

NEW YORK UNIVERSITY — COURANT INSTITUTE, MATH-UA 234

# Mathematical Statistics



*Maximilian Nitzschner*

05/06/2022

**Disclaimer:**

These are lecture notes for the course *Mathematical Statistics (MATH-UA 234)*, given at New York University in Spring 2022.

The primary textbook reference for this course is [6]. For some further reading [2, Chapter 7-12] may also be helpful. Other useful references from which these notes occasionally draw include (but are not limited to) [1, 5]. For more advanced mathematical details, see, for instance [3, 4].

These notes are preliminary and may contain typos. If you see any mistakes or think that the presentation is unclear and could be improved, please send an email to:

[maximilian.nitzschner@cims.nyu.edu](mailto:maximilian.nitzschner@cims.nyu.edu). All comments and suggestions are appreciated.

# Contents

<b>0</b>	<b>Motivation</b>	<b>5</b>
<b>1</b>	<b>Probability Essentials</b>	<b>7</b>
1.1	Probability spaces and random variables . . . . .	7
1.2	Some elementary distributions . . . . .	10
1.3	Joint distribution of random variables and independence . . . . .	13
1.4	Limit theorems . . . . .	18
<b>2</b>	<b>Special distributions: Multivariate normal, <math>\chi^2</math>-, <math>t</math>- and <math>F</math>-distributions</b>	<b>23</b>
2.1	The multivariate normal distributions . . . . .	23
2.2	The $\chi^2$ -, $t$ - and $F$ -distributions . . . . .	25
<b>3</b>	<b>Introduction to inference: Estimators, statistics, sufficiency</b>	<b>28</b>
3.1	Estimators and their elementary properties . . . . .	29
3.2	Sufficient statistics . . . . .	32
<b>4</b>	<b>Construction principles for estimators</b>	<b>36</b>
4.1	The method of moments . . . . .	36
4.2	Maximum-Likelihood estimators . . . . .	37
4.3	UMVU estimators . . . . .	40
4.4	Exponential Families . . . . .	45
<b>5</b>	<b>Asymptotic properties of the Maximum-Likelihood estimators</b>	<b>47</b>
5.1	Consistency of the MLE . . . . .	47
5.2	Fisher information, Cramér-Rao inequality and asymptotic efficiency . . . . .	48
<b>6</b>	<b>Confidence intervals</b>	<b>53</b>
<b>7</b>	<b>Statistical tests</b>	<b>57</b>
7.1	Basic notions of statistical tests . . . . .	57
7.2	The Neyman-Pearson lemma . . . . .	61
7.3	The $Z$ - and $t$ -tests . . . . .	64
7.4	Two-sided tests . . . . .	67
7.5	$p$ -values . . . . .	69
7.6	Pearson's $\chi^2$ -test . . . . .	70
<b>8</b>	<b>A brief introduction to Bayesian statistics</b>	<b>73</b>
8.1	The Bayesian method: Prior and posterior distributions . . . . .	73

8.2	Choice of the prior . . . . .	76
8.3	The Bayes estimator . . . . .	78
<b>9</b>	<b>Linear regression and the method of least squares</b>	<b>79</b>
9.1	The method of least squares and simple linear regression . . . . .	79
9.2	Connection with Maximum-Likelihood estimators . . . . .	80
9.3	The general linear model . . . . .	81

## 0 Motivation

The area of *mathematical statistics* or *statistical inference* is — broadly speaking — concerned with the determination of parameters, trends, regularities etc. from data that exhibit a certain *randomness*.

Consider the following examples:

- ▶ Is a coin of a given currency fair, i.e. is the probability to obtain Heads (H) or Tails (T) equal to  $\frac{1}{2}$ ?
- ▶ What is the size of a certain fundamental physical constant such as the elementary charge  $e$  or the gravitational constant  $\gamma$ ?
- ▶ How widespread is a certain opinion / belief within a population?
- ▶ Is a newly developed treatment against a certain disease more effective than a previous one?

To answer such questions, one ultimately has to perform *measurements*, which are not exact and involve a level of randomness, either due to the inherent inaccuracy of the way the measurement is obtained, or due to limits in the sample size.<sup>1</sup> The purpose of mathematical statistics is then to develop techniques and tools to give an estimate / make a decision / ... with *high accuracy*, or in other words, to *minimize the risk* of obtaining an incorrect value for an unknown quantity under consideration, or to make a wrong decision. Quantifying such a risk is the main underlying theme in most of the material covered in this course, and naturally brings into play the framework of Probability Theory.

To illustrate some ideas, consider the first example above.

*Example 0.1.* Suppose you are given a coin. Let  $p \in (0, 1)$  denote the probability that the coin shows Heads (H) when flipped once.

- (i) To determine  $p$ , you may want to flip the coin a certain number of times  $n$ , say  $n = 100$ , and to record the (random) outcomes as  $X_1, X_2, \dots, X_n$ , where for  $j = 1, 2, \dots, n$  we set

$$X_j = \begin{cases} 1, & \text{if the coin shows H on the } j\text{th toss,} \\ 0, & \text{if the coin shows T on the } j\text{th toss.} \end{cases}$$

---

<sup>1</sup>For instance, in the third example, one could think of the population being that of a large country, and the statistical procedure being a poll of a representative sample of the population of that country.

Giving an approximation of  $p$  by some function  $\hat{p}_n \equiv \hat{p}_n(X_1, \dots, X_n)$  (a so-called *estimator*) is the purpose of (point) estimation. A very natural guess of how to construct  $\hat{p}_n$  would be

$$\hat{p}_n = \frac{1}{n} \sum_{j=1}^n X_j = \frac{\text{number of times H was obtained}}{\text{number of flips}}. \quad (0.1)$$

We will see in this course, why and in what sense  $\hat{p}_n$  is indeed an optimal choice for such an estimator of  $p$ .

- (ii) In the previous example, we may also want to determine some error bars on our estimate of  $p$ . Clearly if we obtained that the coin shows H 49 times when flipped  $n = 100$  times, we would obtain  $\hat{p}_n = 0.49$ , but  $p = \frac{1}{2}$  also seems consistent with such an outcome. We could therefore ask:

$$\text{What is an interval } \hat{I}_n \text{ such that } p \in \hat{I}_n \text{ with a probability of at least (say) 95\%?} \quad (0.2)$$

Note that  $p$  is not random, but unknown, and  $\hat{I}_n \equiv \hat{I}_n(X_1, \dots, X_n)$  depends on the random sample  $X_1, \dots, X_n$ . Such intervals (or more general sets) are called **confidence regions**.

- (iii) Now suppose someone *claims* that the coin is biased, and in fact shows H with probability  $p = 60\%$ . How can one substantiate or disprove such a claim? Again, we could flip the coin  $n = 100$  times, and see whether under the *hypothesis*  $p = 0.6$ , the observed outcome is very unlikely: For instance, if H came up 49 times, this seems at odds with  $p = 0.6$ . To decide whether the hypothesis can or cannot be rejected based on the observed data, we will study **statistical tests**.

Besides the three topics mentioned above, we will also study other fundamental aspects of Statistics, such as **regression**, **Bayesian statistics** and **nonparametric estimation**.

## Probability Prerequisites

As outlined, these notes make use of the language of Probability Theory, and some familiarity with the concepts – at the level of, e.g., *Theory of Probability (MATH-UA 233)* – is helpful. Nevertheless, the notes should be reasonably self-contained when used together with the textbook [6]. In the first chapter, some fundamentals of Probability Theory are recalled without proofs. In some instances, some advanced concepts / notions from Probability Theory will be mentioned, and those will be **marked with the symbol (♠)**: These concepts are *only stated for completeness*, and you may choose to ignore them at first reading.

# 1 Probability Essentials

(Reference: [6, Chapter 1–5])

In this chapter, we recall without proofs some key concepts and results from Probability Theory that will be used throughout this course. We also introduce some special distributions that are relevant for Statistics, such as the multivariate normal,  $\chi^2$ -,  $t$ -, and  $F$ -distributions.

## 1.1 Probability spaces and random variables

**Definition 1.1.** A *probability space* is a triple  $(\Omega, \mathcal{F}, \mathbf{P})$  consisting of

- (i) a non-empty set  $\Omega$  of outcomes, called *sample space*,
- (ii) (♠) a  $\sigma$ -algebra  $\mathcal{F} \subseteq \mathcal{P}(\Omega)$  describing all events<sup>1</sup>,
- (iii) a probability measure  $\mathbf{P} : \mathcal{F} \rightarrow [0, 1]$ , i.e. a map fulfilling

$$(P1) \quad \mathbf{P}[\Omega] = 1,$$

$$(P2) \quad A_1, A_2, \dots \in \mathcal{F} \text{ with } A_i \cap A_j = \emptyset \text{ for } i \neq j \Rightarrow \mathbf{P} \left[ \bigcup_{j=1}^{\infty} A_j \right] = \sum_{j=1}^{\infty} \mathbf{P}[A_j]. \quad (1.1)$$

We recall that the union  $\cup$  of events has the interpretation “or”, the intersection  $\cap$  has the interpretation “and” and the complement stands for the opposite event.

In most cases that are relevant for us, we will in fact not deal with probability spaces themselves, but instead with random variables.

*End of Lecture 1*

**Definition 1.2.** A (*real*) *random variable* is a map<sup>2</sup>  $X : \Omega \rightarrow \mathbb{R}$ . We denote the image of  $X$  by

$$\Omega_X = \{X(\omega) : \omega \in \Omega\} \subseteq \mathbb{R}. \quad (1.2)$$

<sup>1</sup>(♠) A  $\sigma$ -algebra on  $\Omega$  is a subset of the power set  $\mathcal{P}(\Omega)$  that must fulfill the three axioms

$$(S1) \quad \Omega \in \mathcal{F},$$

$$(S2) \quad A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F},$$

$$(S3) \quad A_1, A_2, \dots \in \mathcal{F} \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}.$$

You may simply think of  $\mathcal{F}$  as “all (nice) subsets” of  $\Omega$

<sup>2</sup>(♠) More precisely,  $X$  needs to be a  $\mathcal{F}$ - $\mathcal{B}(\mathbb{R})$ -measurable map, meaning that

$$X^{-1}(B) \in \mathcal{F} \quad \text{for all } B \in \mathcal{B}(\mathbb{R}),$$

with  $\mathcal{B}(\mathbb{R})$  the *Borel- $\sigma$ -algebra* on  $\mathbb{R}$ . This is because  $\{X \in B\}$  should always be an event, i.e. an element of  $\mathcal{F}$ .

Take for instance  $(\Omega = \{1, 2, 3, 4, 5, 6\}^2, \mathcal{P}(\Omega), \mathbf{P})$  with  $\mathbf{P}[A] = \frac{|A|}{36}$  (the probability space modelling two rolls of a fair die), and

$$X_n : \Omega \rightarrow \mathbb{R}, \quad X_n(\omega_1, \omega_2) = \omega_n, \quad n \in \{1, 2\}, \quad (1.3)$$

then  $X_1$  describes the outcome of the first roll, and  $X_2$  describes the outcome of the second roll. Throughout this course, we will most of the time *only* speak about random variables and their (joint) distributions. Thus:

*When considering random variables  $X_1, X_2, \dots$ , we always tacitly assume the existence of a large enough probability space  $(\Omega, \mathcal{F}, \mathbf{P})$ , on which the  $X_n$  are defined.* (1.4)

Let now  $X$  be a random variable. The assignment  $B \mapsto \mathbf{P}[X \in B]$  for a subset<sup>3</sup>  $B \subseteq \mathbb{R}$  yields itself a new probability measure on  $\mathbb{R}$ , called the *law of  $X$*  (or *distribution of  $X$* ), sometimes denoted as  $\mathbf{P}_X$ . Two types of laws are particularly important:

- If  $\Omega_X$  is countable (most typical cases:  $\Omega_X \subseteq \mathbb{Z}$ ), then  $X$  is *discrete* and its law is characterized by the probability mass function

$$p_X(k) = \mathbf{P}[X = k], \quad k \in \Omega_X. \quad (1.5)$$

A probability mass function fulfills  $\sum_{k \in \Omega_X} p_X(k) = 1$  and  $p_X(k) \geq 0$  (in fact,  $p_X(k) \in [0, 1]$ ).

- We say that  $X$  is *continuous* if its law is characterized by a probability density function  $f_X$ , namely

$$\mathbf{P}[X \in [a, b]] = \int_a^b f_X(x) dx, \quad a, b \in \mathbb{R}, a < b. \quad (1.6)$$

A density function is a piecewise continuous function  $f_X : \mathbb{R} \rightarrow [0, \infty)$  with the normalization  $\int_{-\infty}^{\infty} f_X(x) dx = 1$ .

Let us remark that in the continuous case, we have  $\mathbf{P}[X = a] = 0$  for every  $a \in \mathbb{R}$ , and so the integral  $\int_a^b f_X(x) dx$  also equals  $\mathbf{P}[X \in [a, b]]$ ,  $\mathbf{P}[X \in (a, b)]$  and  $\mathbf{P}[X \in (a, b]]$ .

**Warning:** Not every random variable is discrete or continuous! In this course however, most random variables considered will fall into one of these two categories.

**Definition 1.3.** Let  $X$  be a real random variable. For  $x \in \mathbb{R}$ , we set

$$F_X(x) = \mathbf{P}_X[(-\infty, x]] = \mathbf{P}[X \leq x]. \quad (1.7)$$

The function  $F_X$  is called the *cumulative distribution function* (CDF) of (the law of)  $X$ .

We recall some important properties of the cumulative distribution function without proof.

<sup>3</sup> (♠) More precisely: a Borel subset, i.e. an element of  $\mathcal{B}(\mathbb{R})$ .




**Lemma 1.4.** Let  $X$  be a real random variable. Its cumulative distribution function  $F = F_X$  satisfies the following properties:

- (i)  $F(x) \in [0, 1]$  for all  $x \in \mathbb{R}$ .
- (ii)  $F$  is non-decreasing.
- (iii)  $F$  is right continuous, i.e.
$$\lim_{\varepsilon \downarrow 0} F(x + \varepsilon) = F(x). \quad (1.8)$$
- (iv)  $\lim_{x \rightarrow -\infty} F(x) = 0$  and  $\lim_{x \rightarrow \infty} F(x) = 1$ .

In fact, it is also true that every function satisfying (i)–(iv) above is a cumulative distribution function of some random variable  $X$ . Moreover:

**Lemma 1.5.** Two real random variables  $X$  and  $Y$  are equal in distribution (written as  $X \stackrel{d}{=} Y$ ) if and only if their cumulative distribution functions coincide, i.e.  $F_X(x) = F_Y(x)$  for every  $x \in \mathbb{R}$ .

This means we can characterize the law of  $X$  by

- the probability mass function  $p_X$  if  $X$  is discrete,
- the probability density function  $f_X$  if  $X$  is continuous,
- the cumulative distribution function  $F_X$ , for any random variable,
- (the ) characteristic function  $\varphi_X$ , for any random variable).

**Warning:** Keep in mind that equality in law does not mean equality of the random variables as maps. For instance, in (1.3) we have two random variables  $X_1$  and  $X_2$  with  $X_1 \neq X_2$ , but  $X_1 \stackrel{d}{=} X_2$ .

The calculation of the cumulative distribution function is simple in principle, if  $X$  is either discrete or continuous and its probability mass function / probability density function is known:

- If  $X$  is a discrete real random variable, then we can write

$$F_X(x) = \sum_{y \leq x} \mathbf{P}[X = y]. \quad (1.9)$$

- If  $X$  is a continuous random variable, then we have

$$F_X(x) = \int_{-\infty}^x f_X(t) dt, \quad (1.10)$$

Since  $f_X$  is assumed to be piecewise continuous, we have (by the fundamental theorem of Calculus) the important identity

$$f_X(x) = F'_X(x), \quad \text{at all points of continuity of } f_X. \quad (1.11)$$

Before we recall some standard discrete and continuous distributions, we introduce the concept of (mathematical) expectation.

**Definition 1.6.** Let  $X$  be a real random variable.

- (i) The *expected value* or *mean* or *first moment* of  $X$  is given by

$$\mathbf{E}[X] = \begin{cases} \sum_{k \in \Omega_X} k \cdot p_X(k), & X \text{ discrete,} \\ \int_{-\infty}^{\infty} x \cdot f_X(x) dx, & X \text{ continuous,} \end{cases} \quad (1.12)$$

if the expression on the right-hand side is well-defined (otherwise we say that  $X$  does not have an expected value).

- (ii) The *variance* of  $X$  is given by

$$\text{Var}[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2 (\geq 0), \quad (1.13)$$

which exists if  $\mathbf{E}[X^2]$  exists. The *standard deviation* of  $X$  is the square-root of the variance:

$$\sigma(X) = \sqrt{\text{Var}[X]}. \quad (1.14)$$

We also recall the following rule to calculate the expectation of transformed random variables: If  $g : \mathbb{R} \rightarrow \mathbb{R}$  is a (sufficiently nice<sup>4</sup>) function, then

$$\mathbf{E}[g(X)] = \begin{cases} \sum_{k \in \Omega_X} g(k) \cdot p_X(k), & X \text{ discrete,} \\ \int_{-\infty}^{\infty} g(x) \cdot f_X(x) dx, & X \text{ continuous,} \end{cases}$$

if the expression on the right-hand side is well-defined.

## 1.2 Some elementary distributions

Here, we will briefly recall some of the most important discrete and continuous distributions from probability theory.

*Notation:*  $X \sim \mathbf{Q}$  means that the law of  $X$  is given by  $\mathbf{Q}$ .

### The discrete uniform distribution $\mathcal{U}(S)$

Given a finite subset  $S$ , we say that  $X \sim \mathcal{U}(S)$  if

$$\mathbf{P}[X = k] = \frac{1}{|S|}, \quad k \in S. \quad (1.15)$$

The uniform distribution should be used if every possible realization of a random variable  $X$  taking values in  $S$  is equally likely. In principle  $S$  does not even have to be a subset of  $\mathbb{R}$ , but in most cases  $S = \{1, \dots, N\}$  for  $N \in \mathbb{N}$ . If we restrict our attention therefore to  $X \sim \mathcal{U}(\{1, \dots, N\})$ , we have:

---

<sup>4</sup>(♠) i.e. measurable: for instance this is true if  $g$  is piecewise continuous.

- Parameters:  $N \in \mathbb{N}$ , Expectation:  $\mathbf{E}[X] = \frac{N+1}{2}$ , Variance:  $\text{Var}[X] = \frac{N^2-1}{12}$ .

**The Bernoulli distribution**  $\text{Ber}(p)$

We say that  $X \sim \text{Ber}(p)$  if

$$\mathbf{P}[X = 0] = 1 - p, \quad \mathbf{P}[X = 1] = p. \quad (1.16)$$

The Bernoulli distribution describes a binary coin flip with a biased coin, where  $p$  is the “success probability” (for instance obtaining H when flipping a coin once). More formally, we speak of a “Bernoulli experiment”.

- Parameters:  $p \in [0, 1]$ , Expectation:  $\mathbf{E}[X] = p$ , Variance:  $\text{Var}[X] = p(1 - p)$ .

**The Binomial distribution**  $\text{Bin}(n, p)$

We say that  $X \sim \text{Bin}(n, p)$  if

$$\mathbf{P}[X = k] = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k \in \{0, 1, \dots, n\}. \quad (1.17)$$

The Binomial distribution describes the number of successes in an  $n$ -fold repetition of independent Bernoulli experiments<sup>5</sup>, each having success probability  $p$ . Note in particular that  $\text{Bin}(1, p) = \text{Ber}(p)$ .

- Parameters:  $n \in \mathbb{N}$ ,  $p \in [0, 1]$ , Expectation:  $\mathbf{E}[X] = np$ , Variance:  $\text{Var}[X] = np(1 - p)$ .

**The Geometric distribution**  $\text{Geo}(p)$

We say that  $X \sim \text{Geo}(p)$  if

$$\mathbf{P}[X = k] = p(1 - p)^{k-1}, \quad k \in \mathbb{N}. \quad (1.18)$$

A geometrically distributed random variable  $X \sim \text{Geo}(p)$  describes the number of trials needed for the first success in repetitions of independent Bernoulli experiments with success probability  $p$ .

- Parameters:  $p \in (0, 1)$ , Expectation:  $\mathbf{E}[X] = \frac{1}{p}$ , Variance:  $\text{Var}[X] = \frac{1-p}{p^2}$ .

**The Poisson distribution**  $\text{Pois}(\lambda)$

We say that  $X \sim \text{Pois}(\lambda)$  if

$$\mathbf{P}[X = k] = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k \in \mathbb{N}_0. \quad (1.19)$$

A Poisson-distributed random variable  $X \sim \text{Pois}(\lambda)$  typically describes the number of occurrences of certain rare events during a given time period, and  $\lambda$  has the interpretation of “rate  $\times$  time”.

---

<sup>5</sup>Indeed, if  $X_1, \dots, X_n$  are independent (a concept we recall in the next section)  $\text{Ber}(p)$ -random variables, then  $X_1 + \dots + X_n \sim \text{Bin}(n, p)$ .

- Parameters:  $\lambda > 0$ , Expectation:  $\mathbf{E}[X] = \lambda$ , Variance:  $\text{Var}[X] = \lambda$ .

**The continuous uniform distribution**  $\mathcal{U}([a, b])$

We say that  $X \sim \mathcal{U}([a, b])$  if the law of  $X$  has density

$$f_X(x) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x). \quad (1.20)$$

- Parameters:  $a, b \in \mathbb{R}$  with  $a < b$ , Expectation:  $\mathbf{E}[X] = \frac{a+b}{2}$ , Variance:  $\text{Var}[X] = \frac{(b-a)^2}{12}$ .

**The exponential distribution**  $\mathcal{E}(\beta)$

We say that  $X \sim \mathcal{E}(\beta)$  if the law of  $X$  has density<sup>6</sup>

$$f_X(x) = \beta e^{-\beta x} \mathbb{1}_{[0,\infty)}(x). \quad (1.21)$$

This distribution usually describes some waiting time, for instance the life-time of some device, or the elapsed time between two “clicks” in a Geiger-Müller counter etc.

- Parameters:  $\beta > 0$ , Expectation:  $\mathbf{E}[X] = \frac{1}{\beta}$ , Variance:  $\text{Var}[X] = \frac{1}{\beta^2}$ .

**The Gamma distribution**  $\Gamma(\alpha, \beta)$

We say that  $X \sim \Gamma(\alpha, \beta)$  if the law of  $X$  has density

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \mathbb{1}_{[0,\infty)}(x), \quad (1.22)$$

where  $\Gamma$  is the function

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx. \quad (1.23)$$

If  $\alpha \in \mathbb{N}$ , one has  $\Gamma(\alpha) = (\alpha-1)!$ . Note that  $\Gamma(1, \beta) = \mathcal{E}(\beta)$ . For  $\alpha \in \mathbb{N}$ , one can interpret  $X \sim \Gamma(\alpha, \beta)$  as the waiting time for the  $\alpha^{\text{th}}$  consecutive occurrence of independent events, each of which having an  $\mathcal{E}(\beta)$ -distribution.<sup>7</sup>

- Parameters:  $\alpha, \beta > 0$ , Expectation:  $\mathbf{E}[X] = \frac{\alpha}{\beta}$ , Variance:  $\text{Var}[X] = \frac{\alpha}{\beta^2}$ .

**The normal (or Gaussian) distribution**  $\mathcal{N}(\mu, \sigma^2)$

We say that  $X \sim \mathcal{N}(\mu, \sigma^2)$  if the law of  $X$  has density

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}. \quad (1.24)$$

The normal distribution is ubiquitous in nature and is often found when dealing with a continuous quantity with mean  $\mu$  and a certain spread  $\sigma$ . Its universality comes to some extent from the central limit theorem (see later).

<sup>6</sup>**Warning:** This is defined in a non-standard way in [6], with  $\frac{1}{\beta}$  replacing  $\beta$ . In these notes we will stick to the usual convention (1.21).

<sup>7</sup>If  $X_1, \dots, X_n$  are independent  $\mathcal{E}(\beta)$ -distributed random variables, then  $X_1 + \dots + X_n \sim \Gamma(n, \beta)$ .

► Parameters:  $\mu \in \mathbb{R}, \sigma > 0$ , Expectation:  $\mathbf{E}[X] = \mu$ , Variance:  $\text{Var}[X] = \sigma^2$ .

---

*End of Lecture 2*

We will encounter various other important distributions throughout the course, but the elementary ones above will serve as important benchmark cases for certain concepts introduced later. For illustrational purposes, we verify the claims about the expectation and variance of the exponential distribution and also calculate its cumulative distribution function.

*Example 1.7.* Let  $X \sim \mathcal{E}(\beta)$  with  $\beta > 0$ . Then

$$\begin{aligned} \mathbf{E}[X] &\stackrel{(1.12)}{=} \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^{\infty} \beta x e^{-\beta x} dx \\ &= \left[ -x e^{-\beta x} \right]_0^{\infty} + \int_0^{\infty} e^{-\beta x} dx = \left[ -\frac{1}{\beta} e^{-\beta x} \right]_0^{\infty} = \frac{1}{\beta}, \end{aligned} \quad (1.25)$$

using integration by parts in the third equality. For the variance, we first calculate  $\mathbf{E}[X^2]$  (the expectation of the transformed random variable  $g(X)$  with  $g(t) = t^2$ ):

$$\begin{aligned} \mathbf{E}[X^2] &= \int_{-\infty}^{\infty} x^2 f_X(x) dx = \int_0^{\infty} \beta x^2 e^{-\beta x} dx \\ &= \left[ -x^2 e^{-\beta x} \right]_0^{\infty} + \int_0^{\infty} 2x e^{-\beta x} dx = \left[ -\frac{2x}{\beta} e^{-\beta x} \right]_0^{\infty} + \int_0^{\infty} \frac{2}{\beta} e^{-\beta x} dx \\ &= \left[ -\frac{2}{\beta} e^{-\beta x} \right]_0^{\infty} = \frac{2}{\beta^2}. \end{aligned} \quad (1.26)$$

It follows that

$$\text{Var}[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2 = \frac{2}{\beta^2} - \left(\frac{1}{\beta}\right)^2 = \frac{1}{\beta^2}. \quad (1.27)$$

The cumulative distribution function  $F_X$  is given by

$$\begin{aligned} F_X(x) &\stackrel{(1.10)}{=} \int_{-\infty}^x f_X(x) dx = \begin{cases} 0, & x < 0, \\ \int_0^x \beta e^{-\beta t} dt, & x \geq 0 \end{cases} \\ &= (1 - e^{-\beta x}) \mathbb{1}_{[0, \infty)}(x). \end{aligned} \quad (1.28)$$

### 1.3 Joint distribution of random variables and independence

Until now, we considered for the most part a single random variable  $X : \Omega \rightarrow \mathbb{R}$  and its law  $\mathbf{P}_X$  when a probability space  $(\Omega, \mathcal{F}, \mathbf{P})$  is given. We will now consider many random variables  $X_1, \dots, X_n$  defined on the same probability space and their *joint* distribution.

Similarly as in the case of a single random variable, one can consider the assignment  $B \mapsto \mathbf{P}[(X_1, \dots, X_n)^\top \in B]$  for a subset<sup>8</sup>  $B \subseteq \mathbb{R}^n$ . This is the *joint law* (or *joint distribution*) of  $X_1, \dots, X_n$ , written as  $\mathbf{P}_{(X_1, \dots, X_n)}$ . Again – as in the case of a single random variable – the

---

<sup>8</sup> (♠) Again more precisely: a Borel subset, i.e. an element of  $\mathcal{B}(\mathbb{R}^n)$ , the Borel- $\sigma$ -algebra over  $\mathbb{R}^n$ .

random vector (!)  $(X_1, \dots, X_n)^\top$  may be a discrete or continuous random vector. In the first case, we have a joint probability mass function

$$p_{X_1, \dots, X_n}(k_1, \dots, k_n) = \mathbf{P}[X_1 = k_1, \dots, X_n = k_n], \quad (k_1, \dots, k_n) \in \Omega_{X_1} \times \dots \times \Omega_{X_n}, \quad (1.29)$$

whereas in the second case, the law of  $(X_1, \dots, X_n)^\top$  is characterized by the (multivariate) probability density function  $f_{X_1, \dots, X_n}$  which fulfills

$$\mathbf{P}[(X_1, \dots, X_n) \in A] = \iint \dots \int_A f_{X_1, \dots, X_n}(x_1, \dots, x_n) d^n x, \quad A \subseteq \mathbb{R}^n \text{ (Borel subset)}. \quad (1.30)$$

A (multivariate) probability density function must fulfill

$$f_{X_1, \dots, X_n} \geq 0 \text{ and } \iint \dots \int_{\mathbb{R}^n} f_{X_1, \dots, X_n}(x_1, \dots, x_n) d^n x = 1.$$

Note that if  $X_1, \dots, X_n$  are all discrete, then  $(X_1, \dots, X_n)^\top$  is a discrete random vector, but it is *not* necessarily true that  $(X_1, \dots, X_n)^\top$  is a continuous random vector if  $X_1, \dots, X_n$  are continuous. To stress this, we say that  $X_1, \dots, X_n$  are *jointly continuous* if  $(X_1, \dots, X_n)^\top$  is a continuous random vector.

We also have a notion of a cumulative distribution function of a random vector.

**Definition 1.8.** For  $x_1, \dots, x_n \in \mathbb{R}$ , we set

$$\begin{aligned} F_{X_1, \dots, X_n}(x_1, \dots, x_n) &= \mathbf{P}_{(X_1, \dots, X_n)}[(-\infty, x_1] \times \dots \times (-\infty, x_n)] \\ &= \mathbf{P}[X_1 \leq x_1, \dots, X_n \leq x_n]. \end{aligned} \quad (1.31)$$

The function  $F_{X_1, \dots, X_n}$  is called the *joint cumulative distribution function* of  $X_1, \dots, X_n$  / the *cumulative distribution function of the law of  $(X_1, \dots, X_n)^\top$* .

Let us also recall how to infer from the joint probability density function or probability mass function of two random variables  $X$  and  $Y$  those of  $X$  alone.

*Remark 1.9.* Let  $X, Y$  be two real random variables.

