P}_{\theta_X}\left(\boldsymbol{X} \in S_n\right) \xrightarrow[n\to\infty]{} 1.$$

Take now $\boldsymbol{x} \in S_n$. The likelihood $L_n(\theta; \boldsymbol{x})$ as a function of θ takes a higher value at a point θ_X inside the interval $[\theta_X - \varepsilon, \theta_X + \varepsilon]$ than at any of the two points of its boundary. Because of our assumption, the function $L_n(\theta; \boldsymbol{x})$ is differentiable in $\theta \in \Theta_0$, so it must be also continuous in $\theta \in \Theta_0$. Therefore, it must attain a local maximum at some $\widehat{\theta}_n^{\varepsilon}(\boldsymbol{x}) \in (\theta_X - \varepsilon, \theta_X + \varepsilon)$, and as a differentiable function, at $\theta = \widehat{\theta}_n^{\varepsilon}(\boldsymbol{x})$ it must satisfy equation (34). Clearly, $\left|\widehat{\theta}_n^{\varepsilon}(\boldsymbol{x}) - \theta_X\right| < \varepsilon$.

Overall, we have found that for any $\varepsilon > 0$ small enough, there exists a sequence $\left\{\widehat{\theta}_n^{\varepsilon}\left(\boldsymbol{X}\right)\right\}_{n=1}^{\infty}$ such that

$$\mathsf{P}_{\theta_X}\left(\left|\widehat{\theta}_n^{\varepsilon}(\boldsymbol{X}) - \theta_X\right| < \varepsilon\right) \xrightarrow[n\to\infty]{} 1. \tag{35}$$

It remains to show that such a sequence can be found also independent of ε.

Let $\widehat{\theta}_n(\boldsymbol{x})$ be a root of the likelihood equation (34) as above that is the closest to θ_X. In case there is an infinite sequence of roots, denote by $a_n(\boldsymbol{x})$ the infimum of distances $|\theta_X - r_n(\boldsymbol{x})|$ over all roots $r_n(\boldsymbol{x})$ of (34), and take as $\widehat{\theta}_n(\boldsymbol{x})$ any root of (34) satisfying $\left|\theta_X - \widehat{\theta}_n(\boldsymbol{x})\right| \leq a_n(\boldsymbol{x}) + 1/n$. Such a root exists by the definition of an infimum, so $\widehat{\theta}_n(\boldsymbol{x})$ is well defined. Certainly $\widehat{\theta}_n(\boldsymbol{x})$ does not depend on ε, and at the same time $\left|\widehat{\theta}_n(\boldsymbol{x}) - \theta_X\right| \leq \left|\widehat{\theta}_n^{\varepsilon}(\boldsymbol{x}) - \theta_X\right| + 1/n$. Allowing $\boldsymbol{x}$ to be random again, from (35) we get that for any $\varepsilon > 0$ small enough

$$\mathsf{P}_{\theta_X}\left(\left|\widehat{\theta}_n(\boldsymbol{X}) - \theta_X\right| < \varepsilon + 1/n\right) \geq \mathsf{P}_{\theta_X}\left(\left|\widehat{\theta}_n^{\varepsilon}(\boldsymbol{X}) - \theta_X\right| < \varepsilon\right) \xrightarrow[n\to\infty]{} 1.$$

We get that $\left|\widehat{\theta}_n(\boldsymbol{X}) - \theta_X\right| \xrightarrow[n\to\infty]{\mathsf{P}} 0$ as we wanted to show. □

It is important to give several remarks about Theorem 22. It is not claimed that the maximum likelihood estimator is a consistent estimator of θ. The theorem only says that there exists a sequence of local maxima of the likelihood functions that converges to the true value of θ. These local maxima do not have to be global, and in fact it is possible to construct examples where some sequence of local maxima is consistent, but the corresponding sequence of the global maxima (that is, the sequence of the maximum likelihood estimators) is not.

Further, even though Theorem 22 ensures that some sequence of local maxima of the likelihood function is consistent, in practice we do not know which one, because, of course, the true value θ_X is not known. The only exception is the case when there is only a single local maximum of the likelihood function.

Theorem 22 also does not guarantee the existence of a root $\widehat{\theta}_n(\boldsymbol{x})$ as in its statement for all $\boldsymbol{x}$, or even if the sample size n is fixed for any given $\boldsymbol{x} \in \mathbb{R}^{dn}$. It only asserts that as $n \to \infty$, with probability increasing to one, a consistent root can be found.

We are now interested in the distributional asymptotics of the maximum likelihood estimator. Before stating our main result, we need the following definition.

Definition 13. We say that a sequence of random variables $\{Y_n\}_{n=1}^{\infty}$ is *bounded in probability* if for every $\varepsilon > 0$ there exists $K > 0$ and an index $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ we have $\mathsf{P}(|Y_n| > K) < \varepsilon$.

For the proofs of the following theorems, a lemma will be useful.

Lemma 7. *The following is true:*

(i) If the sequence $\{Y_n\}_{n=1}^{\infty}$ of random variables converges in distribution, then it is bounded in probability.

(ii) Let $\{Y_n\}_{n=1}^{\infty}$ and $\{Z_n\}_{n=1}^{\infty}$ be sequences of random variables that satisfy $|Z_n| \leq |Y_n|$ for all $n \in \mathbb{N}$ almost surely. If $\{Y_n\}_{n=1}^{\infty}$ is bounded in probability, then also $\{Z_n\}_{n=1}^{\infty}$ is bounded in probability.

Proof. The notion of boundedness in probability is clearly equivalent with the idea of tightness of the laws of the corresponding random variables. On the other hand, convergence in distribution is equivalent with weak convergence of the underlying laws. It is a well known fact ([7, Theorem 12.8], or [2, Theorem 11.5.4]) that a weakly convergent sequence of measures is tight, which proves part (i) of the lemma.

For part (ii) it is enough to realise that

$$\mathsf{P}(|Z_n| > K) \leq \mathsf{P}(|Y_n| > K) < \varepsilon,$$

and setting the same K for Z_n as for Y_n is enough to show that Z_n are bounded in probability. □

In the following theorem, the Fisher information $J(\theta) = J_1(\theta)$ is the information contained in a single observation X_i.

Theorem 23. *Let $\{f(x;\theta) \colon \theta \in \Theta\}$ be a regular system of densities with the Fisher information $J(\theta)$. Suppose that the assumptions (P_1)–(P_4) are satisfied, and let the following be true:*

1. *The partial derivative $f'''(x;\theta) = \frac{\partial^3 f(x;\theta)}{\partial \theta^3}$ exists for μ-almost all $x \in M$, for all $\theta \in \Theta_0$.*
2. *For all $\theta \in \Theta_0$ we have*
$$\int_M f''(x;\theta)\,\mathrm{d}\,\mu(x) = 0.$$

3. *There exists a function* $H(x) \geq 0$ *so that*

$$\mathsf{E}_{\boldsymbol{\theta}_X} H(X_1) < \infty,$$

and

$$\left| \frac{\partial^3 \log f(x;\theta)}{\partial \theta^3} \right| \leq H(x) \text{ for all } \theta \in \Theta_0 \text{ and } \mu\text{-almost all } x \in M. \tag{36}$$

Then the following holds true:

(i) We have

$$\frac{1}{\sqrt{n}} L_n'(\theta_X) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}\left(0, J(\theta_X)\right). \tag{37}$$

(ii) Any consistent sequence $\widehat{\theta}_n = \widehat{\theta}_n(\boldsymbol{X})$ *of roots of the system of likelihood equations satisfies*

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_X\right) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}\left(0, \frac{1}{J\left(\theta_X\right)}\right). \tag{38}$$

Proof. For part (i) note that[6]

$$\frac{1}{\sqrt{n}} L_n'(\theta_X) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f(X_i;\theta_X)}{\partial \theta}.$$

The random variables $\frac{\partial \log f(X_i;\theta_X)}{\partial \theta}$ are independent and identically distributed for $i = 1, \ldots, n$. We have

$$\mathsf{E}_{\theta_X} \frac{\partial \log f(X_i;\theta_X)}{\partial \theta} = \int_M f'(x;\theta_X)\,\mathrm{d}\,\mu(x) = 0$$

by (R_3) from the regularity of the system of densities of $\boldsymbol{X}$, and

$$\mathsf{var}_{\theta_X} \frac{\partial \log f(X_i;\theta_X)}{\partial \theta} = \int_M \left(\frac{f'(x;\theta_X)}{f(x;\theta_X)}\right)^2 f(x;\theta_X)\,\mathrm{d}\,\mu(x) = J(\theta_X)$$

by (R_5). Thus, it is enough to apply the central limit theorem to conclude.

The proof of part (ii) is more involved. For $\boldsymbol{x}$ fixed, we use Taylor's expansion of $L_n'(\widehat{\theta}_n)$ around θ_X to get

$$L_n'(\widehat{\theta}_n) = L_n'(\theta_X) + (\widehat{\theta}_n - \theta_X) L_n''(\theta_X) + \frac{1}{2}(\widehat{\theta}_n - \theta_X)^2 L_n'''(\theta_n^*),$$

where θ_n^* lies in the interval between $\widehat{\theta}_n$ and θ_X. Because $\widehat{\theta}_n$ is a root of the likelihood equation, $L_n'(\widehat{\theta}_n) = 0$, and we can rewrite the previous formula into

$$\sqrt{n}(\widehat{\theta}_n - \theta_X) = \frac{\frac{L_n'(\theta_X)}{\sqrt{n}}}{-\frac{L_n''(\theta_X)}{n} - \frac{1}{2n}(\widehat{\theta}_n - \theta_X) L_n'''(\theta_n^*)}. \tag{39}$$

[6] For a function $f(x;\theta)$ we write $\frac{\partial f(x;\theta_X)}{\partial \theta}$ for the partial derivative of $f(x;\theta)$ w.r.t. θ, taken at $\theta = \theta_X$. This notation is used throughout this section.

We performed the Taylor's expansion for $\boldsymbol{x}$ fixed, but since it can be done for any $\boldsymbol{x}$, it must be true also for $\boldsymbol{x}$ replaced by the random vector $\boldsymbol{X}$. Thus, at this point it is important to realise that all L_n', L_n'' and L_n''' depend also on the random sample $\boldsymbol{X}$. As $n \to \infty$, the numerator converges in distribution to $\mathsf{N}\left(0, J(\theta_X)\right)$ by part (i). For the first term in the denominator in (39) we have by the law of large numbers and Theorem 1

$$-\frac{L_n''(\theta_X)}{n} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(X_i;\theta_X)}{\partial\theta^2} \xrightarrow[n\to\infty]{\mathsf{P}} -\mathsf{E}_{\theta_X} \frac{\partial^2 \log f(X_1;\theta_X)}{\partial\theta^2} = J(\theta_X). \tag{40}$$

Finally, for the second summand in the denominator we know that because $\widehat{\theta}_n$ is a consistent estimator of $\theta_X \in \Theta_0$ and θ_n^* lies between $\widehat{\theta}_n$ and θ_X, for all n large enough also θ_n^* lies in Θ_0 with high probability. Thus, whenever $\theta_n^* \in \Theta_0$, we can use our assumption (36) and bound

$$\left|\frac{1}{2\,n} L_n'''(\theta_n^*)\right| = \left|\frac{1}{2\,n}\sum_{i=1}^{n} \frac{\partial^3 \log f(X_i;\theta_n^*)}{\partial\theta^3}\right| \le \frac{1}{2\,n}\sum_{i=1}^{n} H(X_i) \xrightarrow[n\to\infty]{\mathsf{P}} \frac{\mathsf{E}_{\theta_X} H(X_1)}{2} < \infty. \tag{41}$$

The last limit follows from the law of large numbers. We now use the last bound to show that the second summand in the denominator of (39) converges to zero in probability. To show that, let $\varepsilon > 0$ be given and compute

$$\begin{aligned}
\mathsf{P}_{\theta_X}\left(\left|-\left(\widehat{\theta}_n - \theta_X\right)\frac{1}{2\,n} L_n'''(\theta_n^*)\right| > \varepsilon\right) &= \mathsf{P}_{\theta_X}\left(\left|\widehat{\theta}_n - \theta_X\right|\left|\frac{1}{2\,n} L_n'''(\theta_n^*)\right| > \varepsilon, \widehat{\theta}_n \in \Theta_0\right) \\
&+ \mathsf{P}_{\theta_X}\left(\left|\widehat{\theta}_n - \theta_X\right|\left|\frac{1}{2\,n} L_n'''(\theta_n^*)\right| > \varepsilon, \widehat{\theta}_n \notin \Theta_0\right) \\
&\le \mathsf{P}_{\theta_X}\left(\left|\widehat{\theta}_n - \theta_X\right|\left|\frac{1}{2\,n} L_n'''(\theta_n^*)\right| > \varepsilon, \theta_n^* \in \Theta_0\right) + \mathsf{P}_{\theta_X}\left(\widehat{\theta}_n \notin \Theta_0\right) \\
&\le \mathsf{P}_{\theta_X}\left(\left|\widehat{\theta}_n - \theta_X\right|\frac{1}{2\,n}\sum_{i=1}^{n} H(X_i) > \varepsilon\right) + \mathsf{P}_{\theta_X}\left(\widehat{\theta}_n \notin \Theta_0\right).
\end{aligned}$$

We wrote the first equality because simply either $\widehat{\theta}_n$ lies in Θ_0, or it does not. In the following inequality we used that because θ_n^* lies between θ_X and $\widehat{\theta}_n$, necessarily $\widehat{\theta}_n \in \Theta_0$ implies $\theta_n^* \in \Theta_0$. In the second inequality we used the bound (41), under the condition that $\theta_n^* \in \Theta_0$.

On the right hand side of the last formula we have the random variable

$$\left|\widehat{\theta}_n - \theta_X\right|\frac{1}{2\,n}\sum_{i=1}^{n} H(X_i) \xrightarrow[n\to\infty]{\mathsf{P}} 0, \tag{42}$$

where the convergence follows because by our assumption, $\widehat{\theta}_n$ is a consistent estimator of θ_X, meaning that $\widehat{\theta}_n(\boldsymbol{X}) \xrightarrow[n\to\infty]{\mathsf{P}} \theta_X$, and the expression with H converges in probability to a finite constant by (41). Thus, by the continuous mapping theorem, also the product of these two random variables converges to zero in probability. From this we have that for any $\varepsilon > 0$

$$\mathsf{P}_{\theta_X}\left(\left|\widehat{\theta}_n - \theta_X\right|\frac{1}{2\,n}\sum_{i=1}^{n} H(X_i) > \varepsilon\right) \xrightarrow[n\to\infty]{} 0,$$

but also because $\widehat{\theta}_n$ is a consistent estimator of $\theta_X \in \Theta_0$

$$\mathsf{P}_{\theta_X}\left(\widehat{\theta}_n \notin \Theta_0\right) \xrightarrow[n\to\infty]{} 0.$$

Combining both results we obtain

$$\mathsf{P}_{\theta_X}\left(\left|-\left(\widehat{\theta}_n - \theta_X\right)\frac{1}{2\,n}L_n'''(\theta_n^*)\right| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0$$

for each $\varepsilon > 0$, or that the second summand in the denominator of (39) vanishes in probability as $n \to \infty$.

It remains to combine all together. We found that

$$\begin{aligned}\frac{L_n'(\theta_X)}{\sqrt{n}} &\xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}\left(0, J(\theta_X)\right),\\ -\frac{L_n''(\theta_X)}{n} &\xrightarrow[n\to\infty]{\mathsf{P}} J(\theta_X),\end{aligned}$$

and

$$-\frac{1}{2\,n}(\widehat{\theta}_n - \theta_X)L_n'''(\theta_n^*) \xrightarrow[n\to\infty]{\mathsf{P}} 0.$$

By the continuous mapping theorem

$$-\frac{L_n''(\theta_X)}{n} - \frac{1}{2\,n}(\widehat{\theta}_n - \theta_X)L_n'''(\theta_n^*) \xrightarrow[n\to\infty]{\mathsf{P}} J(\theta_X),$$

and finally by the Cramér-Slutsky theorem and (39)

$$\sqrt{n}(\widehat{\theta}_n - \theta_X) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}\left(0, \frac{1}{J(\theta_X)}\right).$$

□

2.3 Asymptotically efficient estimation based on MLE

Comparing our main result of Theorem 23 with the Rao-Cramér bound of Theorem 3 we see that the distribution of the consistent root of the likelihood equations $\widehat{\theta}_n$ is approximately

$$\mathsf{N}\left(\theta_X, \left(n\,J(\theta_X)\right)^{-1}\right),$$

meaning that the mean of the asymptotic distribution is the true parameter θ_X and its variance is $\left(n\,J(\theta_X)\right)^{-1} = \left(J_n(\theta_X)\right)^{-1}$. This should be compared with the Rao-Cramér bound that states that among all unbiased estimators of θ, minimum variance that is possible to be attained is $\left(J_n(\theta_X)\right)^{-1}$. These two statements are remarkably close, and suggest that the maximum likelihood estimation may produce consistent estimators with decent properties. In one way this is true. But, one has to be careful when formulating such statements, especially for the following reasons:

1. The maximum likelihood estimators are not asymptotically unbiased in the sense of Definition 2. They only satisfy that the expectation of the asymptotic distribution of $\sqrt{n}\left(\widehat{\theta}_n - \theta_X\right)$ equals 0. But, this does not mean that they have to be unbiased, even under the conditions of Theorems 22 and 23. The convergence in (38) to a centred random variable in distribution does not in general imply that that $\mathsf{E}_{\theta_X}\sqrt{n}\left(\widehat{\theta}_n - \theta_X\right)$ converges to 0, or that the bias of $\widehat{\theta}_n$ converges to 0, as $n \to \infty$. Strictly speaking, the Rao-Cramér bound is therefore not applicable to maximum likelihood estimators.

2. Likewise, the variance of the asymptotic distribution in (38) does not imply that $\mathsf{var}_{\theta_X}\widehat{\theta}_n$ approaches $(J_n(\theta_X))^{-1}$ as $n \to \infty$. There are examples where the variance of $\widehat{\theta}_n$ does not even exists for any $n \in \mathbb{N}$.

Under the assumptions of Theorem 22 we know that a consistent sequence of roots to the likelihood equations exists. We can therefore apply Theorem 23 to this sequence, and obtain the convergence in (38). In practice, it may be however difficult to identify the root of the likelihood equation that results in a consistent estimator of θ_X, especially as the equations can have many different roots. In that case, three standard approaches are possible to be taken.

1. Theorem 22 was formulated for a sequence of local maxima of the likelihood function, and thus not necessarily for the maximum likelihood estimators directly. Under additional technical assumptions such as those given in [12], it is possible to prove that also the sequence of maximum likelihood estimators consistently estimates θ_X. These results are however much more involved than our proofs, and their conditions are not always simple to verify.

2. Suppose that we are given any sequence of consistent estimators $\delta_n = \delta_n(\boldsymbol{X})$ of θ. For each $n \in \mathbb{N}$ pick $\widehat{\theta}_n$ to be the root of the likelihood equation that is closest to δ_n. Such a closest root exists by the proof of Theorem 22. By Theorem 22 we know that there exists a consistent sequence of roots of the likelihood equations $\widetilde{\theta}_n$. Because both δ_n and $\widetilde{\theta}_n$ are consistent, we get $\left|\delta_n - \widetilde{\theta}_n\right| \xrightarrow[n\to\infty]{\mathsf{P}} |\theta_X - \theta_X| = 0$. Hence, because $\widehat{\theta}_n$ is defined as the closest root to δ_n, we have $\left|\delta_n - \widehat{\theta}_n\right| \leq \left|\delta_n - \widetilde{\theta}_n\right| \xrightarrow[n\to\infty]{\mathsf{P}} 0$, and also the sequence $\widehat{\theta}_n$ is consistent for θ. Therefore, Theorem 23 can be applied to the sequence $\widehat{\theta}_n$, and the asymptotic distribution (38) is valid.