(i) Suppose that  $X$  and  $Y$  are discrete. Then

$$p_X(k) = \mathbf{P}[X = k] = \sum_{\ell \in \Omega_Y} \mathbf{P}[X = k, Y = \ell] = \sum_{\ell \in \Omega_Y} p_{X,Y}(k, \ell), \quad k \in \Omega_X. \quad (1.32)$$

(ii) Suppose now that  $X$  and  $Y$  are jointly continuous. Then

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy, \quad x \in \mathbb{R}. \quad (1.33)$$

Some example calculations with joint densities can be found in [6, Sections 2.5–2.6]. We will now turn to the important concept of independence of random variables. Here is a short general reminder on conditional probability and independence:

**Remark 1.10.** Consider a probability space  $(\Omega, \mathcal{F}, \mathbf{P})$ . The *conditional probability* of  $A$  given  $B$  is defined as

$$\mathbf{P}[A|B] = \frac{\mathbf{P}[A \cap B]}{\mathbf{P}[B]}, \quad A, B \in \mathcal{F}, \mathbf{P}[B] > 0. \quad (1.34)$$

The quantity  $\mathbf{P}[A|B]$  gives the probability of  $A$  occurring if we *know* that  $B$  occurred. Recall that  $\mathbf{P}[\cdot|B]$  itself is a probability measure. The following two results are classical:

- The law of total probability: Suppose that  $A_1, \dots, A_n \in \mathcal{F}$  are pairwise disjoint ( $A_i \cap A_j = \emptyset$  for every  $i \neq j$ ),  $\bigcup_{i=1}^n A_i = \Omega$  and have all strictly positive probability. Then for every event  $B \in \mathcal{F}$ :

$$\mathbf{P}[B] = \sum_{i=1}^n \mathbf{P}[A_i] \mathbf{P}[B|A_i]. \quad (1.35)$$

- Bayes' theorem: Under the same conditions as before, and  $\mathbf{P}[B] > 0$ , one has for every  $1 \leq k \leq n$ :

$$\mathbf{P}[A_k|B] = \frac{\mathbf{P}[A_k] \mathbf{P}[B|A_k]}{\sum_{i=1}^n \mathbf{P}[A_i] \mathbf{P}[B|A_i]}. \quad (1.36)$$

We say that the events  $A$  and  $B$  are *independent* if

$$\mathbf{P}[A \cap B] = \mathbf{P}[A] \cdot \mathbf{P}[B]. \quad (1.37)$$

Note that if  $\mathbf{P}[A] > 0$  and  $\mathbf{P}[B] > 0$ , this is equivalent to  $\mathbf{P}[A|B] = \mathbf{P}[A]$  or  $\mathbf{P}[B|A] = \mathbf{P}[B]$ . In other words, the occurrence of  $B$  does not make  $A$  more likely (or vice versa). Events  $C_1, \dots, C_n \in \mathcal{F}$  are (*jointly*) *independent* if for every subset  $\{i_1, \dots, i_k\} \subseteq \{1, \dots, n\}$  with  $1 \leq k \leq n$  and pairwise distinct  $i_j$ , one has

$$\mathbf{P}\left[\bigcap_{j=1}^k C_{i_j}\right] = \prod_{j=1}^k \mathbf{P}[C_{i_j}]. \quad (1.38)$$

**Definition 1.11.** (i) The random variables  $X_1, \dots, X_n : \Omega \rightarrow \mathbb{R}$  are *independent*, if

$$\mathbf{P}[X_1 \in A_1, X_2 \in A_2, \dots, X_n \in A_n] = \prod_{i=1}^n \mathbf{P}[X_i \in A_i], \quad \text{for } A_1, \dots, A_n \subseteq \mathbb{R}. \quad (1.39)$$

- (ii) The random variables  $(X_i; i \in \mathcal{I})$ , where  $X_i : (\Omega, \mathcal{F}) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$  and  $\mathcal{I}$  is an arbitrary set, are *independent*, if for all finite sets  $\{i_1, \dots, i_n\} \subseteq \mathcal{I}$  (with  $i_1, \dots, i_n$  pairwise distinct):

$$\mathbf{P}[X_{i_1} \in A_1, X_{i_2} \in A_2, \dots, X_{i_n} \in A_n] = \prod_{j=1}^n \mathbf{P}[X_{i_j} \in A_j], \quad \text{for } A_1, \dots, A_n \subseteq \mathbb{R}. \quad (1.40)$$

We will often encounter sequences  $X_1, X_2, \dots$  of random variables that are independent and identically distributed (for instance corresponding to independent realizations of the same random experiment). We will abbreviate independent and identically distributed by “*i.i.d.*”.

We now give an effective criterion to check whether  $n$  random variables  $X_1, \dots, X_n$  are independent.

**Proposition 1.12.** *Let  $X_1, \dots, X_n : \Omega \rightarrow \mathbb{R}$  be random variables.*

(i)  $X_1, \dots, X_n$  are independent if and only if

$$F_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n F_{X_i}(x_i), \quad \text{for all } x_1, \dots, x_n \in \mathbb{R}. \quad (1.41)$$

(ii) Assume that  $X_1, \dots, X_n$  are discrete random variables. Then  $X_1, \dots, X_n$  are independent if and only if

$$\mathbf{P}[X_1 = x_1, \dots, X_n = x_n] = \prod_{i=1}^n \mathbf{P}[X_i = x_i], \quad \text{for all } x_i \in \Omega_{X_i}, 1 \leq i \leq n. \quad (1.42)$$

(iii) Assume that  $X_1, \dots, X_n$  are continuous random variables. Then  $X_1, \dots, X_n$  are independent if and only if

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i), \quad \text{for all } x_1, \dots, x_n \in \mathbb{R}. \quad (1.43)$$

---

### End of Lecture 3

A very important set-up later in this course will be to work with a collection of *independent, identically distributed* real random variables. We will abbreviate this as *i.i.d.*: Thus a sequence of real random variables  $(X_n)_{n \geq 1}$  is *i.i.d.* if it is independent and  $X_i \stackrel{d}{=} X_j$  for every  $i, j \in \mathbb{N}$ .

Let us briefly move back to general (dependent) random variables. To quantify how related two random variables  $X$  and  $Y$  are, we introduce the notions of covariance and correlation.

**Definition 1.13.** Let  $X$  and  $Y$  be two real random variables defined on some probability space fulfilling  $\mathbf{E}[X^2] < \infty$  and  $\mathbf{E}[Y^2] < \infty$ .

(i) The *covariance* of  $X$  and  $Y$  is defined as

$$\text{Cov}[X, Y] = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]. \quad (1.44)$$

(ii) The *correlation coefficient* of  $X$  and  $Y$  is defined as

$$\rho(X, Y) = \begin{cases} \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}}, & \text{Var}[X] \neq 0, \text{Var}[Y] \neq 0, \\ 0, & \text{else.} \end{cases} \quad (1.45)$$



Note that  $\text{Cov}[X, X] = \text{Var}[X]$ . Moreover, note that  $\rho(X, Y) \in [-1, 1]$ , and  $|\rho(X, Y)| = 1$  if and only if there is a linear relation between  $X$  and  $Y$ . Let us now summarize some properties of expectation and variance, in particular in the case of independent random variables.

**Lemma 1.14.** *Let  $X_1, \dots, X_n$  be real random variables (we assume below that all expectations exist).*

(i) *Let  $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ . One has*

$$\mathbf{E} \left[ \sum_{j=1}^n \lambda_j X_j \right] = \sum_{j=1}^n \lambda_j \mathbf{E}[X_j]. \quad (1.46)$$

(ii) *Suppose that  $X_1, \dots, X_n$  are independent. Then,*

$$\mathbf{E} \left[ \prod_{j=1}^n X_j \right] = \prod_{j=1}^n \mathbf{E}[X_j]. \quad (1.47)$$

*In particular one has  $\text{Cov}[X_i, X_j] = 0$  for every  $i \neq j$ .*

(iii) *Let  $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ . Then one has*

$$\text{Var} \left[ \sum_{j=1}^n \lambda_j X_j \right] = \sum_{j=1}^n \lambda_j^2 \text{Var}[X_j] + 2 \sum_{1 \leq j < k \leq n} \lambda_j \lambda_k \text{Cov}[X_j, X_k]. \quad (1.48)$$

*In particular, if  $X_1, \dots, X_n$  are independent*

$$\text{Var} \left[ \sum_{j=1}^n \lambda_j X_j \right] = \sum_{j=1}^n \lambda_j^2 \text{Var}[X_j]. \quad (1.49)$$

It is sometimes helpful to put the covariances of  $X_1, \dots, X_n$  in a matrix, this is called the *covariance matrix*:

$$\Sigma = \begin{pmatrix} \text{Var}[X_1] & \text{Cov}[X_1, X_2] & \cdots & \text{Cov}[X_1, X_n] \\ \text{Cov}[X_2, X_1] & \text{Var}[X_2] & \cdots & \text{Cov}[X_2, X_n] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[X_n, X_1] & \text{Cov}[X_n, X_2] & \cdots & \text{Var}[X_n] \end{pmatrix} \quad (1.50)$$

This matrix is symmetric and positive semidefinite (the latter means that for every vector  $v \in \mathbb{R}^n$  one has  $v^\top \Sigma v \geq 0$ ).<sup>9</sup> Finally we quote without proof:

**Lemma 1.15.** *Let  $v \in \mathbb{R}^n$ ,  $A \in \mathbb{R}^{n \times n}$  a matrix,  $X = (X_1, \dots, X_n)^\top$  a random vector with mean vector  $\mu = (\mathbf{E}[X_1], \dots, \mathbf{E}[X_n])^\top$  and covariance matrix  $\Sigma$ . One has*

$$\begin{aligned} \mathbf{E}[v^\top X] &= v^\top \mu, & \text{Var}[v^\top X] &= v^\top \Sigma v, \\ \mathbf{E}[AX] &= A\mu, & \text{Var}[AX] &= A\Sigma A^\top. \end{aligned} \quad (1.51)$$

<sup>9</sup>The covariance matrix is positive definite (meaning that it is positive semidefinite and  $v^\top \Sigma v = 0$  if and only if  $v = 0$ ) if the random vector  $(X_1, \dots, X_n)^\top$  is such  $v^\top X$  is *not* constant unless  $v = 0$ .

## 1.4 Limit theorems

In this short section, we recall two fundamental limit theorems from probability theory, the *(weak) law of large numbers* and the *central limit theorem*. Before we can do this, we need to first define two fundamental modes of convergence for random variables.

**Definition 1.16.** Let  $(Z_n)_{n \in \mathbb{N}}$  a sequence of random variables  $Z_n : \Omega \rightarrow \mathbb{R}$  and  $Z : \Omega \rightarrow \mathbb{R}$  another random variable.

(i) We say that  $(Z_n)_{n \in \mathbb{N}}$  *converges in probability* to  $Z$  if

$$\lim_{n \rightarrow \infty} \mathbf{P}[|Z_n - Z| > \varepsilon] = 0, \quad \text{for all } \varepsilon > 0. \quad (1.52)$$

We write this as

$$Z_n \xrightarrow[n \rightarrow \infty]{\mathbf{P}} Z. \quad (1.53)$$

(ii) We say that  $Z_n$  *converges in law / in distribution* if

$$F_{Z_n}(z) = \mathbf{P}[Z_n \leq z] \xrightarrow[n \rightarrow \infty]{} F_Z(z), \quad (1.54)$$

for all points of continuity  $z \in \mathbb{R}$  of  $F_Z$ . We write this as

$$Z_n \xrightarrow[n \rightarrow \infty]{d} Z. \quad (1.55)$$

Here are some important relations between convergence in probability and convergence in distribution:

**Theorem 1.17.** Let  $(X_n)_{n \in \mathbb{N}}$  and  $(Y_n)_{n \in \mathbb{N}}$  be sequences of real random variables,  $X, Y$  real random variables and  $c \in \mathbb{R}$  a constant.

(i) If  $(X_n)_{n \in \mathbb{N}}$  converges in probability to  $X$ , then it converges also to  $X$  in distribution:

$$X_n \xrightarrow[n \rightarrow \infty]{\mathbf{P}} X \quad \Rightarrow \quad X_n \xrightarrow[n \rightarrow \infty]{d} X. \quad (1.56)$$

(ii) We have that

$$X_n \xrightarrow[n \rightarrow \infty]{\mathbf{P}} X, Y_n \xrightarrow[n \rightarrow \infty]{\mathbf{P}} Y \quad \Rightarrow \quad X_n Y_n \xrightarrow[n \rightarrow \infty]{\mathbf{P}} XY \text{ and } X_n + Y_n \xrightarrow[n \rightarrow \infty]{\mathbf{P}} X + Y \quad (1.57)$$

(iii) We have that

$$X_n \xrightarrow[n \rightarrow \infty]{d} X, Y_n \xrightarrow[n \rightarrow \infty]{\mathbf{P}} 0 \quad \Rightarrow \quad X_n Y_n \xrightarrow[n \rightarrow \infty]{\mathbf{P}} 0. \quad (1.58)$$

(iv) One has Slutsky's theorem:

$$\begin{aligned} X_n \xrightarrow[n \rightarrow \infty]{d} X, Y_n \xrightarrow[n \rightarrow \infty]{\mathbf{P}} c \quad &\Rightarrow \quad X_n + Y_n \xrightarrow[n \rightarrow \infty]{d} X + c, \\ &X_n Y_n \xrightarrow[n \rightarrow \infty]{d} cX, \\ &\frac{X_n}{Y_n} \xrightarrow[n \rightarrow \infty]{d} \frac{X}{c}, \text{ if } c \neq 0. \end{aligned} \quad (1.59)$$

There are more identities of this sort, and we will see some of these moving forward. Next we state the *(weak) law of large numbers*:

**Theorem 1.18.** Let  $(X_n)_{n \in \mathbb{N}}$  a sequence of i.i.d. real random variables with  $\mu = \mathbf{E}[X_1]$ . Then

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow[n \rightarrow \infty]{\mathbf{P}} \mu. \quad (1.60)$$

The (weak) law of large numbers informally states that as  $n$  grows, the probability of observing any deviation of at least  $\varepsilon > 0$  between the arithmetic mean of i.i.d. random variables  $X_1, \dots, X_n$  and their expectation becomes small. It does *not* tell us how this arithmetic mean is (approximately) distributed. For this, we state the *central limit theorem*.

**Theorem 1.19.** Let  $X_1, X_2, \dots$  be a sequence of i.i.d. real random variables with  $\mu = \mathbf{E}[X_1]$  and  $\sigma^2 = \text{Var}[X_1] \in (0, \infty)$  (and  $\sigma > 0$ ). Then,

$$\sqrt{n} \frac{\overline{X}_n - \mu}{\sigma} \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, 1). \quad (1.61)$$

The notation in (1.61) means that

$$\sqrt{n} \frac{\overline{X}_n - \mu}{\sigma} \xrightarrow[n \rightarrow \infty]{d} Z, \text{ where } Z \sim \mathcal{N}(0, 1), \quad (1.62)$$

or in other words

$$\mathbf{P} \left[ \sqrt{n} \frac{\overline{X}_n - \mu}{\sigma} \leq x \right] \xrightarrow[n \rightarrow \infty]{} \Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt. \quad (1.63)$$

---

*End of Lecture 4*

Let us exemplify how the central limit theorem can be used.

*Example 1.20.* Suppose a coin is flipped  $n = 100$  times, and heads shows up 60 times. Is the coin fair?<sup>10</sup>

To determine this, let us assume the coin is fair. Then we can model this experiment with

$$(X_n)_{n=1}^{100} \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(\tfrac{1}{2}).$$

How likely is it that  $\sum_{j=1}^{100} X_j \geq 60$ ? We have  $\mu = \mathbf{E}[X_1] = \frac{1}{2}$  and  $\sigma^2 = \text{Var}[X_1] = \frac{1}{4}$ , and therefore (recall  $n = 100$ )

$$\begin{aligned} \mathbf{P} \left[ \sum_{j=1}^n X_j \geq 60 \right] &= \mathbf{P} \left[ \frac{\sum_{j=1}^n X_j - n \cdot \mathbf{E}[X_1]}{\sqrt{n \text{Var}[X_1]}} \geq \frac{60 - n \cdot \mathbf{E}[X_1]}{\sqrt{n \cdot \text{Var}[X_1]}} \right] \\ &= \mathbf{P} \left[ \frac{\sum_{j=1}^n X_j - n \cdot \mathbf{E}[X_1]}{\sqrt{n \text{Var}[X_1]}} \geq \frac{60 - 100 \cdot \frac{1}{2}}{\sqrt{100 \cdot \frac{1}{4}}} \right] \\ &\approx 1 - \Phi(2) = 0.02275. \end{aligned} \quad (1.64)$$

---

<sup>10</sup>As explained in the introduction, this set-up will be a standard motivational example for *statistical tests*.

The values of  $\Phi$  can be easily found in tables. Note that  $\sum_{j=1}^n X_j \sim \text{Bin}(100, \frac{1}{2})$ , so we could have calculated

$$\mathbf{P} \left[ \sum_{j=1}^n X_j \geq 60 \right] = \sum_{k=60}^{100} \binom{100}{k} \left( \frac{1}{2} \right)^{100} = 0.02844. \quad (1.65)$$

This however becomes much more complicated with increasing  $n$ . As we can see, the value obtained using the central limit theorem is not exact (it only becomes exact in the limit  $n \rightarrow \infty$ ). It nevertheless can be quite useful for getting a decent approximation. In both cases, we see that it is quite *unlikely* to observe an outcome at least as extreme as  $\sum_{j=1}^{100} X_j = 60$  if the coin is fair.

Note that while  $\sum_{j=1}^n X_j$  can only attain values in  $\{0, 1, \dots, 100\}$ , the quantity  $\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma}$  is approximated by  $Z \sim \mathcal{N}(0, 1)$  which takes values in  $\mathbb{R}$ . So we may alternatively to (1.66) try the following:

$$\begin{aligned} \mathbf{P} \left[ \sum_{j=1}^n X_j \geq 60 \right] &= \mathbf{P} \left[ \sum_{j=1}^n X_j > 60 - \frac{1}{2} \right] \\ &= \mathbf{P} \left[ \frac{\sum_{j=1}^n X_j - n \cdot \mathbf{E}[X_1]}{\sqrt{n \text{Var}[X_1]}} > \frac{60 - \frac{1}{2} - 100 \cdot \frac{1}{2}}{\sqrt{100 \cdot \frac{1}{4}}} \right] \\ &\approx 1 - \Phi(1.9) = 0.02872. \end{aligned} \quad (1.66)$$

This leads to a slightly better approximation. In general (using the exact same ideas), for  $Y \sim \text{Bin}(n, p)$ ,  $n \in \mathbb{N}$ ,  $p \in (0, 1)$ , one can approximate

$$\mathbf{P}[Y \leq k] \approx \Phi \left( \frac{k + \frac{1}{2} - np}{\sqrt{np(1-p)}} \right), \quad (1.67)$$

provided that both  $n$  and  $np(1-p)$  are not too small.

If we define in the set-up of the central limit theorem 1.19 the sum  $S_n = \sum_{j=1}^n X_j$ , then we can write its statement as

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \approx \mathcal{N}(0, 1) \quad \text{or} \quad S_n \approx \mathcal{N}(n\mu, n\sigma^2) \quad \text{or} \quad \bar{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \quad (1.68)$$

where  $X \approx \mathbf{Q}$  should be understood as “the law of  $X$  can be approximated by  $\mathbf{Q}$ ”.

The next statement is known as the *continuous mapping theorem*.

**Proposition 1.21.** Consider random variables  $(Z_n)_{n \in \mathbb{N}}$  and a random variable  $Z$ . Furthermore, let  $g : \mathbb{R} \rightarrow \mathbb{R}$  be a (measurable) function such that

$$\mathbf{P}[Z \in D_g] = 0, \quad D_g = \{x \in \mathbb{R}; g \text{ is not continuous in } x\}. \quad (1.69)$$

(i) Suppose that  $Z_n \xrightarrow[n \rightarrow \infty]{\mathbf{P}} Z$ . Then  $g(Z_n) \xrightarrow[n \rightarrow \infty]{\mathbf{P}} g(Z)$ .

(ii) Suppose that  $Z_n \xrightarrow[n \rightarrow \infty]{d} Z$ . Then  $g(Z_n) \xrightarrow[n \rightarrow \infty]{d} g(Z)$ .

Here is an example how to apply the continuous mapping theorem.

*Example 1.22.* Suppose that  $(X_n)_{n=1}^\infty \stackrel{\text{i.i.d.}}{\sim} \mathcal{E}(\lambda)$ ,  $\lambda > 0$ . Since  $\mathbf{E}[X_1] = \frac{1}{\lambda}$  we have therefore  $\lambda = \frac{1}{\mathbf{E}[X_1]}$ . By applying the continuous mapping theorem for the function  $g(x) = \frac{1}{x} \mathbb{1}_{\{x \neq 0\}}$  (which is continuous away from  $x = 0$ ), and since  $\bar{X}_n \xrightarrow[n \rightarrow \infty]{\mathbf{P}} \frac{1}{\lambda}$  by the law of large numbers, we find that  $\frac{1}{\bar{X}_n} \xrightarrow[n \rightarrow \infty]{\mathbf{P}} g(\frac{1}{\lambda}) = \lambda$ .

We finish this chapter with an important result on convergence in distribution, known as the  $\delta$ -method.

**Theorem 1.23.** Suppose that for a sequence  $(Y_n)_{n \in \mathbb{N}}$  of random variables, a random variable  $Y$  and  $c \in \mathbb{R}$ , one has

$$\sqrt{n}(Y_n - c) \xrightarrow[n \rightarrow \infty]{d} Y. \quad (1.70)$$

Let  $g : \mathbb{R} \rightarrow \mathbb{R}$  be differentiable and  $g'(c) \neq 0$ . Then

$$\sqrt{n}(g(Y_n) - g(c)) \xrightarrow[n \rightarrow \infty]{d} g'(c)Y. \quad (1.71)$$

Note that this applies in particular to the case where  $Y_n = \bar{X}_n$  for  $(X_n)_{n \in \mathbb{N}}$  i.i.d. with  $\mathbf{E}[X_1] = \mu (= c)$  and  $\text{Var}[X_1] = \sigma^2$  (i.e. the setting of the central limit theorem). Then we have

$$\sqrt{n} \frac{g(\bar{X}_n) - g(\mu)}{|g'(\mu)|\sigma} \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, 1). \quad (1.72)$$

*Proof of Theorem 1.23 (♠).* Define the function

$$h(x) = \begin{cases} \frac{g(x) - g(c)}{x - c} - g'(c), & \text{if } x \neq c, \\ 0, & \text{if } x = c. \end{cases} \quad (1.73)$$

This function is continuous by the definition of the derivative. Now

$$\sqrt{n}(Y_n - c) \xrightarrow[n \rightarrow \infty]{d} Y \quad \Rightarrow \quad Y_n \xrightarrow[n \rightarrow \infty]{\mathbf{P}} c, \quad (1.74)$$

using Theorem 1.17, (iii). By the continuous mapping theorem (Proposition 1.21), (i), we find that

$$\begin{aligned} h(Y_n) &\xrightarrow[n \rightarrow \infty]{\mathbf{P}} h(c), & \text{meaning that} \\ \frac{g(Y_n) - g(c) - g'(c)(Y_n - c)}{Y_n - c} &\xrightarrow[n \rightarrow \infty]{\mathbf{P}} 0. \end{aligned} \quad (1.75)$$

We can multiply the right-hand side by  $\sqrt{n}(Y_n - c)$ , and obtain that

$$\sqrt{n}(g(Y_n) - g(c)) - g'(c)\sqrt{n}(Y_n - c) \xrightarrow[n \rightarrow \infty]{\mathbf{P}} 0, \quad (1.76)$$

again by Theorem 1.17, (iii). Finally, Slutsky's theorem together with the assumption that  $\sqrt{n}(Y_n - c)$  converges in distribution to  $Y$  shows that in fact

$$\sqrt{n}(g(Y_n) - g(c)) \xrightarrow[n \rightarrow \infty]{d} g'(c)Y. \quad (1.77)$$

This finishes the proof. □

Some other topics from probability that we may occasionally use in this course include

- ▶ Conditional distributions and conditional expectation (see [6, Sections 2.8, 3.5]),
- ▶ Transformations of random variables and convolution (see [6, Sections 2.11–2.12]),
- ▶ Moment generating functions (see [6, Section 3.6]),
- ▶ Markov, Chebyshev, Jensen, Hölder, (...) inequalities (see [6, Chapter 4]).

These will also be recalled in the Problem sets or at later parts of these notes.

---

*End of Lecture 5*

## 2 Special distributions: Multivariate normal, $\chi^2$ -, $t$ - and $F$ -distributions

(Reference: [6, Section 2.10])

In this chapter we introduce some special distributions, which will be relevant for later parts of the course. This may be seen as a sort of *toolbox*, from which we use parts in particular for the theory of estimators and tests.

### 2.1 The multivariate normal distributions

**Definition 2.1.** Let  $\mu = (\mu_1, \dots, \mu_d) \in \mathbb{R}^d$  a vector and  $\Sigma \in \mathbb{R}^{d \times d}$  a symmetric positive definite matrix. Recall that this means that

$$\Sigma_{ij} = \Sigma_{ji}, \quad \text{for all } 1 \leq i, j \leq d \text{ (symmetry)} \quad (2.1)$$

and

$$\begin{aligned} v^\top \Sigma v &= \sum_{i,j=1}^d v_i \Sigma_{ij} v_j \geq 0, & \text{for all } v \in \mathbb{R}^d, \\ v^\top \Sigma v &= 0 & \Leftrightarrow & v = 0 \quad \text{(positive definiteness).} \end{aligned} \quad (2.2)$$

Consider the function  $f : \mathbb{R}^d \rightarrow [0, \infty)$  given by

$$f(x_1, \dots, x_d) = \frac{1}{(2\pi)^{d/2} \sqrt{\det(\Sigma)}} \exp \left( -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right), \quad (2.3)$$

where  $x = (x_1, \dots, x_d)^\top$ . This function is a density function, and if random variables  $X_1, \dots, X_d$  are jointly continuous with probability density function  $f$ , we say that the random vector  $(X_1, \dots, X_d)^\top$  has a *multivariate normal distribution* with mean vector  $\mu$  and covariance matrix  $\Sigma$ , and we denote it by  $(X_1, \dots, X_d)^\top \sim \mathcal{N}_d(\mu, \Sigma)$ .

The following lemma which we state without proof summarizes some properties of multivariate normal distributions.

**Lemma 2.2.** (i) Let  $X = (X_1, \dots, X_d)^\top : \Omega \rightarrow \mathbb{R}^d$  a continuous random vector with  $X \sim \mathcal{N}_d(\mu, \Sigma)$ . Then one has

$$\mathbf{E}[X_i] = \mu_i, \quad \text{Cov}[X_i, X_j] = \Sigma_{ij}, \quad 1 \leq i, j \leq d. \quad (2.4)$$

In particular,  $\text{Var}[X_i] = \Sigma_{ii}$  for every  $1 \leq i \leq d$ .

(ii) Let  $A \in \mathbb{R}^{k \times d}$  with  $k \leq d$  a matrix with rank  $k$  and  $b = (b_1, \dots, b_k)^\top \in \mathbb{R}^k$ , and  $X \sim \mathcal{N}_d(\mu, \Sigma)$ . Then

$$A \cdot X + b \sim \mathcal{N}_k(A\mu + b, A\Sigma A^\top). \quad (2.5)$$

**Remark 2.3.** (i) Note that (i) in the above theorem implies in particular that if  $X_1, \dots, X_k$  are jointly continuous random variables such that their joint law is a multivariate normal distribution (in particular they are then themselves normally distributed, see part (ii) below), then  $X_1, \dots, X_k$  are independent if and only if they are pairwise uncorrelated. Indeed, in the latter case  $\Sigma$  is diagonal, and so is  $\Sigma^{-1}$ , so the density  $f_{X_1, \dots, X_n}$  factorizes. We stress that we need here that  $X_1, \dots, X_n$  are *jointly* normally distributed.

(ii) A special case of part (ii) above is the following: Consider

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad \Sigma_{11} \in \mathbb{R}^{d_1 \times d_1}, \Sigma_{22} \in \mathbb{R}^{(d-d_1) \times (d-d_1)}. \quad (2.6)$$

Then the first  $d_1$  components of  $X$ , denoted  $X_{(1)} = (X_1, \dots, X_{d_1})^\top$  follows  $X_{(1)} \sim \mathcal{N}_{d_1}(\mu_{(1)}, \Sigma_{11})$ , where  $\mu_{(1)} = (\mu_1, \dots, \mu_{d_1})^\top$ .

---

*End of Lecture 6*

Before me move on, we state the multivariate versions of the central limit theorem and the  $\delta$ -method. Here, for a sequence of random vectors  $(Z^{(n)})_{n \in \mathbb{N}}$  and a random vector  $Z$  (both with values in  $\mathbb{R}^d$ ),  $Z^{(n)} \xrightarrow[n \rightarrow \infty]{d} Z$  means that

$$F_{Z^{(n)}}(z_1, \dots, z_d) \xrightarrow[n \rightarrow \infty]{} F_Z(z_1, \dots, z_d), \quad z_1, \dots, z_d \in \mathbb{R},$$

for all continuity points  $(z_1, \dots, z_d)^\top \in \mathbb{R}^d$  of  $F_Z$ .<sup>1</sup>

**Theorem 2.4.** (i) (*Multivariate central limit theorem*) Let  $(X^{(n)})_{n \in \mathbb{N}}$  be a sequence of i.i.d. random vectors with values in  $\mathbb{R}^d$ , such that  $X^{(1)}$  has mean vector  $\mu = (\mathbf{E}[X_1^{(1)}], \dots, \mathbf{E}[X_d^{(1)}])^\top$  and covariance matrix  $\Sigma$ . Then we have

$$\sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n X^{(i)} - \mu \right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}_d(0, \Sigma). \quad (2.7)$$

(ii) (*Multivariate  $\delta$ -method*) Suppose that for a sequence  $(Y_n)_{n \in \mathbb{N}}$  of random vectors with values in  $\mathbb{R}^d$  and

$$\sqrt{n}(Y_n - \mu) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}_d(0, \Sigma). \quad (2.8)$$

Let  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  be differentiable and  $\nabla g(\mu) \neq 0$ . Then

$$\sqrt{n}(g(Y_n) - g(\mu)) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \nabla g(\mu)^\top \Sigma \nabla g(\mu)). \quad (2.9)$$

---

<sup>1</sup>Here, continuity point  $z = (z_1, \dots, z_d)^\top$  of  $F_Z$  is a point for which  $\lim_{(y_1, \dots, y_d) \rightarrow (z_1, \dots, z_d)} F_Z(y_1, \dots, y_d) = F_Z(z_1, \dots, z_d)$ .



## 2.2 The $\chi^2$ -, $t$ - and $F$ -distributions

Let  $X_1, X_2, \dots, X_n$  be i.i.d. random variables with  $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ ,  $\mu \in \mathbb{R}$ ,  $\sigma > 0$ . Therefore, the random vector  $X = (X_1, \dots, X_n)^\top$  has a distribution  $X \sim \mathcal{N}_n(\boldsymbol{\mu}, \sigma^2 I_{n \times n})$ , where  $\boldsymbol{\mu} = (\mu, \dots, \mu)^\top$  and  $I_{n \times n}$  is the  $(n \times n)$  identity matrix. We consider the mean value and the empirical variance

$$\begin{aligned}\bar{X}_n &= \frac{1}{n} \sum_{i=1}^n X_i \\ S_n^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.\end{aligned}\tag{2.10}$$

Notice the crucial factor  $\frac{1}{n-1}$  in the above definition. We have that

$$\bar{X}_n = \frac{1}{n} (1 \quad \dots \quad 1) X \sim \mathcal{N} \left( \mu, \frac{\sigma^2}{n^2} (1 \quad \dots \quad 1) I_{n \times n} \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right) = \mathcal{N} \left( \mu, \frac{\sigma^2}{n} \right).\tag{2.11}$$

In particular, we have that  $\mathbf{E}[\bar{X}_n] = \mu$  and  $\text{Var}[\bar{X}_n] = \frac{\sigma^2}{n}$ . Moreover:

$$\begin{aligned}\mathbf{E}[S_n^2] &= \frac{n}{n-1} \mathbf{E} \left[ \frac{1}{n} \sum_{i=1}^n \left( (X_i - \mu)^2 + 2(X_i - \mu)(\mu - \bar{X}_n) + (\mu - \bar{X}_n)^2 \right) \right] \\ &= \frac{n}{n-1} \mathbf{E} \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - (\mu - \bar{X}_n)^2 \right] = \frac{n}{n-1} \left( \sigma^2 - \frac{\sigma^2}{n} \right) = \sigma^2.\end{aligned}\tag{2.12}$$

Note that  $S_n^2$  has the variance of  $X_1$  as its expectation (this is why the factor  $\frac{1}{n-1}$  is used instead of  $\frac{1}{n}$ ). Next, we try to calculate the laws of  $S_n^2$  and  $\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$ .