3. Another interesting method is the refinement of an initial consistent estimator of θ by means of the maximum likelihood estimation. We begin from the Newton-Raphson iterative procedure for finding the roots of an equation numerically. The procedure for

finding a root of the likelihood equation

$$L_n'(\theta) = 0 \tag{43}$$

proceeds by replacing the left hand side by its Taylor's expansion about some approximate solution $\widetilde{\theta}_n$. If $\widehat{\theta}_n$ denotes the root of (43), we approximate

$$0 = L_n'(\widehat{\theta}_n) \approx L_n'(\widetilde{\theta}_n) + (\widehat{\theta}_n - \widetilde{\theta}_n)L_n''(\widetilde{\theta}_n),$$

which leads to the approximate equality

$$\widehat{\theta}_n \approx \widetilde{\theta}_n - \frac{L_n'(\widetilde{\theta}_n)}{L_n''(\widetilde{\theta}_n)}. \tag{44}$$

The Newton-Raphson procedure for finding a root of (43) proceeds iteratively. It initialises in some value $\theta_0 \in \Theta$, and plugs θ_0 instead of $\widetilde{\theta}_n$ into the right hand side of (44), obtaining θ_1. This procedure is usually iterated until convergence.

In our task of statistical estimation, we use the iterative procedure in (44) to improve an initial estimator $\widetilde{\theta}_n$ of θ. We plug $\widetilde{\theta}_n$ into the right hand side of (44) and obtain the so-called *one-step Newton-Raphson estimator* based on $\widetilde{\theta}_n$ given by

$$\delta_n = \widetilde{\theta}_n - \frac{L_n'(\widetilde{\theta}_n)}{L_n''(\widetilde{\theta}_n)}. \tag{45}$$

It turns out that if the estimator $\widetilde{\theta}_n$ has nice properties, the improved estimator δ_n already satisfies the asymptotic efficiency as in (38).

Theorem 24. *Suppose that the assumptions of Theorem 23 are satisfied. Let $\widetilde{\theta}_n$ be a $\sqrt{n}$-consistent estimator of θ, meaning that $\sqrt{n}\left(\widetilde{\theta}_n - \theta_X\right)$ is bounded in probability. Then the estimator δ_n given by (45) has the property*

$$\sqrt{n}\,(\delta_n - \theta_X) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}\left(0, \frac{1}{J(\theta_X)}\right).$$

Proof. [7] The proof is similar to that of Theorem 23. We apply the Taylor expansion to the function $L_n'(\widetilde{\theta}_n)$ around θ_X to get

$$L_n'(\widetilde{\theta}_n) = L_n'(\theta_X) + (\widetilde{\theta}_n - \theta_X)L_n''(\theta_X) + \frac{1}{2}(\widetilde{\theta}_n - \theta_X)^2 L_n'''(\theta_n^*)$$

for θ_n^* lying between $\widetilde{\theta}_n$ and θ_X. Plug this expression into (45) to get

$$\delta_n = \widetilde{\theta}_n - \frac{L_n'(\theta_X) + (\widetilde{\theta}_n - \theta_X)L_n''(\theta_X) + \frac{1}{2}(\widetilde{\theta}_n - \theta_X)^2 L_n'''(\theta_n^*)}{L_n''(\widetilde{\theta}_n)}.$$

[7] This proof was not done at the lectures.

From here we express

$$\sqrt{n}\,(\delta_n - \theta_X) = \sqrt{n}\left(\widetilde{\theta}_n - \theta_X\right) - \frac{\sqrt{n}\,L_n'(\theta_X)}{L_n''(\widetilde{\theta}_n)} - \sqrt{n}\frac{(\widetilde{\theta}_n - \theta_X)L_n''(\theta_X) + \frac{1}{2}(\widetilde{\theta}_n - \theta_X)^2 L_n'''(\theta_n^*)}{L_n''(\widetilde{\theta}_n)}$$
$$= \underbrace{-\frac{\sqrt{n}\,L_n'(\theta_X)}{L_n''(\widetilde{\theta}_n)}}_{=S_1} + \underbrace{\sqrt{n}\left(\widetilde{\theta}_n - \theta_X\right)\left(1 - \frac{L_n''(\theta_X)}{L_n''(\widetilde{\theta}_n)} - \frac{1}{2}(\widetilde{\theta}_n - \theta_X)\frac{L_n'''(\theta_n^*)}{L_n''(\widetilde{\theta}_n)}\right)}_{=S_2}.$$

We denoted the first summand on the right hand side by S_1 and the second by S_2. We analyse them separately.

For S_1, we have

$$S_1 = -\frac{L_n'(\theta_X)/\sqrt{n}}{L_n''(\theta_X)/n}\frac{L_n''(\theta_X)/n}{L_n''(\widetilde{\theta}_n)/n}.$$

From (37) and (40) we have, using the Cramér-Slutsky theorem,

$$\frac{L_n'(\theta_X)/\sqrt{n}}{L_n''(\theta_X)/n} \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}\left(0, \frac{1}{J(\theta_X)}\right).$$

Next we show that the second factor in S_1 converges in probability to 1. To see this, expand $L_n''(\theta_X)$ around $\widetilde{\theta}_n$ to get

$$L_n''(\theta_X) = L_n''(\widetilde{\theta}_n) + (\theta_X - \widetilde{\theta}_n)L_n'''(\theta_n^{**}),$$

for θ_n^{**} between θ_X and $\widetilde{\theta}_n$. Rewriting the last formula we get

$$\frac{1}{n}L_n''(\theta_X) - \frac{1}{n}L_n''(\widetilde{\theta}_n) = (\theta_X - \widetilde{\theta}_n)\frac{1}{n}L_n'''(\theta_n^{**}).$$

We assumed that $\widetilde{\theta}_n$ is $\sqrt{n}$-consistent, which gives directly that $\widetilde{\theta}_n \xrightarrow[n\to\infty]{\mathsf{P}} \theta_X$. The term $L_n'''(\theta_n^{**})/n$ is dominated by a random variable that converges in probability to a finite constant, as can be proved in the same way as in (41) and we can conclude as in (42) that

$$\frac{1}{n}L_n''(\theta_X) - \frac{1}{n}L_n''(\widetilde{\theta}_n) = (\theta_X - \widetilde{\theta}_n)\frac{L_n'''(\theta_n^{**})}{n} \xrightarrow[n\to\infty]{\mathsf{P}} 0.$$

Because by (40) we have that

$$-\frac{1}{n}L_n''(\theta_X) \xrightarrow[n\to\infty]{\mathsf{P}} J(\theta_X),$$

necessarily also

$$-\frac{1}{n}L_n''(\widetilde{\theta}_n) \xrightarrow[n\to\infty]{\mathsf{P}} J(\theta_X), \tag{46}$$

and we can use the continuous mapping theorem to assert the desired

$$\frac{L_n''(\theta_X)/n}{L_n''(\widetilde{\theta}_n)/n} \xrightarrow[n\to\infty]{\mathsf{P}} 1. \tag{47}$$

This concludes our analysis of S_1, as the Cramér-Slutsky theorem gives

$$S_1 = -\frac{L_n'(\theta_X)/\sqrt{n}}{L_n''(\theta_X)/n}\frac{L_n''(\theta_X)/n}{L_n''(\widetilde{\theta}_n)/n} \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}\left(0, \frac{1}{J(\theta_X)}\right).$$

To finish the proof of the theorem, it remains to show that $S_2 \xrightarrow[n\to\infty]{\mathsf{P}} 0$. By our assumption of $\sqrt{n}$-consistency of $\widetilde{\theta}_n$, we know that $\sqrt{n}\left(\widetilde{\theta}_n - \theta_X\right)$ is bounded in probability. Thus, it is enough to show

$$\frac{L_n''(\theta_X)}{L_n''(\widetilde{\theta}_n)} + \frac{1}{2}(\widetilde{\theta}_n - \theta_X)\frac{L_n'''(\theta_n^*)}{L_n''(\widetilde{\theta}_n)} \xrightarrow[n\to\infty]{\mathsf{P}} 1.$$

By (47) we have that it is enough to show

$$(\widetilde{\theta}_n - \theta_X)\frac{L_n'''(\theta_n^*)/(2\,n)}{L_n''(\widetilde{\theta}_n)/n} \xrightarrow[n\to\infty]{\mathsf{P}} 0. \tag{48}$$

But $\widetilde{\theta}_n - \theta_X \xrightarrow[n\to\infty]{\mathsf{P}} 0$ by our assumption of $\sqrt{n}$-consistency of $\widetilde{\theta}_n$, $L_n'''(\theta_n^*)/(2\,n)$ is dominated by a sequence of random variables that converges in probability to a finite constant as in (41), and by (46) we can write $-L_n''(\widetilde{\theta}_n)/n \xrightarrow[n\to\infty]{\mathsf{P}} J(\theta_X)$. Putting all together we get the needed (48). □

Often, it is not difficult to find estimators $\widetilde{\theta}_n$ of θ that are $\sqrt{n}$-consistent. As initial estimators $\widetilde{\theta}_n$, one could for example use the method of moments. Another important application of Theorem 24 comes with the numerical solution of the likelihood equation. It turns out that often, an explicit solution to the likelihood equation is difficult, or impossible to obtain explicitly. In that case, it is again possible to use maximum likelihood theory in conjunction with an arbitrary initial $\sqrt{n}$-consistent estimator $\widetilde{\theta}_n$ of θ to obtain an estimator (45) that is easier to work with, and is asymptotically as good as the maximum likelihood estimator.

2.4 Extension of MLE — Profile likelihood

Often, the unknown parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} = \mathbb{R}^p$ is p-dimensional, but only one part of $\boldsymbol{\theta}$ is of interest. In that case, we split $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\mathsf{T}$ into two sub-vectors $\boldsymbol{\tau} = (\theta_1, \ldots, \theta_q)^\mathsf{T}$ and $\boldsymbol{\psi} = (\theta_{q+1}, \ldots, \theta_p)^\mathsf{T}$ with $1 \le q < p$, so that $\boldsymbol{\theta}^\mathsf{T} = \left(\boldsymbol{\tau}^\mathsf{T}, \boldsymbol{\psi}^\mathsf{T}\right)$. We are primarily interested in the estimation of $\boldsymbol{\tau}$. The sub-vector $\boldsymbol{\psi}$ is a so-called nuisance parameter whose value is also unknown, but we are not interested in it. In Zehna's invariance result from Theorem 20 we could therefore choose the mapping $\boldsymbol{u}\colon \mathbb{R}^p \to \mathbb{R}^q\colon (\theta_1, \ldots, \theta_p)^\mathsf{T} \mapsto (\theta_1, \ldots, \theta_q)^\mathsf{T}$, and express the induced log-likelihood L_n^* of $\boldsymbol{\tau} = (\theta_1, \ldots, \theta_q)^\mathsf{T}$ from (30) as

$$L_n^*(\boldsymbol{\tau}) = \sup_{\boldsymbol{\psi}\in\mathbb{R}^{p-q}} L_n\left(\left(\boldsymbol{\tau}^\mathsf{T}, \boldsymbol{\psi}^\mathsf{T}\right)^\mathsf{T}\right) \text{ for } \boldsymbol{\tau} \in \mathbb{R}^q.$$

In this expression, we "profiled out" the contribution of the nuisance parameter $\boldsymbol{\psi}$, and obtained only a log-likelihood for our parameter of interest $\boldsymbol{\tau}$; the function L_n^* is therefore called the *profile log-likelihood* of parameter $\boldsymbol{\tau}$. Theorem 20 gives that maximizing the profile log-likelihood in $\boldsymbol{\tau}$ is equivalent with the joint maximization of the log-likelihood $L_n(\boldsymbol{\theta})$ in both $\boldsymbol{\tau}$ and $\boldsymbol{\psi}$. For $q = 1$, results analogous to Theorems 22 and 23 can be derived also for profile log-likelihood. We saw profiling already in Example 2.2, now we consider a more interesting setup of the so-called Box-Cox transforms to normality.

Example 2.3. Let $Y_1, \ldots, Y_n$ form a random sample in $\mathbb{R}$ from a distribution F such that $F(0) = 0$, that is $Y_1 > 0$ almost surely. Because of the assumption of positivity, the distribution F cannot be normal. But, we believe that there might exist a simple transform $g\colon [0, \infty) \to \mathbb{R}$ that will make $g(Y_1)$ normal. Our intent is to find such a transform. We consider the family of Box-Cox transforms given by

$$g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \log y & \text{if } \lambda = 0, \end{cases} \quad \text{for } y > 0.$$

Observe that this family is continuous in λ, because

$$\lim_{\lambda \to 0} g_\lambda(y) = \lim_{\lambda \to 0} \frac{\exp(\lambda \log y) - 1}{\lambda} = \lim_{\lambda \to 0} \sum_{j=0}^{\infty} \left(\frac{1}{\lambda} \frac{(\lambda \log y)^j}{j!} - \frac{1}{\lambda} \right) = \lim_{\lambda \to 0} \sum_{j=1}^{\infty} \lambda^{j-1} \frac{(\log y)^j}{j!} = g_0(y)$$

for all $y > 0$. Also, note that g_λ is essentially just a power transform $y \mapsto y^\lambda$ for $\lambda \neq 0$, properly scaled and shifted to make g_λ satisfy this continuity condition at $\lambda = 0$. We assume that for some $\lambda \in \mathbb{R}$, the transformed distribution of $g_\lambda(Y_1)$ is normal, with parameters $(\mu, \sigma^2)^\mathsf{T} \in \mathbb{R} \times (0, \infty)$. The complete vector of unknown parameters is therefore $\boldsymbol{\theta} = (\lambda, \mu, \sigma^2)^\mathsf{T}$. We want to find $\tau = \lambda$, but are not interested in the values of $\boldsymbol{\psi} = (\mu, \sigma^2)^\mathsf{T}$. To do that we derive the profile log-likelihood of τ. Since we assume that $g_\lambda(Y_1) \sim \mathsf{N}(\mu, \sigma^2)$, we have

$$F(y) = \mathsf{P}(Y_1 \leq y) = \mathsf{P}(g_\lambda(Y_1) \leq g_\lambda(y)) = \Phi\left(\frac{g_\lambda(y) - \mu}{\sigma} \right) \quad \text{for } y > 0,$$

where Φ is the distribution function of $\mathsf{N}(0, 1)$. Taking the derivative of F w.r.t. y we obtain the density of $g_\lambda(Y_1)$

$$f(y) = \frac{\partial}{\partial y} \Phi\left(\frac{g_\lambda(y) - \mu}{\sigma} \right) = \varphi\left(\frac{g_\lambda(y) - \mu}{\sigma} \right) \frac{1}{\sigma} \frac{\partial g_\lambda(y)}{\partial y} = \frac{y^{\lambda - 1}}{\sigma} \varphi\left(\frac{g_\lambda(y) - \mu}{\sigma} \right)$$

for $y > 0$, where φ stands for the density of $\mathsf{N}(0, 1)$. The log-likelihood of $\boldsymbol{\theta}$ is therefore

$$L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (g_\lambda(y_i) - \mu)^2 + (\lambda - 1) \sum_{i=1}^{n} \log y_i.$$

To profile out $\boldsymbol{\psi}$, we fix $\lambda \in \mathbb{R}$ and maximize $L_n(\boldsymbol{\theta})$ in μ and σ^2. We obtain solutions

$$\widetilde{\mu}_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n} g_\lambda(y_i), \quad \text{and} \quad \widetilde{\sigma}_n^2(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \left(g_\lambda(y_i) - \widetilde{\mu}_n(\lambda)\right)^2.$$

Plugging them into $L_n(\boldsymbol{\theta})$ we get the profile log-likelihood of λ in the form

$$L_n^*(\lambda) = -\frac{n}{2}\log\left(2\,\pi\right) - \frac{n}{2}\log\left(\widetilde{\sigma}_n^2(\lambda)\right) - \frac{n}{2} + (\lambda - 1)\sum_{i=1}^{n}\log y_i.$$

This is already a simple function of λ, which can be visualised and inspected. For the resulting transform, we are usually not interested in the exact maximizer $\widehat{\lambda}_n$ of L_n^*. Rather, we choose a "reasonable" and well-interpretable value of λ not too far from $\widehat{\lambda}_n$ — usual choices are $\lambda = 0$ which corresponds to Y_1 being log-normal, $\lambda = 1$ that is equivalent to no transformation, or (half-)integer values $\lambda = -1, 1/2, 2$ etc. △

The Box-Cox transform from Example 2.3 is usually taken as an informal data preprocessing step. Ignoring the nuisance parameters $\left(\mu, \sigma^2\right)^\mathsf{T}$, we first search for a reasonable transform g_λ that takes Y_1 "close to" a normal distribution, then transform all random variables $Y_1, \ldots, Y_n$ into $Z_i = g_\lambda(Y_i)$ for $i = 1, \ldots, n$, and finally work with $Z_1, \ldots, Z_n$ as with a random sample from a normal (or at least more regular) distribution. Note, however, that this procedure is more of a useful heuristic than an exact method of analysis. First, since we assumed that $Y_1 > 0$ almost surely, $Z_i = g_\lambda(Y_i) > -1/\lambda$ almost surely for any $\lambda \neq 0$, and Z_i can thus never be exactly normal if $\lambda \neq 0$. The expression $g_\lambda(Y_1) \sim \mathsf{N}\left(\mu, \sigma^2\right)$ used for the derivation of the log-likelihood in Example 2.3 is therefore only approximate. Second, each Z_i now depends also on the chosen λ, which in turn depends on all $Y_1, \ldots, Y_n$. Thus, the random variables $Z_1, \ldots, Z_n$ are dependent. Both these problems are often ignored in the analysis.

The end of lecture 7 (2.4.2024)

2.5 Properties of MLE — multi-dimensional parameter

Analogously as in the one-dimensional situation, we make the following assumptions.

- (P_1^*) The parameter space $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$ contains the true value of the parameter $\boldsymbol{\theta}_X$ in its interior. In other words, there exists an open ball $\boldsymbol{\Theta}_0 \subseteq \boldsymbol{\Theta}$ such that $\boldsymbol{\theta}_X \in \boldsymbol{\Theta}_0$.
- (P_2^*) The random vector $\boldsymbol{X} = (X_1, \ldots, X_n)^\mathsf{T}$ corresponds to a random sample $X_1, \ldots, X_n$, where the random variable X_i has a density $f(x; \boldsymbol{\theta})$ w.r.t. a σ-finite measure μ on $\mathbb{R}^d$.
- (P_3^*) The support $M = \left\{x \in \mathbb{R}^d \colon f(x; \boldsymbol{\theta}) > 0\right\}$ does not depend on $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.
- (P_4^*) For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$ we have that $f(x; \boldsymbol{\theta}_1) = f(x; \boldsymbol{\theta}_2)$ for μ-almost all $x \in \mathbb{R}^d$ if and only if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

A multivariate version of Theorem 23 is covered by the following statement. Also in this theorem, $\boldsymbol{J}(\boldsymbol{\theta}) = \boldsymbol{J}_1(\boldsymbol{\theta})$ is the Fisher information matrix of $\boldsymbol{\theta}$ contained in a single observation X_1.

Theorem 25. *Let $\{f(x;\boldsymbol{\theta})\colon \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ be a regular system of densities with the Fisher information matrix $\boldsymbol{J}(\boldsymbol{\theta})$. Suppose that the assumptions (P_1^*)–(P_4^*) are satisfied, and let the following be true:*

(I) For μ-almost all $x \in M$ the partial derivative $\frac{\partial^3 f(x;\boldsymbol{\theta})}{\partial\theta_j \partial\theta_k \partial\theta_\ell}$ exists for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, and for all $j, k, \ell = 1, \ldots, p$.