**Trick:** Consider the vector

$$P_n = \begin{pmatrix} \frac{1}{\sqrt{n}} \\ \vdots \\ \frac{1}{\sqrt{n}} \end{pmatrix}.$$

We can find an orthonormal basis  $(P_1, \dots, P_n)$  of  $\mathbb{R}^n$  containing  $P_n$ . Then consider the matrix  $\mathbf{P} = (P_1 \ P_2 \ \dots \ P_n)$ . We have

$$\mathbf{P} \cdot \mathbf{P}^\top = \mathbf{P}^\top \cdot \mathbf{P} = I_{n \times n}.\tag{2.13}$$

We see that

$$\mathbf{P}^\top (X - \boldsymbol{\mu}) \sim \mathcal{N}_n(0, \sigma^2 \mathbf{P}^\top \mathbf{P}) = \mathcal{N}_n(0, \sigma^2 I_{n \times n}).\tag{2.14}$$

This means that the real random variables  $P_i^\top (X - \boldsymbol{\mu})$ ,  $i = 1, \dots, n$ , are independent. Note that

$$\bar{X}_n - \mu = \frac{1}{\sqrt{n}} P_n^\top (X - \boldsymbol{\mu}),\tag{2.15}$$

and

$$\begin{aligned}
 S_n^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{n}{n-1} \left( \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - (\bar{X}_n - \mu)^2 \right) \\
 &= \frac{n}{n-1} \left( \frac{1}{n} (X - \mu)^\top (X - \mu) - (\bar{X}_n - \mu)^2 \right) \\
 &= \frac{n}{n-1} \left( \frac{1}{n} (X - \mu)^\top P P^\top (X - \mu) - \frac{1}{n} (P_n^\top (X - \mu))^2 \right) \\
 &= \frac{1}{n-1} \sum_{i=1}^{n-1} \left( P_i^\top (X - \mu) \right)^2.
 \end{aligned} \tag{2.16}$$

We therefore see

**Proposition 2.5.**  $\bar{X}_n$  and  $S_n^2$  are independent.

We come to the definition of the  $\chi^2$ -,  $t$ - and  $F$ -distributions, and we will see how these distributions relate to  $\bar{X}_n$  and  $S_n^2$ .

**Definition 2.6.** (i) Let  $Y \sim \mathcal{N}(0, 1)$ . Then the law of the random variable  $Z = Y^2$  is called  $\chi^2$ -distribution with one degree of freedom. We write  $Z \sim \chi_1^2$ .

(ii) Let  $Z_1, \dots, Z_n$  be i.i.d. real random variables with  $Z_1 \sim \chi_1^2$ . The law of the random variable  $Z = \sum_{i=1}^n Z_i$  is called  $\chi^2$ -distribution with  $n$  degrees of freedom. We write  $Z \sim \chi_n^2$ .

(iii) Let  $Y \sim \mathcal{N}(0, 1)$  and  $Z \sim \chi_n^2$  be stochastically independent. Then the law of

$$T = \frac{Y}{\sqrt{\frac{Z}{n}}} \tag{2.17}$$

is called  $t$ -distribution with  $n$  degrees of freedom. We write  $T \sim t_n$ .

(iv) Let  $X \sim \chi_m^2$  and  $Y \sim \chi_n^2$  be independent. Then the law of the random variable

$$F = \frac{\frac{1}{m} X}{\frac{1}{n} Y} \tag{2.18}$$

is called  $F$ -distribution with  $(m, n)$  degrees of freedom. We write  $F \sim F_{m,n}$ .

**Theorem 2.7.** Let  $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$  i.i.d. Then

(i)  $(n-1) \frac{S_n^2}{\sigma^2} \sim \chi_{n-1}^2$ , and

(ii)  $\frac{\bar{X}_n - \mu}{S_n / \sqrt{n}} \sim t_{n-1}$ .

*Proof.* We have

$$\underbrace{(n-1)\frac{S_n^2}{\sigma^2}}_{=:Z} = \sum_{i=1}^{n-1} \left( \frac{1}{\sigma} P_i^\top (X - \mu) \right)^2 \sim \chi_{n-1}^2. \quad (2.19)$$

Also, we have

$$\bar{X}_n - \mu \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right) \quad \Rightarrow \quad \underbrace{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}}_{=:Y} \sim \mathcal{N}(0, 1). \quad (2.20)$$

By Proposition 2.5,  $\bar{X}_n$  and  $S_n^2$  are independent, so  $Y$  and  $Z$  are independent. Now

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \bigg/ \sqrt{\frac{S_n^2}{\sigma^2}} = \frac{\bar{X}_n - \mu}{\frac{S_n}{\sqrt{n}}} \sim t_{n-1}. \quad (2.21)$$

□

Note that  $\sqrt{n} \frac{\bar{X}_n - \mu}{S_n}$  is the expression as  $\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma}$ , but with  $S_n$  replacing  $\sigma$ . We will see the relevance of this quantity when we discuss the  $t$ -test.

Let us state without proof the probability density functions of the  $\chi^2$ -,  $t$ - and  $F$ -distributions.

**Proposition 2.8.** (i) *The  $\chi^2$ -distribution with  $n$  degrees of freedom has the probability density function*

$$f_{\chi_n^2}(x) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} e^{-\frac{x}{2}} x^{\frac{n}{2}-1} \mathbb{1}_{(0,\infty)}(x). \quad (2.22)$$

(ii) *The  $t$ -distribution with  $n$  degrees of freedom has the probability density function*

$$f_{t_n}(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi} \Gamma(\frac{n}{2})} \left( 1 + \frac{x^2}{n} \right)^{-\frac{n+1}{2}}. \quad (2.23)$$

(iii) *The  $F$ -distribution with  $(m, n)$  degrees of freedom has the probability density function*

$$f_{F_{m,n}}(x) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \frac{m}{n} \left( \frac{m}{n} x \right)^{\frac{m}{2}-1} \left( 1 + \frac{m}{n} x \right)^{-\frac{m+n}{2}} \mathbb{1}_{(0,\infty)}(x). \quad (2.24)$$

---

*End of Lecture 7*

### 3 Introduction to inference: Estimators, statistics, sufficiency

(Reference: [6, Chapter 6])

Suppose we know that certain independent realizations  $X_1, X_2, \dots$  of an experiment follow a given distribution, but we do not know its parameter. An example would be that  $X_1, X_2, \dots$  are the life-times of some atoms of a radioactive isotope, which we know from physics to be  $\mathcal{E}(\beta)$ -distributed, but perhaps we have not found  $\beta$  yet. In other words

$$\text{Given } X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{E}(\beta), \text{ how do we infer } \beta? \quad (3.1)$$

The above problem is an example of a *parametric statistical model*. In most of the notes, we will consider such models, so let us formalize this.

**Definition 3.1.** A *parametric statistical model* consists of a family  $\mathbf{X} = (X_1, \dots, X_n)$  of random variables  $X_1, \dots, X_n$  defined on  $(\Omega, \mathcal{F}, \mathbf{P})$ ,  $\mathbf{X}$  having values in  $\mathcal{X}$ , where

$$\mathbf{P} \in \mathfrak{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}, \quad (3.2)$$

for some  $\Theta \subseteq \mathbb{R}^d$ .

*Remark 3.2.* (i) The goal will be to gain information on  $\theta$  using the data  $X_1, \dots, X_n$ .

(ii) Typically, we will assume that under  $\mathbf{P}_\theta$ ,  $X_1, \dots, X_n$  are i.i.d. random variables. These correspond to *measurements* of a given quantity.

We illustrate this by giving a simple example.

*Example 3.3.* Consider tossing a (potentially biased) coin 3 times. We denote the (unknown) probability of obtaining heads as an outcome in a single toss by  $\theta \in [0, 1]$ . Let

$$\Omega = \{0, 1\}^3, \quad \mathcal{F} = \mathcal{P}(\Omega), \quad \mathbf{P}_\theta[\{(\omega_1, \omega_2, \omega_3)\}] = \theta^{\sum_{j=1}^3 \omega_j} (1 - \theta)^{3 - \sum_{j=1}^3 \omega_j}. \quad (3.3)$$

Here 1 refers to heads and 0 to tails. The outcome of the  $k$ th toss is then given by

$$X_k : \{0, 1\}^3 \rightarrow \{0, 1\}, \quad X_k(\omega_1, \omega_2, \omega_3) = \omega_k, \quad k \in \{1, 2, 3\}. \quad (3.4)$$

Note that under  $\mathbf{P}_\theta$ , the random variables  $X_1, X_2, X_3$  are i.i.d. with  $X_1 \sim \text{Ber}(\theta)$ . In other words: Our statistical model consists of  $(X_1, X_2, X_3)$ , with  $\mathcal{X} = \{0, 1\}^3$ ,  $\Theta = [0, 1]$  and  $\mathbf{P}_\theta$  is such that  $X_1, X_2, X_3 \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(\theta)$  under  $\mathbf{P}_\theta$ .

We will usually *not* be interested in the exact structure of  $\mathbf{P}_\theta$ , but only in  $\theta$  or functions thereof. For instance, the above problem will simply be stated as

Consider  $X_1, \dots, X_n$  i.i.d. with  $X_1 \sim \text{Ber}(\theta)$  and unknown  $\theta \in \Theta = [0, 1]$ .

To remind ourselves that we consider statements for different possible choices of  $\theta$ , we will still retain the notation  $\mathbf{P}_\theta, \mathbf{E}_\theta, \text{Var}_\theta$  etc. for probabilities, expectation and variance. There are two main tasks to approach problems as above

- Data reduction: Do we need the entirety of  $\mathbf{X} = (X_1, \dots, X_n)$  to infer information on  $\theta$ , or does some function of  $\mathbf{X}$  (such as its average) suffice?
- Point estimation: How should we guess  $\theta$ ?

The first question will lead us to the consideration of a statistic, the second to point estimators. The goal of the next few chapters is to develop a theory of point estimators.

### 3.1 Estimators and their elementary properties

**Definition 3.4.** Consider a parametric statistical model consisting of data  $\mathbf{X} = (X_1, \dots, X_n) \in \mathcal{X}$  and a family of probability measures  $(\mathbf{P}_\theta)_{\theta \in \Theta}$ .

- (i) Let  $T : \mathcal{X} \rightarrow \mathcal{T}$  be a (measurable) function. Then

$$T \circ \mathbf{X} : \Omega \rightarrow \mathcal{T}, \quad \omega \mapsto T(X_1(\omega), \dots, X_n(\omega)) \quad (3.5)$$

is called a *statistic*.

- (ii) Consider a function  $f : \Theta \rightarrow \mathcal{T}$  that assigns each  $\theta \in \Theta$  a characteristic  $f(\theta)$ . Then a statistic  $T \circ \mathbf{X}$  is called an *estimator* for  $f(\theta)$  based on the  $n$  observations  $X_1, \dots, X_n$ . For a given realization  $\omega \in \Omega$ , the quantity  $T(\mathbf{X}(\omega)) = T(X_1(\omega), \dots, X_n(\omega))$  is called an *estimate* for  $f(\theta)$ .

Note that the definition of a statistic and an estimator are the same mathematically, but the difference comes with their interpretation: A statistic is a function of the data with the aim of reducing its complexity, whereas an estimator is constructed to give a “reasonable guess” for the value of some function of the parameter in the model under consideration.

*Example 3.5.* Consider  $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{E}(\theta)$ , where  $\theta \in (0, \infty)$  is unknown. An estimator for  $f(\theta) = \frac{1}{\theta}$  is given by the function

$$T(X_1, \dots, X_n) = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i. \quad (3.6)$$

Our motivation is that  $\mathbf{E}_\theta[X_1] = \frac{1}{\theta}$ , where  $\mathbf{E}_\theta$  denotes the expectation for the probability measure  $\mathbf{P}_\theta$ .

We will now study estimators more systematically, and introduce some of their key properties.

**Definition 3.6.** Suppose  $T(X_1, \dots, X_n)$  and  $S(X_1, \dots, X_n)$  are estimators for  $f(\theta)$ ,  $\theta \in \Theta$  and for  $n \in \mathbb{N}$ ,  $T_n(X_1, \dots, X_n)$  is an estimator for  $f(\theta)$ .

(i) The *bias* of  $T(X_1, \dots, X_n)$

$$\text{Bias}_\theta[T(X_1, \dots, X_n)] = \mathbf{E}_\theta[T(X_1, \dots, X_n)] - f(\theta). \quad (3.7)$$

The estimator  $T(X_1, \dots, X_n)$  for  $f(\theta)$  is called *unbiased* if

$$\mathbf{E}_\theta[T(X_1, \dots, X_n)] = f(\theta) \quad \Leftrightarrow \quad \text{Bias}_\theta[T(X_1, \dots, X_n)] = 0 \quad (3.8)$$

for all  $\theta \in \Theta$ .

(ii) The *mean square error* of  $T(X_1, \dots, X_n)$  is defined by

$$\text{MSE}_\theta[T(X_1, \dots, X_n)] = \mathbf{E}_\theta[(T(X_1, \dots, X_n) - f(\theta))^2]. \quad (3.9)$$

(iii) If  $T(X_1, \dots, X_n)$  and  $S(X_1, \dots, X_n)$  are both unbiased, we say that  $T(X_1, \dots, X_n)$  is *more efficient* than  $S(X_1, \dots, X_n)$  if for all  $\theta \in \Theta$ ,

$$\text{Var}_\theta[T(X_1, \dots, X_n)] \leq \text{Var}_\theta[S(X_1, \dots, X_n)]. \quad (3.10)$$

(iv) The sequence of estimators  $(T_n(X_1, \dots, X_n))_{n \in \mathbb{N}}$  is called *consistent*<sup>1</sup>, if

$$T_n(X_1, \dots, X_n) \xrightarrow[n \rightarrow \infty]{\mathbf{P}_\theta} f(\theta) \quad (3.11)$$

for all  $\theta \in \Theta$ .

---

*End of Lecture 8*

Very often, we will be slightly less diligent in the notation: In fact, most of the time we drop the dependence on  $X_1, \dots, X_n$  and also speak of “an estimator”  $T$  when we really mean a sequence of estimators  $T_n(X_1, \dots, X_n)$  as above.

Here are some examples of estimators including their properties.

*Example 3.7.* Suppose that  $X_1, X_2, \dots \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(p)$ , where  $p \in (0, 1)$  is unknown. Consider the following estimators for  $f(p) = p$ :

$$\begin{aligned} T_1 &= X_n, \\ T_2 &= \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n, \\ T_3 &= \sqrt{\frac{2}{n} \sum_{i=1}^{\frac{n}{2}} X_{2i-1} X_{2i}}, \quad (n \text{ even}), \\ T_4 &= \frac{1}{n+3} \sum_{i=1}^n X_i. \end{aligned} \quad (3.12)$$

---

<sup>1</sup>By abuse of notation, we will sometimes say that “the estimator  $T_n(X_1, \dots, X_n)$  is consistent”, even though we are really speaking about a sequence of estimators.

Note that  $T_1$  is unbiased since  $\mathbf{E}_p[T_1] = p$ , but not consistent. The estimator  $T_2$  is both unbiased (since  $\mathbf{E}_p[T_2] = \mathbf{E}_p[\bar{X}_n] = p$ ) and consistent (by the weak law of large numbers). The estimator  $T_3$  is consistent by the weak law of large numbers and the continuous mapping theorem: Indeed,  $Z_i = X_{2i-1} \cdot X_{2i}$  are independent and  $\mathbf{P}_p[Z_1 = 1] = \mathbf{P}_p[X_1 = 1, X_2 = 1] = p^2$ , so  $Z_1, \dots, Z_{n/2}$  are i.i.d.  $\text{Ber}(p^2)$  and therefore

$$\begin{aligned} & \frac{2}{n} \sum_{i=1}^{\frac{n}{2}} X_{2i-1} \cdot X_{2i} \xrightarrow[n \rightarrow \infty]{\mathbf{P}_p} \mathbf{E}_p[Z_1] = p^2 \text{ by Theorem 1.18,} \\ \Rightarrow & \sqrt{\frac{2}{n} \sum_{i=1}^{\frac{n}{2}} X_{2i-1} \cdot X_{2i}} \xrightarrow[n \rightarrow \infty]{\mathbf{P}_p} \sqrt{\mathbf{E}_p[Z_1]} = p \text{ by Proposition 1.21, (i).} \end{aligned}$$

However,  $T_3$  is not unbiased: Note that  $Z = \sum_{i=1}^{\frac{n}{2}} Z_i \sim \text{Bin}(\frac{n}{2}, p^2)$ , so

$$\mathbf{E}_p[T_3] = \frac{1}{\sqrt{\frac{n}{2}}} \mathbf{E}_p[\sqrt{Z}] < \frac{1}{\sqrt{\frac{n}{2}}} \sqrt{\mathbf{E}_p[Z]} = \frac{1}{\sqrt{\frac{n}{2}}} \sqrt{\frac{n}{2} p^2} = p,$$

using (♠) *Jensen's inequality* for the concave function  $\varphi(x) = \sqrt{x}$ . Considering  $T_4$ , this estimator is again consistent (using the law of large numbers and (1.57)), but not unbiased.

The following lemma explains why the mean value and empirical variance are natural estimators with desirable properties.

**Lemma 3.8.** *Let  $X_1, \dots, X_n$  be i.i.d. random variables with  $\mathbf{E}_\theta[X_1^2] < \infty$ .*

- (i)  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$  is an unbiased estimator for  $f(\theta) = \mathbf{E}_\theta[X_1]$ .
- (ii)  $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$  is an unbiased estimator for  $g(\theta) = \text{Var}_\theta[X_1]$ .
- (iii) Both  $\bar{X}_n$  and  $S_n^2$  are consistent.

*Proof.* We see immediately that

$$\mathbf{E}_\theta[\bar{X}_n] = \frac{1}{n} \sum_{i=1}^n \mathbf{E}_\theta[X_i] = \mathbf{E}_\theta[X_1], \quad (3.13)$$

which shows (i). The unbiasedness of  $S_n^2$  follows exactly as in (2.12). Finally, we know that  $\bar{X}_n$  is consistent by the weak law of large numbers, and  $S_n^2$  is consistent because  $\frac{n}{n-1} S_n^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X}_n)^2$  is consistent, and  $\frac{n}{n-1} \rightarrow 1$  as  $n \rightarrow \infty$ .  $\square$

The following is called the *bias-variance decomposition* of the MSE:

**Lemma 3.9.** *Let  $T(X_1, \dots, X_n)$  be an estimator for  $f(\theta)$ ,  $\theta \in \Theta$ . Then*

$$\text{MSE}_\theta[T(X_1, \dots, X_n)] = \text{Var}_\theta[T(X_1, \dots, X_n)] + \text{Bias}_\theta[T(X_1, \dots, X_n)]^2. \quad (3.14)$$

*Proof.* We abbreviate  $T(X_1, \dots, X_n)$  by  $T$ . Then

$$\begin{aligned}
 \text{MSE}_\theta[T] &= \mathbf{E}_\theta[(T - f(\theta))^2] \\
 &= \mathbf{E}_\theta[(T - \mathbf{E}_\theta[T] + \mathbf{E}_\theta[T] - f(\theta))^2] \\
 &= \mathbf{E}_\theta[(T - \mathbf{E}_\theta[T])^2] + 2 \underbrace{\mathbf{E}_\theta[T - \mathbf{E}_\theta[T]]}_{=\mathbf{E}_\theta[T] - \mathbf{E}_\theta[T]=0} (\mathbf{E}_\theta[T] - f(\theta)) + (\mathbf{E}_\theta[T] - f(\theta))^2 \quad (3.15) \\
 &= \text{Var}_\theta[T] + \text{Bias}_\theta[T]^2.
 \end{aligned}$$

□

There are estimators for which it may be reasonable to allow for a small bias, in order to obtain a smaller mean square error, and we will see such examples later when we study optimality of estimators. Finally, we give a useful criterion relating the mean square error to consistency.

**Proposition 3.10.** *If for a sequence of estimators  $T_n(X_1, \dots, X_n)$  for  $f(\theta)$  one has*

$$\begin{aligned}
 \text{Bias}_\theta[T_n(X_1, \dots, X_n)] &\rightarrow 0, \text{ and} \\
 \text{Var}_\theta[T_n(X_1, \dots, X_n)] &\rightarrow 0
 \end{aligned} \quad (3.16)$$

*as  $n \rightarrow \infty$ , then  $T_n(X_1, \dots, X_n)$  is consistent.*

*Proof.* We abbreviate  $T_n(X_1, \dots, X_n)$  by  $T_n$ . By Lemma 3.9, we see that the conditions imply that  $\text{MSE}_\theta[T_n] \rightarrow 0$ . By Markov's inequality, for any  $\varepsilon > 0$  one has

$$\mathbf{P}_\theta[|T_n - f(\theta)| > \varepsilon] \leq \frac{\mathbf{E}_\theta[(T_n - f(\theta))^2]}{\varepsilon^2} = \frac{\text{MSE}_\theta[T_n]}{\varepsilon^2} \rightarrow 0$$

as  $n \rightarrow \infty$ , which shows that  $T_n$  is consistent. □

---

*End of Lecture 9*

## 3.2 Sufficient statistics

**Definition 3.11.** Consider a parametric statistical model consisting of data  $\mathbf{X} = (X_1, \dots, X_n) \in \mathcal{X}$  and a family of probability measures  $(\mathbf{P}_\theta)_{\theta \in \Theta}$ . For  $T : \mathcal{X} \rightarrow \mathcal{T}$  let  $T(\mathbf{X})$  be a statistic. We say that the statistic  $T(\mathbf{X})$  is *sufficient for  $\theta$*  if the conditional distribution

$$(\mathbf{P}_\theta)_{\mathbf{X}|T(\mathbf{X})=t} \text{ is independent of } \theta \text{ for every } t \in \mathcal{T}. \quad (3.17)$$

Heuristically, a statistic  $T(\mathbf{X})$  is sufficient for  $\theta$  if it contains all information about  $\theta$  that we can obtain from  $\mathbf{X}$ . The procedure of obtaining  $T(\mathbf{X})$  from  $\mathbf{X}$  should be interpreted as *data reduction*: In particular, we often have a situation where  $\mathcal{X}$  is of a much higher dimension than  $\mathcal{T}$ .

For practical purposes, we will only consider the case where

- $\mathbf{X} = (X_1, \dots, X_n)$  is discrete, and therefore also  $T(\mathbf{X})$  is discrete, or



►  $\mathbf{X} = (X_1, \dots, X_n)$  is jointly continuous, and  $T(\mathbf{X})$  is continuous

and where both  $\mathcal{X}$  and  $\mathcal{T}$  are subsets of  $\mathbb{R}^d$  with appropriate  $d$ .

*Example 3.12.* Let  $\mathbf{X} = (X_1, \dots, X_n)$  with  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(p)$  and  $p$  is unknown. Here,  $\mathcal{X} = \{0, 1\}^n$ . Consider the statistic

$$T(X_1, \dots, X_n) = \sum_{i=1}^n X_i \quad (3.18)$$

with values in  $\mathcal{T} = \{0, \dots, n\}$ . Heuristically, we should be able to obtain the same information on  $p$  by knowing  $T(\mathbf{X})$  than we do from knowing  $\mathbf{X}$ . In other words,  $T(\mathbf{X})$  should be a sufficient statistic for  $p$ . Indeed, let  $(x_1, \dots, x_n)^\top \in \{0, 1\}^n$  and  $t \in \{0, \dots, n\}$ , then

$$\mathbf{P}_p \left[ X_1 = x_1, \dots, X_n = x_n \mid \sum_{i=1}^n X_i = t \right] = \begin{cases} 0, & \sum_{i=1}^n x_i \neq t, \\ \frac{1}{\binom{n}{t}}, & \sum_{i=1}^n x_i = t, \end{cases} \quad (3.19)$$

since  $T(\mathbf{X}) = \sum_{i=1}^n X_i \sim \text{Bin}(n, p)$ . In other words, the conditional distribution of  $\mathbf{X}$  given  $T(\mathbf{X}) = t$  does not depend on  $p$ , and therefore  $T(\mathbf{X})$  is indeed sufficient for  $p$ .

In general, it can be hard to verify sufficiency directly from the definition, since it requires to calculate the conditional distribution of  $\mathbf{X}$  under  $T(\mathbf{X})$ . There is a very helpful characterization of sufficiency, known as the *Neyman characterization theorem*:

**Theorem 3.13.** Consider a statistic  $T(\mathbf{X})$  in a parametric statistical model, where  $T : \mathcal{X} \rightarrow \mathcal{T}$ , and  $\mathcal{X} \subseteq \mathbb{R}^n$ ,  $\mathcal{T} \subseteq \mathbb{R}^k$ . Then  $T(\mathbf{X})$  is sufficient for  $\theta$  if and only if for every  $\theta \in \Theta$ , there exists a function  $g_\theta : \mathcal{T} \rightarrow \mathbb{R}$  and a function  $h : \mathcal{X} \rightarrow \mathbb{R}$  (independent of  $\theta$ ) such that

► (discrete case) the joint probability mass function of  $\mathbf{X}$  can be decomposed as

$$p_\theta(x_1, \dots, x_n) = g_\theta(T(x_1, \dots, x_n))h(x_1, \dots, x_n), \quad (x_1, \dots, x_n) \in \mathcal{X}, \quad (3.20)$$

or

► (continuous case) the joint probability density function of  $\mathbf{X}$  can be decomposed as

$$f_\theta(x_1, \dots, x_n) = g_\theta(T(x_1, \dots, x_n))h(x_1, \dots, x_n), \quad (x_1, \dots, x_n) \in \mathcal{X}. \quad (3.21)$$

*Proof* (♠). A proof can be found e.g. in [5, Section 8.8.1, Theorem A], we sketch it here for completeness: Suppose that we are in the discrete case and  $\mathcal{X} = \{x_1, x_2, \dots\}$  and  $\mathcal{T} = \{t_1, t_2, \dots\}$ .

( $\Rightarrow$ ) Let  $T(\mathbf{X})$  be sufficient for  $\theta$ . We define

$$g_\theta(t_i) = \mathbf{P}_\theta[T(\mathbf{X}) = t_i], \quad h(x) = \mathbf{P}_\theta[\mathbf{X} = x \mid T(\mathbf{X}) = T(x)] \quad (3.22)$$

for  $t_i \in \mathcal{T}$ ,  $x \in \mathcal{X}$ . By the definition of sufficiency,  $h$  does not depend on  $\theta$ . And by the definition of conditional probabilities, we have

$$\begin{aligned} \mathbf{P}_\theta[\mathbf{X} = x] &= \mathbf{P}_\theta[\mathbf{X} = x, T(\mathbf{X}) = T(x)] = \mathbf{P}_\theta[T(\mathbf{X}) = T(x)] \cdot \mathbf{P}_\theta[\mathbf{X} = x \mid T(\mathbf{X}) = T(x)] \\ &= g_\theta(T(x)) \cdot h(x). \end{aligned} \quad (3.23)$$

This is exactly (3.20).

( $\Leftarrow$ ) Now suppose that we have (3.20). We need to calculate  $\mathbf{P}_\theta[\mathbf{X} = x | T(\mathbf{X}) = t_i]$ . If  $\mathbf{P}_\theta[T(\mathbf{X}) = t_i] > 0$ , we can write it as

$$\mathbf{P}_\theta[T(\mathbf{X}) = t_i] = \sum_{x \in \mathcal{X} : T(x) = t_i} \mathbf{P}[\mathbf{X} = x] \stackrel{(3.20)}{=} g_\theta(t_i) \sum_{x \in \mathcal{X} : T(x) = t_i} h(x). \quad (3.24)$$

It follows that

$$\begin{aligned} \mathbf{P}_\theta[\mathbf{X} = x | T(\mathbf{X}) = t_i] &= \frac{\mathbf{P}_\theta[\mathbf{X} = x, T(\mathbf{X}) = t_i]}{\mathbf{P}_\theta[T(\mathbf{X}) = t_i]} \\ &= \begin{cases} 0, & T(\mathbf{X}) \neq t_i, \\ \frac{\mathbf{P}_\theta[\mathbf{X} = x]}{\mathbf{P}_\theta[T(\mathbf{X}) = t_i]} = \frac{g_\theta(t_i)h(x)}{g_\theta(t_i) \sum_{y \in \mathcal{X} : T(y) = t_i} h(y)}, & T(\mathbf{X}) = t_i, \end{cases} \quad (3.25) \\ &= \begin{cases} 0, & T(\mathbf{X}) \neq t_i, \\ \frac{h(x)}{\sum_{y \in \mathcal{X} : T(y) = t_i} h(y)}, & T(\mathbf{X}) = t_i. \end{cases} \end{aligned}$$

We see that this expression does not depend on  $\theta$ , and therefore  $T(\mathbf{X})$  is indeed consistent.  $\square$

Here is an example how to apply the Neyman characterization theorem:

*Example 3.14.* (i) Consider  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{E}(\theta)$ , with unknown  $\theta > 0$ . The joint probability density function of  $X_1, \dots, X_n$  is given by

$$f_\theta(x_1, \dots, x_n) = \prod_{i=1}^n \left( \theta e^{-\theta x_i} \mathbb{1}_{(0, \infty)}(x_i) \right) = \theta^{\sum_{i=1}^n x_i} e^{-\theta \sum_{i=1}^n x_i} \prod_{i=1}^n \mathbb{1}_{(0, \infty)}(x_i).$$

We set  $T(\mathbf{X}) = \sum_{i=1}^n X_i$ , and we see that

$$\begin{aligned} f_\theta(x_1, \dots, x_n) &= g_\theta(T(x_1, \dots, x_n))h(x_1, \dots, x_n) \\ g_\theta(t) &= \theta^t e^{-\theta t}, \\ h(x_1, \dots, x_n) &= \prod_{i=1}^n \mathbb{1}_{(0, \infty)}(x_i). \end{aligned}$$

Therefore,  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$ .

(ii) Consider  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{U}([0, \theta])$ , with unknown  $\theta > 0$ . The joint probability density function of  $X_1, \dots, X_n$  is given by

$$f_\theta(x_1, \dots, x_n) = \prod_{i=1}^n \left( \frac{1}{\theta} \mathbb{1}_{[0, \theta]}(x_i) \right).$$

Note that we *cannot* extract a simple quantity as  $\sum_{i=1}^n x_i$  from this expression. However, one can write

$$\begin{aligned} f_{\theta}(x_1, \dots, x_n) &= \frac{1}{\theta^n} \mathbb{1}_{\{0 \leq \min\{x_1, \dots, x_n\} \leq \max\{x_1, \dots, x_n\} \leq \theta\}} = g_{\theta}(T(x_1, \dots, x_n))h(x_1, \dots, x_n), \\ g_{\theta}(t) &= \frac{1}{\theta^n} \mathbb{1}_{\{t \leq \theta\}}, \\ h(x_1, \dots, x_n) &= \mathbb{1}_{\{\min\{x_1, \dots, x_n\} \geq 0\}}, \end{aligned}$$

where  $T(X_1, \dots, X_n) = \max\{X_1, \dots, X_n\}$ . Thus,  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$ . Notice however that this statistic behaves much differently from (say) the sum.

We will later use sufficient statistics to construct certain optimal estimators.

## 4 Construction principles for estimators

(Reference: [6, Sections 9.2–9.3])

In this chapter, we start with a systematic approach to point estimation. We first present two general construction principles for estimators, namely estimators based on the *method of moments* and *maximum-likelihood* estimators. After that, we introduce a certain class of optimal estimators (the uniformly minimal variance unbiased or UMVU estimators) and present theoretical methods to obtain them.

### 4.1 The method of moments

Consider a parametric model with i.i.d. data  $X_1, \dots, X_n$  under  $(\mathbf{P}_\theta)_{\theta \in \Theta}$ , where  $\theta$  is unknown. In many cases, one has access to the *moments*  $\mathbf{E}_\theta[X_1]$ ,  $\mathbf{E}_\theta[X_1^2]$ , ...,  $\mathbf{E}_\theta[X_1^k]$ , but wants to estimate  $\theta$  itself. To do this, we can use the continuous mapping theorem. Let us illustrate this by an example.

*Example 4.1.* Let  $X_1, X_2, \dots$  be i.i.d. random variables with  $X_1 \sim \mathcal{E}(\theta)$ ,  $\theta \in (0, \infty)$  unknown. Since  $\mathbf{E}_\theta[X_1] = \frac{1}{\theta}$  and therefore  $\theta = \frac{1}{\mathbf{E}_\theta[X_1]}$ , a reasonable estimator for  $\theta$  is given by

$$\hat{\theta}_n = \frac{1}{\bar{X}_n} = \frac{n}{\sum_{i=1}^n X_i}. \quad (4.1)$$

This is the exact same situation as Example 1.22, and we see that  $\hat{\theta}_n$  is a consistent estimator for  $\theta$ .

---

*End of Lecture 10*

We will now show a general construction principle for estimators known as the *method of moments*

**Lemma 4.2.** Consider a parametric statistical model where  $X_1, \dots, X_n$  are i.i.d. and  $\theta \in \Theta \subseteq \mathbb{R}^d$  is unknown. Suppose we can write

$$\theta = g(\mathbf{E}_\theta[X_1], \mathbf{E}_\theta[X_1^2], \dots, \mathbf{E}_\theta[X_1^k]), \quad (4.2)$$

for a (measurable) function  $g : \mathbb{R}^k \rightarrow \mathbb{R}^d$  which is continuous in the point

$$(\mathbf{E}_\theta[X_1], \mathbf{E}_\theta[X_1^2], \dots, \mathbf{E}_\theta[X_1^k])^\top,$$

then a consistent estimator (called *method of moments estimator*) for  $\theta$  is given by

$$\hat{\theta}_n = g\left(\frac{1}{n} \sum_{i=1}^n X_i, \frac{1}{n} \sum_{i=1}^n X_i^2, \dots, \frac{1}{n} \sum_{i=1}^n X_i^k\right). \quad (4.3)$$

*Proof.* For simplicity, suppose that  $g : \mathbb{R} \rightarrow \mathbb{R}$  (i.e.  $\hat{\theta}_n = g(\bar{X}_n)$ ). By the weak law of large numbers,  $\bar{X}_n$  converges in probability to  $\mathbf{E}_\theta[X_1]$ . Since  $g$  is continuous in  $\mathbf{E}_\theta[X_1]$ , we can apply the continuous mapping theorem (Proposition 1.21, (i)) and infer that

$$\hat{\theta}_n = g(\bar{X}_n) \xrightarrow[n \rightarrow \infty]{\mathbf{P}_\theta} g(\mathbf{E}_\theta[X_1]) = \theta. \quad (4.4)$$

The case where  $k \geq 2$  follows similarly, but with a slight modification of the continuous mapping theorem (namely if  $g : \mathbb{R}^k \rightarrow \mathbb{R}$  is continuous in  $z \in \mathbb{R}^k$  and  $(Y_1^{(n)}, \dots, Y_k^{(n)})^\top \xrightarrow[n \rightarrow \infty]{\mathbf{P}} z$ , then also  $g(Y_1, \dots, Y_k) \xrightarrow[n \rightarrow \infty]{\mathbf{P}} g(z)$ ).  $\square$

Let us stress again the important fact that estimators based on the method of moments are consistent (but not necessarily unbiased!). Here is another example for the method of moments:

*Example 4.3.* Let  $X_1, X_2, \dots$  be i.i.d. random variables with  $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ ,  $\mu \in \mathbb{R}$  and  $\sigma > 0$ . Suppose we want to estimate both  $\mu$  and  $\sigma^2$ , or in a more compact notation:  $\theta = (\mu, \sigma^2)^\top \in \Theta = \mathbb{R} \times (0, \infty)$ . Note that  $\mu = \mathbf{E}_\theta[X_1]$  and  $\sigma^2 = \text{Var}_\theta[X_1] = \mathbf{E}_\theta[X_1^2] - \mathbf{E}_\theta[X_1]^2$ . In other words, we have

$$\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} \mathbf{E}_\theta[X_1] \\ \mathbf{E}_\theta[X_1^2] - \mathbf{E}_\theta[X_1]^2 \end{pmatrix} = g(\mathbf{E}_\theta[X_1], \mathbf{E}_\theta[X_1^2]), \quad (4.5)$$

where  $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  is the continuous function

$$g(x, y) = \begin{pmatrix} x \\ y - x^2 \end{pmatrix}. \quad (4.6)$$

Then, an estimator for  $(\mu, \sigma^2)^\top$  based on the method of moments is given by

$$\begin{aligned} \begin{pmatrix} \hat{\mu}_n \\ \hat{\sigma}_n^2 \end{pmatrix} &= g\left(\frac{1}{n} \sum_{i=1}^n X_i, \frac{1}{n} \sum_{i=1}^n X_i^2\right) = \left(\frac{1}{n} \sum_{i=1}^n X_i, \frac{1}{n} \sum_{i=1}^n X_i^2 - \left(\frac{1}{n} \sum_{i=1}^n X_i\right)^2\right) \\ &= \begin{pmatrix} \bar{X}_n \\ \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \end{pmatrix}. \end{aligned} \quad (4.7)$$

Note that  $\hat{\sigma}_n^2 = \frac{n-1}{n} S_n^2$ , where  $S_n^2$  is the (unbiased and consistent) sample variance. In particular,  $\hat{\sigma}_n^2$  is *not* unbiased.

## 4.2 Maximum-Likelihood estimators

We now introduce a powerful general estimation principle: the Maximum-Likelihood method:

Consider a parametric statistical model with  $X_1, \dots, X_n$  real random variables under  $(\mathbf{P}_\theta)_{\theta \in \Theta}$  with  $\Theta \subseteq \mathbb{R}^k$  such that  $X_1, \dots, X_n$  have joint probability density function  $f_\theta^{(n)}(x)$  (or joint probability mass function  $p_\theta^{(n)}(x)$ ) under  $\mathbf{P}_\theta$ , depending on a parameter  $\theta \in \Theta$ .

<sup>1</sup>Convergence in probability of a sequence of random vector is defined by all of the components converging in probability.

As before, we want to find an estimator  $\hat{\theta}_n$  (depending on  $X_1, \dots, X_n$ ) for  $\theta$ . We define the *Maximum-likelihood estimator (MLE)*

$$\begin{aligned}\hat{\theta}_n &= \operatorname{argsup}_{\theta \in \Theta} f_{\theta}^{(n)}(X_1, \dots, X_n) \\ &= \operatorname{argsup}_{\theta \in \Theta} \prod_{i=1}^n f_{\theta}(X_i) \quad (\text{if } X_i \text{ are i.i.d.})\end{aligned}\tag{4.8}$$

Importantly,  $f_{\theta}(x)$  can be the probability density function of a continuous distribution or the probability mass function  $f_{\theta}(x) = p_{\theta}(x)$  of a discrete distribution. All following results are valid in both cases.

*Example 4.4.* (i)  $X_1, \dots, X_n$  i.i.d.,  $X_1 \sim \text{Ber}(p)$  ( $\theta = p$ ). We have

$$\begin{aligned}p_{\theta}^{(n)}(x_1, \dots, x_n) &= \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i} \\ \Rightarrow \frac{\partial}{\partial \theta} p_{\theta}^{(n)}(x_1, \dots, x_n) &= \left( \sum_{i=1}^n x_i \right) \theta^{\sum_{i=1}^n x_i - 1} (1 - \theta)^{n - \sum_{i=1}^n x_i} \\ &\quad - \theta^{\sum_{i=1}^n x_i} \left( n - \sum_{i=1}^n x_i \right) (1 - \theta)^{n - \sum_{i=1}^n x_i - 1} \stackrel{!}{=} 0 \\ \Leftrightarrow \theta &= \frac{\sum_{i=1}^n x_i}{n},\end{aligned}\tag{4.9}$$

so  $\hat{\theta}_n = \bar{X}_n$  is the MLE.

(ii)  $X_1, \dots, X_n$  i.i.d.,  $X_1 \sim \mathcal{E}(\lambda)$  ( $\theta = \lambda$ ). We have

$$\begin{aligned}f_{\theta}^{(n)}(x_1, \dots, x_n) &= \prod_{i=1}^n \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_{i=1}^n x_i} \\ \Rightarrow \frac{\partial}{\partial \theta} f_{\theta}^{(n)}(x_1, \dots, x_n) &= n\theta^{n-1} e^{-\theta \sum_{i=1}^n x_i} - \theta^n \left( \sum_{i=1}^n x_i \right) e^{-\theta \sum_{i=1}^n x_i} \stackrel{!}{=} 0 \\ \Leftrightarrow n - \theta \sum_{i=1}^n x_i &= 0 \quad \Leftrightarrow \quad \theta = \frac{1}{\frac{1}{n} \sum_{i=1}^n x_i},\end{aligned}\tag{4.10}$$

so  $\hat{\theta}_n = \hat{\lambda}_n = \frac{1}{\bar{X}_n}$  is the MLE.

---

*End of Lecture 11*

Since log is monotonically increasing,  $\hat{\theta}_n$  also maximizes  $\log f_{\theta}^{(n)}(X_1, \dots, X_n)$ , or minimizes

$$L_n(\theta) = -\frac{1}{n} \log f_{\theta}^{(n)}(X_1, \dots, X_n) \left( = -\frac{1}{n} \sum_{i=1}^n \log f_{\theta}(X_i) \text{ if } X_i \text{ are i.i.d.} \right).\tag{4.11}$$

$L_n(\theta)$  is called the *log-Likelihood function*. Note that in the case of i.i.d. random variables,  $L_n(\theta)$  (under  $\mathbf{P}_{\theta_0}$  for some fixed  $\theta_0 \in \Theta$ ) converges in probability to

$$L(\theta) = -\mathbf{E}_{\theta_0} [\log f_{\theta}(X_1)].\tag{4.12}$$

Here  $\theta_0$  should be understood as the “true” parameter of the model. We state the following general fact:

**Proposition 4.5.**  $\theta_0$  is the unique minimum of the function  $L(\theta)$ .

*Proof.* We can write

$$L(\theta) = -\mathbf{E}_{\theta_0} [\log f_{\theta_0}(X_1)] + \mathbf{E}_{\theta_0} \left[ \log \left( \frac{f_{\theta_0}(X_1)}{f_{\theta}(X_1)} \right) \right] \quad (4.13)$$

Using that  $\log(y) \leq y - 1$  for all  $y > 0$ , we have that  $\log(\frac{1}{y}) \geq 1 - y$  for all  $y > 0$  and therefore (in the continuous case, the discrete case is analogous)

$$\mathbf{E}_{\theta_0} \left[ \log \left( \frac{f_{\theta_0}(X_1)}{f_{\theta}(X_1)} \right) \right] = \int \log \left( \frac{f_{\theta_0}(x)}{f_{\theta}(x)} \right) f_{\theta_0}(x) dx \geq \int \left( 1 - \frac{f_{\theta}(x)}{f_{\theta_0}(x)} \right) f_{\theta_0}(x) dx = 0. \quad (4.14)$$

In other words:  $L(\theta) = -\mathbf{E}_{\theta_0} [\log f_{\theta}(X_1)] \geq -\mathbf{E}_{\theta_0} [\log f_{\theta_0}(X_1)] = L(\theta_0)$ . We omit the proof that  $\theta_0$  is unique.<sup>2</sup>  $\square$

We will later study asymptotic properties of the MLE in more detail. For now, let us also observe the following simple but important property of the MLE.

**Proposition 4.6.** Suppose that for  $T : \mathcal{X} \rightarrow \mathcal{T}$  with  $\mathcal{T} \subseteq \mathbb{R}^\ell$ ,  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$ . If a unique MLE for  $\theta$  exists, it must factorize over  $T(\mathbf{X})$ . In other words, there exists a function  $r : \mathcal{T} \rightarrow \Theta$  with

$$\hat{\theta}_n = r(T(\mathbf{X})). \quad (4.15)$$

*Proof.* By the Neyman characterization of sufficiency (Theorem 3.13), we have functions  $g_\theta$  and  $h$  with

$$f_{\theta}^{(n)}(X_1, \dots, X_n) = g_{\theta}(T(X_1, \dots, X_n))h(X_1, \dots, X_n).$$

From this representation it is immediately clear, that the maximum in  $\theta \in \Theta$  only depends on  $T(X_1, \dots, X_n)$ .  $\square$

There are examples where the MLE does not exist. Here is one example due to Kiefer and Wolfowitz (1956):

*Example 4.7.* Consider i.i.d. random variables  $X_1, \dots, X_n$ , with the distribution of  $X_1$  being a mixture of two normals: each observation is either  $\mathcal{N}(\mu, 1)$ -distributed or  $\mathcal{N}(\mu, \sigma^2)$ -distributed with probability  $\frac{1}{2}$  each. The unknown parameter is  $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$ , and  $X_1$  has density

$$f_{\theta}(x) = \frac{1}{2}\varphi(x - \mu) + \frac{1}{2\sigma}\varphi((x - \mu)/\sigma),$$

<sup>2</sup>In fact, we need a mild technical condition on the model: Namely, that equality in (4.14) only holds when  $f_{\theta_0}(x) = f_{\theta}(x)$  for every  $x$ . In particular, different  $\theta_0$  must correspond to different distributions ( $\rightsquigarrow$  *identifiability*, see, e.g., [6, Section 9.5]).

where  $\varphi(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$  is the density of a standard normal distribution. The MLE maximizes

$$f_{\theta}^{(n)}(X_1, \dots, X_n) = \prod_{i=1}^n \left( \frac{1}{2} \varphi(X_i - \mu) + \frac{1}{2\sigma} \varphi((X_i - \mu)/\sigma) \right).$$

over  $\mu$  and  $\sigma^2$ . If we take  $\mu = X_1$ , the right-hand side above gives

$$\frac{1}{\sqrt{2\pi}} \left( \frac{1}{2} + \frac{1}{2\sigma} \right) \prod_{i=2}^n \left( \frac{1}{2} \varphi(X_i - X_1) + \frac{1}{2\sigma} \varphi((X_i - X_1)/\sigma) \right).$$

Note that for all  $z \neq 0$ , we have  $\lim_{\sigma \downarrow 0} \frac{1}{\sigma} \varphi(z/\sigma) = 0$ , but also

$$\lim_{\sigma \downarrow 0} \prod_{j=2}^n \left( \frac{1}{2} \varphi(X_j - X_1) + \frac{1}{2\sigma} \varphi((X_j - X_1)/\sigma) \right) = \prod_{j=2}^n \left( \frac{1}{2} \varphi(X_j - X_1) \right) > 0.$$

In other words, by taking  $\mu = X_1$  and  $\sigma \downarrow 0$ , we can make  $f_{\theta}^{(n)}(X_1, \dots, X_n)$  arbitrarily large. Notice that  $(X_1, 0) \notin \Theta$ , so in this example the MLE does not exist.

### 4.3 UMVU estimators

In the previous two sections we saw general methods to construct estimators. Natural questions arise from here:

- How can we judge the quality of an estimator?
- Is there a “best” estimator?

We already defined in the previous section the desirable concept of *unbiasedness* and considered the *mean square error* as a metric to determine whether an estimator is usually close to the quantity it is supposed to estimate. In this section we shall see that if we restrict to unbiased estimators, there is a way to identify an *optimal* estimator.

**Definition 4.8.** Consider a parametric statistical model with data  $\mathbf{X} = (X_1, \dots, X_n)$  and a family of probability measures  $(\mathbf{P}_{\theta})_{\theta \in \Theta}$ . Denote by  $K$  the set of estimators<sup>3</sup> for some real quantity  $f(\theta)$  and suppose that

$$\emptyset \neq \tilde{K} \subseteq K. \quad (4.16)$$

We say that an estimator  $\hat{\gamma}^*$  ( $\equiv \hat{\gamma}^*(X_1, \dots, X_n)$ ) is the *uniformly best estimator over  $\tilde{K}$*  for  $f(\theta)$ , if  $\hat{\gamma}^* \in \tilde{K}$  and

$$\text{MSE}_{\theta}[\hat{\gamma}^*] = \inf_{\hat{\gamma} \in \tilde{K}} \text{MSE}_{\theta}[\hat{\gamma}] \quad \text{for all } \theta \in \Theta. \quad (4.17)$$

If  $\tilde{K} = K$ , we say that  $\hat{\gamma}^*$  is the *uniformly best estimator*.

<sup>3</sup>Recall that these are simply *all* functions of the form  $\hat{\gamma} \circ \mathbf{X}$  with  $\hat{\gamma} : \mathcal{X} \rightarrow \mathbb{R}$



In many cases, a uniformly best estimator (i.e. one that minimizes  $\text{MSE}_\theta[\hat{\gamma}]$  over all  $\hat{\gamma} \in K$ ) does not exist. If there is one, it may turn out to be biased.<sup>4</sup> By restricting to a subset  $\tilde{K}$ , one can in some instances guarantee the existence of an optimal estimator. Clearly, a particularly important subset are unbiased estimators.

**Definition 4.9.** Consider in the set-up of Definition 4.8 the subset of unbiased estimators

$$\tilde{K} = K_{\text{unbiased}} = \{\hat{\gamma} \in K : \mathbf{E}_\theta[\hat{\gamma}] = f(\theta)\}. \quad (4.18)$$

If it exists, the uniformly best estimator  $\hat{\gamma}^*$  for  $f(\theta)$  over  $K_{\text{unbiased}}$  is called *uniformly minimal variance unbiased estimator* (or *UMVU estimator* or *UMVUE*).

The terminology comes from the fact that if  $\text{Bias}_\theta[\hat{\gamma}] = 0$ , then  $\text{MSE}_\theta[\hat{\gamma}] = \text{Var}_\theta[\hat{\gamma}]$  by the bias-variance decomposition (Lemma 3.9), and so minimizing the mean squared error over  $K_{\text{unbiased}}$  simply corresponds to minimizing the variance. In the terminology of Definition 3.6, (iii), we can also say

$$\hat{\gamma}^* \text{ is the UMVUE} \Leftrightarrow \hat{\gamma}^* \text{ is more efficient than every unbiased estimator } T \text{ for all } \theta \in \Theta. \quad (4.19)$$

---

*End of Lecture 12*

Our goal is to construct the UMVUE, and we need some preparations for that. The following result tells us how to decrease the mean square error of an estimator, given a sufficient statistic. It is known as the *Rao-Blackwell theorem*.

**Theorem 4.10.** Let  $\mathbf{X} = (X_1, \dots, X_n)$ ,  $T(\mathbf{X})$  with  $T : \mathcal{X} \rightarrow \mathcal{T}$  a sufficient statistic for  $\theta \in \Theta$ , and  $\hat{\gamma} \equiv \hat{\gamma}(\mathbf{X})$  an estimator for  $f(\theta)$ . For the estimator<sup>5</sup>

$$\hat{\gamma}^*(\mathbf{X}) = \mathbf{E}_\theta[\hat{\gamma}(\mathbf{X})|T(\mathbf{X})], \quad (4.20)$$

we have that

$$\text{MSE}_\theta[\hat{\gamma}^*(\mathbf{X})] \leq \text{MSE}_\theta[\hat{\gamma}(\mathbf{X})]. \quad (4.21)$$

If in addition,  $\text{Var}_\theta[\hat{\gamma}(\mathbf{X})] < \infty$ , then the inequality is strict, except when ( $\mathbf{P}_\theta$ -almost surely)  $\hat{\gamma}^*(\mathbf{X}) = \hat{\gamma}(\mathbf{X})$ .

*Proof.* See [5, Section 8.8.2, Theorem A]. □

Note that it is important that  $T(\mathbf{X})$  is sufficient, so that the right-hand side of (4.20) does not depend on  $\theta$ . Heuristically the Rao-Blackwell theorem states that an estimator can be improved (in the sense of the mean squared error) if it depends on “more information” than is contained in a sufficient statistic. Let us exemplify this:

---

<sup>4</sup>Note the terminology “uniformly best estimator” is non-standard, but used here for simplicity.

<sup>5</sup>Being more precise, we should note that conditional expectations are only defined ( $\spadesuit$ )  $\mathbf{P}_\theta$ -almost surely: This means we identify random variables  $Z$  and  $Z'$  if there is  $\Omega_0 \in \mathcal{F}$  with  $\mathbf{P}_\theta[\Omega_0] = 1$  and  $Z(\omega) = Z'(\omega)$  for all  $\omega \in \Omega_0$ . We will most of the time be sloppy and ignore this subtlety.

*Example 4.11.* Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(p)$  with unknown  $p$  (the standard coin-flip example). Suppose that  $n \geq 2$ . We saw in Example 3.12 that  $T(X_1, \dots, X_n) = \sum_{i=1}^n X_i$  is a sufficient statistic for  $p$ . Consider the estimator

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

This is an unbiased estimator with  $\text{MSE}_\theta[\hat{\gamma}] = p(1-p)$ . To improve it, we will look at the estimator

$$\hat{\gamma}^*(X_1, \dots, X_n) = \mathbf{E}_p[\hat{\gamma}(X_1, \dots, X_n) | T(X_1, \dots, X_n)] = \mathbf{E}_p\left[X_1 \middle| \sum_{i=1}^n X_i\right].$$

To calculate the expression on the right-hand side, note that  $T(X_1, \dots, X_n) = \sum_{i=1}^n X_i \sim \text{Bin}(n, p)$  can take values in  $\{0, 1, \dots, n\}$ . For fixed  $t \in \{0, 1, \dots, n\}$ , we see that

$$\begin{aligned} \mathbf{P}_p[X_1 = x | T(X_1, \dots, X_n) = t] &= \frac{\mathbf{P}_p[X_1 = x, \sum_{j=1}^n X_j = t]}{\mathbf{P}_p[\sum_{j=1}^n X_j = t]} \\ &= \frac{\mathbf{P}_p[\sum_{j=2}^n X_j = t-x] \cdot \mathbf{P}_p[X_1 = x]}{\mathbf{P}_p[\sum_{j=1}^n X_j = t]} \\ &= \frac{\binom{n-1}{t-x} p^{t-x} (1-p)^{n-1-t+x} p^x (1-p)^{1-x}}{\binom{n}{t} p^t (1-p)^{n-t}} = \frac{\binom{n-1}{t-x}}{\binom{n}{t}}, \end{aligned}$$

where  $x \in \{0, 1\}$ . From here we obtain that

$$\begin{aligned} \mathbf{E}_p[X_1 | T(X_1, \dots, X_n) = t] &= \mathbf{P}_p[X_1 = 1 | T(X_1, \dots, X_n) = t] = \frac{\binom{n-1}{t-1}}{\binom{n}{t}} \\ &= \frac{(n-1)!}{(t-1)!(n-t)!} \cdot \frac{t!(n-t)!}{n!} = \frac{t}{n}. \end{aligned}$$

This finally means (by replacing  $t$  by  $T(X_1, \dots, X_n)$ ) that

$$\hat{\gamma}_n^*(X_1, \dots, X_n) = \mathbf{E}_p\left[X_1 \middle| \sum_{i=1}^n X_i\right] = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}_n.$$

This estimator has mean square error equal to  $\frac{1}{n}p(1-p) < p(1-p) = \text{MSE}_\theta[\hat{\gamma}]$ .

In the example above,  $\hat{\gamma}(\mathbf{X})$  was unbiased, and so was  $\hat{\gamma}^*(\mathbf{X})$ . This is true in general:

**Lemma 4.12.** *Suppose the estimator  $\hat{\gamma}(\mathbf{X})$  in the statement of the Rao-Blackwell theorem (Theorem 4.10) is unbiased. Then also  $\hat{\gamma}^*(\mathbf{X})$  is unbiased.*

*Proof.* This follows from the fact that

$$\mathbf{E}_\theta[\hat{\gamma}^*(\mathbf{X})] = \mathbf{E}_\theta[\mathbf{E}_\theta[\hat{\gamma}(\mathbf{X}) | T(\mathbf{X})]] = \mathbf{E}_\theta[\hat{\gamma}(\mathbf{X})] = f(\theta), \quad (4.22)$$

using that (♠) for any random variables  $U, V$  in any probability space  $(\Omega, \mathcal{F}, \mathbf{P})$ , one has  $\mathbf{E}[\mathbf{E}[U|V]] = \mathbf{E}[U]$  (provided all expectations exist).  $\square$

To recap, we see that once an unbiased estimator for  $f(\theta)$  and a sufficient statistic for  $\theta$  is found, one can construct a more efficient (unbiased) estimator using the Rao-Blackwell theorem. Notice however that this does *not* guarantee that the so obtained estimator is optimal (i.e. that it is the UMVU estimator). To answer this question, we need another concept.

**Definition 4.13.** Consider a parametric statistical model consisting of data  $\mathbf{X} = (X_1, \dots, X_n) \in \mathcal{X}$  and a family of probability measures  $(\mathbf{P}_\theta)_{\theta \in \Theta}$ . For  $T : \mathcal{X} \rightarrow \mathcal{T}$  let  $T(\mathbf{X})$  be a statistic. We say that the statistic  $T(\mathbf{X})$  is *complete* for  $\theta$  if for every  $g : \mathcal{T} \rightarrow \mathbb{R}$ ,

$$\mathbf{E}_\theta[g(T(\mathbf{X}))] = 0 \text{ for all } \theta \in \Theta \quad \Rightarrow \quad \mathbf{P}_\theta[g(T(\mathbf{X})) = 0] = 1 \text{ for all } \theta \in \Theta. \quad (4.23)$$

---

*End of Lecture 13*

Here is an example:

*Example 4.14.* Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Pois}(\theta)$  with unknown  $\theta > 0$ . We claim that

$$T(\mathbf{X}) = \sum_{i=1}^n X_i \text{ is a complete statistic for } \theta.$$

Note that  $T(\mathbf{X}) \sim \text{Pois}(n\theta)$ . Let us assume that for some function  $g : \mathbb{N}_0 \rightarrow \mathbb{R}$ , one has  $\mathbf{E}_\theta[g(T(\mathbf{X}))] = 0$  for all  $\theta > 0$ . This means that

$$\mathbf{E}_\theta[g(T(\mathbf{X}))] = e^{-n\theta} \sum_{k=0}^{\infty} \frac{(n\theta)^k}{k!} g(k) = 0, \quad \text{for all } \theta > 0.$$

Note that  $e^{-n\theta} \neq 0$ , so we have a power series  $\sum_{k=0}^{\infty} \frac{n^k g(k)}{k!} \theta^k$  which is identically 0. This can only be true, if all coefficients vanish, and therefore

$$\frac{n^k}{k!} g(k) = 0 \text{ for all } k \in \mathbb{N}_0 \quad \Rightarrow \quad g(k) = 0 \text{ for all } k \in \mathbb{N}_0.$$

We are finally able to state the main result of this section, which allows us to construct UMVU estimators. This result is known as the *Lehmann-Scheffé theorem*.

**Theorem 4.15.** Let  $\mathbf{X} = (X_1, \dots, X_n)$ ,  $T(\mathbf{X})$  with  $T : \mathcal{X} \rightarrow \mathcal{T}$  a sufficient and complete statistic for  $\theta \in \Theta$ , and  $\hat{\gamma} \equiv \hat{\gamma}(\mathbf{X})$  an unbiased estimator for  $f(\theta)$ . Then the estimator

$$\hat{\gamma}^*(\mathbf{X}) = \mathbf{E}_\theta[\hat{\gamma}(\mathbf{X}) | T(\mathbf{X})], \quad (4.24)$$

is an UMVU estimator for  $f(\theta)$ . If in addition,  $\text{Var}_\theta[\hat{\gamma}(\mathbf{X})] < \infty$ , then the UMVU estimator is unique.

*Proof.* The unbiasedness of  $\hat{\gamma}^*(\mathbf{X})$  follows from Lemma 4.12. Set  $\hat{\gamma}_1(\mathbf{X}) = \hat{\gamma}(\mathbf{X})$  and assume that  $\hat{\gamma}_2(\mathbf{X})$  is another unbiased estimator for  $f(\theta)$ . We write

$$\hat{\gamma}_j^*(\mathbf{X}) = \mathbf{E}_\theta[\hat{\gamma}_j(\mathbf{X}) | T(\mathbf{X})] =: h_j(T(\mathbf{X})), \quad j \in \{1, 2\}. \quad (4.25)$$

From this we see immediately that (since  $\hat{\gamma}_j^*(\mathbf{X})$  is unbiased)

$$\mathbf{E}_\theta[(h_1 - h_2)(T(\mathbf{X}))] = 0, \quad \forall \theta \in \Theta. \quad (4.26)$$

But since  $T(\mathbf{X})$  is complete, we see that  $\mathbf{P}_\theta[h_1(T(\mathbf{X})) = h_2(T(\mathbf{X}))] = 1$ , which means that  $\hat{\gamma}_1^*(\mathbf{X})$  and  $\hat{\gamma}_2^*(\mathbf{X})$  coincide ( $\mathbf{P}_\theta$ -almost surely) and therefore

$$(\star) \quad \text{Var}_\theta[\hat{\gamma}^*(\mathbf{X})] = \text{Var}_\theta[\hat{\gamma}_2^*(\mathbf{X})] \stackrel{\text{Theorem 4.10}}{\leq} \text{Var}_\theta[\hat{\gamma}_2(\mathbf{X})] \quad \forall \theta \in \Theta,$$

where we used the Rao-Blackwell theorem in the last step (the MSE of  $\hat{\gamma}_2^*(\mathbf{X})$  is smaller or equal to the MSE of  $\hat{\gamma}_2(\mathbf{X})$ ). Therefore  $\hat{\gamma}(\mathbf{X})$  has a variance smaller or equal to the variance of *any* unbiased estimator for  $f(\theta)$ .<sup>6</sup> The uniqueness of the UMVU estimator in the case of finite variance follows from the second part of the Rao-Blackwell theorem. Indeed, if another unbiased estimator  $\hat{\gamma}_2(\mathbf{X})$  also minimizes the MSE among all unbiased estimators, then in  $(\star)$ , we must have an equality, but by the uniqueness statement in the Rao-Blackwell theorem, this can only be true if  $\hat{\gamma}^*(\mathbf{X}) = \hat{\gamma}_2^*(\mathbf{X}) = \hat{\gamma}_2(\mathbf{X})$  ( $\mathbf{P}_\theta$ -almost surely).  $\square$

Let us provide two examples for UMVU estimators.