(II) For all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ we have

$$\int_M f''_{j,k}(x;\boldsymbol{\theta})\,\mathrm{d}\,\mu(x) = 0 \text{ for all } j,k = 1,\ldots,p.$$

(III) For all $j, k, \ell = 1, \ldots, p$ there exist functions $M_{j,k,\ell}(x) \geq 0$ so that

$$\mathsf{E}_{\boldsymbol{\theta}_X} M_{j,k,\ell}(X_1) < \infty,$$

and

$$\left|\frac{\partial^3 \log f(x;\boldsymbol{\theta})}{\partial\theta_j \partial\theta_k \partial\theta_\ell}\right| \leq M_{j,k,\ell}(x) \text{ for } \mu\text{-almost all } x \in M \text{ and all } \boldsymbol{\theta} \in \boldsymbol{\Theta}_0.$$

Then the following holds true:

(i) There exists a solution $\widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n(\boldsymbol{X})$ to the likelihood equations that converges in probability to $\boldsymbol{\theta}_X$ as $n \to \infty$.

(ii) For the vector of scores

$$\boldsymbol{U}_n(\boldsymbol{\theta}) = \left(\frac{\partial L_n(\boldsymbol{\theta})}{\partial\theta_1}, \ldots, \frac{\partial L_n(\boldsymbol{\theta})}{\partial\theta_p}\right)^\mathsf{T}$$

it holds true that

$$\frac{1}{\sqrt{n}}\boldsymbol{U}_n(\boldsymbol{\theta}_X) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}_p\left(\mathbf{0}, \boldsymbol{J}(\boldsymbol{\theta}_X)\right).$$

(iii) Any consistent sequence $\widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n(\boldsymbol{X})$ of roots of the system of likelihood equations satisfies

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}_p\left(\mathbf{0}, \left(\boldsymbol{J}\left(\boldsymbol{\theta}_X\right)\right)^{-1}\right). \tag{49}$$

Proof. **Part (i).** For the proof of part (i), we proceed in analogy to the proof of Theorem 22 that established the existence of a consistent solution to the likelihood equations for $p = 1$. Since in the case of a multivariate parameter the neighbourhood of $\boldsymbol{\theta}_X$ is a p-dimensional

ball, we cannot argue directly as for $p = 1$; we need to control the behaviour of the likelihood in the whole neighbourhood of $\boldsymbol{\theta}_X$ uniformly.

Take $\varepsilon > 0$ small enough so that the sphere $S_\varepsilon = \{\boldsymbol{\theta} \in \mathbb{R}^p \colon \|\boldsymbol{\theta} - \boldsymbol{\theta}_X\| = \varepsilon\}$ centred at $\boldsymbol{\theta}_X$ with radius ε is contained in $\boldsymbol{\Theta}_0$. We show that if $\varepsilon > 0$ is small enough, we get

$$L_n(\boldsymbol{\theta}) < L_n(\boldsymbol{\theta}_X) \text{ for all } \boldsymbol{\theta} \in S_\varepsilon \text{ with probability converging to 1.} \tag{50}$$

If this is true, by the same argument as in the proof of Theorem 22 we get that there must exist a local maximum of $L_n(\boldsymbol{\theta})$ inside the open ball centred at $\boldsymbol{\theta}_X$ with radius ε, and this solution to the likelihood equations will be taken for the construction of our estimator. The rest of the proof follows analogously as that of Theorem 22.

Denote by $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\mathsf{T}$ the elements of $\boldsymbol{\theta}$, and analogously write $\boldsymbol{\theta}_X = (\theta_{X,1}, \dots, \theta_{X,p})^\mathsf{T}$. To show (50), we employ the multivariate Taylor expansion of the log-likelihood $L_n(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in S_\varepsilon$ around the true value $\boldsymbol{\theta}_X$ [11, Theorem 11.3.28]. We use the expansion for a fixed value $\boldsymbol{x} \in \mathbb{R}^{dn}$, and in the notation emphasize that the quantities depend on both $\boldsymbol{\theta}$ and $\boldsymbol{x}$. After dividing by n, we get

$$\begin{aligned}
\frac{1}{n} L_n(\boldsymbol{\theta}) - \frac{1}{n} L_n(\boldsymbol{\theta}_X) &= \frac{1}{1!\, n} \sum_{j=1}^{p} A_j(\boldsymbol{x}, \boldsymbol{\theta}_X)\, (\theta_j - \theta_{X,j}) \\
&\quad + \frac{1}{2!\, n} \sum_{j=1}^{p} \sum_{k=1}^{p} B_{j,k}(\boldsymbol{x}, \boldsymbol{\theta}_X)\, (\theta_j - \theta_{X,j})\, (\theta_k - \theta_{X,k}) \\
&\quad + \frac{1}{3!\, n} \sum_{j=1}^{p} \sum_{k=1}^{p} \sum_{\ell=1}^{p} C_{j,k,\ell}(\boldsymbol{x}, \boldsymbol{\theta}_n^*)\, (\theta_j - \theta_{X,j})\, (\theta_k - \theta_{X,k})\, (\theta_\ell - \theta_{X,\ell}) \\
&= S_1(\boldsymbol{x}) + S_2(\boldsymbol{x}) + S_3(\boldsymbol{x})
\end{aligned}$$

where

$$\begin{aligned}
A_j(\boldsymbol{x}, \boldsymbol{\theta}_X) &= \frac{\partial L_n(\boldsymbol{\theta}_X)}{\partial \theta_j} = \sum_{i=1}^{n} \frac{\partial \log f(x_i; \boldsymbol{\theta}_X)}{\partial \theta_j}, \\
B_{j,k}(\boldsymbol{x}, \boldsymbol{\theta}_X) &= \frac{\partial^2 L_n(\boldsymbol{\theta}_X)}{\partial \theta_j \partial \theta_k} = \sum_{i=1}^{n} \frac{\partial^2 \log f(x_i; \boldsymbol{\theta}_X)}{\partial \theta_j \partial \theta_k}, \\
C_{j,k,\ell}(\boldsymbol{x}, \boldsymbol{\theta}_n^*) &= \frac{\partial^3 L_n(\boldsymbol{\theta}_n^*)}{\partial \theta_j \partial \theta_k \partial \theta_\ell} = \sum_{i=1}^{n} \frac{\partial^3 \log f(x_i; \boldsymbol{\theta}_n^*)}{\partial \theta_j \partial \theta_k \partial \theta_\ell},
\end{aligned}$$

for $\boldsymbol{\theta}_n^*$ on the straight line between $\boldsymbol{\theta} \in S_\varepsilon$ and $\boldsymbol{\theta}_X$, meaning that $\boldsymbol{\theta}_n^* \in \boldsymbol{\Theta}_0$. We will show that in the Taylor expansion above, with high probability the summands $S_1(\boldsymbol{X})$ and $S_3(\boldsymbol{X})$ are small compared to $S_2(\boldsymbol{X})$, and that the maximum of $S_2(\boldsymbol{X})$ over $\boldsymbol{\theta} \in S_\varepsilon$ is negative. Thus, the sum $L_n(\boldsymbol{\theta}) - L_n(\boldsymbol{\theta}_X)$ is negative for all $\boldsymbol{\theta} \in S_\varepsilon$ with probability converging to 1, and (50) can be applied.

As usual, we use laws of large numbers to control terms A_j and $B_{j,k}$ when considered as functions of the random variable $\boldsymbol{X}$, as we have

$$\begin{aligned}\frac{1}{n}A_j\left(\boldsymbol{X},\boldsymbol{\theta}_X\right) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f(X_i;\boldsymbol{\theta}_X)}{\partial\theta_j} \xrightarrow[n\to\infty]{\mathsf{P}} \mathsf{E}_{\boldsymbol{\theta}_X}\frac{\partial \log f(X_i;\boldsymbol{\theta}_X)}{\partial\theta_j} = 0,\\ \frac{1}{n}B_{j,k}\left(\boldsymbol{X},\boldsymbol{\theta}_X\right) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(X_i;\boldsymbol{\theta}_X)}{\partial\theta_j\partial\theta_k} \xrightarrow[n\to\infty]{\mathsf{P}} \mathsf{E}_{\boldsymbol{\theta}_X}\frac{\partial^2 \log f(X_i;\boldsymbol{\theta}_X)}{\partial\theta_j\partial\theta_k} = -J_{j,k}(\boldsymbol{\theta}_X)\end{aligned} \tag{51}$$

The first convergence follows by (R_4), the second by our assumption (II) and Theorem 7. Here, of course, $J_{j,k}$ is the (j,k)-th element of the matrix $\boldsymbol{J}$.

As in the previous proofs, the third summand is dominated by a convergent sequence of random variables, as by (III) we have

$$\left|\frac{1}{n}C_{j,k,\ell}(\boldsymbol{X},\boldsymbol{\theta}_n^*)\right| \le \frac{1}{n}\sum_{i=1}^{n}M_{j,k,\ell}(X_i) \xrightarrow[n\to\infty]{\mathsf{P}} \mathsf{E}_{\boldsymbol{\theta}_X}\,M_{j,k,\ell}(X_1) < \infty \tag{52}$$

for each $j,k,\ell = 1,\ldots,p$.

We have now all ready to bound the three terms $S_1(\boldsymbol{X})$, $S_2(\boldsymbol{X})$, $S_3(\boldsymbol{X})$ in probability; for simplicity, we write S_j instead of $S_j(\boldsymbol{X})$. We use that since $\boldsymbol{\theta}\in S_\varepsilon$, necessarily $|\theta_j-\theta_{X,j}|\le\varepsilon$ for each $j=1,\ldots,p$. For S_1, we have

$$|S_1| \le \varepsilon\sum_{j=1}^{p}\frac{|A_j(\boldsymbol{X},\boldsymbol{\theta}_X)|}{n}.$$

Because $A_j(\boldsymbol{X};\boldsymbol{\theta}_X)/n \xrightarrow[n\to\infty]{\mathsf{P}} 0$ if and only if $|A_j(\boldsymbol{X};\boldsymbol{\theta}_X)|/n \xrightarrow[n\to\infty]{\mathsf{P}} 0$, we know that

$$\begin{aligned}\mathsf{P}_{\boldsymbol{\theta}_X}\left(|S_1|\ge p\,\varepsilon^3\right) &\le \mathsf{P}_{\boldsymbol{\theta}_X}\left(\varepsilon\sum_{j=1}^{p}\frac{|A_j(\boldsymbol{X},\boldsymbol{\theta}_X)|}{n}\ge p\,\varepsilon^3\right)\\ &\le \sum_{j=1}^{p}\mathsf{P}_{\boldsymbol{\theta}_X}\left(\frac{|A_j(\boldsymbol{X},\boldsymbol{\theta}_X)|}{n}\ge\varepsilon^2\right)\xrightarrow[n\to\infty]{} 0.\end{aligned}$$

Equivalently,

$$\mathsf{P}_{\boldsymbol{\theta}_X}\left(|S_1|<p\,\varepsilon^3\right)\xrightarrow[n\to\infty]{} 1. \tag{53}$$

As for S_2 we write

$$\begin{aligned}2\,S_2 &= \sum_{j=1}^{p}\sum_{k=1}^{p}\left(-J_{j,k}(\boldsymbol{\theta}_X)\right)\left(\theta_j-\theta_{X,j}\right)\left(\theta_k-\theta_{X,k}\right)\\ &\quad+\sum_{j=1}^{p}\sum_{k=1}^{p}\left(\frac{1}{n}B_{j,k}\left(\boldsymbol{X},\boldsymbol{\theta}_X\right)-\left(-J_{j,k}(\boldsymbol{\theta}_X)\right)\right)\left(\theta_j-\theta_{X,j}\right)\left(\theta_k-\theta_{X,k}\right)\\ &= S_{2,1}+S_{2,2}.\end{aligned}$$

The second summand $S_{2,2}$ above can be bounded using (51) in the same way as we did for S_1 before, and we can write

$$\mathsf{P}_{\boldsymbol{\theta}_X}\left(|S_{2,2}| < p^2\,\varepsilon^3\right) \xrightarrow[n\to\infty]{} 1. \tag{54}$$

As for $S_{2,1}$, note that this is in fact a non-random quadratic form of the Fisher information matrix $\boldsymbol{J}$

$$S_{2,1} = -\left(\boldsymbol{\theta}-\boldsymbol{\theta}_X\right)^{\mathsf{T}} \boldsymbol{J}(\boldsymbol{\theta}_X)\left(\boldsymbol{\theta}-\boldsymbol{\theta}_X\right),$$

taken with $\boldsymbol{\theta} \in S_\varepsilon$. The matrix $\boldsymbol{J}(\boldsymbol{\theta}_X)$ symmetric, and assumed to be positive definite by (R_6). Therefore, it can be diagonalised to the form $\boldsymbol{J}(\boldsymbol{\theta}_X) = \boldsymbol{U}^{\mathsf{T}}\boldsymbol{\Lambda}\boldsymbol{U}$ for $\boldsymbol{U} \in \mathbb{R}^{p\times p}$ orthogonal and $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1,\ldots,\lambda_p)$ a diagonal matrix whose diagonal entries are the ordered eigenvalues $0 < \lambda_p \le \cdots \le \lambda_1$. Using this decomposition we get

$$\begin{aligned} S_{2,1} &= -\left(\boldsymbol{\theta}-\boldsymbol{\theta}_X\right)^{\mathsf{T}}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{\Lambda}\boldsymbol{U}\left(\boldsymbol{\theta}-\boldsymbol{\theta}_X\right) = -\left(\boldsymbol{U}\left(\boldsymbol{\theta}-\boldsymbol{\theta}_X\right)\right)^{\mathsf{T}}\boldsymbol{\Lambda}\boldsymbol{U}\left(\boldsymbol{\theta}-\boldsymbol{\theta}_X\right) \\ &= -\sum_{j=1}^{p}\lambda_j\zeta_j^2 \le -\lambda_p\sum_{j=1}^{p}\zeta_j^2, \end{aligned}$$

where $(\zeta_1,\ldots,\zeta_p)^{\mathsf{T}} = \boldsymbol{U}\left(\boldsymbol{\theta}-\boldsymbol{\theta}_X\right)$ is the image of $\boldsymbol{\theta}-\boldsymbol{\theta}_X$ by the orthogonal transformation given by $\boldsymbol{U}$. Since $\boldsymbol{\theta} \in S_\varepsilon$, we have $\|\boldsymbol{\theta}-\boldsymbol{\theta}_X\| = \varepsilon$, and

$$\begin{aligned} \sum_{j=1}^{p}\zeta_j^2 &= \left\|\boldsymbol{U}\left(\boldsymbol{\theta}-\boldsymbol{\theta}_X\right)\right\|^2 = \left(\boldsymbol{U}\left(\boldsymbol{\theta}-\boldsymbol{\theta}_X\right)\right)^{\mathsf{T}}\boldsymbol{U}\left(\boldsymbol{\theta}-\boldsymbol{\theta}_X\right) \\ &= \left(\boldsymbol{\theta}-\boldsymbol{\theta}_X\right)^{\mathsf{T}}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U}\left(\boldsymbol{\theta}-\boldsymbol{\theta}_X\right) = \left\|\boldsymbol{\theta}-\boldsymbol{\theta}_X\right\|^2 = \varepsilon^2, \end{aligned}$$

where we used the orthogonality of the matrix $\boldsymbol{U}$. In other words, for any $\boldsymbol{\theta} \in S_\varepsilon$ we can bound

$$S_{2,1} \le -\lambda_p\,\varepsilon^2. \tag{55}$$

Finally, for S_3 we have by (52)

$$\mathsf{P}_{\boldsymbol{\theta}_X}\left(\left|\frac{1}{n}C_{j,k,\ell}(\boldsymbol{X},\boldsymbol{\theta}_n^*)\right| \le 2\,\mathsf{E}_{\boldsymbol{\theta}_X}\,M_{j,k,\ell}(X_1)\right) \xrightarrow[n\to\infty]{} 1, \tag{56}$$

and we can bound

$$|S_3| \le \frac{\varepsilon^3}{6}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{\ell=1}^{p}\left|\frac{1}{n}C_{j,k,\ell}(\boldsymbol{X},\boldsymbol{\theta}_n^*)\right|,$$

which together with (56) gives

$$\mathsf{P}_{\boldsymbol{\theta}_X}\left(|S_3| \le \frac{\varepsilon^3}{3}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathsf{E}_{\boldsymbol{\theta}_X}\,M_{j,k,\ell}(X_1)\right) \xrightarrow[n\to\infty]{} 1. \tag{57}$$

It remains to combine all the bounds that we obtained. From (53), (54), (55), and (57) we get that

$$\mathsf{P}_{\boldsymbol{\theta}_X}\left(\sup_{\boldsymbol{\theta}\in S_\varepsilon}(S_1+S_2+S_3)\leq p\,\varepsilon^3+\frac{p^2}{2}\varepsilon^3-\frac{\lambda_p}{2}\varepsilon^2+b\,\varepsilon^3\right)\xrightarrow[n\to\infty]{}1,$$

where we denoted

$$b=\frac{1}{3}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathsf{E}_{\boldsymbol{\theta}_X}\,M_{j,k,\ell}(X_1)\in(0,\infty).$$

For

$$\varepsilon<\frac{\lambda_p}{p(2+p)+2\,b},$$

which is equivalent with $p\,\varepsilon^3+\frac{p^2}{2}\varepsilon^3-\frac{\lambda_p}{2}\varepsilon^2+b\,\varepsilon^3<0$, we thus get

$$\mathsf{P}_{\boldsymbol{\theta}_X}\left(\sup_{\boldsymbol{\theta}\in S_\varepsilon}(S_1+S_2+S_3)<0\right)=\mathsf{P}_{\boldsymbol{\theta}_X}\left(\sup_{\boldsymbol{\theta}\in S_\varepsilon}L_n(\boldsymbol{\theta})<L_n(\boldsymbol{\theta}_X)\right)\xrightarrow[n\to\infty]{}1,$$

as we needed to show in (50). This concludes the proof of part (i).

Part (ii). Part (ii) follows by a direct application of the central limit theorem, analogously as in the proof of Theorem 23.