*Example 4.16.* (i) Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(p)$  with unknown  $p$ . We have seen that the statistic  $T(X_1, \dots, X_n) = \sum_{i=1}^n X_i$  is a sufficient statistic (see Example 3.12), and one can show similarly to Example 4.14 that it is also complete. Moreover, the estimator  $\hat{\gamma}(X_1, \dots, X_n) = X_1$  is an unbiased estimator for  $p$ . By the Lehmann-Scheffé theorem (Theorem 4.15), we see that

$$\hat{\gamma}_n^*(X_1, \dots, X_n) = \mathbf{E}_p[X_1 | T(X_1, \dots, X_n)] \stackrel{\text{Example 4.11}}{=} \bar{X}_n \quad (4.27)$$

is *the* UMVU estimator for  $p$  (since its variance is finite, it is unique). In other words: There is no unbiased estimator for  $p$  with a variance less than  $\bar{X}_n$ .

(ii) The following example should serve as a caveat to using UMVU estimators: Suppose that the arrival of customers in a bank within 10 minutes can be modelled by  $X \sim \text{Pois}(\lambda)$ , with unknown  $\lambda > 0$ . We observe  $X$  and try to find an estimator for the quantity

$$f(\lambda) = \mathbf{P}_\lambda[\text{“There are no customers in the following 20 minutes”}] = e^{-2\lambda},$$

based on the single observation  $X$ . Note that it is immediately clear from the definition that  $T(X) = X$  is a sufficient and complete statistic for  $\lambda$ . Moreover, for the estimator

$$\hat{\gamma}(X) = (-1)^X \quad \text{we have} \quad \mathbf{E}_\lambda[\hat{\gamma}(X)] = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} (-1)^k = e^{-2\lambda}.$$

In other words,  $\hat{\gamma}(X)$  is unbiased. By the Lehmann-Scheffé theorem, we have that  $\mathbf{E}_\lambda[\hat{\gamma}(X) | X] = \hat{\gamma}(X) = (-1)^X$  is *already the UMVU estimator* (its variance is finite, so it is unique). However, this estimator  $(-1)^X$  **does not make any sense!** Its values are limited to  $-1$  and  $+1$ , but nevertheless it is the best unbiased estimator for  $f(\lambda)$ . Here it becomes quite clear that restricting to unbiased estimators can be problematic.

<sup>6</sup>Note that the proof shows even more, namely that  $\hat{\gamma}^*(\mathbf{X})$  does not depend on the choice of  $\hat{\gamma}$  apart from sets of  $\mathbf{P}_\theta$ -probability 0, if we condition on the same complete sufficient statistic.

Note that the method we presented requires to calculate a conditional expectation, which can be difficult. If we want to *show* that an estimator is the UMVU estimator, this may be avoided:

**Remark 4.17.** If a  $T(\mathbf{X})$  is a sufficient and complete statistic for  $\theta \in \Theta$  and  $\hat{\gamma} = g(T(\mathbf{X}))$  is an unbiased estimator for  $f(\theta)$  depending on  $T(\mathbf{X})$ , then it is already the UMVU estimator. This follows from the Lehmann-Scheffé theorem and the fact that for any random variables  $Y, Z$  one has  $\mathbf{E}[\mathbf{E}[Z|Y]|Y] = \mathbf{E}[Z|Y]$ .

## 4.4 Exponential Families

As we have seen, it can be quite technical to find a complete sufficient statistic. In this short section we give a certain class of parametric distributions, for which these properties can be easily checked.

**Definition 4.18.** A family  $\mathcal{Q} = \{\mathbf{Q}_\theta; \theta \in \Theta\}$  of probability measures is called an *k-parameter exponential family*, if every  $\mathbf{Q}_\theta$  has a probability mass function or probability density function of the form

$$f_\theta(x) = C(\theta) \exp \left( \sum_{j=1}^k \eta_j(\theta) T_j(x) \right) h(x), \quad (4.28)$$

where  $\eta_j(\theta)$  and  $C(\theta)$  are continuous, real-valued functions.

**Example 4.19.** Most of the standard distributions form exponential families, for instance:

(i) The family  $\{Pois(\lambda), \lambda > 0\}$  is a 1-parameter exponential family, since

$$p_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda} = \underbrace{e^{-\lambda}}_{C(\lambda)} \exp \left( \underbrace{\log(\lambda)}_{\eta_1(\lambda)} \cdot \underbrace{k}_{T_1(k)} \right) \cdot \underbrace{\frac{1}{k!}}_{h(k)}.$$

(ii) The family  $\{\mathcal{N}(\mu, \sigma^2), \mu > 0, \sigma^2 > 0\}$  is a 2-parameter exponential family, since

$$\begin{aligned} f_{(\mu, \sigma^2)}(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2\sigma^2} (x - \mu)^2 \right) \\ &= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{\mu^2}{2\sigma^2} \right)}_{C(\mu, \sigma^2)} \exp \left( \underbrace{-\frac{1}{2\sigma^2}}_{\eta_1(\mu, \sigma^2)} \cdot \underbrace{x^2}_{T_1(x)} + \underbrace{\frac{\mu}{\sigma^2}}_{\eta_2(\mu, \sigma^2)} \cdot \underbrace{x}_{T_2(x)} \right), \end{aligned}$$

and  $h(x) = 1$ .

Note that the uniform distributions  $\mathcal{U}([a, b])$ , for  $a < b$ , do *not* form an exponential family.

**Lemma 4.20.** (i) If the distributions of  $X$  under  $\mathbf{P}_\theta$  form an exponential family as in (4.28), then  $S(X) = (T_1(X), \dots, T_k(X))$  is a sufficient statistic for  $\theta$ .

(ii) Consider  $\mathbf{X} = (X_1, \dots, X_n)$ . If  $X_1, \dots, X_n$  are i.i.d. and the distributions of  $X_1$  under  $\mathbf{P}_\theta$  form an exponential family, then

$$S(\mathbf{X}) = (\bar{T}_1(\mathbf{X}), \dots, \bar{T}_k(\mathbf{X})), \quad \bar{T}_j(\mathbf{x}) = \sum_{i=1}^n T_j(x_i)$$

is a sufficient statistic for  $\theta$ .

(iii) Consider the set

$$\mathcal{C} = \{(\eta_1(\theta), \dots, \eta_k(\theta))^\top : \theta \in \Theta\} \subseteq \mathbb{R}^k$$

with the  $\eta_j$  as in (4.28). Suppose that  $\mathcal{C}$  is truly  $k$ -dimensional (meaning that it contains an open ball of  $\mathbb{R}^k$ ), then  $S(X) = (T_1(X), \dots, T_k(X))$  is complete for  $\theta$ . Similarly in the case of i.i.d. data  $\mathbf{X} = (X_1, \dots, X_n)$ ,  $S(\mathbf{X}) = (\bar{T}_1(\mathbf{X}), \dots, \bar{T}_k(\mathbf{X}))$  from (ii) is complete for  $\theta$ .

*Proof.* Claim (i) follows immediately from the Neyman-characterization. For claim (ii), note that if  $X_1, \dots, X_n$  are i.i.d., then the family of joint distributions of  $(X_1, \dots, X_n)$  under  $\mathbf{P}_\theta$  is itself also a  $k$ -parameter exponential family. Indeed:

$$\begin{aligned} f_\theta^{(n)}(x_1, \dots, x_n) &= C(\theta)^n \exp \left( \sum_{i=1}^n \sum_{j=1}^k \eta_j(\theta) T_j(x_i) \right) \underbrace{\prod_{i=1}^n h(x_i)}_{=h(\mathbf{x})} \\ &= C(\theta)^n \exp \left( \sum_{j=1}^k \eta_j(\theta) \bar{T}_j(\mathbf{x}) \right) h(\mathbf{x}), \end{aligned} \quad (4.29)$$

the claim then follows from (i). We omit the (harder) proof of (iii).  $\square$

---

#### End of Lecture 14

The lemma above is particularly effective in finding statistics which are both sufficient and complete.

*Example 4.21.* Consider  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$  where  $(\mu, \sigma^2)^\top \in \mathbb{R} \rightarrow (0, \infty)$  is unknown. We have seen in Example 4.19, (ii) above that the normal distributions form an exponential family with  $T_1(x) = x^2$  and  $T_2(x) = x$ . Moreover, we have  $\mathcal{C} = \{(-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2})^\top : \mu \in \mathbb{R}, \sigma^2\} \subseteq \mathbb{R}^2$ , which is truly 2-dimensional. Therefore, we see that

$$S(\mathbf{X}) = \left( \sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i \right) \quad (4.30)$$

is a complete sufficient statistic for  $\theta$ . Recall that  $\bar{X}_n$  and  $S_n^2$  are unbiased estimators for  $\mu$  and  $\sigma^2$ . Since conditioning on  $S(\mathbf{X})$  does not change the values of  $\bar{X}_n$  or  $S_n^2$ , we conclude by the Lehmann-Scheffé theorem that

$$\text{For i.i.d. } \mathcal{N}(\mu, \sigma^2) \text{ data, } \bar{X}_n \text{ and } S_n^2 \text{ are UMVU estimators for } \mu \text{ and } \sigma^2, \text{ respectively.} \quad (4.31)$$

## 5 Asymptotic properties of the Maximum-Likelihood estimators

In this chapter, we study asymptotic properties of the MLE in more detail. In particular, we establish the consistency of the MLE (recall that estimators based on the method of moments are always consistent) and study its limiting distribution, in a sufficiently general setup.

### 5.1 Consistency of the MLE

Suppose that  $X_1, \dots, X_n$  are i.i.d. data in a parametric statistical model with a family  $(\mathbf{P}_\theta)_{\theta \in \Theta}$ . In what follows, we will denote by  $\theta_0 \in \Theta$  a fixed parameter (the “true” parameter of the model). Recall the notation from (4.11) and (4.11). As we saw earlier, the (random) log-Likelihood function  $L_n(\theta)$  converges (in  $\mathbf{P}_{\theta_0}$ -probability) to the deterministic function  $L(\theta)$ , and we also saw that  $\theta_0$  is the unique minimum of  $L(\theta)$  (see Proposition 4.5). Can we show the convergence of

$$\hat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} L_n(\theta),$$

towards

$$\theta_0 = \operatorname{argmin}_{\theta \in \Theta} L(\theta)?$$

The following establishes exactly that and gives the consistency of  $\hat{\theta}_n$ .

**Proposition 5.1.** *Let  $L_n(\theta)$  be a sequence of random variables depending on  $\theta$ , and  $L(\theta)$  a deterministic function with*

- (i)  *$L$  is continuous and  $\theta_0 = \operatorname{argmin}_{\theta \in \Theta} L(\theta)$  is unique,*
- (ii)  *$\Theta$  is bounded and closed, and  $\theta_0 \in \overset{\circ}{\Theta}$ ,*
- (iii)  *$\sup_{\theta \in \Theta} |L_n(\theta) - L(\theta)| \xrightarrow[n \rightarrow \infty]{\mathbf{P}_{\theta_0}} 0$ .*

*Then, for  $\hat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} L_n(\theta)$ , one has*

$$\hat{\theta}_n \xrightarrow[n \rightarrow \infty]{\mathbf{P}_{\theta_0}} \theta_0. \quad (5.1)$$

---

*End of Lecture 15*

When are the conditions of Proposition 5.1 fulfilled? It turns out that exponential families are helpful for this as well:

**Theorem 5.2.** *Let  $X_1, \dots, X_n$  be i.i.d. real random variables under  $\mathbf{P}_\theta$ . Furthermore, suppose that  $\mathcal{Q} = \{(\mathbf{P}_\theta)_{X_1}; \theta \in \Theta\}$  is an exponential family. If assumptions (i) and (ii) of Proposition 5.1 are fulfilled, then also assumption (iii) is fulfilled, and the Maximum-Likelihood estimator for  $\theta$  is consistent.*

*Proof.* Note that

$$\log f_\theta(x) = \log C(\theta) + \sum_{j=1}^k \eta_j(\theta) T_j(x) + \log h(x). \quad (5.2)$$

We see that

$$\begin{aligned} & \sup_{\theta \in \Theta} |L_n(\theta) - L(\theta)| \\ &= \sup_{\theta \in \Theta} \left| \sum_{j=1}^k \eta_j(\theta) \left( \frac{1}{n} \sum_{i=1}^n T_j(X_i) - \mathbf{E}_{\theta_0}[T_j(X_1)] \right) + \left( \frac{1}{n} \sum_{i=1}^n \log h(X_i) - \mathbf{E}_{\theta_0}[\log h(X_1)] \right) \right| \\ &\leq \sum_{j=1}^k \underbrace{\sup_{\theta \in \Theta} |\eta_j(\theta)|}_{\leq K < \infty} \left| \frac{1}{n} \sum_{i=1}^n T_j(X_i) - \mathbf{E}_{\theta_0}[T_j(X_1)] \right| + \left| \frac{1}{n} \sum_{i=1}^n \log h(X_i) - \mathbf{E}_{\theta_0}[\log h(X_1)] \right| \\ &\xrightarrow[n \rightarrow \infty]{\mathbf{P}_{\theta_0}} 0. \end{aligned} \quad (5.3)$$

We used the compactness of  $\Theta$  (to bound uniformly  $\sup_{\theta \in \Theta} |\eta_j(\theta)|$ , since  $\Theta$  is bounded and closed), and the weak law of large numbers.  $\square$

The assumptions (i) and (ii) are often not checked in practice. In fact, in many relevant cases, (ii) is not fulfilled (often  $\Theta$  is unbounded). One can however show that this assumption may also be dropped for exponential families (we omit further details here).

## 5.2 Fisher information, Cramér-Rao inequality and asymptotic efficiency

**Theorem 5.3.** *Let  $X_1, \dots, X_n$  be i.i.d. real random variables under  $\mathbf{P}_\theta$ . Furthermore, suppose that  $\mathcal{Q} = \{(\mathbf{P}_\theta)_{X_1}; \theta \in \Theta\}$ , where  $\Theta \subseteq \mathbb{R}$ , is an exponential family with  $\eta_j(\theta)$  and  $C(\theta)$  twice continuously differentiable. Moreover, let assumptions (i) and (ii) of Proposition 5.1 be fulfilled, and suppose that  $L$  and  $L_n$  are twice continuously differentiable in  $\theta$  and  $L''(\theta_0) > 0$ . Then we have (with  $\hat{\theta}_n$  denoting the MLE)*

$$\sqrt{n} \left( \hat{\theta}_n - \theta_0 \right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N} \left( 0, \frac{1}{I(\theta_0)} \right), \quad (5.4)$$

where  $I(\theta_0)$  is the Fisher-information, defined by

$$I(\theta_0) = \mathbf{E}_{\theta_0} \left[ \left( \frac{d}{d\theta} \log f_\theta(X_1) \Big|_{\theta=\theta_0} \right)^2 \right]. \quad (5.5)$$

The significance of this result comes from the fact that the variance of the asymptotic normal distribution in (5.4) is optimal. This is the statement of the *Cramér-Rao inequality*.



**Proposition 5.4.** Let  $X_1, \dots, X_n$  be i.i.d. real random variables under  $\mathbf{P}_\theta$ . Let  $T(X_1, \dots, X_n)$  be an estimator for  $\theta$  and  $b(\theta) = \text{Bias}_\theta(T(X_1, \dots, X_n))$  the bias of  $T(X_1, \dots, X_n)$ . Then

$$\text{Var}_\theta(T(X_1, \dots, X_n)) \geq \frac{(1 + b'(\theta))^2}{nI(\theta)}, \quad (5.6)$$

where we assume that (♠)

(A1)  $\mathcal{M}_f = \{x \in \mathbb{R} : f_\theta(x) > 0\}$  does not depend on  $\theta \in \Theta$ ,

(A2) For all  $x \in \mathcal{M}_f$ ,  $\frac{d}{d\theta} \log f_\theta(x)$  exists and  $I(\theta) \in (0, \infty)$ ,

(A3)  $\mathbf{E}_\theta[\frac{d}{d\theta} \log f_\theta(X_1)] = 0$ , and  $\frac{d}{d\theta} \mathbf{E}_\theta[T(X_1, \dots, X_n)] = \mathbf{E}_\theta[T(X_1, \dots, X_n) \frac{\partial}{\partial \theta} \log f_\theta^{(n)}(X_1, \dots, X_n)]$  for all  $\theta \in \Theta$ .

---

End of Lecture 16

Before we start the proof, let us briefly comment on the conditions (A3): These *regularity conditions* essentially correspond to “interchanging differentiation and integration”, and are usually not checked in practice. In fact, in (5.7) and (5.8), we will explain why the conditions (A3) are reasonable.

*Proof.* Since the  $X_1, \dots, X_n$  are i.i.d., their joint probability density function / joint probability mass function is  $f_\theta^{(n)}(x) = \prod_{i=1}^n f_\theta(x_i)$ . We set  $X = (X_1, \dots, X_n)$

$$\begin{aligned} \mathbf{E}_\theta[T(X)] &= \int T(x) f_\theta^{(n)}(x) dx \\ \Rightarrow \quad \frac{\partial}{\partial \theta} \mathbf{E}_\theta[T(X)] &= \int T(x) \frac{\partial}{\partial \theta} f_\theta^{(n)}(x) dx = \int T(x) \left( \frac{\partial}{\partial \theta} \log f_\theta^{(n)}(x) \right) f_\theta^{(n)}(x) dx. \end{aligned} \quad (5.7)$$

Moreover we have

$$\begin{aligned} 1 &= \int f_\theta^{(n)}(x) dx \\ \Rightarrow \quad 0 &= \int \frac{\partial}{\partial \theta} f_\theta^{(n)}(x) dx = \int \left( \frac{\partial}{\partial \theta} \log f_\theta^{(n)}(x) \right) f_\theta^{(n)}(x) dx \\ \Rightarrow \quad \frac{\partial}{\partial \theta} \mathbf{E}_\theta[T(X)] &= \int (T(x) - \mathbf{E}_\theta[T(X)]) \left( \frac{\partial}{\partial \theta} \log f_\theta^{(n)}(x) \right) f_\theta^{(n)}(x) dx. \end{aligned} \quad (5.8)$$

Using the Cauchy-Schwarz inequality (recall  $(\int fg)^2 \leq (\int f^2)(\int g^2)$ ), we have

$$\begin{aligned} \left( \frac{\partial}{\partial \theta} \mathbf{E}_\theta[T(X)] \right)^2 &\leq \int (T(x) - \mathbf{E}_\theta[T(X)])^2 f_\theta^{(n)}(x) dx \int \left( \frac{\partial}{\partial \theta} \log f_\theta^{(n)}(x) \right)^2 f_\theta^{(n)}(x) dx \\ &= (\text{Var}_\theta[T(X)]) \cdot \text{Var}_\theta \left[ \frac{\partial}{\partial \theta} \log f_\theta^{(n)}(X) \right]. \end{aligned} \quad (5.9)$$

Now we see that

$$\text{Var}_\theta \left[ \frac{\partial}{\partial \theta} \log f_\theta^{(n)}(X) \right] = \text{Var}_\theta \left[ \frac{\partial}{\partial \theta} \sum_{i=1}^n \log f_\theta(X_i) \right] = n \text{Var}_\theta \left[ \frac{\partial}{\partial \theta} \log f_\theta(X_1) \right] = nI(\theta). \quad (5.10)$$

In the last equation, we used

$$\text{Var}_\theta \left[ \frac{\partial}{\partial \theta} \log f_\theta(X_1) \right] = \mathbf{E}_\theta \left[ \left( \frac{\partial}{\partial \theta} \log f_\theta(X_1) \right)^2 \right] - \underbrace{\mathbf{E}_\theta \left[ \frac{\partial}{\partial \theta} \log f_\theta(X_1) \right]^2}_{=0} = I(\theta). \quad (5.11)$$

The claim then follows by combining (5.9) and (5.10).  $\square$

*Example 5.5.* Consider  $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$  i.i.d. with fixed  $\sigma^2 > 0$ . We have

$$\log f_\mu(x) = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x - \mu)^2. \quad (5.12)$$

Thus, we see

$$\frac{\partial}{\partial \mu} \log f_\mu(x) = \frac{1}{\sigma^2}(x - \mu) \quad \Rightarrow \quad \mathbf{E}_\mu \left[ \left( \frac{\partial}{\partial \mu} \log f_\mu(X_1) \right)^2 \right] = \frac{\mathbf{E}_\mu[(X - \mu)^2]}{\sigma^4} = \frac{1}{\sigma^2}. \quad (5.13)$$

This means that  $I(\mu) = \frac{1}{\sigma^2}$ . Note that the MLE for  $\mu$  is given by

$$\hat{\mu}_n = \bar{X}_n \quad \Rightarrow \quad \text{Var}_\mu[\bar{X}_n] = \frac{\sigma^2}{n} = \frac{1}{nI(\mu)}. \quad (5.14)$$

Also  $b(\mu) = 0$ , so we see that in fact the variance of the MLE saturates the Cramér-Rao lower bound.

Note that if the MLE fulfills  $\frac{\partial}{\partial \theta} \text{Bias}_\theta[\hat{\theta}_n] \rightarrow 0$ , then the variance of the MLE fulfills the Cramér-Rao lower bound by (5.4). One also says that the MLE is *asymptotically (Fisher) efficient*.

*Remark 5.6.* (i)  $\spadesuit$  The conditions of Theorem 5.3 that guarantee the asymptotic efficiency of the MLE were chosen such that the conditions (A1)–(A3) in Proposition 5.4 are fulfilled.<sup>1</sup>

(ii) To calculate  $I(\theta_0)$  from (5.5) can be somewhat tedious due to the square. Under regularity conditions, one has the alternative representation:

$$I(\theta_0) = \mathbf{E}_{\theta_0} \left[ -\frac{d^2}{d\theta^2} \log f_\theta(X_1) \Big|_{\theta=\theta_0} \right]. \quad (5.15)$$

<sup>1</sup>Indeed, (A1) follows immediately from the fact that we are considering an exponential family. The first part in (A2) follows from the fact that  $C(\theta)$  and  $\eta_j(\theta)$  are assumed to be twice continuously differentiable, and the second part of (A2) can be checked via the assumption that  $L''(\theta_0)$  exists and is strictly positive. The regularity assumptions in (A3) can also be verified like this.

- (iii) We have chosen in Proposition 5.4 to state the Cramér-Rao inequality for estimators for  $\theta$ . If instead we are trying to estimate *any* other real quantity  $\gamma(\theta)$ , we have instead (under the same conditions)

$$\text{Var}_\theta(T(X_1, \dots, X_n)) \geq \frac{\left(\frac{d}{d\theta} \mathbf{E}_\theta[T(X_1, \dots, X_n)]\right)^2}{nI(\theta)}. \quad (5.16)$$

Finally, we state a multivariate version of the Cramér-Rao inequality. For this, let  $X_1, \dots, X_n$  be i.i.d. real random variables under  $\mathbf{P}_\theta$ , where  $\theta \in \Theta \subseteq \mathbb{R}^k$ . We can then consider the *score vector*

$$s(\theta; X_1) = s(\theta) = \left( \frac{\partial}{\partial \theta_1} \log f_\theta(X_1), \dots, \frac{\partial}{\partial \theta_k} \log f_\theta(X_1) \right)^\top \in \mathbb{R}^k \quad (5.17)$$

and the *Fisher information matrix*

$$\mathbf{I}(\theta) = \mathbf{E}_\theta \left[ s(\theta) \cdot s(\theta)^\top \right] \in \mathbb{R}^{k \times k}. \quad (5.18)$$

(this is the covariance matrix of the random vector  $s(\theta)$ ). Here is the *multivariate Cramér-Rao inequality*:

**Proposition 5.7.** *Let  $T(X_1, \dots, X_n)$  be an estimator for  $\theta$  and consider the “bias vector”*

$$\mathbf{b}(\theta) = \begin{pmatrix} \mathbf{E}_\theta[T_1(X_1, \dots, X_n)] - \theta_1 \\ \vdots \\ \mathbf{E}_\theta[T_k(X_1, \dots, X_n)] - \theta_k \end{pmatrix}.$$

*Then, we have for the covariance matrix  $\Sigma_\theta$  of  $T_1(X_1, \dots, X_n), \dots, T_k(X_1, \dots, X_n)$  (which we also denote as  $\text{Cov}_\theta(T(X_1, \dots, X_n))$ ) the lower bound (under some regularity conditions):*

$$\Sigma_\theta = \text{Cov}_\theta(T(X_1, \dots, X_n)) \geq_L \frac{1}{n} (I_{k \times k} + \nabla \mathbf{b}(\theta)) \mathbf{I}(\theta)^{-1} (I_{k \times k} + \nabla \mathbf{b}(\theta))^\top. \quad (5.19)$$

Here by  $A \geq_L B$  for symmetric  $k \times k$  matrices we mean the Loewner order, namely that  $A - B$  is positive semidefinite, and the notation  $\nabla \mathbf{b}(\theta)$  denotes the Jacobian, i.e.

$$\nabla \mathbf{b}(\theta) = \begin{pmatrix} \frac{\partial}{\partial \theta_1} b_1(\theta) & \dots & \frac{\partial}{\partial \theta_k} b_1(\theta) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial \theta_1} b_k(\theta) & \dots & \frac{\partial}{\partial \theta_k} b_k(\theta) \end{pmatrix}.$$

**Theorem 5.8.** *Let  $X_1, \dots, X_n$  be i.i.d. real random variables under  $\mathbf{P}_\theta$ . Furthermore, suppose that  $\mathcal{Q} = \{(\mathbf{P}_\theta)_{X_1}; \theta \in \Theta\}$ , where  $\Theta \subseteq \mathbb{R}^k$ , is an exponential family with  $\eta_j(\theta)$  and  $C(\theta)$  twice continuously differentiable. Moreover, let assumptions (i) and (ii) of Proposition 5.1 be fulfilled, and suppose that  $L$  and  $L_n$  are twice continuously differentiable in  $\theta$  and  $\nabla^2 L(\theta_0)$  (the Hessian) is positive definite. Then,*

$$\sqrt{n} \left( \hat{\theta}_n - \theta_0 \right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}_k \left( 0, \mathbf{I}(\theta_0)^{-1} \right). \quad (5.20)$$

*Remark 5.9.* As we have seen, the MLE in exponential families is asymptotically efficient, i.e. its variance saturates the Cramér-Rao lower bound in the limit. However, it may be the case that the Cramér-Rao lower bound is *not* attained for the MLE when  $n$  is a fixed number: In fact, one can show that the Fisher information matrix in  $(\mu, \sigma^2)^\top \in \Theta = \mathbb{R} \times (0, \infty)$  for the 2-parameter exponential family  $\{\mathcal{N}(\mu, \sigma^2) : (\mu, \sigma^2)^\top \in \Theta\}$  is given by

$$\mathbf{I}(\mu, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}, \quad (5.21)$$

but the MLE  $(\hat{\mu}_n, \hat{\sigma}_n^2)^\top$  does *not* have  $\mathbf{I}(\mu, \sigma^2)^{-1}$  as a covariance matrix. This raises the question:

*When do we have equality in the Cramér-Rao inequality?*

In fact one has the following in the case of a scalar parameter: Let  $\Theta \subseteq \mathbb{R}$  and let  $T(X_1, \dots, X_n)$  be an unbiased estimator for  $\gamma(\theta)$  based on i.i.d. observations  $X_1, \dots, X_n$ . Then the Cramér-Rao inequality (5.16) is an equality if and only if  $\{(\mathbf{P}_\theta)_{X_1} : \theta \in \Theta\}$  forms a 1-parameter exponential family.

---

*End of Lecture 17*

## 6 Confidence intervals

In previous chapters, we constructed estimators for unknown parameters. In applications, it is reasonable to give not only an estimate for a quantity, but rather an interval, where we expect the parameter to be. This leads to the following definition.

**Definition 6.1.** Consider a parametric statistical model with data  $\mathbf{X} = (X_1, \dots, X_n)$  under  $(\mathbf{P}_\theta)_{\theta \in \Theta}$  with  $\Theta \subseteq \mathbb{R}^k$ . For  $\alpha \in (0, 1)$ , a subset  $S(\mathbf{X}) \subseteq \Theta$  is called  $(1 - \alpha)$ -confidence region for  $\theta$ , if

$$\mathbf{P}_\theta[\theta \in S(\mathbf{X})] \geq 1 - \alpha \quad \text{for all } \theta \in \Theta. \quad (6.1)$$

If  $\Theta$  is a subset of  $\mathbb{R}$  and  $S(\mathbf{X}) \subseteq \Theta$  is an interval, we also call  $S(\mathbf{X})$  a  $(1 - \alpha)$ -confidence interval.

In essence, confidence intervals retain aspects of both estimators and statistical tests: A confidence interval is a (random) interval around an estimator based on data  $X_1, \dots, X_n$ , which has the property that the it contains the true parameter with at least a certain probability. The latter property is related to the “type I error” in a statistical test, as we will see.

*Remark 6.2.* (i) Recall that an estimator for  $\theta$  was simply a function  $T \circ \mathbf{X}$  with values in  $\Theta$ . A confidence region is similarly a function  $S \circ \mathbf{X}$ , but with values in  $\mathcal{P}(\Theta)$  (the power set of  $\Theta$ , with

$$\mathcal{P}(\Theta) = \{A : A \subseteq \Theta\}.$$

In particular, note that the interval  $S(\mathbf{X})$  is itself random and depends on the realization  $\omega \in \Omega$ .

- (ii) In line with the previous remark, we stress again that in formula (6.1),  $\theta$  is *not* the random quantity, but  $S(\mathbf{X})$  is. The interpretation of  $S(\mathbf{X})$  is therefore that the probability that  $\theta$  is contained in a realization of  $S(\mathbf{X})$  is at least  $1 - \alpha$ . This does *not* mean that if  $x = X(\omega)$  is observed,  $S(x)$  contains  $\theta$  with probability  $\geq 1 - \alpha$  (the previous statement does not make sense:  $\theta$  is not random!). Instead, if the experiment is repeated very often,  $\theta$  will be contained in the (different) realizations of  $S(\mathbf{X})$  in approximately at least  $(1 - \alpha) \cdot 100\%$  of the cases.

*Goal:* We want to make both  $\alpha$  and  $S(\mathbf{X})$  small. Since simultaneous minimization is not possible, we will *fix*  $\alpha$  and then look for small sets  $S(\mathbf{X})$  where “ $\geq$ ” in (6.1) is essentially replaced by “ $=$ ”.

*Example 6.3.* Assume that under  $X_1, \dots, X_n$  are i.i.d. with  $X_1 \sim \mathcal{N}(\mu, \sigma^2)$  and *known* (!) variance  $\sigma^2 > 0$ . We want to construct a  $(1 - \alpha)$ -confidence interval for  $\mu$ , based on the observations  $X_1, \dots, X_n$ . Recall that in this situation

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \quad \Rightarrow \quad \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \sim \mathcal{N}(0, 1). \quad (6.2)$$

Let  $u_{1-\gamma}$  denote the  $(1-\gamma)$ -quantile of the  $\mathcal{N}(0, 1)$ -distribution, i.e. the unique number  $u_{1-\gamma} \in \mathbb{R}$  such that  $\Phi(u_{1-\gamma}) = 1 - \gamma$ . Note that

$$\mathbf{P}_\mu \left[ u_{\frac{\alpha}{2}} \leq \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \leq u_{1-\frac{\alpha}{2}} \right] = \Phi(u_{1-\frac{\alpha}{2}}) - \Phi(u_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha. \quad (6.3)$$

Since  $u_{\frac{\alpha}{2}} = -u_{1-\frac{\alpha}{2}}$  by the symmetry of  $\mathcal{N}(0, 1)$ , we have

$$\mathbf{P}_\mu \left[ \mu \in \left[ \bar{X}_n - \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}} \right] \right] = 1 - \alpha. \quad (6.4)$$

Therefore, the interval  $S(\mathbf{X}) = \left[ \bar{X}_n - \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}} \right]$  is a  $(1 - \alpha)$ -confidence interval for  $\mu$ .

This example shows that typically, confidence intervals are constructed around estimators (the empirical mean  $\bar{X}_n$  is an estimator for  $\mu$ ).

The choice of the confidence interval in question depends on the nature of the question: If we simply want to avoid the real parameter to be outside the confidence interval, we choose  $S(X) = \left[ \bar{X}_n - \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}} \right]$ . However we could also have chosen the interval  $S(X) = \left[ \bar{X}_n - \frac{\sigma}{\sqrt{n}} u_{1-\alpha}, \infty \right)$  corresponds to the test in which the more problematic error is to erroneously overestimate  $\mu$ , so we have

$$\begin{aligned} S_{\text{two-sided}}(\mathbf{X}) &= \left[ \bar{X}_n - \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}} \right], \\ S_{\text{right-sided}}(\mathbf{X}) &= \left[ \bar{X}_n - \frac{\sigma}{\sqrt{n}} u_{1-\alpha}, \infty \right), \\ S_{\text{left-sided}}(\mathbf{X}) &= \left( -\infty, \bar{X}_n + \frac{\sigma}{\sqrt{n}} u_{1-\alpha} \right]. \end{aligned} \quad (6.5)$$

Similarly, we can give confidence intervals for normally distributed random variables with *unknown* variance.

*Example 6.4.* Assume that under  $X_1, \dots, X_n$  are i.i.d. with  $X_1 \sim \mathcal{N}(\mu, \sigma^2)$  and both the mean  $\mu$  and the variance  $\sigma^2 > 0$  are *unknown*. The following expressions are all  $(1 - \alpha)$ -confidence intervals for  $\mu$ , based on the observations  $X_1, \dots, X_n$ :

$$\begin{aligned} S_{\text{two-sided}}(\mathbf{X}) &= \left[ \bar{X}_n - \frac{S_n}{\sqrt{n}} t_{n-1, 1-\frac{\alpha}{2}}, \bar{X}_n + \frac{S_n}{\sqrt{n}} t_{n-1, 1-\frac{\alpha}{2}} \right], \\ S_{\text{right-sided}}(\mathbf{X}) &= \left[ \bar{X}_n - \frac{S_n}{\sqrt{n}} t_{n-1, 1-\alpha}, \infty \right), \\ S_{\text{left-sided}}(\mathbf{X}) &= \left( -\infty, \bar{X}_n + \frac{S_n}{\sqrt{n}} t_{n-1, 1-\alpha} \right]. \end{aligned} \quad (6.6)$$

Here,  $t_{n-1, 1-\gamma}$  is the  $(1 - \gamma)$ -quantile of the  $t_{n-1}$ -distribution. To see this, recall that

$$\sqrt{n} \frac{\bar{X}_n - \mu}{S_n} \sim t_{n-1} \quad \text{by Theorem 2.7, (ii).} \quad (6.7)$$

Thus, for every  $\mu, \sigma^2$  one has:

$$\begin{aligned} \mathbf{P}_{(\mu, \sigma^2)} \left[ \mu \in \left[ \bar{X}_n - \frac{S_n}{\sqrt{n}} t_{n-1, 1-\frac{\alpha}{2}}, \bar{X}_n + \frac{S_n}{\sqrt{n}} t_{n-1, 1-\frac{\alpha}{2}} \right] \right] \\ = \mathbf{P}_{(\mu, \sigma^2)} \left[ t_{n-1, \frac{\alpha}{2}} \leq \sqrt{n} \frac{\bar{X}_n - \mu}{S_n} \leq t_{n-1, 1-\frac{\alpha}{2}} \right] = 1 - \alpha, \end{aligned} \quad (6.8)$$

where we use that  $t_{n-1, \frac{\alpha}{2}} = -t_{n-1, 1-\frac{\alpha}{2}}$  (by the symmetry of the  $t_{n-1}$ , similar as for the normal distribution). The second and third lines in (6.6) follow similarly.

---

*End of Lecture 18*

As we saw above:

- Confidence intervals are typically constructed around estimators,
- to find confidence intervals explicitly, the distribution of the estimator used is often required.

Especially the last point can be problematic in general, but for large enough  $n$ , we can use our knowledge on the asymptotic distribution of estimators *approximate* confidence intervals. We need to introduce a new concept first, to make this idea mathematically sound.

**Definition 6.5.** Consider a parametric statistical model with data  $\mathbf{X} = (X_1, \dots, X_n)$  under  $(\mathbf{P}_\theta)_{\theta \in \Theta}$  with  $\Theta \subseteq \mathbb{R}^k$ . For  $\alpha \in (0, 1)$ , a sequence of subsets  $(S_n(\mathbf{X}))_{n \in \mathbb{N}} \subseteq \Theta$  is called *asymptotic  $(1 - \alpha)$ -confidence region* for  $\theta$ , if

$$\liminf_{n \rightarrow \infty} \mathbf{P}_\theta[\theta \in S_n(\mathbf{X})] \geq 1 - \alpha \quad \text{for all } \theta \in \Theta. \quad (6.9)$$

If  $\Theta$  is a subset of  $\mathbb{R}$  and  $S(\mathbf{X}) \subseteq \Theta$  is an interval, we also call  $(S_n(\mathbf{X}))_{n \in \mathbb{N}}$  an *asymptotic  $(1 - \alpha)$ -confidence interval*.

In most relevant cases,  $\liminf_{n \rightarrow \infty}$  is replaced by  $\lim_{n \rightarrow \infty}$  and “ $\geq$ ” is replaced by “ $=$ ”.

*Example 6.6.* Suppose that  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(p)$  with unknown  $p$ . We know that the MLE and moment estimators for  $p$  are both equal to  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ . Note that for  $\text{Var}_p[X_1] = p(1-p)$ , a consistent estimator is given by  $\bar{X}_n(1 - \bar{X}_n)$ . By the central limit theorem and Slutsky’s theorem (Theorem 1.17, (iv)), we have

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - p)}{\sqrt{\bar{X}_n(1 - \bar{X}_n)}} \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, 1). \quad (6.10)$$

Therefore an asymptotic  $(1 - \alpha)$ -confidence interval for  $p$  is given by

$$S(\mathbf{X}) = \left[ \bar{X}_n - \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n}} \sqrt{\bar{X}_n(1 - \bar{X}_n)}, \bar{X}_n + \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n}} \sqrt{\bar{X}_n(1 - \bar{X}_n)} \right]. \quad (6.11)$$

Let us finally give a general construction principle for confidence intervals, based on the MLE.

**Theorem 6.7.** Let  $X_1, \dots, X_n$  be i.i.d. under  $\mathbf{P}_\theta$ , where  $\theta \in \Theta \subseteq \mathbb{R}$  is unknown. Suppose that the assumptions of Theorem 5.3 are fulfilled and assume furthermore that (with  $\hat{\theta}_n$  denoting the MLE for  $\theta$ ), one has

$$I(\hat{\theta}_n) \xrightarrow[n \rightarrow \infty]{\mathbf{P}_\theta} I(\theta) \quad \text{for all } \theta \in \Theta, \quad (6.12)$$

with  $I$  denoting the Fisher information (recall (5.5)). Then

$$S(\mathbf{X}) = \left[ \hat{\theta}_n - \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{nI(\hat{\theta}_n)}}, \hat{\theta}_n + \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{nI(\hat{\theta}_n)}} \right] \quad (6.13)$$

is an asymptotic  $(1 - \alpha)$ -confidence interval for  $\theta$ . Here,  $u_{1-\gamma}$  denotes again the  $(1 - \gamma)$ -quantile of the  $\mathcal{N}(0, 1)$ -distribution.

*Proof.* By Theorem 5.3, (6.12), and Slutsky's theorem (Theorem 1.17, (iv)), we have

$$\sqrt{nI(\hat{\theta}_n)}(\hat{\theta}_n - \theta) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, 1). \quad (6.14)$$

With this, we see that

$$\lim_{n \rightarrow \infty} \mathbf{P}_\theta [\theta \in S(\mathbf{X})] = \lim_{n \rightarrow \infty} \mathbf{P}_\theta \left[ u_{\frac{\alpha}{2}} \leq \sqrt{nI(\hat{\theta}_n)}(\hat{\theta}_n - \theta) \leq u_{1-\frac{\alpha}{2}} \right] = 1 - \alpha. \quad (6.15)$$

□



## 7 Statistical tests

(Reference: [6, Chapter 10, in particular sections 10.1–10.4, 10.10])

In this section, we introduce the notion of statistical tests and prove the *Neyman-Pearson lemma*. We motivate this using an example.

### 7.1 Basic notions of statistical tests

*Example 7.1.* We consider a certain drug  $A$  that has a known efficacy of 60 %. We want to evaluate the

*Claim:* A new (more expensive) drug  $B$  has an efficacy of 70 %. To see whether the claim is valid, the drug  $B$  is tested with 100 persons.

We choose the model:

$$X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(\theta), \quad (7.1)$$

and the outcome  $\mathbf{X} = (X_1, \dots, X_n) \in \mathcal{X}$  is the “vector” of persons out of the 100 with a positive reaction (1) to the administered drug  $B$  or no reaction (0). For simplicity, we assume that the new drug either has the same efficacy 60% as  $A$  or 70% efficacy.

Question: Do we have

$$\begin{array}{ll} \mathbf{P}_{\theta_0} & (\mathbf{P}_{\theta_1})_{X_1} = \text{Ber}(\theta_0), \quad \theta_0 = 0.6 \quad (\text{null hypothesis } H_0), \\ \text{or } \mathbf{P}_{\theta_1} & (\mathbf{P}_{\theta_1})_{X_1} = \text{Ber}(\theta_1), \quad \theta_1 = 0.7 \quad (\text{alternative hypothesis } H_1)? \end{array} \quad (7.2)$$

The challenge is to determine a *statistical test* that decides between  $H_0$  and  $H_1$ , based on the test result  $\phi(\mathbf{X})$ , where

$$\phi : \mathcal{X} \rightarrow \{0, 1\}, \quad (7.3)$$

which should be “optimal” in a way. Here,  $\phi(\mathbf{X}(\omega)) = 0$  models keeping the null hypothesis with the observed outcome  $\omega$ , and  $\phi(\mathbf{X}(\omega)) = 1$  models rejecting the null hypothesis with the observed outcome  $\omega$ .

---

*End of Lecture 19*

**Definition 7.2.** Consider a parametric statistical model with data  $\mathbf{X}$  and probability measures  $(\mathbf{P}_\theta)_{\theta \in \Theta}$ . We assume that  $\Theta$  can be written as a disjoint union

$$\Theta = \Theta_0 \cup \Theta_1, \quad \Theta_0 \cap \Theta_1 = \emptyset. \quad (7.4)$$

The region  $\Theta_0$  corresponds to the *null hypothesis*, the region  $\Theta_1$  to the *alternative hypothesis*. A *(non-randomized) statistical test* is a map  $\phi \circ \mathbf{X}$ , where  $\phi : \mathcal{X} \rightarrow \{0, 1\}$ . A *type I error occurs*, if we falsely reject the null hypothesis  $H_0$ , i.e.  $\phi(\mathbf{X}(\omega)) = 1$  when  $H_0$  is correct. A *type II error occurs*, if we falsely do not reject the null hypothesis  $H_0$ , i.e.  $\phi(\mathbf{X}(\omega)) = 0$  when  $H_1$  is correct.

*The null hypothesis  $H_0$  is chosen such that falsely rejecting it (type I errors) should be unlikely.* (7.5)

This means: When we set up statistical tests, a false rejection of  $H_0$ , i.e. a type  $I$  error is considered the *more problematic error* than a false non-rejection, i.e. a type  $II$  error. We illustrate this in two examples:

- Suppose we are unsure whether certain collected wild mushrooms are edible or poisonous. Then

$$\begin{aligned} H_0 : & \quad \text{mushrooms are poisonous,} \\ H_1 : & \quad \text{mushrooms are edible.} \end{aligned}$$

We have chosen  $H_0$  in such a way, that falsely rejecting it is the more severe error!

- Let us consider the previous Example 7.1. We want to avoid a situation, in which a new and more expensive drug is authorized, while being not more effective than the standard one used. According to this, we should define

$$\begin{aligned} H_0 : & \quad \text{drug } B \text{ has the same efficacy 0.6 as drug } A, \\ H_1 : & \quad \text{drug } B \text{ has an higher efficacy of 0.7 than drug } A. \end{aligned}$$

This is of course nothing else than (7.2); here  $\Theta_0 = \{0.6\}$  and  $\Theta_1 = \{0.7\}$ .

How likely are the errors of type  $I$  and  $II$ ? We consider the case of *simple hypotheses* where  $H_0$  corresponds to  $\Theta_0 = \{\theta_0\}$  and  $H_1$  to  $\Theta_1 = \{\theta_1\}$ . This is in line with our Example 7.1. The probability of a type  $I$  error is given by

$$\mathbf{P}_{\theta_0}[\phi(\mathbf{X}) = 1]. \quad (7.6)$$

The probability of a type  $II$  error on the other hand is

$$\mathbf{P}_{\theta_1}[\phi(\mathbf{X}) = 0]. \quad (7.7)$$

If we consider the extreme case

$$\phi_{\text{extreme}}(x) = 0 \quad \text{for all } x \in \mathcal{X}, \quad (7.8)$$

where we never reject  $H_0$ , we see that the type  $I$  error has probability

$$\mathbf{P}_{\theta_0}[\phi_{\text{extreme}}(\mathbf{X}) = 1] = \mathbf{P}_{\theta_0}[\emptyset] = 0. \quad (7.9)$$

On the other hand, the type  $II$  error of course has probability

$$\mathbf{P}_{\theta_1}[\phi_{\text{extreme}}(\mathbf{X}) = 0] = \mathbf{P}_{\theta_1}[\Omega] = 1. \quad (7.10)$$

For a meaningful test, we need to do the following:

- Specify an upper bound  $\alpha \in (0, 1)$  for the type  $I$  error, that is, require

$$\mathbf{P}_{\theta_0}[\phi(\mathbf{X}) = 1] \leq \alpha. \quad (7.11)$$

The value  $\alpha$  is called the *significance level* for the test. Typical levels in practice are  $\alpha = 0.05$  or  $\alpha = 0.01$ .

- Minimize the value of  $\mathbf{P}_{\theta_1}[\phi(\mathbf{X}) = 0]$  under the constraint  $\mathbf{P}_{\theta_0}[\phi(\mathbf{X}) = 1] \leq \alpha$ . If a solution  $\phi^* : \mathcal{X} \rightarrow \{0, 1\}$  exists, this is called the *best / most powerful test at level  $\alpha$* .

We return to Example 7.1: Heuristically, given  $\alpha \in (0, 1)$ , we should try to look for a test  $\phi \circ \mathbf{X}$  with  $\phi : \mathcal{X} \rightarrow \{0, 1\}$  of the form

$$\phi(x) = \begin{cases} 1, & \sum_{i=1}^n x_i \in C_\alpha = \{k_\alpha, k_\alpha + 1, \dots, 100\} \Leftrightarrow \sum_{i=1}^n x_i \geq k_\alpha, \\ 0, & \sum_{i=1}^n x_i \in C_\alpha^c = \{0, 1, \dots, k_\alpha - 1\} \Leftrightarrow \sum_{i=1}^n x_i < k_\alpha, \end{cases} \quad (7.12)$$

for some  $k_\alpha \in \{0, \dots, 100\}$  to be determined. The range

$$C_\alpha = \{x \in \mathcal{X} : \sum_{i=1}^n x_i \in \{k_\alpha, k_\alpha + 1, \dots, 100\}\} \quad (7.13)$$

is called the *critical region* for the test  $\phi \circ \mathbf{X}$ .

After this discussion of *simple hypotheses*, we now formulate some general definitions.

**Definition 7.3.** Suppose that  $\phi \circ \mathbf{X}$  is a test for

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

where  $\Theta = \Theta_0 \cup \Theta_1$  with  $\Theta_0 \cap \Theta_1 = \emptyset$ . Let  $\Phi$  be the set of all tests<sup>1</sup>.

- (i) The function

$$G_\phi : \begin{cases} \Theta \rightarrow [0, 1], \\ \theta \mapsto G_\phi(\theta) = \mathbf{P}_\theta[\phi(\mathbf{X}) = 1] \end{cases} \quad (7.14)$$

is called the *power function* of the test  $\phi \circ \mathbf{X}$ .

- (ii) We say that  $\phi \circ \mathbf{X}$  is a *test at (significance) level  $\alpha \in [0, 1]$* , if

$$\sup_{\theta \in \Theta_0} G_\phi(\theta) = \sup_{\theta \in \Theta_0} \mathbf{P}_\theta[\phi(\mathbf{X}) = 1] \leq \alpha. \quad (7.15)$$

The set of all tests at level  $\alpha$  is denoted by  $\Phi_\alpha$ .

- (iii) A test  $\phi \circ \mathbf{X}$  is *unbiased* at level  $\alpha$ , if  $\phi \in \Phi_\alpha$  and

$$\inf_{\theta \in \Theta_1} G_\phi(\theta) = \inf_{\theta \in \Theta_1} \mathbf{P}_\theta[\phi(\mathbf{X}) = 1] \geq \alpha. \quad (7.16)$$

We write  $\Phi_{\alpha\alpha}$  for the set of unbiased tests at level  $\alpha$ .

- (iv) A test  $\phi^* \circ \mathbf{X}$  with  $\phi^* \in \Phi_\alpha$  is a *uniformly most powerful (UMP) test at level  $\alpha$*  if

$$G_{\phi^*}(\theta) = \sup_{\phi \in \Phi_\alpha} G_\phi(\theta) \quad \text{for all } \theta \in \Theta_1. \quad (7.17)$$

<sup>1</sup>Slight abuse of notation:  $\phi \in \Phi$  is a function  $\phi : \mathcal{X} \rightarrow \{0, 1\}$ , and the corresponding test is  $\phi \circ \mathbf{X}$ .

- (iv) A test  $\phi^* \circ \mathbf{X}$  with  $\phi^* \in \Phi_{\alpha\alpha}$  is a *uniformly most powerful unbiased (UMPU) test at level  $\alpha$*  if

$$G_{\phi^*}(\theta) = \sup_{\phi \in \Phi_{\alpha\alpha}} G_{\phi}(\theta) \quad \text{for all } \theta \in \Theta_1. \quad (7.18)$$

---

*End of Lecture 20*

Let us comment on this definition and compare it to the informal discussion above.

**Remark 7.4.** (i) The function  $G_{\phi}(\theta)$  is the probability to reject  $H_0$  and decide for  $H_1$ , if  $\theta$  is the true parameter. In other words:

- For  $\theta \in \Theta_0$ ,  $G_{\phi}(\theta)$  is the probability of a type I error.
- For  $\theta \in \Theta_1$ ,  $1 - G_{\phi}(\theta)$  is the probability of a type II error.

The definition of a test at level  $\alpha$  is in line with the requirement (7.5) that the type I error should be smaller than  $\alpha$ . If  $\Theta = \{\theta_0\}$  is a simple hypothesis, we already saw this in (7.11), but for general  $H_0$  (i.e. general  $\Theta_0$ ), we require this condition uniformly over  $\Theta_0$ .

- (ii) For a test  $\phi \circ \mathbf{X}$  to be unbiased means that the probability to decide for  $H_1$  should never be smaller if  $\theta \in \Theta_1$ , than for  $\theta \in \Theta_0$ .
- (iii) The UMP property means that the test under consideration *minimizes the probability of a type II error* uniformly over  $\Theta_1$ , as we discussed previously informally with  $\Theta_0 = \{\theta_0\}$  and  $\Theta_1 = \{\theta_1\}$ .

Recall that a sufficient statistic  $T(\mathbf{X})$  essentially encodes all relevant information contained in  $\mathbf{X}$  about  $\theta$ . It should therefore not be surprising, that statistical tests only depend on  $\mathbf{X}$  through  $T(\mathbf{X})$  if the latter is a sufficient statistic for  $\theta$ . We have the following:

**Proposition 7.5.** Let  $\mathbf{X} = (X_1, \dots, X_n) \in \mathcal{X}$  and  $(\mathbf{P}_{\theta})_{\theta \in \Theta}$  be a parametric statistical model. Suppose that  $\phi \circ \mathbf{X}$  is a statistical test for

$$H_0 : \theta \in \Theta_0, \quad \text{against} \quad \theta \in \Theta_1, \quad (7.19)$$

with  $\Theta = \Theta_0 \cup \Theta_1$  and  $\Theta_0 \cap \Theta_1 = \emptyset$ . Moreover, suppose that  $T \circ \mathbf{X}$  is a sufficient statistic for  $\theta$ . Then, there exists a test  $\psi \circ T(\mathbf{X})$  with

$$G_{\psi \circ T}(\theta) = G_{\phi}(\theta). \quad (7.20)$$

*Proof.* We set

$$\psi \circ T(\mathbf{X}) = \mathbf{E}_{\theta}[\phi(\mathbf{X})|T(\mathbf{X})] = \mathbf{P}_{\theta}[\phi(\mathbf{X}) = 1|T(\mathbf{X})]. \quad (7.21)$$

Since  $T(\mathbf{X})$  is sufficient for  $\theta$ , this expression does not depend on  $\theta$ . Now we use that

$$G_{\psi \circ T}(\theta) = \mathbf{E}_{\theta}[\mathbf{E}_{\theta}[\phi(\mathbf{X})|T(\mathbf{X})]] = \mathbf{E}_{\theta}[\phi(\mathbf{X})] = \mathbf{P}_{\theta}[\phi(\mathbf{X}) = 1] = G_{\phi}(\theta). \quad (7.22)$$

□

## 7.2 The Neyman-Pearson lemma

In this section, we show the existence of UMP tests for two simple hypotheses against each other. To this end, let  $\mathbf{X} = (X_1, \dots, X_n) \in \mathcal{X}$  be the data and consider the simple hypotheses

$$H_0 : \Theta_0 = \{\theta_0\}, \quad H_1 : \Theta_1 = \{\theta_1\}, \quad \theta_0 \neq \theta_1. \quad (7.23)$$

**Definition 7.6.** We denote by  $f_{\theta_0}^{(n)}(x_1, \dots, x_n)$  and  $f_{\theta_1}^{(n)}(x_1, \dots, x_n)$  the joint probability mass function or joint probability density function of  $X_1, \dots, X_n$  under  $\mathbf{P}_{\theta_0}$  or  $\mathbf{P}_{\theta_1}$ , respectively. The expression

$$L_{\theta_0, \theta_1}(x_1, \dots, x_n) = \frac{f_{\theta_1}^{(n)}(x_1, \dots, x_n)}{f_{\theta_0}^{(n)}(x_1, \dots, x_n)}, \quad (x_1, \dots, x_n) \in \mathcal{X} \quad (7.24)$$

is called *likelihood quotient* (where  $\frac{x}{0} := \infty$  for  $x \geq 0$ ).

A test  $\phi^* \circ \mathbf{X}$  is called *Neyman-Pearson test* (or *likelihood quotient test*) if there exists  $c \in [0, \infty]$  such that

$$\phi^*(x) = \begin{cases} 1, & L_{\theta_0, \theta_1}(x) \geq c \Leftrightarrow f_{\theta_1}^{(n)}(x) \geq c f_{\theta_0}^{(n)}(x), \\ 0, & L_{\theta_0, \theta_1}(x) < c \Leftrightarrow f_{\theta_1}^{(n)}(x) < c f_{\theta_0}^{(n)}(x), \end{cases} \quad x = (x_1, \dots, x_n) \in \mathcal{X}. \quad (7.25)$$

We now state the *Neyman-Pearson lemma*.

**Lemma 7.7.** *The Neyman-Pearson test  $\phi^* \circ \mathbf{X}$  is a UMP test for (7.23) at significance level  $\alpha = \mathbf{P}_{\theta_0}[\phi^*(\mathbf{X}) = 1]$ .*

*Proof.* Let  $\phi^* \circ \mathbf{X}$  be the Neyman-Pearson test and  $c_\alpha^*$  the corresponding constant. For  $\phi \in \Phi_\alpha$ , we have

$$G_\phi(\theta_0) = \mathbf{P}_{\theta_0}[\phi(\mathbf{X}) = 1] \leq \alpha = \mathbf{P}_{\theta_0}[\phi^*(\mathbf{X}) = 1] = G_{\phi^*}(\theta_0). \quad (7.26)$$

Now suppose that the  $f_{\theta_0}^{(n)}$  and  $f_{\theta_1}^{(n)}$  are probability density functions, then

$$\begin{aligned} G_{\phi^*}(\theta_1) - G_\phi(\theta_1) &= \int_{\mathbb{R}^n} (\phi^*(x) - \phi(x)) f_{\theta_1}^{(n)}(x) dx \\ &= \int_{\mathbb{R}^n} \underbrace{\left( f_{\theta_1}^{(n)}(x) - c_\alpha^* f_{\theta_0}^{(n)}(x) \right) \cdot (\phi^*(x) - \phi(x))}_{\geq 0} dx \\ &\quad + c_\alpha^* \int_{\mathbb{R}^n} f_{\theta_0}^{(n)}(x) (\phi^*(x) - \phi(x)) dx \\ &\geq c_\alpha^* G_{\phi^*}(\theta_0) - G_\phi(\theta_0) \stackrel{(7.26)}{\geq} 0. \end{aligned} \quad (7.27)$$

We have used the fact that  $\mathbf{P}_\theta[\phi^*(\mathbf{X}) = 1] = \mathbf{E}_\theta[\phi^*(\mathbf{X})]$  for  $\theta \in \{\theta_0, \theta_1\}$ . □

*Example 7.8.* We continue Example 7.1. We have for  $x = (x_1, \dots, x_n) \in \{0, 1\}^n$ :

$$\begin{aligned}
 H_0 : f_{\theta_0}^{(n)}(x_1, \dots, x_n) &= \prod_{i=1}^n \theta_0^{x_i} \prod_{i=1}^n (1 - \theta_0)^{1-x_i}, \quad \theta_0 = 0.6, \\
 H_1 : f_{\theta_1}^{(n)}(x_1, \dots, x_n) &= \prod_{i=1}^n \theta_1^{x_i} \prod_{i=1}^n (1 - \theta_1)^{1-x_i}, \quad \theta_1 = 0.7 \\
 \Rightarrow L_{\theta_0, \theta_1}(x_1, \dots, x_n) &= \left( \frac{\theta_1}{\theta_0} \right)^{\sum_{i=1}^n x_i} \left( \frac{1 - \theta_1}{1 - \theta_0} \right)^{n - \sum_{i=1}^n x_i} \\
 &= \left( \frac{\theta_1}{1 - \theta_1} / \frac{\theta_0}{1 - \theta_0} \right)^{\sum_{i=1}^n x_i} \left( \frac{1 - \theta_1}{1 - \theta_0} \right)^n \geq c_{\alpha}^*.
 \end{aligned} \tag{7.28}$$

Since  $\theta_1 > \theta_0$ , we have  $\frac{\theta_1}{1 - \theta_1} / \frac{\theta_0}{1 - \theta_0} > 1$ , and so

$$L_{\theta_0, \theta_1}(x_1, \dots, x_n) \geq c_{\alpha}^* \quad \Leftrightarrow \quad \sum_{i=1}^n x_i \geq k_{\alpha}^*. \tag{7.29}$$

Then choose  $k_{\alpha}^*$  such that

$$\mathbf{P}_{\theta_0}[\phi^*(\mathbf{X}) = 1] = \mathbf{P}_{\theta_0}[L_{\theta_0, \theta_1}(\mathbf{X}) \geq c_{\alpha}^*] = \mathbf{P}_{\theta_0}\left[\sum_{i=1}^n X_i \geq k_{\alpha}^*\right] = \alpha. \tag{7.30}$$

We consider  $n = 100$ ,  $\theta_0 = 0.6$  and  $\alpha \lesssim 0.01$ .<sup>2</sup> We require

$$\mathbf{P}_{\theta_0}\left[\sum_{i=1}^n X_i \geq k_{\alpha}^*\right] = \sum_{k=k_{\alpha}^*}^n \binom{n}{k} \theta_0^k (1 - \theta_0)^{n-k} \stackrel{!}{=} \alpha. \tag{7.31}$$

We compute that for  $k_{\alpha}^* = 72$ , we have  $\alpha = 0.0084$ . The UMP test at this level is given by

$$\phi^* \circ \mathbf{X} = \begin{cases} 1, & \sum_{i=1}^n X_i \geq 72 \text{ } (H_0 \text{ is rejected}) \\ 0, & \sum_{i=1}^n X_i < 72 \text{ } (H_0 \text{ is not rejected}). \end{cases} \tag{7.32}$$

For the *type I error*, we have the probability

$$\mathbf{P}_{\theta_0}[\phi^*(\mathbf{X}) = 1] = \mathbf{P}_{\theta_0}\left[\sum_{i=1}^n X_i \geq 72\right] = 0.0084. \tag{7.33}$$

The *type II error* is given by

$$\mathbf{P}_{\theta_1}[\phi^*(\mathbf{X}) = 0] = \mathbf{P}_{\theta_1}\left[\sum_{i=1}^n X_i < 72\right] = \sum_{k=0}^{k_{\alpha}^*-1} \binom{100}{k} \theta_1^k (1 - \theta_1)^{100-k} \approx 0.62. \tag{7.34}$$

Note that the *type II error* is very large, but cannot be decreased, since the test is optimal by the Neyman-Pearson Lemma 7.7. To decrease the error, we need to *increase*  $n$ .

<sup>2</sup>Not every  $\alpha \in (0, 1)$  can be chosen, see also Remark 7.9, (i) below.

---

End of Lecture 21

**Remark 7.9.** (i) Let us stress again that in the discrete set-up, not all significance levels  $\alpha \in (0, 1)$  can be chosen in the statement of the Neyman-Pearson lemma. The reason is that the probability  $\alpha = \mathbf{P}_{\theta_0}[\phi(\mathbf{X}) = 1]$  can only attain countably many values. Again in the above example where  $\sum_{i=1}^n X_i \sim \text{Bin}(100, 0.6)$  under  $\mathbf{P}_{\theta_0}$ , we have

$$\begin{aligned}
 & \vdots \\
 \mathbf{P}_{\theta_0} \left[ \sum_{i=1}^n X_i \in \{70, 71, \dots, 100\} \right] &= 0.0248, \\
 \mathbf{P}_{\theta_0} \left[ \sum_{i=1}^n X_i \in \{71, 72, \dots, 100\} \right] &= 0.0148, \\
 \mathbf{P}_{\theta_0} \left[ \sum_{i=1}^n X_i \in \{72, 73, \dots, 100\} \right] &= 0.0084, \\
 \mathbf{P}_{\theta_0} \left[ \sum_{i=1}^n X_i \in \{73, 74, \dots, 100\} \right] &= 0.0046, \\
 & \vdots
 \end{aligned} \tag{7.35}$$

The values on the right-hand side in (7.35) are some of the possible choices for  $\alpha$ . In practice, if we are looking for a test at significance level  $\alpha = 0.01$ , we look for the *largest*  $\alpha \leq 0.01$ , for which a test exist. From (7.35), we see that in the case of Examples 7.1 and 7.8, this means we take  $\alpha = 0.0084$  and  $k_\alpha^* = 72$ .