Part (iii). A detailed proof of part (iii) is quite similar to that of Theorem 23, and involves multivariate Taylor's expansions in the same spirit as in the proof of part (i) above. We denote

$$L'_{n,j}(\boldsymbol{\theta})=\frac{\partial L_n(\boldsymbol{\theta})}{\partial\theta_j},$$
$$L''_{n,j,k}(\boldsymbol{\theta})=\frac{\partial^2 L_n(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k},$$
$$L'''_{n,j,k,\ell}(\boldsymbol{\theta})=\frac{\partial^3 L_n(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k\partial\theta_\ell},$$

for $j,k,\ell=1,\ldots,p$. We start from the system of likelihood equations

$$0=L'_{n,j}(\widehat{\boldsymbol{\theta}}_n)\quad\text{for } j=1,\ldots,p,\tag{58}$$

and denote $\widehat{\boldsymbol{\theta}}_n=\left(\widehat{\theta}_{n,1},\ldots,\widehat{\theta}_{n,p}\right)^{\mathsf{T}}$. Performing a multivariate Taylor's expansion to the j-th equation from this system, we get

$$\begin{aligned}0&=L'_{n,j}(\boldsymbol{\theta}_X)\\&\quad+\frac{1}{1!}\sum_{k=1}^{p}L''_{n,j,k}(\boldsymbol{\theta}_X)\left(\widehat{\theta}_{n,k}-\theta_{X,k}\right)+\frac{1}{2!}\sum_{k=1}^{p}\sum_{\ell=1}^{p}L'''_{n,j,k,\ell}(\boldsymbol{\theta}^*_{n,j})\left(\widehat{\theta}_{n,k}-\theta_{X,k}\right)\left(\widehat{\theta}_{n,\ell}-\theta_{X,\ell}\right)\end{aligned}$$

for some $\boldsymbol{\theta}^*_{n,j}$ on the line segment between $\boldsymbol{\theta}_X$ and $\widehat{\boldsymbol{\theta}}_n$, for all $j=1,\ldots,p$. All these equations can be rewritten in a matrix form, if we use the notation from part (ii) of this theorem

$$\boldsymbol{U}_n(\boldsymbol{\theta})=\left(L'_{n,1}(\boldsymbol{\theta}),\ldots,L'_{n,p}(\boldsymbol{\theta})\right)^{\mathsf{T}}\in\mathbb{R}^p,$$

and write $\boldsymbol{L}_n''(\boldsymbol{\theta})$ for the $(p\times p)$-matrix whose (j,k)-th element is $L_{n,j,k}''(\boldsymbol{\theta})$, and finally denote by $\boldsymbol{R}_{n,j}$ the $(p\times p)$-matrix whose (k,ℓ)-th element is $\frac{1}{2}L_{n,j,k,\ell}'''(\boldsymbol{\theta}_{n,j}^*)$. The system of equations (58) then becomes

$$\begin{aligned}\mathbf{0} &= \boldsymbol{U}_n(\boldsymbol{\theta}_X) + \boldsymbol{L}_n''(\boldsymbol{\theta}_X)\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) + \begin{pmatrix} \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right)^{\mathsf{T}} \boldsymbol{R}_{n,1}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X) \\ \dots \\ \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right)^{\mathsf{T}} \boldsymbol{R}_{n,p}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X) \end{pmatrix} \\ &= \boldsymbol{U}_n(\boldsymbol{\theta}_X) + \boldsymbol{L}_n''(\boldsymbol{\theta}_X)\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) + \begin{pmatrix} \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right)^{\mathsf{T}} \boldsymbol{R}_{n,1} \\ \dots \\ \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right)^{\mathsf{T}} \boldsymbol{R}_{n,p} \end{pmatrix} (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X) \\ &= \boldsymbol{U}_n(\boldsymbol{\theta}_X) + \left(\boldsymbol{L}_n''(\boldsymbol{\theta}_X) + \boldsymbol{R}_n\right)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X),\end{aligned} \tag{59}$$

where

$$\boldsymbol{R}_n = \begin{pmatrix} \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right)^{\mathsf{T}} \boldsymbol{R}_{n,1} \\ \dots \\ \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right)^{\mathsf{T}} \boldsymbol{R}_{n,p} \end{pmatrix}.$$

Rearranging (59), we get

$$\left(-\frac{1}{n}\boldsymbol{L}_n''(\boldsymbol{\theta}_X) - \frac{1}{n}\boldsymbol{R}_n\right)\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X) = \frac{1}{\sqrt{n}}\boldsymbol{U}_n(\boldsymbol{\theta}_X), \tag{60}$$

which is the multivariate analogue of (39). By part (ii) of this theorem, we know that the right hand side of the expression above converges in distribution to $\mathsf{N}_p\left(\mathbf{0}, \boldsymbol{J}(\boldsymbol{\theta}_X)\right)$. Just as for $p = 1$, it is also easy to see that on the left hand side of (60),

$$-\frac{1}{n}\boldsymbol{L}_n''(\boldsymbol{\theta}_X) \xrightarrow[n\to\infty]{\mathsf{P}} \boldsymbol{J}(\boldsymbol{\theta}_X).$$

Finally, thanks to the consistency of $\widehat{\boldsymbol{\theta}}_n$ and our condition (III), we can in the same way as for $p = 1$ find that each element of the random matrix $\frac{1}{n}\boldsymbol{R}_n$ converges to zero in probability, that is

$$\left(-\frac{1}{n}\boldsymbol{L}_n''(\boldsymbol{\theta}_X) - \frac{1}{n}\boldsymbol{R}_n\right) \xrightarrow[n\to\infty]{\mathsf{P}} \boldsymbol{J}(\boldsymbol{\theta}_X).$$

Multiplying both sides of (60) by the matrix $\left(\boldsymbol{J}(\boldsymbol{\theta}_X)\right)^{-1}$ and applying the Cramér-Slutsky theorem, we get the final representation

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) = \frac{1}{\sqrt{n}}\left(\boldsymbol{J}(\boldsymbol{\theta}_X)\right)^{-1}\boldsymbol{U}_n(\boldsymbol{\theta}_X) + \widetilde{\boldsymbol{R}}_n, \tag{61}$$

where the remainder term $\widetilde{\boldsymbol{R}}_n$ is negligible, meaning that it satisfies

$$\left\|\widetilde{\boldsymbol{R}}_n\right\| \xrightarrow[n\to\infty]{\mathsf{P}} 0.$$

The final expression for the distributional asymptotics of $\widehat{\boldsymbol{\theta}}_n$ then follows by a combination of part (ii) and (61). □

The end of lecture 8 (9.4.2024)

3 Theory of statistical hypotheses testing

In the final part of the lecture, we are concerned with statistical hypotheses testing. Similarly as for the theory of point estimation, we intend to find a way to construct either optimal tests, or at least tests that are almost optimal in a certain sense. Before doing so, we recall some basic facts about testing hypotheses.

Suppose that a random vector[8] $\boldsymbol{X} = (X_1, \dots, X_n)^\mathsf{T} \in \mathbb{R}^n$ has a distribution that depends on the unknown parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. We know that the parameter space $\boldsymbol{\Theta}$ can be an arbitrary set; typically it is a subset of $\mathbb{R}^p$, but if the parameter $\boldsymbol{\theta}$ is, e.g., the whole density of $\boldsymbol{X}$, it can be also a space of functions.

Our intention is to infer about the true value of the parameter $\boldsymbol{\theta}$ from which $\boldsymbol{X}$ was sampled. We denote this true value by $\boldsymbol{\theta}_X \in \boldsymbol{\Theta}$. More specifically, for a given subset $\boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$, based on our dataset $\boldsymbol{X}$ we want to decide whether the true value of the parameter $\boldsymbol{\theta}_X$ belongs to $\boldsymbol{\Theta}_0$ or not. To do this, we distinguish between the *null hypothesis* $H_0 \colon \boldsymbol{\theta}_X \in \boldsymbol{\Theta}_0$ and the *alternative* $H_1 \colon \boldsymbol{\theta}_X \notin \boldsymbol{\Theta}_0$. We can also write $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta} \setminus \boldsymbol{\Theta}_0$, in which case we have $H_1 \colon \boldsymbol{\theta}_X \in \boldsymbol{\Theta}_1$. If $\boldsymbol{\Theta}_0$ or $\boldsymbol{\Theta}_1$ is a single point set, we say that the hypothesis or the alternative is *simple*, respectively. Otherwise, the hypothesis or alternative are called *composite*.

Formally speaking, a *statistical test* ϕ is a function that to the values $\boldsymbol{x}$ of $\boldsymbol{X}$ assigns a decision of either rejecting, or not rejecting H_0. The set $W \subset \mathbb{R}^n$ defined by those values of $\boldsymbol{X}$ that reject H_0 using a test ϕ is called the *critical region* of the test ϕ. It will be convenient to denote the decision of the test ϕ numerically. From now on, the test ϕ takes value 0 if it does not reject the null hypothesis, and 1 if it does reject H_0. In this notation, a test ϕ is a measurable function from $\mathbb{R}^n$ to $\{0, 1\}$.

Using any test ϕ we can arrive at the correct decision, or to make one of the two errors:

(i) the *error of the first kind* of rejecting H_0 when in fact H_0 is true; or

(ii) the *error of the second kind* of not rejecting H_0 when H_1 is true.

We search for tests that give small probabilities of both these errors. In non-trivial situations, for the sample size n fixed, both these errors are impossible to be controlled simultaneously.

[8] In this section we do not assume that the elements of $\boldsymbol{X}$ form a random sample. Therefore, there is no need to consider $\boldsymbol{X} \in \mathbb{R}^{dn}$ as in the previous sections, and we simplify our argumentation to the equivalent $\boldsymbol{X} \in \mathbb{R}^n$.

We therefore fix a small number $\alpha \in (0,1)$ called the *level of significance* of the test, and under the condition that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} \mathsf{P}_{\boldsymbol{\theta}}\left(\phi(\boldsymbol{X})=1\right) \leq \alpha \tag{62}$$

we want to minimize the probability of the error of the second kind

$$\mathsf{P}_{\boldsymbol{\theta}}\left(\phi(\boldsymbol{X})=0\right) \text{ for all } \boldsymbol{\theta}\in\boldsymbol{\Theta}_1. \tag{63}$$

The supremum on the left hand side of (62) is called the *size* of the test ϕ. It determines the largest probability of the error of the first kind that we are willing to accept. The task of minimizing the function of $\boldsymbol{\theta}\in\boldsymbol{\Theta}_1$ in (63) is equivalent with the problem of maximizing the probability of rejecting H_0 for all $\boldsymbol{\theta}\in\boldsymbol{\Theta}_1$

$$\mathsf{P}_{\boldsymbol{\theta}}\left(\phi(\boldsymbol{X})=1\right), \tag{64}$$

which is called the *power* of the test ϕ against the alternative $\boldsymbol{\theta}\in\boldsymbol{\Theta}_1$. Considered as a function of $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, the probability of rejection of H_0 is called the *power function* of the test ϕ. Overall, we search for tests whose power function does not exceed α for $\boldsymbol{\theta}\in\boldsymbol{\Theta}_0$, and whose power function is as large as possible for $\boldsymbol{\theta}\in\boldsymbol{\Theta}_1$.

When dealing with discrete distributions, it is often impossible to find tests of size exactly α. For that reason, and for the reason of simplification of the subsequent mathematical arguments, we therefore expand the previous setting by considering also *randomized tests*. We saw that a usual (non-randomized) test is a measurable function $\phi\colon \mathbb{R}^n \to \{0,1\}$, with ϕ taking value 1 if and only if H_0 is rejected. For randomized tests we allow, instead of a hard decision of not rejecting ($\phi=0$) or rejecting ($\phi=1$) H_0, for ϕ to take any value in the interval $[0,1]$. The value $\phi(\boldsymbol{x})\in[0,1]$ can then be interpreted as the probability of rejecting H_0. If $\phi(\boldsymbol{x})\in(0,1)$ for $\boldsymbol{x}$ the observed value of $\boldsymbol{X}$, we interpret the outcome of the test as a Bernoulli random variable with the probability of rejecting H_0 being $\phi(\boldsymbol{x})$, and the probability $1-\phi(\boldsymbol{x})$ of not rejecting H_0. Thus, formally, a *randomized test* is any measurable function $\phi\colon \mathbb{R}^n \to [0,1]$. The function ϕ is also called the *critical function* of the test. Setting ϕ as the indicator of the critical region of a (non-randomized) test, we return to the usual setting.

Randomized tests are quite convenient to work with. Using randomization it is always possible to find a test of the exact size $\alpha\in(0,1)$. One such test is the trivial test $\phi(\boldsymbol{x})=\alpha$ for all $\boldsymbol{x}\in\mathbb{R}^n$. In practice, it is of course not acceptable that a result of a test is random. When possible, randomization is therefore avoided, and non-randomized tests are used. If, for discrete distributions, an exact test of size α is not possible to be constructed without resorting to randomization, the level of significance α is usually slightly changed so that a non-randomized test with that level of significance exists.

For randomized tests, it is simple to extend the notion of power function of ϕ from (64) by considering

$$\mathsf{E}_{\boldsymbol{\theta}}\,\phi(\boldsymbol{X}) = \int_{\mathbb{R}^n} \phi(\boldsymbol{x})\,\mathrm{d}\,\mathsf{P}_{\boldsymbol{\theta}}(\boldsymbol{x}).$$

In analogy to the standard setting, we search for a test ϕ that maximizes the power

$$\mathsf{E}_{\boldsymbol{\theta}}\,\phi(\boldsymbol{X}) \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}_1,$$

while keeping its size bounded by the given level $\alpha \in (0,1)$

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} \mathsf{E}_{\boldsymbol{\theta}}\,\phi(\boldsymbol{X}) \le \alpha. \tag{65}$$

The problem of finding tests that maximize power for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$ while satisfying (65) is frequently not well posed, and there does not exist a single test ϕ of size α that maximizes power uniformly in $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$. If, however, such a test exists, we call it the *uniformly most powerful* test of size α in our testing problem. In the first part of this section we are interested in constructing uniformly most powerful tests when possible.

3.1 Simple hypothesis and alternative: Neyman-Pearson theorem

An important situation when a uniformly most powerful test can always be found is that of a simple hypothesis and a simple alternative. Then, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1\}$, and $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta}_0\}$, $\boldsymbol{\Theta}_1 = \{\boldsymbol{\theta}_1\}$. For stating the fundamental theorem of Neyman and Pearson that gives the form of the uniformly most powerful test, a simple lemma will be useful.

Lemma 8. *Let $f\colon S \to (0,\infty)$, and let μ be a σ-finite measure. Then $\int_S f(x)\,\mathrm{d}\,\mu(x) = 0$ implies $\mu(S) = 0$.*

Proof. Let $S_n = \{x \in S\colon f(x) \ge 1/n\}$. Then $S = \bigcup_{n\in\mathbb{N}} S_n$, and therefore $\mu(S) \le \sum_{n=1}^{\infty} \mu(S_n)$. Also,

$$\mu(S_n) = \int_{\{x\in S\colon f(x)\ge 1/n\}} 1\,\mathrm{d}\,\mu(x) \le n\int_{S_n} f(x)\,\mathrm{d}\,\mu(x) \le n\int_S f(x)\,\mathrm{d}\,\mu(x) = 0,$$

for each $n \in \mathbb{N}$, which gives $\mu(S) = 0$ as we needed. □

Theorem 26 (Neyman-Pearson)**.** *Let P_0 and P_1 be probability distributions that correspond to the values of the parameter $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$, respectively. Suppose that P_0 and P_1 have densities p_0 and p_1 respectively w.r.t. a σ-finite measure μ on $\mathbb{R}^n$.*

(i) For testing

$$H_0\colon \boldsymbol{\theta}_X = \boldsymbol{\theta}_0 \quad \textit{against} \quad H_1\colon \boldsymbol{\theta}_X = \boldsymbol{\theta}_1 \tag{66}$$

at level $\alpha \in (0,1)$ *there exists a test* ϕ *and a constant* $k \geq 0$ *such that*

$$\mathsf{E}_{\boldsymbol{\theta}_0}\, \phi(\boldsymbol{X}) = \alpha, \tag{NP$_1$}$$

and

$$\phi(\boldsymbol{x}) = \begin{cases} 1 & \text{when } p_1(\boldsymbol{x}) > k\, p_0(\boldsymbol{x}), \\ 0 & \text{when } p_1(\boldsymbol{x}) < k\, p_0(\boldsymbol{x}). \end{cases} \tag{NP$_2$}$$

The test given by (NP$_1$) *and* (NP$_2$) *is almost surely unique.*

(ii) If a test satisfies (NP$_1$) *and* (NP$_2$) *for some* $k \geq 0$*, then it is the most powerful for testing* (66) *at level* α*.*

(iii) If ϕ *is the most powerful test for testing* (66) *at level* α*, then for some* $k \geq 0$ *it satisfies* (NP$_2$) μ*-almost everywhere. If there does not exist a test of size smaller than* α *with power* 1*, it also satisfies* (NP$_1$)*.*

Proof. In the proof we write $f(x-)$ for the limit of the function $f\colon \mathbb{R} \to \mathbb{R}$ at $x \in \mathbb{R}$ from the left.

Part (i). Let $\alpha(c) = P_0\left(p_1(\boldsymbol{X}) > c\, p_0(\boldsymbol{X})\right)$ for $c \geq 0$. Since this probability is computed w.r.t. the measure P_0, the probability in $\alpha(c)$ needs to be considered only in the set where p_0 is positive, meaning that $P_0\left(p_0(\boldsymbol{X}) = 0\right) = 0$. Thus, $\alpha(c)$ is the probability that the almost surely non-negative random variable $p_1(\boldsymbol{X})/p_0(\boldsymbol{X})$ exceeds $c \geq 0$. The expression

$$1 - \alpha(c) = P_0\left(\frac{p_1(\boldsymbol{X})}{p_0(\boldsymbol{X})} \leq c\right)$$

is therefore the cumulative distribution function of the random variable $p_1(\boldsymbol{X})/p_0(\boldsymbol{X})$, and $\alpha(c)$ must therefore be non-increasing, continuous from the right, satisfy

$$\alpha(c-) - \alpha(c) = P_0\left(\frac{p_1(\boldsymbol{X})}{p_0(\boldsymbol{X})} = c\right),$$

$\alpha(0-) = 1$ and $\lim_{c\to\infty} \alpha(c) = 0$. For any $\alpha \in (0,1)$ let $c_0 \geq 0$ be any number such that

$$\alpha(c_0) \leq \alpha \leq \alpha(c_0-). \tag{67}$$

Consider the test defined by

$$\phi(\boldsymbol{x}) = \begin{cases} 1 & \text{when } p_1(\boldsymbol{x}) > c_0\, p_0(\boldsymbol{x}), \\ \frac{\alpha - \alpha(c_0)}{\alpha(c_0-) - \alpha(c_0)} & \text{when } p_1(\boldsymbol{x}) = c_0\, p_0(\boldsymbol{x}), \\ 0 & \text{when } p_1(\boldsymbol{x}) < c_0\, p_0(\boldsymbol{x}). \end{cases}$$

The middle expression remains undefined if $\alpha(c_0-) = \alpha(c_0)$. But, the test $\phi(\boldsymbol{X})$ is well defined both P_0-almost everywhere, and P_1-almost everywhere. To see that, observe that if $\alpha(c_0-) = \alpha(c_0)$ then

$$P_0\left(p_1(\boldsymbol{X}) = c_0\, p_0(\boldsymbol{X})\right) = \alpha(c_0-) - \alpha(c_0) = 0, \tag{68}$$

and the test is defined P_0-almost everywhere. For measure P_1 write

$$\begin{aligned} P_1\left(p_1(\boldsymbol{X}) = c_0\, p_0(\boldsymbol{X})\right) = P_1\left(p_1(\boldsymbol{X}) = c_0\, p_0(\boldsymbol{X}), p_0(\boldsymbol{X}) > 0\right) \\ + P_1\left(p_1(\boldsymbol{X}) = c_0\, p_0(\boldsymbol{X}), p_0(\boldsymbol{X}) = 0\right). \end{aligned} \tag{69}$$

For the second summand on the right hand side of (69) we have

$$P_1\left(p_1(\boldsymbol{X}) = c_0\, p_0(\boldsymbol{X}), p_0(\boldsymbol{X}) = 0\right) \le P_1\left(p_1(\boldsymbol{X}) = 0\right) = \int_{\{\boldsymbol{y}\colon\, p_1(\boldsymbol{y})=0\}} p_1(\boldsymbol{x})\,\mathrm{d}\,\mu(\boldsymbol{x}) = 0.$$

For the first summand in (69), we denote

$$S = \left\{\boldsymbol{x}\colon p_1(\boldsymbol{x}) = c_0\, p_0(\boldsymbol{x}), p_0(\boldsymbol{x}) > 0\right\}.$$

By (68) we have

$$\int_S p_0(\boldsymbol{x})\,\mathrm{d}\,\mu(\boldsymbol{x}) = P_0\left(\boldsymbol{X} \in S\right) \le P_0\left(p_1(\boldsymbol{X}) = c_0\, p_0(\boldsymbol{X})\right) = 0,$$

and setting $f = p_0$ in Lemma 8 we get $\mu(S) = 0$. Because P_1 is absolutely continuous w.r.t. μ, it therefore must be also that $P_1(S) = 0$. Putting the two bounds together, we get that the probability in (69) equals zero, and we have indeed proved that the test ϕ is well defined also P_1-almost everywhere.

For the size of the test ϕ we have

$$\mathsf{E}_{\boldsymbol{\theta}_0}\,\phi(\boldsymbol{X}) = P_0\left(\frac{p_1(\boldsymbol{X})}{p_0(\boldsymbol{X})} > c_0\right) + \frac{\alpha - \alpha(c_0)}{\alpha(c_0-) - \alpha(c_0)} P_0\left(\frac{p_1(\boldsymbol{X})}{p_0(\boldsymbol{X})} = c_0\right) = \alpha(c_0) + (\alpha - \alpha(c_0)) = \alpha.$$

We can therefore choose k to be c_0, and we obtain the test from part (i) of the theorem.