(ii) We used  $\theta_1 = 0.7$  to establish the equivalence (7.29), but in fact the only information we used was that  $\theta_1 > \theta_0$ . In other words, we could actually strengthen the test by not testing  $H_0 : \Theta_0 = \{\theta_0\}$  against  $H_1 : \Theta_1 = \{\theta_1\}$ , but in fact test

$$\begin{aligned}
 & \Theta_0 = \{\theta_0\}, \quad (\text{null hypothesis } H_0), \\
 & \text{against } \Theta_1 = (\theta_0, 1], \quad (\text{alternative hypothesis } \tilde{H}_1).
 \end{aligned} \tag{7.36}$$

The decisive factor here is that the likelihood ratio  $L_{\theta_0, \theta_1}(x)$  is monotone in  $\sum_{i=1}^n x_i$ , which is why (7.29) is valid. Therefore the test  $\phi^* \circ \mathbf{X}$  is a UMP test for (7.36).

Let us formalize the second part of the remark above.

**Definition 7.10.** Consider a statistical model with i.i.d. data  $\mathbf{X} = (X_1, \dots, X_n) \in \mathcal{X} \subseteq \mathbb{R}^n$  and probability measures  $(\mathbf{P}_\theta)_{\theta \in \Theta}$ ,  $\Theta \subseteq \mathbb{R}$  and  $T : \mathcal{X} \rightarrow \mathbb{R}$ . The family  $\{(\mathbf{P}_\theta)_{X_1} : \theta \in \Theta\}$  of distributions of  $X_1$  is a *class with monotone likelihood ratio in  $T$*  if for every  $\theta_0, \theta_1 \in \Theta$  with  $\theta_0 < \theta_1$ , there exists a monotone increasing function  $H_{\theta_0, \theta_1} : \mathbb{R} \rightarrow [0, \infty]$  such that

$$L_{\theta_0, \theta_1}^{(n)}(x_1, \dots, x_n) = \frac{f_{\theta_1}^{(n)}(x_1, \dots, x_n)}{f_{\theta_0}^{(n)}(x_1, \dots, x_n)} = H_{\theta_0, \theta_1}(T(x)). \tag{7.37}$$

As we saw earlier, the family  $\{Ber(\theta) : \theta \in (0, 1)\}$  has is a class with monotone likelihood ratio in  $T$  for  $T(x) = \sum_{i=1}^n x_i$ .

Another general class of examples are 1-parameter exponential families: If the distributions of  $X_1$  under  $\mathbf{P}_\theta$  form a 1-parameter exponential family, i.e.

$$f_\theta(x) = C(\theta) \exp(\eta(\theta)T(x)) h(x),$$

and  $\eta(\cdot)$  is monotone increasing, then we have this family has a monotone likelihood ratio in  $\bar{T}(x) = \sum_{i=1}^n T(x_i)$ . We can now obtain a more specific version of the Neyman-Pearson lemma in the situation of classes with a monotone likelihood ratio.

**Theorem 7.11.** *Suppose that  $\mathbf{X} = (X_1, \dots, X_n)$  are i.i.d. data and  $T(\mathbf{X})$  is a statistic. Assume that the family  $\{(\mathbf{P}_\theta)_{X_1} : \theta \in \Theta\}$  forms a class with monotone likelihood ratio in  $T$  with  $\Theta \subseteq \mathbb{R}$ . Then the test*

$$\phi^*(\mathbf{X}) = \mathbb{1}_{\{T(\mathbf{X}) \geq k_\alpha^*\}} \quad (7.38)$$

is a UMP test at level  $\alpha \in (0, 1)$  where  $\alpha = \mathbf{P}_{\theta_0}[T(\mathbf{X}) \geq k_\alpha^*]$ , for

$$H_0 : \Theta_0 = (-\infty, \theta_0] \cap \Theta, \quad \text{against} \quad H_1 : \Theta_1 = (\theta_0, \infty) \cap \Theta. \quad (7.39)$$

### 7.3 The $Z$ - and $t$ -tests

We will now give an application of Theorem 7.11, giving the so called  $Z$ -test:

**Theorem 7.12.** *Let  $X_1, \dots, X_n$  be i.i.d. real random variables with  $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ . For the hypothesis*

$$\begin{aligned} H_0 : \mu &= \mu_0, \text{ against} \\ H_1 : \mu &> \mu_0, \end{aligned} \quad (7.40)$$

(with  $\sigma^2 \in (0, \infty)$  fixed!) the UMP test at level  $\alpha \in (0, 1)$  is given by

$$\phi^*(\mathbf{X}) = \begin{cases} 1, & \bar{X}_n \geq k_\alpha^*, \\ 0, & \bar{X}_n < k_\alpha^*, \end{cases} \quad (7.41)$$

with  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ , where  $k_\alpha^*$  is given by

$$k_\alpha^* = \mu_0 + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}, \quad \Phi(u_{1-\alpha}) = 1 - \alpha, \quad (7.42)$$

(i.e.  $u_{1-\alpha}$  is the  $(1 - \alpha)$ -quantile of the  $\mathcal{N}(0, 1)$ -distribution).

*Proof.* We want to apply Theorem 7.11. Note that under  $\mathbf{P}_{\mu_0}$ , we have

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \sim \mathcal{N}(\mu_0, \frac{\sigma^2}{n}), \quad (7.43)$$



and under  $\mathbf{P}_\mu$ , we have

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}). \quad (7.44)$$

We now calculate the likelihood quotient:

$$\begin{aligned} L_{\mu_0, \mu}^{(n)}(x_1, \dots, x_n) &= \frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right)}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu_0)^2\right)} \\ &= \exp\left(\frac{1}{\sigma^2}(\mu - \mu_0) \sum_{i=1}^n x_i - \frac{1}{2\sigma^2}(\mu^2 - \mu_0^2)n\right) \\ &= K(\mu, \mu_0, \sigma^2) \exp\left(\frac{n}{\sigma^2}(\mu - \mu_0)\bar{x}_n\right). \end{aligned} \quad (7.45)$$

Since  $\mu > \mu_0$ , we can write  $L_{\mu_0, \mu}^{(n)}(x_1, \dots, x_n) = H_{\mu_0, \mu}^{(n)}(T(x_1, \dots, x_n))$  with  $T(x_1, \dots, x_n) = \bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$  where the function

$$H_{\mu_0, \mu}^{(n)}(t) = K(\mu, \mu_0, \sigma^2) \exp\left(\frac{n}{\sigma^2}(\mu - \mu_0)t\right), \quad (7.46)$$

is monotone increasing in  $t$ , and we can apply Theorem 7.11. Therefore we can consider the test

$$\phi^*(X_1, \dots, X_n) = \begin{cases} 1 & \bar{X}_n \geq k_\alpha^*, \\ 0 & \bar{X}_n < k_\alpha^*, \end{cases} \quad (7.47)$$

where we set

$$\begin{aligned} \mathbf{P}_{\mu_0}[\bar{X}_n \geq k_\alpha^*] &= \alpha, \text{ which implies} \\ \mathbf{P}_{\mu_0}[\bar{X}_n \geq k_\alpha^*] &= \mathbf{P}_{\mu_0}\left[\sqrt{n} \frac{\bar{X}_n - \mu_0}{\sigma} \geq \sqrt{n} \frac{k_\alpha^* - \mu_0}{\sigma}\right] = 1 - \Phi\left(\sqrt{n} \frac{k_\alpha^* - \mu_0}{\sigma}\right) = \alpha. \end{aligned} \quad (7.48)$$

The latter condition of course means that

$$\sqrt{n} \frac{k_\alpha^* - \mu_0}{\sigma} = \Phi^{-1}(1 - \alpha) = u_{1-\alpha} \quad \Leftrightarrow \quad k_\alpha^* = \mu_0 + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}. \quad (7.49)$$

□

*Example 7.13.* Suppose that the maximal concentration of a certain chemical substance in a water reservoir is 1100 ppm. A company is discharging water into the reservoir, and it is suspected that the concentration exceeds the threshold. To investigate this, a test is performed on 8 different days (far enough apart for the measurements to be independent). The measuring device has a known error of  $\sigma = 55$  ppm. We can therefore assume the measurements to be normally distributed with parameter  $\mu$  and  $\sigma = 55$ . Assuming that the observed measurements in ppm are: 1090, 1150, 1170, 1080, 1210, 1230, 1180 and 1140, can we reject

$$H_0 : \mu = \mu_0 = 1100,$$

in favor of

$$H_1 : \mu > \mu_0 = 1100,$$

at a 5% level of significance? For this, we calculate

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i = 1156.25,$$

and

$$k_{0.05}^* = \mu_0 + \frac{\sigma}{\sqrt{n}} u_{0.95} = 1100 + \frac{55}{\sqrt{8}} \cdot 1.645 = 1131.99,$$

where we have used that  $u_{0.95} = 1.645$ . Since  $\bar{X}_n > k_{0.05}^*$ , we can indeed reject  $H_0$ .

In the above example, we assumed that  $\sigma$  is known. This is not the case in most applications. The idea is then to replace  $\sigma^2$  by the estimator  $S_n^2$ . For this reason, we state the (one-sample) *t*-test.

**Theorem 7.14.** *Let  $X_1, \dots, X_n$  be i.i.d. real random variables with  $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ . A test at level  $\alpha \in (0, 1)$  for the hypothesis*

$$\begin{aligned} H_0 : \mu &= \mu_0, \text{ against} \\ H_1 : \mu &> \mu_0, \end{aligned} \tag{7.50}$$

is given by

$$\phi^*(\mathbf{X}) = \begin{cases} 1, & \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \geq k_\alpha^*, \\ 0, & \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} < k_\alpha^*, \end{cases} \tag{7.51}$$

with  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$  and  $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ , and  $k_\alpha^*$  is given by

$$k_\alpha^* = t_{n-1, 1-\alpha}, \tag{7.52}$$

where  $t_{k, \beta}$  is the  $\beta$ -quantile of the  $t_k$ -distribution.

*Proof.* Note that  $\frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \sim t_{n-1}$  under  $\mathbf{P}_{(\mu_0, \sigma^2)}$  (independently of the value of  $\sigma^2 > 0$ ) by Theorem 2.7, (ii). Therefore,  $\mathbf{P}_{(\mu_0, \sigma^2)}[\frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \geq k_\alpha^*] = 1 - \mathbf{P}_{(\mu_0, \sigma^2)}[\frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} < k_\alpha^*] = \alpha$  if and only if  $k_\alpha^* = t_{n-1, 1-\alpha}$ .  $\square$

In other words: If  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ , and we test  $H_0 : \mu = \mu_0$  against  $H_1 : \mu > \mu_0$ , then

- if  $\sigma^2$  is known, we should reject  $H_0$  if  $\bar{X}_n \geq \mu_0 + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}$  (Theorem 7.12),
- if  $\sigma^2$  is unknown, we should reject  $H_0$  if  $\bar{X}_n \geq \mu_0 + \frac{S_n}{\sqrt{n}} t_{n-1, 1-\alpha}$  (Theorem 7.14).

Note that if we were to test  $H_0 : \mu = \mu_0$  against  $H_1 : \mu < \mu_0$ , the corresponding conditions would become  $\bar{X}_n \leq \mu_0 - \frac{\sigma}{\sqrt{n}} u_{1-\alpha}$  (if  $\sigma^2$  is known) or  $\bar{X}_n \leq \mu_0 - \frac{S_n}{\sqrt{n}} u_{1-\alpha}$ , respectively.

---

*End of Lecture 22*

## 7.4 Two-sided tests

Consider a situation where  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(\theta)$ , where  $\theta \in (0, 1)$  is unknown and we want to test

$$H_0 : \theta = \theta_0, \quad \text{against} \quad H_1 : \theta \neq \theta_0, \quad (7.53)$$

for some  $\theta_0 \in (0, 1)$ . This formally corresponds to  $\Theta_0 = \{\theta_0\}$  and  $\Theta_1 = (0, 1) \setminus \{\theta_0\}$ . One can show that in this situation, a UMP test *does not exist*<sup>3</sup>.

Nevertheless, we can still construct reasonable tests for (7.53), by “splitting” the critical region into a part of large values of  $\sum_{i=1}^n X_i$  and a part of small values of  $\sum_{i=1}^n X_i$ , each having probability  $\lesssim \frac{\alpha}{2}$ .

*Example 7.15.* Consider again the set-up of Example 7.1, but now suppose we want to test

$$\begin{aligned} H_0 : & \quad \text{drug } B \text{ has the same efficacy 0.6 as drug } A, \\ H_1 : & \quad \text{drug } B \text{ has a different efficacy than drug } A, \end{aligned}$$

which puts us into the framework of (7.53), and we again suppose that  $n = 100$  and  $\alpha \lesssim 0.01$ . We now look for a test of the form

$$\phi_{\text{two-sided}}(X_1, \dots, X_n) = \begin{cases} 1, & \sum_{i=1}^n X_i \in \{0, \dots, k_{\alpha,L}^*\} \cup \{k_{\alpha,R}^*, \dots, n\}, \\ 0, & \sum_{i=1}^n X_i \in \{k_{\alpha,L}^* + 1, \dots, k_{\alpha,R}^* - 1\}. \end{cases} \quad (7.54)$$

To achieve a test at level  $\alpha \lesssim 0.01$ , we want

$$\alpha = \mathbf{P}_{\theta_0}[\phi_{\text{two-sided}}(X_1, \dots, X_n) = 1] = \sum_{k=0}^{k_{\alpha,L}^*} \binom{n}{k} \theta_0^k (1 - \theta_0)^{n-k} + \sum_{k=k_{\alpha,R}^*}^n \binom{n}{k} \theta_0^k (1 - \theta_0)^{n-k}. \quad (7.55)$$

We now look for values  $k_{\alpha,L}^*$  and  $k_{\alpha,R}^*$ , such that both parts of the sum are just below  $\frac{0.01}{2}$ . We find (setting  $T = \sum_{i=1}^n X_i$ ):

$$\begin{aligned} & \vdots \\ \mathbf{P}_{\theta_0} [T \in \{0, 1, \dots, 46\}] &= 0.0032, \\ \mathbf{P}_{\theta_0} [T \in \{0, 1, \dots, 47\}] &= 0.0058, \\ & \vdots \\ \mathbf{P}_{\theta_0} [T \in \{72, 73, \dots, 100\}] &= 0.0084, \\ \mathbf{P}_{\theta_0} [T \in \{73, 74, \dots, 100\}] &= 0.0046, \\ & \vdots \end{aligned} \quad (7.56)$$

<sup>3</sup>Essentially, a UMP test has to be a Neyman-Pearson test: if there were a UMP test for  $H_0$  against  $H_1$ , it would necessarily also be a UMP test for  $H_0 : \theta = \theta_0$  against any  $\tilde{H}_1 : \theta = \theta_1$  whenever  $\theta_1 \neq \theta_0$ . This test would then reject at high values for  $\sum_{i=1}^n X_i$  if  $\theta_1 > \theta_0$ , but also at low values for  $\sum_{i=1}^n X_i$  if  $\theta_1 < \theta_0$ , and thus cannot have the form of a Neyman-Pearson test.

so with the choice  $k_{\alpha,L}^* = 46$  and  $k_{\alpha,R}^* = 73$ , we find  $\mathbf{P}_{\theta_0}[T \in \{0, \dots, k_{\alpha,L}^*\} \cup \{k_{\alpha,R}^*, \dots, n\}] = 0.0032 + 0.0047 = 0.0079 \lesssim \alpha$ .

Let us also state the corresponding two-sided version of the  $Z$ - and  $t$ -tests.

**Example 7.16.** Let  $X_1, \dots, X_n$  be i.i.d. real random variables with  $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ . Consider testing the hypothesis

$$\begin{aligned} H_0 : \mu &= \mu_0, \text{ against} \\ H_1 : \mu &\neq \mu_0. \end{aligned} \tag{7.57}$$

(i) If with  $\sigma^2 \in (0, \infty)$  is known, a two-sided test at level  $\alpha \in (0, 1)$  is given by

$$\phi^*(\mathbf{X}) = \begin{cases} 1, & |\bar{X}_n - \mu_0| \geq k_\alpha^*, \\ 0, & |\bar{X}_n - \mu_0| < k_\alpha^*, \end{cases} \tag{7.58}$$

with  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ , where  $k_\alpha^*$  is given by

$$k_\alpha^* = \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}, \quad \Phi(u_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}, \tag{7.59}$$

(ii) If with  $\sigma^2 \in (0, \infty)$  is unknown, a two-sided test at level  $\alpha \in (0, 1)$  is given by

$$\phi^*(\mathbf{X}) = \begin{cases} 1, & |\bar{X}_n - \mu_0| \geq K_\alpha^*, \\ 0, & |\bar{X}_n - \mu_0| < K_\alpha^*, \end{cases} \tag{7.60}$$

with  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ ,  $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ , where  $K_\alpha^*$  is given by

$$K_\alpha^* = \frac{S_n}{\sqrt{n}} t_{n-1, 1-\frac{\alpha}{2}}, \quad t_{n-1, 1-\frac{\alpha}{2}} \text{ is the } (1 - \frac{\alpha}{2})\text{-quantile of } t_{n-1}. \tag{7.61}$$

To conclude this section, we present a general class of *asymptotic tests*. We start with a definition.

**Definition 7.17.** Consider a parametric statistical model with data  $\mathbf{X} = (X_1, \dots, X_n) \in \mathcal{X}_n$  and probability measures  $(\mathbf{P}_\theta)_{\theta \in \Theta}$ . We say that for a sequence  $\phi_n : \mathcal{X}_n \rightarrow \{0, 1\}$ , the expression  $\phi_n \circ \mathbf{X}$  is an *asymptotic test at level  $\alpha$*  for

$$H_0 : \theta \in \Theta_0, \quad \text{against} \quad H_1 : \theta \in \Theta_1, \tag{7.62}$$

if we have

$$\limsup_{n \rightarrow \infty} \sup_{\theta \in \Theta_0} \mathbf{P}_\theta[\phi_n(\mathbf{X}) = 1] = \limsup_{n \rightarrow \infty} \sup_{\theta \in \Theta_0} G_{\phi_n}(\theta) \leq \alpha. \tag{7.63}$$

The sequence of tests  $(\phi_n \circ \mathbf{X})_{n \in \mathbb{N}}$  is called *consistent* if

$$\lim_{n \rightarrow \infty} G_{\phi_n}(\theta) = 1, \quad \text{for all } \theta \in \Theta_1. \tag{7.64}$$

We now give a general two-sided test based on an approximation by the normal distribution, which asymptotically attains level  $\alpha$ .

**Definition 7.18.** Consider a parametric statistical model with data  $\mathbf{X} = (X_1, \dots, X_n) \in \mathcal{X}_n$  and probability measures  $(\mathbf{P}_\theta)_{\theta \in \Theta}$ . We assume that  $\hat{\theta}_n$  is an asymptotically normal estimator for  $\theta_0$ , i.e. under  $\mathbf{P}_{\theta_0}$  one has

$$\frac{\hat{\theta}_n - \theta_0}{\hat{\sigma}_n} \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, 1), \quad (7.65)$$

with some estimators  $\hat{\sigma}_n$ . Then the *Wald test* for

$$H_0 : \theta = \theta_0, \quad \text{against } H_1 : \theta \neq \theta_0 \quad (7.66)$$

at level  $\alpha \in (0, 1)$  is given by

$$\phi_W(\mathbf{X}) = \begin{cases} 1, & \left| \frac{\hat{\theta}_n - \theta_0}{\hat{\sigma}_n} \right| \geq u_{1-\frac{\alpha}{2}}, \\ 0, & \left| \frac{\hat{\theta}_n - \theta_0}{\hat{\sigma}_n} \right| < u_{1-\frac{\alpha}{2}} \end{cases} \quad (7.67)$$

where  $u_{1-\frac{\alpha}{2}}$  is the  $(1 - \frac{\alpha}{2})$ -quantile of the  $\mathcal{N}(0, 1)$ -distribution.

A typical situation in the setting above is that  $\hat{\theta}_n$  is the MLE,  $\hat{\sigma}_n = \frac{1}{\sqrt{nI(\hat{\theta}_n)}}$  (with  $I(\hat{\theta}_n)$  the Fisher information) and the conditions of Theorem 6.7 are fulfilled.

**Lemma 7.19.** *The Wald test is an asymptotic test at level  $\alpha$ .*

*Proof.* We have that

$$\mathbf{P}_{\theta_0} \left[ \left| \frac{\hat{\theta}_n - \theta_0}{\hat{\sigma}_n} \right| \geq u_{1-\frac{\alpha}{2}} \right] \xrightarrow[n \rightarrow \infty]{} 1 - \left( \Phi(u_{1-\frac{\alpha}{2}}) - \Phi(-u_{1-\frac{\alpha}{2}}) \right) = 1 - \left( 1 - \frac{\alpha}{2} - \frac{\alpha}{2} \right) = \alpha. \quad (7.68)$$

□

In most cases,  $\hat{\sigma}_n$  converges to 0 in probability, and one can show that the Wald test is also consistent.

## 7.5 $p$ -values

So far, we have seen how to construct tests and to decide whether or not a hypothesis  $H_0$  can be rejected at level  $\alpha$  based on the observed data  $\mathbf{X}$ . However, one can give more information than the decision of the test based on observing the data.

**Definition 7.20.** Suppose that for every  $\alpha \in \mathcal{A} \subseteq (0, 1)$ , we have a test  $\phi_\alpha \circ \mathbf{X}$  at level  $\alpha$  such that

$$\phi_\alpha(\mathbf{X}) = 1 \quad \Leftrightarrow \quad T(\mathbf{X}) \in C_\alpha, \quad (7.69)$$

where  $T(\mathbf{X})$  is some statistic. We say that the *p-value* is given by

$$p\text{-value} = \inf \{ \alpha \in \mathcal{A} : T(\mathbf{X}) \in C_\alpha \}. \quad (7.70)$$

Let us briefly discuss this notion: The  $p$ -value can be interpreted as the smallest value for  $\alpha$ , such that we can reject  $H_0$  based on the observation  $\mathbf{X}$ . This is best illustrated in an example.

*Example 7.21.* Consider again Example 7.1. We computed in (7.35) some values for  $\mathbf{P}_{\theta_0}[T(\mathbf{X}) \geq k]$  where  $T(\mathbf{X}) = \sum_{i=1}^n X_i \sim \text{Bin}(n, \theta_0)$  with  $n = 100$  and  $\theta_0 = 0.6$ . In particular, we have (UMP) tests

$$\begin{aligned}\phi_{0.0248}(\mathbf{X}) &= \mathbb{1}_{\{T(\mathbf{X}) \geq 70\}}, \\ \phi_{0.0148}(\mathbf{X}) &= \mathbb{1}_{\{T(\mathbf{X}) \geq 71\}}, \\ \phi_{0.0084}(\mathbf{X}) &= \mathbb{1}_{\{T(\mathbf{X}) \geq 72\}}, \\ \phi_{0.0046}(\mathbf{X}) &= \mathbb{1}_{\{T(\mathbf{X}) \geq 73\}},\end{aligned}\tag{7.71}$$

at levels  $\mathcal{A} = \{\dots, 0.0248, 0.0148, 0.0084, 0.0046, \dots\}$  and so on. Suppose now we observe  $T(\mathbf{X}) = \sum_{i=1}^n X_i = 73$ . Then

$$p\text{-value} = 0.0046 = \mathbf{P}_{\theta_0}[T(\mathbf{X}) \geq 73].\tag{7.72}$$

This means, if we observe  $\sum_{i=1}^n X_i = 73$ , we can clearly reject  $H_0$  at significance level  $\alpha = 0.01$ , but in fact we could also reject it at level 0.0046. In other words: The  $p$ -value describes the probability that an outcome at least as extreme as the observed one is found if  $H_0$  is true.

---

*End of Lecture 23*

## 7.6 Pearson's $\chi^2$ -test

So far we have only mainly parametric statistical tests in which the null and alternative hypotheses concerned a single scalar parameter. Here we give an important example of an asymptotic test in which the hypotheses concern multinomial distributions.

*Example 7.22.* Suppose that  $N_j$  stand for the number of observations of a given “type”  $j$ , where  $j = 1, \dots, r$ , in which we expect that type  $j$  occurs with a certain probability  $p_j$  (with  $\sum_{j=1}^r p_j = 1$ ). Concrete examples:

- *Mendel's peas:* In his famous experiment, Georg Mendel studied the relative frequencies of certain phenotypes of peas, which in shape could be either round ( $R$ ) or wrinkled ( $r$ ), and in color could be either yellow ( $Y$ ) or green ( $y$ ). Since both round and yellow are dominant alleles (and the information for shape and color is contained in different chromosomes), the F2 generation of purely round-yellow and wrinkled-green peas should have a distribution

$$\begin{array}{cccc} 9 & : & 3 & : & 3 & : & 1 \\ RY & : & Ry & : & rY & : & ry \end{array}$$

- *Literature analysis:* To study whether some text can be attributed to a certain author, one can investigate the frequency at which certain words occur.

In the first situation, one could investigate whether the distribution of a certain number of peas does indeed follow the distribution with probabilities  $p_1 = \frac{9}{16}, p_2 = \frac{3}{16}, p_3 = \frac{3}{16}, p_4 = \frac{1}{16}$ . In the second example, assume that it is known that a certain author uses two certain words with frequencies  $\frac{1}{200}$  and  $\frac{1}{500}$ , then one could test whether a given text follows a distribution with  $p_1 = \frac{1}{100}, p_2 = \frac{1}{500}$  and  $p_3 = \frac{594}{600}$  (all other words).

To make this precise, we let  $X_1, \dots, X_n$  be i.i.d. vectors in with values in  $\{0, 1\}^r$ ,  $r \in \mathbb{N}$  (standing for the number of “types”), and

$$X_i = (X_{i,1}, \dots, X_{i,r})^\top, \quad \mathbf{P}[(X_{i,1}, \dots, X_{i,r})^\top = (0, \dots, 0, \underbrace{1}_{\text{place } j}, 0, \dots, 0)^\top] = p_j, \quad (7.73)$$

where  $\sum_{j=1}^r p_j = 1$ . Consider

$$\sum_{i=1}^n X_i =: (N_1, \dots, N_r)^\top. \quad (7.74)$$

One can easily show that for every  $n_1, \dots, n_r \in \mathbb{N}_0$  with  $\sum_{j=1}^r n_j = n$ :

$$\mathbf{P}[N_1 = n_1, \dots, N_r = n_r] = \frac{n!}{n_1! \dots n_r!} p_1^{n_1} \cdot \dots \cdot p_r^{n_r} \quad (7.75)$$

(Multinomial distribution). We want to test the hypothesis

$$H_0 : p_j = \pi_j (j = 1, \dots, r), \text{ where the } \pi_j \text{ are known and } \sum_{j=1}^r \pi_j = 1, \text{ against} \quad (7.76)$$

$$H_1 : p_j \neq \pi_j \text{ for some } j = 1, \dots, r,$$

(this is a parametric setup where  $\Theta = \{(p_j)_{j=1, \dots, r} \in [0, 1]^r : \sum_{j=1}^r p_j = 1\}$  and  $\Theta_0 = \{(\pi_1, \dots, \pi_r)^\top\}$ ,  $\Theta_1 = \Theta \setminus \Theta_0$ ).

The following theorem gives *Pearson's  $\chi^2$ -test*.

**Theorem 7.23.** For  $\alpha \in (0, 1)$ , the test

$$\phi(\mathbf{X}) = \begin{cases} 1, & \sum_{i=1}^r \frac{(N_i - n\pi_i)^2}{n\pi_i} \geq \chi_{r-1, 1-\alpha}^2, \\ 0, & \sum_{i=1}^r \frac{(N_i - n\pi_i)^2}{n\pi_i} < \chi_{r-1, 1-\alpha}^2, \end{cases} \quad (7.77)$$

where  $\chi_{k, \beta}^2$  denotes the  $\beta$ -quantile of the  $\chi^2$ -distribution with  $k$  degrees of freedom is an asymptotic test at level  $\alpha$  for (7.76).

*Proof.* We set  $U_i = \frac{N_i - n\pi_i}{\sqrt{n\pi_i}}$  and consider the random vector  $U = (U_1, \dots, U_r)^\top$  as well as the deterministic vector  $P_1 = (\sqrt{\pi_1}, \dots, \sqrt{\pi_r})^\top$  (note that  $P_1 \neq 0$ ). Now extend  $P_1$  to an orthonormal basis  $P_1, \dots, P_r$  of  $\mathbb{R}^r$ . We have

$$P_1^\top U = \sum_{i=1}^r \frac{N_i - n\pi_i}{\sqrt{n}} = 0, \quad (7.78)$$

therefore

$$\sum_{i=1}^r \frac{(N_i - n\pi_i)^2}{n\pi_i} = U^\top U = U^\top P P^\top U = \sum_{j=2}^r (P_j^\top U)^2, \quad (7.79)$$

where we have

$$(P_j^\top U)_{j=2,\dots,r} = \sqrt{n} \frac{1}{n} \sum_{i=1}^n \underbrace{\left( P_j^\top \left( \frac{X_{i,1} - \pi_1}{\sqrt{\pi_1}}, \dots, \frac{X_{i,r} - \pi_r}{\sqrt{\pi_r}} \right)^\top \right)}_{=: Y_i} \quad (7.80)$$

Note that this expression is a vector (in  $\mathbb{R}^{r-1}$ ). Now under  $H_0$ , we have  $\mathbf{E}[Y_i] = 0$ , and moreover

$$\begin{aligned} \text{Cov}[X_{ij}, X_{ik}] &= \mathbf{E}[X_{ij} X_{ik}] - \pi_j \pi_k = \pi_j \delta_{jk} - \pi_j \pi_k, \quad \text{so} \\ \Sigma(Y_i)_{j,k} &= P_{j+1}^\top \Sigma \left( \left( \frac{X_{i,1}}{\sqrt{\pi_1}}, \dots, \frac{X_{i,r}}{\sqrt{\pi_r}} \right)^\top \right) P_{k+1} \\ &= P_{j+1}^\top (I_{r \times r} - P_1 P_1^\top) P_{k+1} = \delta_{jk}. \end{aligned} \quad (7.81)$$

Now set  $C = \{x \in \mathbb{R}^{r-1} : \sum_{j=1}^{r-1} x_j^2 \leq c\}$  (the closed ball in the Euclidean norm in  $\mathbb{R}^{r-1}$ ), then

$$\begin{aligned} \mathbf{P} \left[ \sum_{i=1}^r \frac{(N_i - n\pi_i)^2}{n\pi_i} \leq c \right] &= \mathbf{P} \left[ \sum_{j=2}^r (P_j^\top U)^2 \leq c \right] = \mathbf{P} \left[ (P_2^\top U, \dots, P_r^\top U)^\top \in C \right] \\ &\xrightarrow[n \rightarrow \infty]{(\star)} \mathbf{P} \left[ (Z_1, \dots, Z_{r-1})^\top \in C \right] = \mathbf{P} \left[ \sum_{j=1}^{r-1} Z_j^2 \leq c \right], \end{aligned} \quad (7.82)$$

where  $(Z_1, \dots, Z_{r-1}) \sim \mathcal{N}_{r-1}(0, I_{(r-1) \times (r-1)})$ , and we used in  $(\star)$  the multivariate central limit theorem (Theorem 2.4) for the random vectors  $(Y_i)_{i=1}^n$ . Now remember that  $\sum_{j=1}^{r-1} Z_j^2 \sim \chi_{r-1}^2$ , so that (7.82) gives us

$$\sum_{i=1}^r \frac{(N_i - n\pi_i)^2}{n\pi_i} \xrightarrow[n \rightarrow \infty]{d} \chi_{r-1}^2. \quad (7.83)$$

In particular, we have that

$$\mathbf{P}[\phi(\mathbf{X}) = 1] = \mathbf{P} \left[ \sum_{i=1}^r \frac{(N_i - n\pi_i)^2}{n\pi_i} \geq \chi_{r-1, 1-\alpha}^2 \right] \xrightarrow[n \rightarrow \infty]{} \alpha. \quad (7.84)$$

□

---

End of Lecture 24



## 8 A brief introduction to Bayesian statistics

(Reference: [6, Chapter 11])

In the previous chapters, we discussed classical (or “frequentist”) statistics. Here we want to give a very brief introduction to a method known as *Bayesian statistics*. In essence, this will correspond to treating parameters themselves as random variables which change after observing data (“Bayesian updating”).