It remains to show that the test given by ϕ is P_0-almost surely unique, and P_1-almost surely unique. The only situation where there could be more values of c_0 that satisfy (67) is when there is an interval of values $(c', c'') \subset [0, \infty)$ such that $\alpha(c) = \alpha$ for all $c \in (c', c'')$. In that case, denote

$$C = \left\{\boldsymbol{x}\colon p_0(\boldsymbol{x}) > 0 \text{ and } c' < \frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} < c''\right\}.$$

Then we have $P_0(C) = \alpha(c') - \alpha(c''-) = \alpha - \alpha = 0$, and by Lemma 8 we have that also $\mu(C) = 0$. Because we assumed that P_1 is absolutely continuous w.r.t. μ, it must be that also $P_1(C) = 0$. Thus, the sets corresponding to two different values of c differ only in a set

which has probability 0 under both P_0 and P_1, and for all practical purposes, the test ϕ is almost surely unique.

Part (ii). Let ϕ be a test that satisfies both (NP_1) and (NP_2), and let ϕ^* be any other test with $\mathsf{E}_{\boldsymbol{\theta}_0}\, \phi^*(\boldsymbol{X}) \leq \alpha$. Denote

$$\begin{aligned} S^+ &= \{\boldsymbol{x} \colon \phi(\boldsymbol{x}) > \phi^*(\boldsymbol{x})\}\,, \\ S^- &= \{\boldsymbol{x} \colon \phi(\boldsymbol{x}) < \phi^*(\boldsymbol{x})\}\,. \end{aligned} \tag{70}$$

If $\boldsymbol{x} \in S^+$, then $\phi(\boldsymbol{x}) > 0$ and thus $p_1(\boldsymbol{x}) \geq k\, p_0(\boldsymbol{x})$. In the same way if $\boldsymbol{x} \in S^-$, we have $\phi(\boldsymbol{x}) < 1$ and $p_1(\boldsymbol{x}) \leq k\, p_0(\boldsymbol{x})$. We can therefore write

$$\begin{aligned} &\int_{\mathbb{R}^n} (\phi(\boldsymbol{x}) - \phi^*(\boldsymbol{x}))\, (p_1(\boldsymbol{x}) - k\, p_0(\boldsymbol{x}))\, \mathrm{d}\, \mu(\boldsymbol{x}) \\ &\qquad = \int_{S^+ \cup S^-} (\phi(\boldsymbol{x}) - \phi^*(\boldsymbol{x}))\, (p_1(\boldsymbol{x}) - k\, p_0(\boldsymbol{x}))\, \mathrm{d}\, \mu(\boldsymbol{x}) \geq 0. \end{aligned}$$

Rewriting this expression, we get that the difference in power of these tests is

$$\begin{aligned} \mathsf{E}_{\boldsymbol{\theta}_1}\, \phi(\boldsymbol{X}) - \mathsf{E}_{\boldsymbol{\theta}_1}\, \phi^*(\boldsymbol{X}) &= \int_{\mathbb{R}^n} (\phi(\boldsymbol{x}) - \phi^*(\boldsymbol{x}))\, p_1(\boldsymbol{x})\, \mathrm{d}\, \mu(\boldsymbol{x}) \\ &\geq k \int_{\mathbb{R}^n} (\phi(\boldsymbol{x}) - \phi^*(\boldsymbol{x}))\, p_0(\boldsymbol{x})\, \mathrm{d}\, \mu(\boldsymbol{x}) \\ &= k\, (\mathsf{E}_{\boldsymbol{\theta}_0}\, \phi(\boldsymbol{X}) - \mathsf{E}_{\boldsymbol{\theta}_0}\, \phi^*(\boldsymbol{X})) \geq k\, (\alpha - \alpha) = 0, \end{aligned}$$

as we wanted to prove. We get that the test ϕ^* cannot have a greater power than ϕ.

Part (iii). Let ϕ^* be the most powerful test at level α for testing (66), and let ϕ satisfy both (NP_1) and (NP_2). First, we want to show that ϕ^* must satisfy (NP_2) μ-almost everywhere. Take the sets S^+ and S^- defined in (70), and let $S = (S^+ \cup S^-) \cap \{\boldsymbol{x} \colon p_1(\boldsymbol{x}) \neq k\, p_0(\boldsymbol{x})\}$. It is enough to prove that $\mu(S) = 0$. For a contradiction, suppose that $\mu(S) > 0$. Analysing by cases as in part (ii) of this proof, we get that

- for $\boldsymbol{x} \in S^+$, necessarily $\phi(\boldsymbol{x}) > 0$ and thus $p_1(\boldsymbol{x}) \geq k\, p_0(\boldsymbol{x})$, and
- for $\boldsymbol{x} \in S_-$, necessarily $\phi(\boldsymbol{x}) < 1$, and thus $p_1(\boldsymbol{x}) \leq k\, p_0(\boldsymbol{x})$.

Because for $\boldsymbol{x} \in S$ we also assume that $p_1(\boldsymbol{x}) \neq k\, p_0(\boldsymbol{x})$, we get that the function $(\phi - \phi^*)\, (p_1 - k\, p_0)$ is strictly positive on S. From Lemma 8 it therefore follows that

$$\begin{aligned} &\int_{S^+ \cup S^-} (\phi(\boldsymbol{x}) - \phi^*(\boldsymbol{x}))\, (p_1(\boldsymbol{x}) - k\, p_0(\boldsymbol{x}))\, \mathrm{d}\, \mu(\boldsymbol{x}) \\ &\qquad = \int_S (\phi(\boldsymbol{x}) - \phi^*(\boldsymbol{x}))\, (p_1(\boldsymbol{x}) - k\, p_0(\boldsymbol{x}))\, \mathrm{d}\, \mu(\boldsymbol{x}) > 0, \end{aligned}$$

meaning that ϕ is more powerful against P_1 than ϕ^*. That can be asserted in the same way as in part (ii) of this proof. This is a contradiction with ϕ^* being the most powerful test, and thus $\mu(S) = 0$ as we wanted to prove.

Finally, if ϕ^* does not satisfy (NP_1), i.e. if $\mathsf{E}_{\boldsymbol{\theta}_0}\,\phi^*(\boldsymbol{X}) < \alpha$ and if the power of ϕ^* is less than 1, then it would be possible to expand the critical region of ϕ^* (more precisely, construct a test ϕ^{**} such that $\phi^{**}(\boldsymbol{x}) \geq \phi^*(\boldsymbol{x})$). That would increase both the size and the power of the test ϕ^*, until either the size would be exactly α, or the power would be exactly 1. Thus for the uniformly most powerful test ϕ^* it must be either $\mathsf{E}_{\boldsymbol{\theta}_0}\,\phi^*(\boldsymbol{X}) = \alpha$ or $\mathsf{E}_{\boldsymbol{\theta}_1}\,\phi^*(\boldsymbol{X}) = 1$ as we wanted to show. □

The assumption of both P_0 and P_1 having a density w.r.t. μ is not restrictive. One can, for example, always take for μ the sum of the measures P_0 and P_1, w.r.t. which both P_0 and P_1 are absolutely continuous.

Theorem 27. *The power β of the most powerful test at level $\alpha \in (0,1)$ for testing (66) satisfies $\beta \geq \alpha$, with equality only if $P_0 = P_1$.*

Proof. Comparing the uniformly most powerful test ϕ from (26) with the trivial test $\phi^*(\boldsymbol{x}) = \alpha$ for all $\boldsymbol{x} \in \mathbb{R}^n$ we have

$$\beta = \mathsf{E}_{\boldsymbol{\theta}_1}\,\phi(\boldsymbol{X}) \geq \mathsf{E}_{\boldsymbol{\theta}_1}\,\phi^*(\boldsymbol{X}) = \alpha.$$

If $\alpha = \beta$, also the test ϕ^* is the most powerful, and so by part (iii) of Theorem 26 it must satisfy (NP_2) μ-almost everywhere. But, since $\alpha \in (0,1)$, in (NP_2) we necessarily get that $\{\boldsymbol{x} \in \mathbb{R}^n \colon p_1(\boldsymbol{x}) \neq k\, p_0(\boldsymbol{x})\}$ must be of null μ-measure, and hence $p_1(\boldsymbol{x}) = k\, p_0(\boldsymbol{x})$ for μ-almost all $\boldsymbol{x} \in \mathbb{R}^n$. Because both p_0 and p_1 integrate to one, it must be that $k = 1$, and the measures P_0 and P_1 are the same. □

Theorem 26 of Neyman and Pearson gives the form of the (uniformly) most powerful test of any simple hypothesis against a simple alternative, and Theorem 27 asserts that the power of this test is always larger than its size.

Example 3.1. Let $\boldsymbol{X} = (X_1, \ldots, X_n)^{\mathsf{T}}$ with $X_1, \ldots, X_n$ a random sample from an exponential distribution with density $f(x;\lambda) = \lambda \exp(-\lambda\, x)\, \mathbb{I}(x > 0)$, where $\lambda \in (0,\infty)$ is unknown. We want to test

$$H_0 \colon \lambda_X = \lambda_0 \text{ against } H_1 \colon \lambda_X = \lambda_1,$$

where $0 < \lambda_0 < \lambda_1$ are given. We use Theorem 26 to find the most powerful test. The test statistic takes the form

$$\frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} = \frac{\prod_{i=1}^n f(x_i;\lambda_1)}{\prod_{i=1}^n f(x_i;\lambda_0)} = \frac{\lambda_1^n \exp(-\lambda_1 \sum_{i=1}^n x_i)}{\lambda_0^n \exp(-\lambda_0 \sum_{i=1}^n x_i)} = \left(\frac{\lambda_1}{\lambda_0}\right)^n \exp\left((\lambda_0 - \lambda_1)\sum_{i=1}^n x_i\right).$$

According to part (iii) of Theorem 26, the uniformly most powerful test rejects H_0 if $p_1(\boldsymbol{x})/p_0(\boldsymbol{x})$ is too large. Since in our case $\lambda_0 - \lambda_1 < 0$, that is equivalent with rejection of H_0 if $\sum_{i=1}^n x_i$

is too small. To determine the threshold value k in the test, we must therefore find the exact distribution of the test statistic $\sum_{i=1}^n X_i$ under H_0. Under H_0, each X_i is distributed exponentially with parameter λ_0, which is the same as the gamma distribution $\Gamma(1, 1/\lambda_0)$ (in the shape-scale parametrization). A sum of independent gamma distributions is gamma distributed, and we have $\sum_{i=1}^n X_i \sim \Gamma(n, 1/\lambda_0)$. To obtain the desired size of the test $\alpha \in (0,1)$, we must therefore have

$$\phi(\boldsymbol{x}) = \begin{cases} 1 \text{ if } \sum_{i=1}^n x_i < q_\alpha, \\ 0 \text{ if } \sum_{i=1}^n x_i \geq q_\alpha, \end{cases} \tag{71}$$

where $q_\alpha > 0$ is the α-quantile of the distribution $\Gamma(n, 1/\lambda_0)$. Since the distribution of the test statistic $\sum_{i=1}^n X_i$ is continuous, we do not have to worry about the situation $\sum_{i=1}^n x_i = q_\alpha$, which occurs with null probability under both H_0 and H_1. Therefore, we do not have to randomize our test. △

Note that in Example 3.1 we had the luck to know the exact distribution of the test statistic $\sum_{i=1}^n X_i$ under H_0. This is often not the case. Nevertheless, even if we did not know the exact distribution of $\sum_{i=1}^n X_i$ under H_0, we could still use the central limit theorem to find at least the asymptotic distribution of the equivalent test statistic $(1/n)\sum_{i=1}^n X_i$ under H_0, and use that distribution to approximate the quantile q_α. Yet another alternative is to approximate the quantile q_α by simulating independently random variables $\sum_{i=1}^n X_i$ under H_0; think about how this could be done.

3.2 Simple hypothesis and composite alternative

The theorem of Neyman and Pearson is a quite valuable result, but the situation with a two-points parameter space $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1\}$ is very specific. We are therefore in the sequel concerned with tests for one-dimensional parameters $\theta \in \Theta \subseteq \mathbb{R}$ of two more common types: for $\theta_0 \in \Theta$ given we consider either one-sided tests of the type

$$H_0 \colon \theta_X \leq \theta_0 \text{ against } H_1 \colon \theta_X > \theta_0, \tag{72}$$

or two-sided tests that take the form

$$H_0 \colon \theta_X = \theta_0 \text{ against } H_1 \colon \theta_X \neq \theta_0. \tag{73}$$

The one-sided test in (72) is, of course, equivalent to the test of the other form

$$H_0 \colon \theta_X \geq \theta_0 \text{ against } H_1 \colon \theta_X < \theta_0,$$

because of the obvious symmetry of the problem — one just needs to reparametrize $\theta \mapsto -\theta$. For simplicity, we therefore consider only the tests of the form (72). Without additional

assumptions, it turns out that both testing problems can also be looked upon as tests with simple hypotheses. In (72), we can restrict our parameter space $\boldsymbol{\Theta}$ only to the interval $[\theta_0, \infty)$, and search for a test of

$$H_0 \colon \theta_X = \theta_0 \text{ against } H_1 \colon \theta_X > \theta_0.$$

This reduction is not completely equivalent with (72) as in the computation of the size of the test (62) we must control $\sup_{\theta \leq \theta_0} \mathsf{E}_\theta \, \phi(\boldsymbol{X}) \leq \alpha$ in (72), but only $\mathsf{E}_{\theta_0} \, \phi(\boldsymbol{X}) \leq \alpha$ in the restricted formulation. But, we will see that for the tests that we are going to construct, both formulations are equivalent.

For general distributions, the uniformly most powerful test does not exist for neither the one-sided (72) nor the two-sided (73) alternative. An exception are families with monotone likelihood ratio.

Definition 14. A family of densities $\{p_\theta \colon \theta \in \Theta \subseteq \mathbb{R}\}$ w.r.t. a σ-finite measure μ on $\mathbb{R}^n$ is said to have a *monotone likelihood ratio* if there exists a measurable function $T \colon \mathbb{R}^n \to \mathbb{R}$ such that for any $\theta < \theta'$ in Θ the densities p_θ and $p_{\theta'}$ correspond to different distributions, and the ratio $p_{\theta'}(\boldsymbol{x})/p_\theta(\boldsymbol{x})$ is a non-decreasing function of $T(\boldsymbol{x})$.

For families of distributions with monotone likelihood ratio, Theorem 26 of Neyman and Pearson is simple to be extended to the situation with one-sided tests.

Theorem 28. *Let $\theta \in \Theta \subseteq \mathbb{R}$ and let the random vector $\boldsymbol{X}$ have probability density $p_\theta(\boldsymbol{x})$ with monotone likelihood ratio in $T(\boldsymbol{x})$.*

(i) For the one-sided test (72) with $\theta_0 \in \Theta$ given there exists a uniformly most powerful test, which is given by

$$\phi(\boldsymbol{x}) = \begin{cases} 1 & \text{when } T(\boldsymbol{x}) > C, \\ \gamma & \text{when } T(\boldsymbol{x}) = C, \\ 0 & \text{when } T(\boldsymbol{x}) < C, \end{cases} \tag{74}$$

where C and γ are determined by

$$\mathsf{E}_{\theta_0} \phi(\boldsymbol{X}) = \alpha. \tag{75}$$

(ii) The power function

$$\beta(\theta) = \mathsf{E}_\theta \, \phi(\boldsymbol{X})$$

of this test is strictly increasing for all points θ for which $0 < \beta(\theta) < 1$.

(iii) For all $\theta' \in \boldsymbol{\Theta}$, the test determined by (74) and (75) is uniformly most powerful for testing

$$H_0 \colon \theta_X \leq \theta' \text{ against } H_1 \colon \theta_X > \theta'$$

at level $\alpha' = \beta(\theta')$.

(iv) For any $\theta < \theta_0$ *this test minimizes* $\beta(\theta)$*, that is the probability of the error of the fist kind, among all tests satisfying* (75).

Proof. Parts (i) and (ii). Take first the test of a simple hypothesis $H_0 \colon \theta_X = \theta_0$ against the simple alternative $H_1 \colon \theta_X = \theta_1$ for $\theta_1 > \theta_0$ fixed. By Theorem 26 we know that the best possible critical region for this test is the set of those $\boldsymbol{x} \in \mathbb{R}^n$ for which the ratio $r(\boldsymbol{x}) = p_{\theta_1}(\boldsymbol{x})/p_{\theta_0}(\boldsymbol{x}) = g\left(T(\boldsymbol{x})\right)$ is large enough. Here $g \colon \mathbb{R} \to \mathbb{R}$ is a non-decreasing function from the definition of the monotone likelihood ratio. If $\boldsymbol{x}' \in \mathbb{R}^d$ is such that $T(\boldsymbol{x}') > T(\boldsymbol{x})$, then $r(\boldsymbol{x}') \geq r(\boldsymbol{x})$. Therefore, if $\boldsymbol{x}$ lies in the critical region of this test, then also $\boldsymbol{x}'$ does. Thus, the test which rejects H_0 for large values of $T(\boldsymbol{x})$ is the most powerful test. In the same way as in the proof of part (i) of Theorem 26, it is seen that there exist constants C and $\gamma \in (0, 1)$ such that (74) and (75) hold. Now, by part (ii) of Theorem 26 we also know that this test is the most powerful for testing $H_0 \colon \theta_X = \theta'$ against $H_1 \colon \theta_X = \theta''$ at level $\alpha' = \beta(\theta')$ for any $\theta' < \theta''$. Part (ii) of the present theorem now follows from Theorem 27. Because we found that $\beta(\theta)$ is non-decreasing, we have that

$$\beta(\theta) = \mathsf{E}_\theta \, \phi(\boldsymbol{X}) \leq \alpha \text{ for } \theta \leq \theta_0. \tag{76}$$

The family of tests that satisfy condition (76) is a subset of the family of tests with size $\mathsf{E}_{\theta_0} \, \phi(\boldsymbol{X}) \leq \alpha$. Since by the Neyman's and Pearson's Theorem 26 this test maximizes $\beta(\theta_1)$ among all tests within this larger class of tests, it must maximize $\beta(\theta_1)$ also among all tests that satisfy (76). And, because the test does not depend on the alternative $\theta_1 > \theta_0$, it must be the uniformly most powerful test for testing (72).

Part (iii) follows completely analogously. For part (iv) it is enough apply Theorem 26 again to the test $H_0 \colon \theta_X = \theta_0$ against $H_1 \colon \theta = \theta_1$, and to realise that a test that minimizes the power $\beta(\theta_1)$ is the test from Theorem 26 with all inequalities reversed. □

An important class of distributions that satisfies the conditions of Theorem 28 is the *one-parameter exponential family* of densities, defined in the next result.

Theorem 29. *Let* $\theta \in \Theta \subseteq \mathbb{R}$*, and let* $\boldsymbol{X}$ *have a density of the form*

$$p_\theta(\boldsymbol{x}) = C(\theta) \exp\left(Q(\theta) \, T(\boldsymbol{x})\right) h(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \mathbb{R}^n, \tag{77}$$

w.r.t. a σ*-finite measure* μ *on* $\mathbb{R}^n$*, where* $Q \colon \mathbb{R} \to \mathbb{R}$ *is strictly monotone. Then there exists a uniformly most powerful test* ϕ *for testing* (72) *for* $\theta_0 \in \Theta$ *given. If* Q *is increasing,*

$$\phi(\boldsymbol{x}) = \begin{cases} 1 & \text{when } T(\boldsymbol{x}) > C, \\ \gamma & \text{when } T(\boldsymbol{x}) = C, \\ 0 & \text{when } T(\boldsymbol{x}) < C, \end{cases} \tag{78}$$

where C and γ are determined by $\mathsf{E}_{\theta_0}\,\phi(\boldsymbol{X}) = \alpha$. If Q is decreasing, the inequalities in (78) are reversed.

Proof. The system of densities satisfies for $\theta < \theta'$

$$\frac{p_{\theta'}(\boldsymbol{x})}{p_{\theta}(\boldsymbol{x})} = \frac{C(\theta')\exp\left(Q(\theta')\,T(\boldsymbol{x})\right)}{C(\theta)\exp\left(Q(\theta)\,T(\boldsymbol{x})\right)} = \frac{C(\theta')}{C(\theta)}\exp\left(T(\boldsymbol{x})\left(Q(\theta') - Q(\theta)\right)\right).$$

If Q is increasing, $Q(\theta') - Q(\theta) > 0$ and the likelihood ratio is a non-decreasing function of $T(\boldsymbol{x})$. It remains to apply Theorem 28. □

The family of densities of exponential type turns out to be the only important class of distributions which verifies the conditions of Theorem 28. It can be shown that, under additional weak regularity conditions, only for densities of exponential type, uniformly most powerful one-sided tests exist for all values $n \in \mathbb{N}$.

Example 3.2. In the setup of testing with exponential distributions from Example 3.1, suppose that we now want to test

$$H_0\colon \lambda_X \leq \lambda_0 \text{ against } H_1\colon \lambda_X > \lambda_0,$$

where $0 < \lambda_0$ is given. Using Theorems 26 and 28 we find that the test from (71) is the uniformly most powerful test also in the present situation. This is, of course, possible because the test ϕ was in Example 3.1 constructed in a way not depending on the value of λ_1, as long as $\lambda_1 > \lambda_0$. △

For testing against a two-sided alternative (73) the situation is even more complicated. It can be shown that even for densities of exponential type, uniformly most powerful tests do not exist. To proceed, one needs to impose an additional restriction, and search for a *uniformly most powerful unbiased test*, meaning that in addition to the given size $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0}\mathsf{E}_{\boldsymbol{\theta}}\,\phi(\boldsymbol{X}) \leq \alpha$, we require also that under the alternative, the power of the test is at least α, meaning that $\mathsf{E}_{\boldsymbol{\theta}}\,\phi(\boldsymbol{X}) \geq \alpha$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$. Here, of course, $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ are the subsets partitioning the parameter space $\boldsymbol{\Theta}$ into the hypothesis and the alternative, respectively. For the special case of exponential families of distributions (77), there exist uniformly most powerful unbiased tests for the two-sided alternative (73). As one could expect, under reasonable assumptions, the tests take the form

$$\phi(\boldsymbol{x}) = \begin{cases} 1 & \text{when } T(\boldsymbol{x}) < K_1 \text{ or } T(\boldsymbol{x}) > K_2, \\ \gamma & \text{when } T(\boldsymbol{x}) = K_1 \text{ or } T(\boldsymbol{x}) = K_2, \\ 0 & \text{when } T(\boldsymbol{x}) \in (K_1, K_2). \end{cases}$$

Here, $K_1 < K_2$ and $\gamma \in (0,1)$ are constants that need to be determined. The proof of this result is not difficult, yet it is a rather technical extension of the theory we provided. For a comprehensive account of these tests, one can consult [1, Section 8] and [9].

The end of lecture 9 (16.4.2024)

3.3 Asymptotic tests based on the likelihood

We now turn to the practical problem of exploiting the advances we gathered on the maximum likelihood estimation, and use them to construct useful statistical tests. Similarly as in the theory of point estimation, the likelihood based tests will usually not have to be the best, or optimal in any specific sense. They are, however, extremely useful in practice as they are simple to design, and widely used in all kinds of scenarios. We will deal with two situations separately. First, we consider the multi-dimensional parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ and design tests for the hypotheses of the type

$$H_0 \colon \boldsymbol{\theta}_X = \boldsymbol{\theta}_0 \text{ against } H_1 \colon \boldsymbol{\theta}_X \neq \boldsymbol{\theta}_0, \tag{79}$$

for $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ given. In the second step, we will be concerned with tests only about a sub-vector of the whole vector $\boldsymbol{\theta}$, where the rest of the parameter $\boldsymbol{\theta}$ is not of interest, and remains unspecified.

3.3.1 Tests without nuisance parameters

In what follows we construct three families of asymptotic tests for the hypothesis (79) based on the log-likelihood function $L_n(\boldsymbol{\theta})$: (i) the *Rao score test* based on the score function $\boldsymbol{U}_n(\boldsymbol{\theta}) = \frac{\partial L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$; (ii) the *Wald test* defined using the asymptotic distribution of $\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)$ under H_0 for $\widehat{\boldsymbol{\theta}}_n$ the maximum likelihood estimator of $\boldsymbol{\theta}$; and (iii) the *likelihood ratio test* which can be seen as an extension of the idea of the Neyman-Pearson testing procedure from Theorems 26 and 28.

For simplicity, in our first theorem we begin with the case of a one-dimensional parameter $\theta \in \Theta \subseteq \mathbb{R}$.

Theorem 30. *Suppose that all conditions of Theorem 23 are satisfied, and let $\theta_0 \in \Theta$ be given. Then, if the true value of the parameter θ_X is θ_0, we have the following.*

1. ***The Rao score test.***
$$R_n = \frac{L_n'(\theta_0)}{\sqrt{n\, J(\theta_0)}} \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}(0,1),$$
$$R_n^2 = \frac{\left(L_n'(\theta_0)\right)^2}{n\, J(\theta_0)} \xrightarrow[n\to\infty]{\mathrm{d}} \chi_1^2.$$

2. ***Wald test.*** *If the Fisher information $J(\theta)$ is continuous in θ_0, then*
$$W_n = \sqrt{n\, J(\widehat{\theta}_n)}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}(0,1),$$
$$W_n^2 = n\, J(\widehat{\theta}_n)\left(\widehat{\theta}_n - \theta_0\right)^2 \xrightarrow[n\to\infty]{\mathrm{d}} \chi_1^2.$$

3. ***Likelihood ratio test.*** *If the Fisher information $J(\theta)$ is continuous in θ_0, then*

$$LR_n = 2\left(L_n(\widehat{\theta}_n) - L_n(\theta_0)\right) \xrightarrow[n\to\infty]{\mathrm{d}} \chi^2_1.$$

Proof. The asymptotic distribution of R_n and R_n^2 follows directly from the asymptotic distribution of the score statistic $L'_n(\theta_0) = L'_n(\theta_X)$ established in (37) in Theorem 23.

For the Wald statistics W_n and W_n^2, we know from (38) that

$$\sqrt{n\,J(\theta_0)}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}(0,1).$$

Because we assume that the Fisher information is continuous in θ_0, and by Theorem 23 we also know that $\widehat{\theta}_n \xrightarrow[n\to\infty]{\mathsf{P}} \theta_0$, the continuous mapping theorem gives $J(\widehat{\theta}_n) \xrightarrow[n\to\infty]{\mathsf{P}} J(\theta_0)$. An application of the Cramér-Slutsky theorem gives the asymptotic result for W_n, and the continuous mapping theorem concludes also for W_n^2.

We need to prove only the statement about the likelihood ratio test. We apply the Taylor expansion to the log-likelihood function

$$L_n(\theta_0) = L_n(\widehat{\theta}_n) + (\theta_0 - \widehat{\theta}_n)L'_n(\widehat{\theta}_n) + \frac{1}{2}(\theta_0 - \widehat{\theta}_n)^2 L''_n(\widehat{\theta}_n) + \frac{1}{6}(\theta_0 - \widehat{\theta}_n)^3 L'''_n(\theta_n^*),$$

where θ_n^* lies in the interval between θ_0 and $\widehat{\theta}_n$. Since $\widehat{\theta}_n$ solves $L'_n(\widehat{\theta}_n) = 0$, we can reorganize the expression above and write

$$2\left(L_n(\widehat{\theta}_n) - L_n(\theta_0)\right) = -\left(\widehat{\theta}_n - \theta_0\right)^2 L''_n(\widehat{\theta}_n) + \frac{1}{3}(\widehat{\theta}_n - \theta_0)^3 L'''_n(\theta_n^*). \tag{80}$$

By (40) from the proof of Theorem 23 we know that

$$-\frac{L''_n(\theta_0)}{n} \xrightarrow[n\to\infty]{\mathsf{P}} J(\theta_0). \tag{81}$$

Further, we can use Taylor's expansion again and get

$$\frac{1}{n}L''_n(\theta_0) - \frac{1}{n}L''_n(\widehat{\theta}_n) = \left(\theta_0 - \widehat{\theta}_n\right)\frac{1}{n}L'''_n(\theta_n^{**}),$$

for some θ_n^{**} between θ_0 and $\widehat{\theta}_n$. In (41) in the proof of Theorem 23 we established that $L'''_n(\theta_n^{**})/n$ is bounded in probability. We also know by Theorem 23 that $\widehat{\theta}_n \xrightarrow[n\to\infty]{\mathsf{P}} \theta_0$. Thus, the Cramér-Slutsky theorem gives

$$\frac{1}{n}L''_n(\theta_0) - \frac{1}{n}L''_n(\widehat{\theta}_n) \xrightarrow[n\to\infty]{\mathsf{P}} 0. \tag{82}$$

From (81) and (82) we obtain

$$-\frac{1}{n}L''_n(\widehat{\theta}_n) \xrightarrow[n\to\infty]{\mathsf{P}} J(\theta_0).$$

We therefore see that for the first summand in (80) we have

$$\left(\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right)\right)^2 \left(-\frac{L_n''(\widehat{\theta}_n)}{n}\right) \xrightarrow[n\to\infty]{\mathrm{d}} \chi_1^2.$$

We used (38). Plugging this into (80) we see that it remains to show that the second summand in (80) vanishes in probability as $n \to \infty$. To see this, we write

$$\frac{1}{3}(\widehat{\theta}_n - \theta_0)^3 L_n'''(\theta_n^*) = \frac{1}{3}(\widehat{\theta}_n - \theta_0)\left(\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right)\right)^2 \frac{1}{n} L_n'''(\theta_n^*).$$

By (41) from the proof of Theorem 23 again, we know that $L_n'''(\theta_n^*)/n$ is bounded in probability. Also, by (38), $\left(\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right)\right)^2$ is bounded in probability, and finally by the consistency of $\widehat{\theta}_n$ we know that $(\widehat{\theta}_n - \theta_0) \xrightarrow[n\to\infty]{\mathsf{P}} 0$. Altogether, the remainder in (80) does indeed converge to zero in probability

$$\frac{1}{3}(\widehat{\theta}_n - \theta_0)^3 L_n'''(\theta_n^*) \xrightarrow[n\to\infty]{\mathsf{P}} 0,$$

and the asymptotic distribution of LR_n is χ_1^2 as we wanted to show. □

The result of Theorem 30 can be used for the construction of asymptotic tests about hypotheses on θ. For the both-sided tests $H_0\colon \theta_X = \theta_0$ against $H_1\colon \theta_X \neq \theta_0$ for $\theta_0 \in \Theta$ given, any of the test statistics in Theorem 30 can be used. For tests using R_n or W_n, we reject H_0 at level α if the observed absolute value of the test statistic exceeds $u_{1-\alpha/2}$, the $(1-\alpha/2)$-quantile of the standard normal distribution $\mathsf{N}(0,1)$. For tests based on R_n^2, W_n^2, or LR_n the asymptotic critical region is $[F_1^{-1}(1-\alpha), \infty)$, for F_1^{-1} the quantile function of the χ_1^2 distribution. The test statistics R_n and W_n allow us also to test against one-sided alternatives; their squares or the likelihood ratio test do not permit this.

Note that the Fisher information $J(\theta_0)$ is in Theorem 30 estimated using different approaches — for the Rao score tests we use directly $J(\theta_0)$, but for the Wald tests we plug in $J(\widehat{\theta}_n)$ and assume that the function J is continuous in θ_0. In principle, we could in both tests use $J(\theta_0)$, or also $J(\widehat{\theta}_n)$ if J is continuous. The asymptotic distributions of the statistics would not be affected by this change. Our choice in Theorem 30 is a customary one. Note that in the Rao score test, we do not have to have access to the estimator $\widehat{\theta}_n$, as in this way, the test works also without being able to find $\widehat{\theta}_n$ explicitly.

Example 3.3. Consider a random sample $X_1, \ldots, X_n$ from an exponential distribution with density $f(x;\lambda) = \lambda \exp(-\lambda x)\,\mathbb{I}(x > 0)$ as in Example 3.1. The parameter $\lambda \in (0, \infty)$ is unknown and we want to test a hypothesis

$$H_0\colon \lambda_X = \lambda_0 \text{ against } H_1\colon \lambda_X \neq \lambda_0,$$

for $\lambda_0 > 0$ given. The log-likelihood of $\boldsymbol{X} = (X_1, \ldots, X_n)^\mathsf{T}$ is

$$L_n(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i + \log \left(\mathbb{I} \left(\min_{i=1,\ldots,n} x_i > 0 \right) \right),$$

for $x_1, \ldots, x_n$ the observed values of $\boldsymbol{X}$. The maximum likelihood estimator $\widehat{\lambda}_n$ satisfies the likelihood equation

$$L_n'(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0,$$

and clearly $\widehat{\lambda}_n = \left(\bar{X}_n\right)^{-1}$. For the Fisher information of λ we have

$$-L_n''(\lambda) = \frac{n}{\lambda^2} = J_n(\lambda).$$

Applying Theorem 30 we obtain three test statistics

$$\begin{aligned}
R_n^2 &= \frac{\left(n/\lambda_0 - \sum_{i=1}^{n} X_i\right)^2}{n/\lambda_0^2} = \left(\frac{\sqrt{n}\left(\bar{X}_n - 1/\lambda_0\right)}{1/\lambda_0} \right)^2, \\
W_n^2 &= \frac{n}{\widehat{\lambda}_n^2} \left(\widehat{\lambda}_n - \lambda_0 \right)^2 = \left(\sqrt{n}\, \bar{X}_n \left(1/\bar{X}_n - \lambda_0 \right) \right)^2, \\
LR_n &= 2\, n \left(\log\left(\widehat{\lambda}_n / \lambda_0 \right) - \bar{X}_n \left(\widehat{\lambda}_n - \lambda_0 \right) \right).
\end{aligned}$$

Each test is different, but they all reject H_0 if and only if the observed value of the test statistic exceeds the $(1-\alpha)$-quantile of the distribution χ_1^2. Note that the test statistic R_n^2 can be interpreted using the central limit theorem directly, as under H_0 we have

$$\sqrt{n}\left(\bar{X}_n - 1/\lambda_0\right) = \sqrt{n}\left(\bar{X}_n - \mathsf{E}\, X_1\right) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}\left(0, \mathsf{var}\, X_1\right) = \mathsf{N}\left(0, 1/\lambda_0^2\right).$$

Analogously, the form of the test statistic W_n^2 follows from the previous formula by applying the delta theorem with function $g(t) = 1/t$, giving

$$\sqrt{n}\left(1/\bar{X}_n - \lambda_0\right) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}\left(0, \lambda_0^2\right).$$

In this case, the parameter λ in the variance term is approximated by $\widehat{\lambda}_n$ in W_n^2, which is allowed by the Cramér-Slutsky theorem. △

An extension of Theorem 30 to a multi-dimensional parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ is straightforward.

Theorem 31. *Suppose that all conditions of Theorem 25 are satisfied, and let $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ be given. Then, if the true value of the parameter $\boldsymbol{\theta}_X$ is $\boldsymbol{\theta}_0$, we have the following.*

1. ***The Rao score test.*** *Writing* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\mathsf{T}$ *and denoting*

$$\boldsymbol{U}_n(\boldsymbol{\theta}) = \left(\frac{\partial L_n(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial L_n(\boldsymbol{\theta})}{\partial \theta_p} \right)^\mathsf{T} = \frac{\partial L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

we have that

$$R_n = \frac{1}{n} \left(\boldsymbol{U}_n(\boldsymbol{\theta}_0) \right)^\mathsf{T} \left(\boldsymbol{J}\left(\boldsymbol{\theta}_0 \right) \right)^{-1} \boldsymbol{U}_n(\boldsymbol{\theta}_0) \xrightarrow[n\to\infty]{\mathrm{d}} \chi^2_p.$$

2. ***Wald test.*** *If the Fisher information matrix* $\boldsymbol{J}(\boldsymbol{\theta})$ *is continuous in* $\boldsymbol{\theta}_0$*, then*

$$W_n = n \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \right)^\mathsf{T} \boldsymbol{J}(\widehat{\boldsymbol{\theta}}_n) \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \right) \xrightarrow[n\to\infty]{\mathrm{d}} \chi^2_p.$$

3. ***Likelihood ratio test.*** *If the Fisher information matrix* $\boldsymbol{J}(\boldsymbol{\theta})$ *is continuous in* $\boldsymbol{\theta}_0$*, then*

$$LR_n = 2 \left(L_n(\widehat{\boldsymbol{\theta}}_n) - L_n(\boldsymbol{\theta}_0) \right) \xrightarrow[n\to\infty]{\mathrm{d}} \chi^2_p.$$

Proof. The proof is based on Theorem 25. For the Rao score test and the Wald test it follows directly. The derivation of the asymptotic distribution for the likelihood ratio statistic is more technical, but completely analogous to that from the proof of Theorem 30. One starts from the multivariate Taylor series for $L_n(\boldsymbol{\theta}_0)$ around $\widehat{\boldsymbol{\theta}}_n$ and gets

$$\begin{aligned} L_n(\boldsymbol{\theta}_0) = L_n(\widehat{\boldsymbol{\theta}}_n) &+ \frac{1}{1!} \sum_{j=1}^{p} L'_{n,j}(\widehat{\boldsymbol{\theta}}_n) \left(\theta_{0,j} - \widehat{\theta}_{n,j} \right) \\ &+ \frac{1}{2!} \sum_{j=1}^{p} \sum_{k=1}^{p} L''_{n,j,k}(\widehat{\boldsymbol{\theta}}_n) \left(\theta_{0,j} - \widehat{\theta}_{n,j} \right) \left(\theta_{0,k} - \widehat{\theta}_{n,k} \right) \\ &+ \frac{1}{3!} \sum_{j=1}^{p} \sum_{k=1}^{p} \sum_{\ell=1}^{p} L'''_{n,j,k,\ell}(\boldsymbol{\theta}^*_n) \left(\theta_{0,j} - \widehat{\theta}_{n,j} \right) \left(\theta_{0,k} - \widehat{\theta}_{n,k} \right) \left(\theta_{0,\ell} - \widehat{\theta}_{n,\ell} \right), \end{aligned}$$

where $\boldsymbol{\theta}_0 = (\theta_{0,1}, \ldots, \theta_{0,p})^\mathsf{T}$, $\widehat{\boldsymbol{\theta}}_n = \left(\widehat{\theta}_{n,1}, \ldots, \widehat{\theta}_{n,p} \right)^\mathsf{T}$, $\boldsymbol{\theta}^*_n$ lies in the line segment between $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}_n$, and

$$\begin{aligned} L'_{n,j}(\widehat{\boldsymbol{\theta}}_n) &= \frac{\partial L_n(\widehat{\boldsymbol{\theta}}_n)}{\partial \theta_j}, \\ L''_{n,j,k}(\widehat{\boldsymbol{\theta}}_n) &= \frac{\partial^2 L_n(\widehat{\boldsymbol{\theta}}_n)}{\partial \theta_j \partial \theta_k}, \\ L'''_{n,j,k,\ell}(\widehat{\boldsymbol{\theta}}_n) &= \frac{\partial^3 L_n(\widehat{\boldsymbol{\theta}}_n)}{\partial \theta_j \partial \theta_k \partial \theta_\ell}. \end{aligned}$$