### 8.1 The Bayesian method: Prior and posterior distributions

So far, we have considered parametric statistical models (see Definition 3.1) and our goal was

*Obtain information on the unknown (but deterministic) quantity  $\theta \in \Theta$  using data  $\mathbf{X}$ .* (8.1)

This is plausible if we treat probabilities as relative frequencies (hence “frequentist” statistics) and parameters as fixed unknown constants. For instance, if a coin has a probability  $\theta \in (0, 1)$  to land on heads, we have by the law of large numbers

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow[n \rightarrow \infty]{\mathbf{P}_\theta} \theta. \quad (8.2)$$

So, if we observe “sufficiently many data points”, we will eventually come close to the true parameter  $\theta$ .

We now adopt a different point of view, which may be summarized as follows:

- we treat probabilities not as relative frequencies, but instead of “degrees of belief”;
- parameters are random variables themselves, which change their value after observing more data.

The way we update our belief about the value of the parameter is given by Bayes’ theorem.

**Definition 8.1.** Let  $(\Omega, \mathcal{F}, \mathbf{P})$  be a probability space. We consider a random variable  $\theta$  with values in  $\Theta \subseteq \mathbb{R}^d$  and random variables  $X_1, X_2, \dots, X_n$ , where  $\mathbf{X} = (X_1, \dots, X_n) \in \mathcal{X}$ .

- (i) The distribution of  $\theta$  is called the *prior distribution*.
- (ii) The conditional distribution of  $\mathbf{X}$  given  $\theta$  is called the *likelihood*.
- (iii) The conditional distribution of  $\theta$  given  $\mathbf{X}$  is called the *posterior distribution*.

Suppose for simplicity that both  $\theta$  and  $\mathbf{X}$  are discrete random variables. Then we have by Bayes' theorem (1.36)

$$\mathbf{P}[\theta = \vartheta | \mathbf{X} = x] = \frac{\mathbf{P}[\mathbf{X} = x | \theta = \vartheta] \cdot \mathbf{P}[\theta = \vartheta]}{\sum_{\vartheta' \in \Theta} \mathbf{P}[\mathbf{X} = x | \theta = \vartheta'] \cdot \mathbf{P}[\theta = \vartheta']}, \quad \theta \in \Theta. \quad (8.3)$$

Note that the expression in the denominator is nothing else than  $\mathbf{P}[\mathbf{X} = x]$  (in particular, it does not depend on  $\vartheta$ ). We see that

$$\underbrace{\mathbf{P}[\theta = \vartheta | \mathbf{X} = x]}_{\text{posterior}} \propto \underbrace{\mathbf{P}[\mathbf{X} = x | \theta = \vartheta]}_{\text{likelihood}} \cdot \underbrace{\mathbf{P}[\theta = \vartheta]}_{\text{prior}}. \quad (8.4)$$

This last observation is very helpful in calculations. Note that the proportionality constant is a normalization factor that can be recovered if needed (recall that  $\mathbf{P}[\cdot | \mathbf{X} = x]$  must be a probability measure, so it must sum up to one).

*Remark 8.2.* (i) In the case that the prior has a continuous distribution, one has to replace the sums by integrals in the above calculation. The formula

$$f_{\theta | \mathbf{X} = x}(\vartheta) \propto f_{\mathbf{X} | \theta = \vartheta}(x) \cdot f_{\theta}(\vartheta), \quad (8.5)$$

which is the general version of (8.4) remains true in all cases (here  $f$  stands either for a probability mass function or a probability density function).

(ii) Various notations for the prior, likelihood and posterior. A particularly simple notation is to write densities / probability mass functions as  $[\cdot]$  and use  $|$  for conditioning in this notation. For instance, equation (8.5) becomes

$$[\vartheta | x] \propto [x | \vartheta] \cdot [\vartheta]. \quad (8.6)$$

Bayes theorem reads

$$[\vartheta | x] = \frac{[x | \vartheta] \cdot [\vartheta]}{[x]}. \quad (8.7)$$

We introduce a distribution that is relevant for a later example.

**Definition 8.3.** The *Beta-distribution* with parameters  $\alpha, \beta > 0$  is a continuous distribution characterized by the density

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \mathbb{1}_{(0,1)}(x), \quad (8.8)$$

where the normalization is given by

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \quad (8.9)$$

If  $X$  follows a Beta-distribution with parameters  $\alpha, \beta$ , we write  $X \sim Be(\alpha, \beta)$ . Note that in this case

$$\mathbf{E}[X] = \frac{\alpha}{\alpha + \beta}. \quad (8.10)$$

The Beta-distribution often comes up in Bayesian statistics. Here is a concrete example.

*Example 8.4.* A lightbulb manufacturer claims that at most 1% of their produced lightbulbs are defective. We try to calculate the probability that this claim is incorrect based on the observation that  $X = x$  lightbulbs out of  $n$  are defective. For this, we choose the prior

$$p \sim \mathcal{U}([0, 1]) \quad (8.11)$$

(note that this coincides with  $Be(1, 1)$ ). If the lightbulbs tested are independent, we have  $X|p \sim Bin(n, p)$ . With our notation in Remark 8.2, (ii), we have that for any  $\alpha, \beta > 0$ :

$$[p|x] \propto [x|p] \cdot [p] \propto p^x(1-p)^{n-x} \cdot p^{\alpha-1}(1-p)^{\beta-1} = p^{(\alpha+x)-1}(1-p)^{(\beta+n-x)-1}. \quad (8.12)$$

Since this is proportional to the density of the  $Be(\alpha + x, \beta + n - x)$ -distribution, we see that the posterior distribution of  $p$  is

$$p|x \sim Be(\alpha + x, \beta + n - x). \quad (8.13)$$

Also note that the posterior expectation is given by

$$\mathbf{E}[p|X = x] = \frac{\alpha + x}{\alpha + \beta + n} = \frac{\alpha + \beta}{\alpha + \beta + n} \cdot \frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \cdot \frac{x}{n}. \quad (8.14)$$

This can be interpreted as a weighted average of the priori expectation  $\mathbf{E}[p] = \frac{\alpha}{\alpha + \beta}$  and the MLE (based on the observation  $x$ ). If  $n$  is large, the latter part dominates. In our specific case ( $\alpha = \beta = 1$ ) we see that

$$\mathbf{E}[p|X = x] = \frac{x + 1}{n + 2}. \quad (8.15)$$

The probability that  $p > 0.01$  based on the observation is given by

$$\mathbf{P}[p > 0.01|X = x] = \int_{0.01}^1 [p|x]dp = \int_{0.01}^1 \frac{1}{B(1+x, 1+n-x)} p^x(1-p)^{n-x} dp, \quad (8.16)$$

which has to be calculated numerically. For instance, if  $n = 100$  and  $x = 1$ , we obtain

$$\mathbf{P}[p > 0.01|X = x] \approx 0.7321. \quad (8.17)$$

Suppose now that we observe “new” data, i.e. observations  $X_1$  and  $X_2$  are sequentially available and independent given  $\theta^1$ . We can then “update” our posterior distribution  $[\theta|x_1]$  by treating it as a new prior distribution. This is known as *Bayesian updating*:

$$\text{The posterior } [\theta|x_1] \text{ becomes the prior when calculating } [\theta|x_1, x_2]. \quad (8.18)$$

More formally:

$$[\theta|x_1, x_2] \propto [x_1, x_2|\theta] \cdot [\theta] = [x_2|\theta][x_1|\theta][\theta] \propto [x_2|\theta] \cdot [\theta|x_1]. \quad (8.19)$$

For instance in the example before, assume that another  $n = 100$  lightbulbs are taken (independently from the first 100), and a new observation  $X_2$  is made. We see

$$[p|x_1, x_2] \propto [x_2|p][p|x_1] \propto p^{x_2}(1-p)^{n-x_2} \cdot p^{(\alpha+x_1)-1}(1-p)^{(\beta+n-x_1)-1}. \quad (8.20)$$

We find that

$$p|x_1, x_2 \sim Be(\alpha + x_1 + x_2, \beta + 2n - x_1 - x_2). \quad (8.21)$$

---

*End of Lecture 25*

<sup>1</sup>This means that  $\mathbf{P}[X_1 \in A, X_2 \in B|\theta = \vartheta] = \mathbf{P}[X_1 \in A|\theta = \vartheta]\mathbf{P}[X_2 \in B|\theta = \vartheta]$

## 8.2 Choice of the prior

**Definition 8.5.** A class  $\mathcal{P}$  of prior distributions is called *conjugate* with respect to a family of likelihoods  $\mathcal{F} = \{[x|\theta] : \theta \in \Theta\}$  if for all  $[\theta] \in \mathcal{P}$  also  $[\theta|x] \in \mathcal{P}$ .

One can often obtain the family of conjugate priors from observing the product of likelihood and prior and requiring the posterior to have the same form.

*Example 8.6.* (i) Binomial distribution: Suppose that  $X|p \sim \text{Bin}(n, p)$ , then

$$[x|p] = \binom{n}{x} p^x (1-p)^{n-x} \propto p^x (1-p)^{n-x},$$

where  $\propto$  means proportionality in  $p$ . If we choose a density of the form  $[p] \propto p^\alpha (1-p)^\beta$ , it will conjugate to this likelihood, and this corresponds to the  $Be(\alpha, \beta)$ -distributions. In other words: The family of priors  $\{Be(\alpha, \beta) : \alpha, \beta > 0\}$  for  $p$  is conjugate to the likelihoods  $\{\text{Bin}(n, p) : p \in (0, 1)\}$ .

(ii) Geometric distribution: Suppose  $X|p \sim \text{Geo}(p)$ , so

$$[x|p] \sim p(1-p)^{x-1}.$$

Again the Beta-distributions are conjugate priors for  $p$  to this class of likelihoods.

In many situations, we would like a prior distribution which is “uniform on  $\mathbb{R}$ ” (think of  $X|\mu \sim \mathcal{N}(\mu, 1)$ ) etc., and this leads to the following definition.

**Definition 8.7.** A prior distribution  $[\theta]$  is *improper* if  $\int [\theta] d\theta = \infty$  resp.  $\sum_\theta [\theta] = \infty$ .

Note that such distributions do not actually exist in a probabilistic sense, but in some cases one can make sense of the posterior distribution, since there we only require some proportionality. We sketch an example.

*Example 8.8.* Let  $X_1, \dots, X_n | \mu \sim \mathcal{N}(\mu, 1)$  i.i.d. and take the improper prior  $[\mu] \propto 1$ . We can still try to calculate

$$[\mu|x_1, \dots, x_n] \propto [x_1, \dots, x_n|\mu][\mu] = \prod_{i=1}^n [x_i|\mu]. \quad (8.22)$$

Writing out the densities gives

$$\prod_{i=1}^n [x_i|\mu] \propto \exp\left(-\frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2\right) \propto \exp\left(-\frac{n}{2} \left(\mu - \frac{1}{n} \sum_{i=1}^n x_i\right)^2\right), \quad (8.23)$$

meaning that  $[\mu|x_1, \dots, x_n] \sim \mathcal{N}(\bar{x}_n, \frac{1}{n})$ .

So improper priors are not a problem as long as we can make sense of the posterior.

How do we construct a prior that has “no initial assumption”? It would seem that flat priors (possibly improper) would be a good choice. This is however *not* the case. Indeed, suppose that

$[\theta] \propto 1$  and consider the function  $\Psi = g(\theta)$ , where  $g$  is a bijective and smooth map. One can show that

$$f_{\Psi}(\psi) = f_{\theta}(g^{-1}(\psi)) \left| \frac{\partial g^{-1}(\psi)}{\partial \psi} \right|. \quad (8.24)$$

In other words: The distribution of a function  $\Psi = g(\theta)$  is generally *not* uniform, even if the distribution of  $\theta$  is. But if a prior  $[\theta]$  having “no information” about  $\theta$  would mean that it is flat, this should also be the case for  $\Psi = g(\theta)$ . So flat priors do *not* necessarily describe a situation in which one has no information. To solve this problem, one can consider a different prior.

**Definition 8.9.** Suppose that  $X$  is a random variable (or random vector) with density / probability mass function  $[x|\theta]$  and unknown parameter  $\theta \in \mathbb{R}$ . The *Jeffreys-prior* has the form

$$f_{\theta}(\vartheta) \propto \sqrt{I(\vartheta)}, \quad (8.25)$$

where  $I(\vartheta)$  is the Fisher information in  $\vartheta$ .

**Theorem 8.10.** The Jeffreys prior is invariant under bijective transformations of  $\theta$ . More precisely, let  $g : \mathbb{R} \rightarrow \mathbb{R}$  be bijective, then one has for  $\Psi = g(\theta)$  that

$$f_{\theta}(\vartheta) \propto \sqrt{I_{\theta}(\vartheta)} \quad \Rightarrow \quad f_{\Psi}(\psi) \propto \sqrt{I_{\Psi}(\psi)}. \quad (8.26)$$

*Proof.* For simplicity, we assume that we work in a continuum framework and that  $g$  is smooth. Let  $f_{X|\theta=\vartheta}(x)$  be the likelihood, then we can calculate the Fisher information in  $\Psi$  using the chain rule:

$$\begin{aligned} I_{\Psi}(\psi) &= \mathbf{E} \left[ \left( \frac{\partial}{\partial \psi} \log f_{X|\theta=g^{-1}(\psi)}(X) \right)^2 \right] = \mathbf{E} \left[ \left( \frac{\partial}{\partial \vartheta} \log f_{X|\theta=\vartheta}(X) \right)^2 \right] \cdot \left( \frac{\partial g^{-1}(\psi)}{\partial \psi} \right)^2 \\ &= I_{\theta}(\vartheta) \left( \frac{\partial g^{-1}(\psi)}{\partial \psi} \right)^2. \end{aligned} \quad (8.27)$$

If we choose  $f_{\theta}(\vartheta) \propto \sqrt{I_{\theta}(\vartheta)}$ , we find that

$$f_{\Psi}(\psi) = f_{\theta}(\vartheta) \left| \frac{\partial g^{-1}(\psi)}{\partial \psi} \right| \propto \left( I_{\theta}(\vartheta) \left( \frac{\partial g^{-1}(\psi)}{\partial \psi} \right)^2 \right)^{\frac{1}{2}} = I_{\Psi}(\psi), \quad (8.28)$$

where we used the transformation rule for densities.  $\square$

**Example 8.11.** (i) Let  $X|\mu \sim \mathcal{N}(\mu, \sigma^2)$  with  $\sigma^2 > 0$  known. Recall that  $I(\mu) = \frac{1}{\sigma^2}$  (see Remark 5.9). Therefore, the Jeffreys prior is

$$[\mu] \propto 1, \quad (8.29)$$

i.e. the flat (improper) prior.

(ii) Now suppose that  $X|\sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$  with  $\mu \in \mathbb{R}$  known. Now we have  $I(\sigma^2) = \frac{1}{2\sigma^4}$  (see again Remark 5.9), and therefore the Jeffreys prior is given by

$$[\sigma^2] \propto \sigma^{-2}. \quad (8.30)$$

### 8.3 The Bayes estimator

To finalize this short introduction to Bayesian statistics, we formulate a theorem that justifies why the posterior expectation is a “good estimator” for the parameter  $\theta$ .

**Definition 8.12.** Suppose that  $\theta \in \Theta \subseteq \mathbb{R}$ . For  $a \in \mathbb{R}$ , we call the function

$$L(a, \theta) = (a - \theta)^2 \quad (8.31)$$

the (quadratic) loss function.

The interpretation is simple: If  $a$  is the true parameter, the loss function increases if  $\theta$  is far away from  $a$ .

**Theorem 8.13.** The posterior expectation  $\mathbf{E}[\theta|\mathbf{X}]$  minimizes the expected loss given the data  $\mathbf{X} = x$ , i.e. minimizes

$$\mathbf{E}[L(a, \theta)|\mathbf{X} = x] = \int_{\Theta} L(a, \theta)[\theta|x]d\theta. \quad (8.32)$$

The expression  $\mathbf{E}[\theta|\mathbf{X}]$  is known as the Bayes estimator (for the quadratic loss).

*Proof.* Set

$$\frac{\partial}{\partial a} \mathbf{E}[L(a, \theta)|\mathbf{X} = x] = \frac{\partial}{\partial a} \int (a - \theta)^2 [\theta|x]d\theta = 2 \int (a - \theta) [\theta|x]d\theta = 0.$$

Solving this gives

$$a \int [\theta|x]d\theta = \int \theta [\theta|x]d\theta \quad \Leftrightarrow \quad a = \mathbf{E}[\theta|\mathbf{X} = x].$$

□

*Example 8.14.* Let  $X_1, \dots, X_n | \mu \sim \mathcal{N}(\mu, \sigma^2)$  where  $\sigma^2 > 0$  is known. Use  $\mu \sim \mathcal{N}(a, b^2)$  as a prior for  $\mu$ . Then the Bayes estimator (for the quadratic loss) is given by

$$\mathbf{E}[\mu|\mathbf{X}] = \frac{b^2}{b^2 + \frac{\sigma^2}{n}} \bar{X}_n + \frac{\frac{\sigma^2}{n}}{b^2 + \frac{\sigma^2}{n}} a. \quad (8.33)$$

Indeed, the posterior distribution has density

$$\begin{aligned} [\mu|x] &\propto [\mu] \cdot [x|\mu] \propto e^{-\frac{1}{2b^2}(\mu-a)^2} \cdot e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2} \\ &\propto e^{-\frac{1}{2} \left( \frac{1}{b^2} + \frac{n}{\sigma^2} \right) \mu^2 + \frac{a\mu}{b^2} + \frac{n}{\sigma^2} \mu \bar{x}_n} \\ &\propto e^{-\frac{1}{2} \left( \frac{1}{b^2} + \frac{n}{\sigma^2} \right) \left( \mu - \frac{a/b^2 + n/\sigma^2 \bar{x}_n}{1/b^2 + n/\sigma^2} \right)^2} = e^{-\frac{1}{2} \left( \frac{1}{b^2} + \frac{n}{\sigma^2} \right) \left( \mu - \frac{\sigma^2 a + nb^2 \bar{x}_n}{\sigma^2 + b^2 n} \right)^2}. \end{aligned}$$

Therefore  $\mu|\mathbf{X}$  has a  $\mathcal{N} \left( \frac{b^2}{b^2 + \frac{\sigma^2}{n}} \bar{X}_n + \frac{\frac{\sigma^2}{n}}{b^2 + \frac{\sigma^2}{n}} a, \left( \frac{1}{b^2} + \frac{n}{\sigma^2} \right)^{-1} \right)$ -distribution and the claim follows.

---

*End of Lecture 26*

## 9 Linear regression and the method of least squares

(Reference: [6, Chapter 13])

### 9.1 The method of least squares and simple linear regression

Consider the following set-up: Let  $(x_1, y_1), \dots, (x_n, y_n)$  be a bivariate data set. We will assume that this data set comes from a linear relation with some random errors. For instance, the  $(x_j, y_j)$  may be viewed as realizations of measurements of physical quantities which are proportional, say

$$x_j = \text{electrical current } (= I), \quad Y_j = \text{voltage } (= U),$$

and we have  $Y_j = R \cdot x_j + \varepsilon_j$ , where the  $\varepsilon_j$  are the i.i.d. errors, which come from the measurements of these quantities. The question is now to estimate the proportionality constant (the resistance  $R$  in the above example).

Let us be more precise: Let  $\varepsilon_1, \dots, \varepsilon_n$  be i.i.d. real random variables with  $\mathbf{E}[\varepsilon_1] = 0$  and  $\sigma^2 = \text{Var}[\varepsilon_1]$ . We consider the random variables

$$Y_j = \alpha + \beta x_j + \varepsilon_j, \quad j = 1, \dots, n, \quad (9.1)$$

with (unknown) parameters  $\alpha, \beta \in \mathbb{R}$  and  $\sigma^2$ .

**Definition 9.1.** The *least-squares estimator* for  $\alpha$  and  $\beta$  is given by the solution to the minimization problem

$$S(\alpha, \beta) = \sum_{j=1}^n (y_j - \alpha - \beta x_j)^2 \rightarrow \min! \quad (9.2)$$

upon insertion of the values  $(x_j, Y_j)$ .

To solve (9.2), we calculate

$$\begin{aligned} \frac{\partial}{\partial \alpha} S(\alpha, \beta) = 0 & \Leftrightarrow \sum_{j=1}^n (y_j - \alpha - \beta x_j) = 0, \\ \frac{\partial}{\partial \beta} S(\alpha, \beta) = 0 & \Leftrightarrow \sum_{j=1}^n (y_j - \alpha - \beta x_j) x_j = 0. \end{aligned} \quad (9.3)$$

This is equivalent to

$$\begin{aligned} n\alpha + \beta \sum_{j=1}^n x_j &= \sum_{j=1}^n y_j, \\ \alpha \sum_{j=1}^n x_j + \beta \sum_{j=1}^n x_j^2 &= \sum_{j=1}^n x_j y_j. \end{aligned} \quad (9.4)$$

Solving this for  $\alpha$  and  $\beta$ , we have the announced least-square estimators:

$$\begin{aligned} \hat{\beta}_n &= \frac{n \sum_{j=1}^n x_j Y_j - \left( \sum_{j=1}^n x_j \right) \left( \sum_{j=1}^n Y_j \right)}{n \sum_{j=1}^n x_j^2 - \left( \sum_{j=1}^n x_j \right)^2}, \\ \hat{\alpha}_n &= \frac{1}{n} \sum_{j=1}^n Y_j - \hat{\beta}_n \frac{1}{n} \sum_{j=1}^n x_j. \end{aligned} \quad (9.5)$$

These estimators are unbiased. One can also try to estimate the variance  $\sigma^2$  of the errors  $\varepsilon_j$ . An unbiased estimator for  $\sigma^2$  is given by

$$\hat{\sigma}_n^2 = \frac{1}{n-2} \sum_{j=1}^n (Y_j - \hat{\alpha}_n - \hat{\beta}_n x_j)^2. \quad (9.6)$$

*Remark 9.2.* Under appropriate conditions, one can show that the estimators  $\hat{\alpha}_n$  and  $\hat{\beta}_n$  are consistent and asymptotically normal, and it is also elementary to calculate the variances of  $\hat{\alpha}_n$  and  $\hat{\beta}_n$  as well as their covariance, see [6, Section 13.3].

## 9.2 Connection with Maximum-Likelihood estimators

Consider again the set-up of the linear regression problem (9.1), but we specify the laws of the i.i.d. random variables  $\varepsilon_1, \dots, \varepsilon_n$  to  $\varepsilon_1 \sim \mathcal{N}(0, \sigma^2)$  (again,  $\sigma^2$  is unknown). For physical measurements, this is a natural assumption. Note that

$$Y_j \sim \mathcal{N}(\alpha + \beta x_j, \sigma^2), \quad j = 1, \dots, n. \quad (9.7)$$

We want to estimate  $(\alpha, \beta)$  by the Maximum-Likelihood method. Note that the joint density of  $(Y_1, \dots, Y_n)$ , evaluated at  $(Y_1, \dots, Y_n)$  is given by

$$f_{(\alpha, \beta)}^{(n)}(Y_1, \dots, Y_n) = \prod_{j=1}^n \left( \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(Y_j - \alpha - \beta x_j)^2} \right). \quad (9.8)$$

To optimize, we take the logarithm

$$\log f_{(\alpha, \beta)}^{(n)}(Y_1, \dots, Y_n) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^n (Y_j - \alpha - \beta x_j)^2, \quad (9.9)$$



and take the derivatives with respect to  $\alpha$  and  $\beta$ , so

$$\frac{\partial}{\partial \alpha} \log f_{(\alpha, \beta)}^{(n)}(Y_1, \dots, Y_n) = 0, \quad \frac{\partial}{\partial \beta} \log f_{(\alpha, \beta)}^{(n)}(Y_1, \dots, Y_n) = 0. \quad (9.10)$$

Solving for  $\alpha$  and  $\beta$  gives the same estimators as in (9.5).

### 9.3 The general linear model

So far we treated simple linear regression, i.e. models with a linear relation of the form (9.1). We will now consider a more general form, known as the *general linear model*. It is given by

$$\begin{aligned} Y &= X\beta + \varepsilon, & \text{with} \\ Y &\in \mathbb{R}^{n \times 1} - \text{observation, stochastic, known,} \\ X &\in \mathbb{R}^{n \times k} - \text{design matrix, deterministic, known,} \\ \beta &\in \mathbb{R}^k - \text{parameter vector, deterministic, unknown,} \\ \varepsilon &\in \mathbb{R}^n - \text{error vector, stochastic, unknown.} \end{aligned} \quad (9.11)$$

Our assumptions will always be

$$\mathbf{E}[\varepsilon_i] = 0, \quad \text{Var}[\varepsilon_i] = \sigma^2. \quad (9.12)$$

*Remark 9.3.* The simple linear model (9.1) corresponds to the choice

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad (9.13)$$

i.e.  $k = 2$ .

We can also consider the least-squares error in this case. To this end, consider

$$R^2(\beta) = (Y - X\beta)^\top (Y - X\beta) = Y^\top Y - 2Y^\top X\beta + \beta^\top X^\top X\beta. \quad (9.14)$$

**Definition 9.4.** The *least-squares estimator* for  $\beta$  is given by the solution (not necessarily unique) of the minimization problem

$$R^2(\beta) \rightarrow \min! \quad (9.15)$$

Upon insertion of the values of  $X$  and  $Y$ .

To find the minimum, we set the gradient of  $R^2$  equal to the 0 vector:

$$\nabla R^2(\beta) = -2X^\top Y + 2X^\top X\beta \stackrel{!}{=} 0 \quad \Leftrightarrow \quad X^\top X\beta = X^\top Y. \quad (9.16)$$

These equations are called the *normal equations*.

The following observation is obvious:

**Lemma 9.5.** *The matrix  $X^\top X \in \mathbb{R}^{k \times k}$  is invertible if and only if  $\text{rank}(X) = k$ .*

The latter condition means that the  $k$  columns of  $X$  must be linearly independent. In this case we have

**Theorem 9.6.** *The least-squares estimator  $\hat{\beta}$  is the unique solution to the normal equations, i.e.*

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y. \quad (9.17)$$

Note that for the matrix-vector formulation of the simple linear model, this recovers the least-square estimators (9.5). Let us finally give without proof the following result, known as *Gauss-Markov theorem*:

**Theorem 9.7.** *Suppose that  $\text{rank}(X) = k$ . Then:*

- (i)  $\hat{\beta}$  is unbiased, i.e.  $\mathbf{E}[\hat{\beta}] = \beta$ .
- (ii) The covariance matrix of  $\hat{\beta}$  is  $\sigma^2(X^\top X)^{-1}$ .
- (iii)  $\hat{\beta}$  is the best linear unbiased estimator (BLUE) for  $\beta$ : For any estimator  $\tilde{\beta} = LY$  with some matrix  $L \in \mathbb{R}^{k \times n}$  fulfilling  $\mathbf{E}[\tilde{\beta}] = \beta$ , we have

$$\Sigma_{\tilde{\beta}} \geq_L \Sigma_{\hat{\beta}}. \quad (9.18)$$

Finally, we mention an important tool to for tests in the context of the general linear model, the so-called *F-test*.

**Theorem 9.8.** *Consider the model  $Y = X\beta + \varepsilon$  and assume that  $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_{n \times n})$ . Furthermore, suppose that  $K \in \mathbb{R}^{k \times \ell}$  has rank  $\ell$ ,  $X$  has rank  $k$  and  $\text{span}(K) \subseteq \text{span}(X^\top)$ <sup>1</sup>. A test at level  $\alpha \in (0, 1)$  for*

$$H_0 : K^\top \beta = 0 \quad \text{against} \quad K^\top \beta \neq 0$$

is then given by

$$\phi(Y) = \begin{cases} 1, & F \geq F_{\ell, n-k, 1-\alpha}, \\ 0, & F < F_{\ell, n-k, 1-\alpha}, \end{cases} \quad (9.19)$$

with  $F_{m, n, \gamma}$  denoting the  $\gamma$ -quantile of the  $F$  distribution with  $(m, n)$  degrees of freedom and

$$F = \frac{\frac{1}{\ell}(R_1^2 - R_0^2)}{\frac{1}{n-k}R_0^2} \quad (9.20)$$

with  $R_0^2 = R^2(\hat{\beta})$ , whereas  $R_1^2 = \min\{(Y - X\beta)^\top(Y - X\beta) : K^\top \beta = 0\}$ .

---

*End of Lecture 27*

---

<sup>1</sup>This is because we want that  $X\beta_1 = X\beta_2$  implies that  $K^\top \beta_1 = K^\top \beta_2$ .

# Bibliography

- [1] G. Casella and R. L. Berger, *Statistical Inference, 2nd ed.* Duxbury Advanced Series, 2001.
- [2] H.-O. Georgii, *Stochastics: Introduction to Probability and Statistics, 2nd ed.* De Gruyter, 2012.
- [3] E. L. Lehmann, G. Casella, *Theory of Point Estimation, 2nd ed.* Springer Texts in Statistics, 1998.
- [4] E. L. Lehmann, J. P. Romano, *Testing Statistical Hypotheses, 3rd ed.* Springer Texts in Statistics, 2005.
- [5] J. A. Rice, *Mathematical statistics and data analysis, 3rd ed.* Duxbury Advanced Series, 2006.
- [6] L. Wasserman, *All of Statistics. A Concise Course in Statistical Inference.* Springer Texts in Statistics, 2004.