For the term with the first derivatives we know that $L'_{n,j}(\widehat{\boldsymbol{\theta}}_n) = 0$ for all $j = 1, \ldots, p$. Analogously as in the proof of Theorem 30 we derive that

$$-\frac{L''_{n,j,k}(\widehat{\boldsymbol{\theta}}_n)}{n} \xrightarrow[n\to\infty]{\mathsf{P}} J_{j,k}(\boldsymbol{\theta}_0), \tag{83}$$

where $J_{j,k}(\boldsymbol{\theta})$ is the (j,k)-th element of the Fisher information matrix $\boldsymbol{J}(\boldsymbol{\theta})$, for all $j,k = 1,\dots,p$. That gives

$$-\sum_{j=1}^{p}\sum_{k=1}^{p} L''_{n,j,k}(\widehat{\boldsymbol{\theta}}_n)\left(\theta_{0,j}-\widehat{\theta}_{n,j}\right)\left(\theta_{0,k}-\widehat{\theta}_{n,k}\right) = -\left(\sqrt{n}\left(\boldsymbol{\theta}_0-\widehat{\boldsymbol{\theta}}_n\right)\right)^{\mathsf{T}}\frac{\boldsymbol{L}''_n(\widehat{\boldsymbol{\theta}}_n)}{n}\left(\sqrt{n}\left(\boldsymbol{\theta}_0-\widehat{\boldsymbol{\theta}}_n\right)\right),$$

where by $\boldsymbol{L}''_n(\widehat{\boldsymbol{\theta}}_n)$ we denoted the matrix whose (j,k)-the element is $L''_{n,j,k}(\widehat{\boldsymbol{\theta}}_n)$. By Theorem 25 we know that

$$\sqrt{n}\left(\boldsymbol{\theta}_0-\widehat{\boldsymbol{\theta}}_n\right)\xrightarrow[n\to\infty]{\mathrm{d}}\mathsf{N}_p\left(\mathbf{0},\left(\boldsymbol{J}(\boldsymbol{\theta}_0)\right)^{-1}\right),$$

while from (83) we know that

$$-\frac{\boldsymbol{L}''_n(\widehat{\boldsymbol{\theta}}_n)}{n}\xrightarrow[n\to\infty]{\mathsf{P}}\boldsymbol{J}(\boldsymbol{\theta}_0). \tag{84}$$

Combining all together we get using the Cramér-Slutsky theorem that

$$-\sum_{j=1}^{p}\sum_{k=1}^{p} L''_{n,j,k}(\widehat{\boldsymbol{\theta}}_n)\left(\theta_{0,j}-\widehat{\theta}_{n,j}\right)\left(\theta_{0,k}-\widehat{\theta}_{n,k}\right)\xrightarrow[n\to\infty]{\mathrm{d}}\chi^2_p.$$

It remains to show that the last term in the Taylor expansion

$$\frac{1}{3!}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{\ell=1}^{p} L'''_{n,j,k,\ell}(\boldsymbol{\theta}^*_n)\left(\theta_{0,j}-\widehat{\theta}_{n,j}\right)\left(\theta_{0,k}-\widehat{\theta}_{n,k}\right)\left(\theta_{0,\ell}-\widehat{\theta}_{n,\ell}\right)$$

vanishes in probability as $n\to\infty$. This is shown exactly in the same way as in the proof of Theorem 30. □

The three tests based on the use of the likelihood have different pros and cons:

- In the Rao score test, we do not need to know the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n$, but need to find an inverse of the Fisher information matrix. This does not have to be simple.
- For the Wald test, we need to know the maximum likelihood estimator explicitly, and need to have access to the Fisher information matrix.
- For the likelihood ratio test, we do not need to know the Fisher information matrix at all, but need to have an explicit maximum likelihood estimator.

The tests in the multi-dimensional case are, of course, well suited only for hypotheses of the type (79); no reasonable one-sided alternatives in the situation $p>1$ exist. In all cases, we reject $H_0\colon \boldsymbol{\theta}_X=\boldsymbol{\theta}_0$ if and only if the observed value of the test statistic exceeds $F_p^{-1}(1-\alpha)$, for F_p^{-1} the quantile function of the χ^2_p distribution.

In both Theorems 30 and 31 we assumed by (P_2) or (P_2^*) that the elements of $\boldsymbol{X}$ correspond to a random sample of size n. Also, recall that the Fisher information $J(\theta)$ or $\boldsymbol{J}(\boldsymbol{\theta})$ in this

situation corresponds to the Fisher information contained in a single observation $X_1 \in \mathbb{R}^d$. Then, we know by Theorem 2 and an analogous result for the Fisher information matrix that the Fisher information contained in the whole sample $\boldsymbol{X} = (X_1, \ldots, X_n)^\mathsf{T}$ equals $J_n(\theta) = n\, J(\theta)$ or $\boldsymbol{J}_n(\boldsymbol{\theta}) = n\, \boldsymbol{J}(\boldsymbol{\theta})$. Thus, in all test statistics, in both Rao score tests and Wald tests the factor n in fact corresponds to the Fisher information being based on n observations.

Finally, it may happen that the computation of the Fisher information matrix $\boldsymbol{J}(\boldsymbol{\theta})$ is difficult. Sometimes it is therefore easier to plug in instead of $\boldsymbol{J}(\boldsymbol{\theta}_0)$ or $\boldsymbol{J}(\widehat{\boldsymbol{\theta}}_n)$ the so-called *observed Fisher information matrix*, which is an estimator of the Fisher information matrix given by

$$\widehat{\boldsymbol{J}}_n(\widehat{\boldsymbol{\theta}}_n) = -\frac{1}{n} \frac{\partial \boldsymbol{U}_n(\widehat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}^\mathsf{T}} = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f(X_i; \widehat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\mathsf{T}}.$$

Here, the derivative $\partial \boldsymbol{\theta}$ gives $\boldsymbol{U}_n(\widehat{\boldsymbol{\theta}}_n)$ being a column vector of length p. The additional derivative $\partial \boldsymbol{\theta}^\mathsf{T}$ takes the p partial derivatives of all the elements of $\boldsymbol{U}_n(\boldsymbol{\theta})$, taken as rows. It results in the expression on the right hand side being indeed a $p \times p$ matrix of second partial derivatives of $\log f(X_i; \boldsymbol{\theta})$, evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_n$. Based on Theorem 7, under mild and obvious integrability assumptions and the continuity of $\boldsymbol{J}$ in $\boldsymbol{\theta}_0$, the law of large numbers guarantees that

$$\widehat{\boldsymbol{J}}_n(\widehat{\boldsymbol{\theta}}_n) \xrightarrow[n\to\infty]{\mathsf{P}} \boldsymbol{J}(\boldsymbol{\theta}_0).$$

Therefore, also the observed Fisher information matrix could be used in the asymptotic tests instead of $\boldsymbol{J}(\boldsymbol{\theta}_0)$ or $\boldsymbol{J}(\widehat{\boldsymbol{\theta}}_n)$. This is useful, especially if the explicit Fisher information matrix is difficult to obtain. Note that since $\widehat{\boldsymbol{J}}_n(\widehat{\boldsymbol{\theta}}_n)$ estimates only the Fisher information matrix $\boldsymbol{J}(\boldsymbol{\theta}_0)$ contained in a single observation X_i, the factor n in the tests must still be used.

The end of lecture 10 (23.4.2024)

3.3.2 Tests with nuisance parameters

We now turn to the problem of hypotheses testing with the presence of the so-called nuisance parameters. Suppose that we are in the situation when $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ is the complete parameter determining the distribution of a random vector $\boldsymbol{X}$. We split this p-dimensional vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\mathsf{T}$ into two parts: a q-dimensional sub-vector $\boldsymbol{\tau}$ of the parameters of interest with $1 \leq q < p$, and the remaining $(p-q)$-dimensional sub-vector $\boldsymbol{\psi}$ of parameters that we are not interested in. Overall, we have

$$\boldsymbol{\tau} = (\theta_1, \ldots, \theta_q)^\mathsf{T} \text{ and } \boldsymbol{\psi} = (\theta_{q+1}, \ldots, \theta_p)^\mathsf{T},$$

and $\boldsymbol{\theta}^\mathsf{T} = (\boldsymbol{\tau}^\mathsf{T}, \boldsymbol{\psi}^\mathsf{T})$. The true value of the parameter is likewise denoted by $\boldsymbol{\theta}_X^\mathsf{T} = (\boldsymbol{\tau}_X^\mathsf{T}, \boldsymbol{\psi}_X^\mathsf{T})$. We have rearranged the elements of the vector $\boldsymbol{\theta}$ so that its first q elements are the parameters of interest — this is of course without loss of generality. Our intention is to perform a test of

the hypothesis

$$H_0 \colon \boldsymbol{\tau}_X = \boldsymbol{\tau}_0 \text{ against } H_1 \colon \boldsymbol{\tau}_X \neq \boldsymbol{\tau}_0,$$

for $\boldsymbol{\tau}_0 \in \mathbb{R}^q$ given. For the construction of likelihood-based tests, we split also the maximum likelihood estimator of $\boldsymbol{\theta}$ into

$$\widehat{\boldsymbol{\theta}}_n^{\mathsf{T}} = \left(\widehat{\boldsymbol{\tau}}_n^{\mathsf{T}}, \widehat{\boldsymbol{\psi}}_n^{\mathsf{T}}\right).$$

In the asymptotic tests about the whole parameter $\boldsymbol{\theta}$ in Section 3.3.1 we compared $\widehat{\boldsymbol{\theta}}_n$ with its hypothesized value $\boldsymbol{\theta}_0$ under H_0. In the present situation, it is natural also to compare $\widehat{\boldsymbol{\theta}}_n$, but this time with the maximum likelihood estimator of $\boldsymbol{\theta}$ under $H_0 \colon \boldsymbol{\tau}_X = \boldsymbol{\tau}_0$. That is, for the null hypothesis our intention is to maximize the likelihood of $\boldsymbol{\theta}$, under the restriction $\boldsymbol{\tau} = \boldsymbol{\tau}_0$, as a function of the remaining parameters $\boldsymbol{\psi}$. We thus denote by

$$\widetilde{\boldsymbol{\psi}}_n = \arg\max_{\boldsymbol{\psi} \in \mathbb{R}^{p-q}} L_n\left(\left(\boldsymbol{\tau}_0^{\mathsf{T}}, \boldsymbol{\psi}^{\mathsf{T}}\right)^{\mathsf{T}}\right)$$

the maximum likelihood estimator of $\boldsymbol{\psi}$ given $\boldsymbol{\tau} = \boldsymbol{\tau}_0$, which is basically the profile likelihood from Section 2.4. This way, we can also formally denote by

$$\widetilde{\boldsymbol{\theta}}_n^{\mathsf{T}} = \left(\boldsymbol{\tau}_0^{\mathsf{T}}, \widetilde{\boldsymbol{\psi}}_n^{\mathsf{T}}\right)$$

the complete estimator of $\boldsymbol{\theta}$ under H_0.

In what follows we need to consider the asymptotic distributions of $\widehat{\boldsymbol{\tau}}_n$ and $\widehat{\boldsymbol{\theta}}_n$ separately. According to the asymptotic distribution of $\widehat{\boldsymbol{\theta}}_n$ from Theorem 25, we will therefore need to deal with the partitioned score function

$$\boldsymbol{U}_n(\boldsymbol{\theta})^{\mathsf{T}} = \left(\boldsymbol{U}_1(\boldsymbol{\theta})^{\mathsf{T}}, \boldsymbol{U}_2(\boldsymbol{\theta})^{\mathsf{T}}\right),$$

where

$$\begin{aligned} \boldsymbol{U}_1(\boldsymbol{\theta}) &= \frac{\partial L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\tau}} = \left(\frac{\partial L_n(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial L_n(\boldsymbol{\theta})}{\partial \theta_q}\right)^{\mathsf{T}}, \\ \boldsymbol{U}_2(\boldsymbol{\theta}) &= \frac{\partial L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\psi}} = \left(\frac{\partial L_n(\boldsymbol{\theta})}{\partial \theta_{q+1}}, \ldots, \frac{\partial L_n(\boldsymbol{\theta})}{\partial \theta_p}\right)^{\mathsf{T}}. \end{aligned} \tag{85}$$

The functions $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$, of course, also depend on n and we should formally write $\boldsymbol{U}_{n,1}$ and $\boldsymbol{U}_{n,2}$. We omit the index n for brevity. To use Theorem 25 appropriately we also need to partition the inverse Fisher information matrix $(\boldsymbol{J}(\boldsymbol{\theta}_X))^{-1}$ from (49) into parts corresponding to $\boldsymbol{\tau}$ and $\boldsymbol{\psi}$ given by

$$(\boldsymbol{J}(\boldsymbol{\theta}))^{-1} = \begin{pmatrix} \boldsymbol{J}^{1,1}(\boldsymbol{\theta}) & \boldsymbol{J}^{1,2}(\boldsymbol{\theta}) \\ \boldsymbol{J}^{2,1}(\boldsymbol{\theta}) & \boldsymbol{J}^{2,2}(\boldsymbol{\theta}) \end{pmatrix},$$

where the blocks are of dimensions $q \times q$ for $\boldsymbol{J}^{1,1}$, $q \times (p-q)$ for $\boldsymbol{J}^{1,2}$, $(p-q) \times q$ for $\boldsymbol{J}^{2,1}$, and $(p-q) \times (p-q)$ for $\boldsymbol{J}^{2,2}$.

Because $\widehat{\boldsymbol{\theta}}_n$ is the maximum likelihood estimator of $\boldsymbol{\theta}$, we know that

$$\boldsymbol{U}_n(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0}.$$

Also, because of the definition of $\widetilde{\boldsymbol{\theta}}_n$ we have that

$$\boldsymbol{U}_n(\widetilde{\boldsymbol{\theta}}_n)^{\mathsf{T}} = \left(\boldsymbol{U}_1(\widetilde{\boldsymbol{\theta}}_n)^{\mathsf{T}}, \mathbf{0}^{\mathsf{T}}\right). \tag{86}$$

Using our Lemma 3 on the inverse of a matrix partitioned into blocks, we can also express all blocks $\boldsymbol{J}^{i,j}$ of the inverse $(\boldsymbol{J}(\boldsymbol{\theta}))^{-1}$ in terms of the blocks of the original Fisher information matrix $\boldsymbol{J}(\boldsymbol{\theta})$. The blocks of the inverse Fisher information matrix are of great importance since, e.g., from (49) it follows that

$$\sqrt{n}\begin{pmatrix}\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_X \\ \widehat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}_X\end{pmatrix} \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}_p\left(\mathbf{0}, (\boldsymbol{J}(\boldsymbol{\theta}_X))^{-1}\right),$$

meaning that

$$\sqrt{n}\left(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_X\right) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}_q\left(\mathbf{0}, \boldsymbol{J}^{1,1}(\boldsymbol{\theta}_X)\right). \tag{87}$$

In our final theorem we derive the variants of the asymptotic likelihood based tests under the presence of nuisance parameters.

Theorem 32. *Suppose that all conditions of Theorem 25 are satisfied, vector $\boldsymbol{\theta}$ is split into $\boldsymbol{\tau}$ and $\boldsymbol{\psi}$ as described above, and let $\boldsymbol{\tau}_0 \in \mathbb{R}^q$ be given. Suppose that the Fisher information matrix $\boldsymbol{J}(\boldsymbol{\theta})$ is continuous in the true value of the parameter $\boldsymbol{\theta}_X^{\mathsf{T}} = \left(\boldsymbol{\tau}_0^{\mathsf{T}}, \boldsymbol{\psi}_X^{\mathsf{T}}\right)$. Then we can write the following.*

1. ***The Rao score test.***

$$R_n^* = \frac{1}{n}\left(\boldsymbol{U}_1(\widetilde{\boldsymbol{\theta}}_n)\right)^{\mathsf{T}} \boldsymbol{J}^{1,1}\left(\widetilde{\boldsymbol{\theta}}_n\right) \boldsymbol{U}_1(\widetilde{\boldsymbol{\theta}}_n) \xrightarrow[n\to\infty]{\mathrm{d}} \chi_q^2.$$

2. ***Wald test.***

$$W_n^* = n\left(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0\right)^{\mathsf{T}} \left(\boldsymbol{J}^{1,1}(\widehat{\boldsymbol{\theta}}_n)\right)^{-1} \left(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0\right) \xrightarrow[n\to\infty]{\mathrm{d}} \chi_q^2.$$

3. ***Likelihood ratio test.***

$$LR_n^* = 2\left(L_n(\widehat{\boldsymbol{\theta}}_n) - L_n(\widetilde{\boldsymbol{\theta}}_n)\right) \xrightarrow[n\to\infty]{\mathrm{d}} \chi_q^2.$$

Proof. We begin with the simplest case of the Wald test. From (87) we know the asymptotic distribution of $\widehat{\boldsymbol{\tau}}_n$. Because the matrix $\boldsymbol{J}^{1,1}(\boldsymbol{\theta}_X)$ is assumed to be square, positive

definite, and symmetric, there exists a unique square, positive definite, and symmetric matrix $\left(\boldsymbol{J}^{1,1}\left(\boldsymbol{\theta}_X\right)\right)^{1/2}$ that satisfies

$$\left(\boldsymbol{J}^{1,1}\left(\boldsymbol{\theta}_X\right)\right)^{1/2} \cdot \left(\left(\boldsymbol{J}^{1,1}\left(\boldsymbol{\theta}_X\right)\right)^{1/2}\right)^{\mathsf{T}} = \boldsymbol{J}^{1,1}(\boldsymbol{\theta}_X).$$

This follows from the spectral decomposition of a symmetric positive definite matrix. Applying the inverse $\left(\boldsymbol{J}^{1,1}\left(\boldsymbol{\theta}_X\right)\right)^{-1/2}$ of this square root matrix to (87) we get

$$\boldsymbol{V}_n = \sqrt{n}\left(\boldsymbol{J}^{1,1}\left(\boldsymbol{\theta}_X\right)\right)^{-1/2}\left(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_X\right) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}_q\left(\mathbf{0}, \boldsymbol{I}\right)$$

for $\boldsymbol{I}$ the $q \times q$ identity matrix. Transforming the last expression into a quadratic form and using the continuous mapping theorem, we get

$$\boldsymbol{V}_n^{\mathsf{T}}\boldsymbol{V}_n = n\,\left(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0\right)^{\mathsf{T}}\left(\boldsymbol{J}^{1,1}(\boldsymbol{\theta}_X)\right)^{-1}\left(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0\right) \xrightarrow[n\to\infty]{\mathrm{d}} \chi^2_q.$$

It remains to substitute $\boldsymbol{J}^{1,1}(\boldsymbol{\theta}_X)$ by its consistent estimator $\boldsymbol{J}^{1,1}(\widetilde{\boldsymbol{\theta}}_n)$, and note that this consistency holds true because we assumed the continuity of $\boldsymbol{J}(\boldsymbol{\theta})$, and $\boldsymbol{J}^{1,1}(\boldsymbol{\theta})$ can be written by Lemma 3 as a continuous function of the blocks of $\boldsymbol{J}(\boldsymbol{\theta})$. The proof for the Wald test is concluded.

Analogously to the first part of this proof, for the asymptotic distribution of the Rao score test it is enough to find that

$$\frac{1}{\sqrt{n}}\boldsymbol{U}_1\left(\widetilde{\boldsymbol{\theta}}_n\right) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}_q\left(\mathbf{0}, \left(\boldsymbol{J}^{1,1}(\boldsymbol{\theta}_X)\right)^{-1}\right).$$

This is done using Taylor's expansion of the q-dimensional function $\boldsymbol{U}_1(\widetilde{\boldsymbol{\theta}}_n)$ around $\boldsymbol{\theta}_X$. This proof is not difficult, but somewhat lengthy and tedious. We omit those technical derivations; they can be found, e.g. in [1, Theorem 8.24].

We prove the result about the asymptotics of the likelihood ratio test. The proof begins very similarly to that of Theorem 31 for the likelihood ratio test without nuisance parameters. We start by expanding the log-likelihood $L_n(\widetilde{\boldsymbol{\theta}}_n)$ around $\widehat{\boldsymbol{\theta}}_n$ into a Taylor series

$$\begin{aligned}
L_n(\widetilde{\boldsymbol{\theta}}_n) &= L_n(\widehat{\boldsymbol{\theta}}_n) + \frac{1}{1!}\sum_{j=1}^{p} L'_{n,j}(\widehat{\boldsymbol{\theta}}_n)\left(\widetilde{\theta}_{n,j} - \widehat{\theta}_{n,j}\right) \\
&\quad + \frac{1}{2!}\sum_{j=1}^{p}\sum_{k=1}^{p} L''_{n,j,k}(\widehat{\boldsymbol{\theta}}_n)\left(\widetilde{\theta}_{n,j} - \widehat{\theta}_{n,j}\right)\left(\widetilde{\theta}_{n,k} - \widehat{\theta}_{n,k}\right) \\
&\quad + \frac{1}{3!}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{\ell=1}^{p} L'''_{n,j,k,\ell}(\boldsymbol{\theta}^*_n)\left(\widetilde{\theta}_{n,j} - \widehat{\theta}_{n,j}\right)\left(\widetilde{\theta}_{n,k} - \widehat{\theta}_{n,k}\right)\left(\widetilde{\theta}_{n,\ell} - \widehat{\theta}_{n,\ell}\right) \\
&= L_n(\widehat{\boldsymbol{\theta}}_n) + \mathbf{0} - \frac{1}{2}\left(\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n\right)\right)^{\mathsf{T}}\left(-\frac{\boldsymbol{L}''_n(\widehat{\boldsymbol{\theta}}_n)}{n}\right)\left(\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n\right)\right) + \boldsymbol{R}_{1,n}.
\end{aligned} \tag{88}$$

Here $\widetilde{\boldsymbol{\theta}}_n = \left(\widetilde{\theta}_{n,1}, \ldots, \widetilde{\theta}_{n,p}\right)^{\mathsf{T}}$, $\boldsymbol{\theta}_n^*$ lies in the line segment between $\widetilde{\boldsymbol{\theta}}_n$ and $\widehat{\boldsymbol{\theta}}_n$, and

$$\boldsymbol{R}_{1,n} = \frac{1}{3!} \sum_{j=1}^{p} \sum_{k=1}^{p} \sum_{\ell=1}^{p} L'''_{n,j,k,\ell}(\boldsymbol{\theta}_n^*) \left(\widetilde{\theta}_{n,j} - \widehat{\theta}_{n,j}\right) \left(\widetilde{\theta}_{n,k} - \widehat{\theta}_{n,k}\right) \left(\widetilde{\theta}_{n,\ell} - \widehat{\theta}_{n,\ell}\right)$$

is a remainder term that converges to zero in probability because $\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n \xrightarrow[n\to\infty]{\mathsf{P}} \boldsymbol{\theta}_X - \boldsymbol{\theta}_X = \mathbf{0}$, exactly as in the proof of Theorem 31. All the remaining notations above are from the proof of Theorem 31. Since we will need to operate with several more remainder terms, in the rest of this proof whenever we write $\boldsymbol{R}_{j,n}$ for $j = 1, 2, \ldots$, we always mean a sequence of random vectors that vanish in probability, i.e. $\boldsymbol{R}_{j,n} \xrightarrow[n\to\infty]{\mathsf{P}} \mathbf{0}$.

Using (84) we know that we can rewrite (88) into

$$2\left(L_n(\widehat{\boldsymbol{\theta}}_n) - L_n(\widetilde{\boldsymbol{\theta}}_n)\right) = \left(\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \widetilde{\boldsymbol{\theta}}_n\right)\right)^{\mathsf{T}} \boldsymbol{J}(\boldsymbol{\theta}_X) \left(\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \widetilde{\boldsymbol{\theta}}_n\right)\right) + \boldsymbol{R}_{2,n}. \tag{89}$$

To prove part (iii) of Theorem 25 we have found in (61) that

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X\right) = \frac{1}{\sqrt{n}} \left(\boldsymbol{J}(\boldsymbol{\theta}_X)\right)^{-1} \boldsymbol{U}_n(\boldsymbol{\theta}_X) + \boldsymbol{R}_{3,n}. \tag{90}$$

Exactly in the same way, an analogous result can be derived for $\widetilde{\boldsymbol{\theta}}_n$, or more precisely for its non-trivial part $\widetilde{\boldsymbol{\psi}}_n$. It takes the form

$$\sqrt{n}\left(\widetilde{\boldsymbol{\psi}}_n - \boldsymbol{\psi}_X\right) = \frac{1}{\sqrt{n}} \left(\boldsymbol{J}_{2,2}(\boldsymbol{\theta}_X)\right)^{-1} \boldsymbol{U}_{n,2}(\boldsymbol{\theta}_X) + \boldsymbol{R}_{4,n} \tag{91}$$

where $\boldsymbol{J}_{2,2}(\boldsymbol{\theta}_X)$ is the $(p-q) \times (p-q)$ block of the matrix

$$\boldsymbol{J}(\boldsymbol{\theta}_X) = \begin{pmatrix} \boldsymbol{J}_{1,1}(\boldsymbol{\theta}_X) & \boldsymbol{J}_{1,2}(\boldsymbol{\theta}_X) \\ \boldsymbol{J}_{2,1}(\boldsymbol{\theta}_X) & \boldsymbol{J}_{2,2}(\boldsymbol{\theta}_X) \end{pmatrix} \tag{92}$$

that corresponds to its last rows and columns. In (91) we wrote $\boldsymbol{U}_{n,2}$ for what we denoted by $\boldsymbol{U}_2$ in (85) to emphasize the dependence on the sample size n. Putting (90) and (91) together we obtain

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \widetilde{\boldsymbol{\theta}}_n\right) = \boldsymbol{A}(\boldsymbol{\theta}_X) \frac{1}{\sqrt{n}} \boldsymbol{U}_n(\boldsymbol{\theta}_X) + \boldsymbol{R}_{5,n}.$$

The symmetric matrix $\boldsymbol{A}(\boldsymbol{\theta}_X)$ takes the form

$$\boldsymbol{A}(\boldsymbol{\theta}_X) = \left(\boldsymbol{J}(\boldsymbol{\theta}_X)\right)^{-1} - \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \left(\boldsymbol{J}_{2,2}(\boldsymbol{\theta}_X)\right)^{-1} \end{pmatrix}.$$

We can therefore rewrite (89) to

$$LR_n^* = \frac{1}{\sqrt{n}} \left(\boldsymbol{U}_n(\boldsymbol{\theta}_X)\right)^{\mathsf{T}} \boldsymbol{A}(\boldsymbol{\theta}_X) \boldsymbol{J}(\boldsymbol{\theta}_X) \boldsymbol{A}(\boldsymbol{\theta}_X) \frac{1}{\sqrt{n}} \boldsymbol{U}_n(\boldsymbol{\theta}_X) + \boldsymbol{R}_{6,n}.$$

By part (iii) of Theorem 25 we know that

$$\frac{1}{\sqrt{n}}\boldsymbol{U}_n(\boldsymbol{\theta}_X) \xrightarrow[n\to\infty]{\mathrm{d}} \mathsf{N}_p\left(\mathbf{0}, \boldsymbol{J}(\boldsymbol{\theta}_X)\right).$$

The distribution of LR_n^* is therefore basically a quadratic form of an asymptotically normal p-dimensional random vector. Thus, by the Cramér-Slutsky theorem we have that LR_n^* converges in distribution to the quadratic form $\boldsymbol{Z}^\mathsf{T}\boldsymbol{A}(\boldsymbol{\theta}_X)\boldsymbol{J}(\boldsymbol{\theta}_X)\boldsymbol{A}(\boldsymbol{\theta}_X)\boldsymbol{Z}$, where $\boldsymbol{Z} \sim \mathsf{N}_p\left(\mathbf{0}, \boldsymbol{J}\left(\boldsymbol{\theta}_X\right)\right)$. It remains to specify its distribution; note that while the random vector $\boldsymbol{Z}$ is p-dimensional, the matrix $\boldsymbol{A}(\boldsymbol{\theta}_X)\boldsymbol{J}(\boldsymbol{\theta}_X)\boldsymbol{A}(\boldsymbol{\theta}_X)$ is singular, and the distribution of the quadratic form must be examined with care. To do that, we use the following lemma, proved in e.g. [6, Lemma A.4] or [1, Theorem 4.16]. For the statement of the lemma recall that a square matrix $\boldsymbol{C}$ is called idempotent if $\boldsymbol{C}\boldsymbol{C} = \boldsymbol{C}$.

Lemma 9. *Let $\boldsymbol{Z} \sim \mathsf{N}_p\left(\mathbf{0}, \boldsymbol{J}\right)$ and let $\boldsymbol{B}$ be a positive definite $p \times p$ matrix such that the matrix $\boldsymbol{B}\boldsymbol{J}$ is non-zero and idempotent with trace q. Then*

$$\boldsymbol{Z}^\mathsf{T}\boldsymbol{B}\boldsymbol{Z} \sim \chi^2_q.$$

To use Lemma 9 with $\boldsymbol{J} = \boldsymbol{J}(\boldsymbol{\theta}_X)$ and $\boldsymbol{B} = \boldsymbol{A}(\boldsymbol{\theta}_X)\boldsymbol{J}(\boldsymbol{\theta}_X)\boldsymbol{A}(\boldsymbol{\theta}_X)$ notice that

$$\boldsymbol{A}(\boldsymbol{\theta}_X)\boldsymbol{J}(\boldsymbol{\theta}_X) = \boldsymbol{I}_p - \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \left(\boldsymbol{J}_{2,2}(\boldsymbol{\theta}_X)\right)^{-1}\boldsymbol{J}_{2,1}\left(\boldsymbol{\theta}_X\right) & \boldsymbol{I}_{p-q} \end{pmatrix} \tag{93}$$

for $\boldsymbol{I}_p$ the $p \times p$ identity matrix and $\boldsymbol{J}_{2,1}$ a block of $\boldsymbol{J}$ from (92). This gives

$$\boldsymbol{B}\boldsymbol{J} = \boldsymbol{A}(\boldsymbol{\theta}_X)\boldsymbol{J}(\boldsymbol{\theta}_X)\boldsymbol{A}(\boldsymbol{\theta}_X)\boldsymbol{J}(\boldsymbol{\theta}_X) = \boldsymbol{I}_p - \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \left(\boldsymbol{J}_{2,2}(\boldsymbol{\theta}_X)\right)^{-1}\boldsymbol{J}_{2,1}\left(\boldsymbol{\theta}_X\right) & \boldsymbol{I}_{p-q} \end{pmatrix} = \boldsymbol{A}(\boldsymbol{\theta}_X)\boldsymbol{J}(\boldsymbol{\theta}_X),$$

which means that

$$\begin{aligned}(\boldsymbol{B}\boldsymbol{J})(\boldsymbol{B}\boldsymbol{J}) &= (\boldsymbol{A}(\boldsymbol{\theta}_X)\boldsymbol{J}(\boldsymbol{\theta}_X)\boldsymbol{A}(\boldsymbol{\theta}_X)\boldsymbol{J}(\boldsymbol{\theta}_X))(\boldsymbol{A}(\boldsymbol{\theta}_X)\boldsymbol{J}(\boldsymbol{\theta}_X)\boldsymbol{A}(\boldsymbol{\theta}_X)\boldsymbol{J}(\boldsymbol{\theta}_X)) \\ &= \boldsymbol{A}(\boldsymbol{\theta}_X)\boldsymbol{J}(\boldsymbol{\theta}_X)\boldsymbol{A}(\boldsymbol{\theta}_X)\boldsymbol{J}(\boldsymbol{\theta}_X) = \boldsymbol{B}\boldsymbol{J} = \boldsymbol{A}(\boldsymbol{\theta}_X)\boldsymbol{J}(\boldsymbol{\theta}_X),\end{aligned}$$

that is $\boldsymbol{B}\boldsymbol{J}$ is idempotent. Its trace (equal to its rank) is from (93) equal to $p - (p - q) = q$, concluding the proof. □

Due to (86), the Rao score test statistic can be written equivalently in a perhaps more elegant form

$$R_n^* = \frac{1}{n}\left(\boldsymbol{U}_n(\widetilde{\boldsymbol{\theta}}_n)\right)^\mathsf{T}\left(\boldsymbol{J}\left(\widetilde{\boldsymbol{\theta}}_n\right)\right)^{-1}\boldsymbol{U}_n(\widetilde{\boldsymbol{\theta}}_n).$$

It is however important to observe that even though we see that R_n^* has been written as a quadratic form of a p-dimensional vector $\boldsymbol{U}_n(\widetilde{\boldsymbol{\theta}}_n)$, its asymptotic distribution is χ^2_q with $q < p$

degrees of freedom. The reason for this is that the $(p-q)$-dimensional sub-vector $\boldsymbol{U}_2(\widetilde{\boldsymbol{\theta}}_n)$ of $\boldsymbol{U}_n(\widetilde{\boldsymbol{\theta}}_n)$ is equal to the constant zero vector.

Just as for the likelihood-based tests about the complete vector $\boldsymbol{\theta}$, also in the test statistics R_n^* and W_n^* can the matrices $\boldsymbol{J}^{1,1}\left(\widetilde{\boldsymbol{\theta}}_n\right)$ and $\boldsymbol{J}^{1,1}\left(\widehat{\boldsymbol{\theta}}_n\right)$ respectively be replaced by any estimator of the block $\boldsymbol{J}^{1,1}(\boldsymbol{\theta}_X)$ that is consistent under H_0.

We conclude our exposition by giving an example of the use of Theorem 32.

Example 3.4. For a random sample $X_1, \ldots, X_n \sim \mathsf{N}\left(\mu, \sigma^2\right)$ with $\boldsymbol{\theta} = \left(\mu, \sigma^2\right)^\mathsf{T} \in \boldsymbol{\Theta} = \mathbb{R} \times (0, \infty)$ we want to test

$$H_0 \colon \mu_X = \mu_0 \text{ against } H_1 \colon \mu_X \neq \mu_0,$$

for $\mu_0 \in \mathbb{R}$ given. We have $p = 2$ and $q = 1$; from Example 2.1 we know that

$$\widehat{\boldsymbol{\theta}}_n = \left(\widehat{\tau}_n, \widehat{\psi}_n\right)^\mathsf{T} = \left(\bar{X}_n, \frac{n-1}{n} S_n^2\right)^\mathsf{T} = \left(\bar{X}_n, \widehat{\sigma_n^2}\right)^\mathsf{T},$$

the log-likelihood takes the form

$$L_n(\boldsymbol{\theta}) = -\frac{n}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{i=1}^n \left(X_i - \mu\right)^2,$$

and the score function is

$$\boldsymbol{U}_n(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{U}_1(\boldsymbol{\theta}) \\ \boldsymbol{U}_2(\boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} \frac{\partial L_n(\boldsymbol{\theta})}{\partial \mu} \\ \frac{\partial L_n(\boldsymbol{\theta})}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} \sum_{i=1}^n \left(X_i - \mu\right) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n \left(X_i - \mu\right)^2 \end{pmatrix}.$$

To compute $\widetilde{\boldsymbol{\theta}}_n$ we maximize the log-likelihood under H_0, that is under the condition $\mu = \mu_0$. We obtain

$$\widetilde{\psi}_n = \widetilde{\sigma_n^2} = \arg\max_{\sigma^2 > 0} L_n\left(\left(\mu_0, \sigma^2\right)^\mathsf{T}\right) = \frac{1}{n} \sum_{i=1}^n \left(X_i - \mu_0\right)^2$$

and

$$\widetilde{\boldsymbol{\theta}}_n = \left(\mu_0, \widetilde{\psi}_n\right)^\mathsf{T} = \left(\mu_0, \frac{1}{n} \sum_{i=1}^n \left(X_i - \mu_0\right)^2\right)^\mathsf{T}.$$

In Example 1.11 we found that

$$\boldsymbol{J}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\,\sigma^4} \end{pmatrix}.$$

To get $\boldsymbol{J}^{1,1}(\boldsymbol{\theta})$ we need to invert $\boldsymbol{J}$, and take its first diagonal term. We get

$$\boldsymbol{J}^{1,1}(\boldsymbol{\theta}) = \sigma^2.$$

Putting all the elements together, we obtain three test statistics

$$R_n^* = \frac{1}{n}\left(\boldsymbol{U}_1(\widetilde{\boldsymbol{\theta}}_n)\right)^\mathsf{T} \boldsymbol{J}^{1,1}\left(\widetilde{\boldsymbol{\theta}}_n\right) \boldsymbol{U}_1(\widetilde{\boldsymbol{\theta}}_n) = n\frac{\left(\bar{X}_n - \mu_0\right)^2}{\widetilde{\sigma_n^2}},$$

$$W_n^* = n\left(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0\right)^\mathsf{T}\left(\boldsymbol{J}^{1,1}(\widehat{\boldsymbol{\theta}}_n)\right)^{-1}\left(\widehat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0\right) = n\frac{\left(\bar{X}_n - \mu_0\right)^2}{\widehat{\sigma_n^2}},$$

$$\begin{aligned} LR_n^* &= 2\left(L_n(\widehat{\boldsymbol{\theta}}_n) - L_n(\widetilde{\boldsymbol{\theta}}_n)\right) = -n\log\left(\frac{\widehat{\sigma_n^2}}{\widetilde{\sigma_n^2}}\right) - \frac{1}{\widehat{\sigma_n^2}}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2 + \frac{1}{\widetilde{\sigma_n^2}}\sum_{i=1}^n \left(X_i - \mu_0\right)^2 \\ &= -n\log\left(\frac{\widehat{\sigma_n^2}}{\widetilde{\sigma_n^2}}\right) - n + n = n\log\left(\frac{\sum_{i=1}^n \left(X_i - \mu_0\right)^2}{\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2}\right). \end{aligned}$$

In all cases, we reject H_0 at level $\alpha \in (0,1)$ if and only if the observed value of the test statistic exceeds the $(1-\alpha)$-quantile of the χ_1^2 distribution. Note the similarity of both R_n^* and W_n^* to the square of the usual t-test statistic in this setup. △

References

[1] Jiří Anděl. *Základy matematické statistiky*. MatfyzPress, 2011.

[2] R. M. Dudley. *Real analysis and probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original.

[3] Rick Durrett. *Probability—theory and examples*, volume 49 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, fifth edition, 2019.

[4] Robert V. Hogg and Allen T. Craig. *Introduction to mathematical statistics*. Macmillan Publishing Co., Inc., New York; Collier Macmillan Publishers, London, fourth edition, 1978.

[5] Bent Jørgensen and Rodrigo Labouriau. Exponential families and theoretical inference, 2012. https://impa.br/wp-content/uploads/2017/04/Mon_52.pdf. Accessed: 2022-03-08.

[6] Michal Kulich and Marek Omelka. NMSA331 Matematická statistika 1. Poznámky k přednášce. Univerzita Karlova, 2022. https://www.karlin.mff.cuni.cz/~komarek/vyuka/2022_23/nmsa331/ms1.pdf. Accessed: 2024-04-30.

[7] Petr Lachout. *Teorie pravděpodobnosti*. Karolinum, 1998.

[8] E. L. Lehmann and George Casella. *Theory of point estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 1998.

[9] E. L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.

[10] Athanasios Papoulis and S. Unnikrishna Pillai. *Probability, random variables, and stochastic processes*. McGraw-Hill Book Co., New York, fourth edition, 2002.

[11] L. Pick, S. Hencl, J. Spurný, and M. Zelený. Matematická analýza 1. https://www2.karlin.mff.cuni.cz/~spurny/doc/ma1/analyza.pdf, 2020. Accessed: 2022-03-28.

[12] Abraham Wald. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statistics*, 20:595–601, 1949.