Statistics

Lecture notes, 2025 Fall

Introduction

In this paper, we discuss the english version of a hungarian lecture note of the statistic part, which was written by András Tóbiás.

Contents

1 Basic Concepts of Mathematical Statistics

In the previous part of the lecture, the distribution of the random variables was assumed to be known. In contrast, in mathematical statistics, random variables correspond to measurement results, and therefore their distributions are not known precisely. It often happens that, based on theoretical considerations, we can assume that the measurements follow some type of distribution quite accurately, but this distribution depends on an unknown *parameter* ϑ, whose possible values form a *parameter domain* $\theta \subseteq \mathbb{R}^d$ for some $d \geq 1$.[1] For example:

[1] ϑ and θ are two lowercase forms of the Greek letter "theta."

- If we find a coin on the street and do not know whether it is fair, then it shows heads with an unknown probability $\vartheta \in \theta = [0,1] \subset \mathbb{R}^1$. By tossing the coin several times, we can try to estimate ϑ, or verify or reject the hypothesis that the coin is fair. The indicator variable

$$\mathbb{1}_{\{\text{the outcome of a given toss is head}\}}$$

  thus has the unknown parameter ϑ.

- The number of accidents at railway crossings in Hungary during a given month can be assumed to follow a Poisson distribution quite accurately, since there are many drivers, each having a small probability of an accident, and the events are more or less independent. The Poisson distribution has an *unknown* parameter $\vartheta \in \theta = (0, \infty) \subset \mathbb{R}^1$.

- The height of a randomly selected female student at BME can be modeled by a normal distribution.[2] In this case, both the mean $\mu \in \mathbb{R}$ and the variance $\sigma^2 > 0$ are unknown, so the parameter domain is

$$\theta = \{(\mu, \sigma^2) \in \mathbb{R}^2 \colon \sigma^2 > 0\} \subset \mathbb{R}^2.$$

  (Of course, in practice μ must be positive, but we ignore this restriction for generality.)

In mathematical statistics, it is typical that in order to study the unknown parameter, we *take a sample*, that is, we "generate" independent, identically distributed random variables $X_1, \dots, X_n$ following the distribution with the unknown parameter. For example:

- We toss the coin n times, and for each $i = 1, \dots, n$, let $X_i = 1$ if the i-th toss results in heads, and $X_i = 0$ otherwise. Then $X_1, \dots, X_n$ are independent, identically distributed indicator variables with parameter ϑ.

- We observe the number of railway crossing accidents in Hungary for n consecutive months, and let X_i denote the number of accidents in month i. Then $X_1, \dots, X_n$ are (approximately) independent and (approximately) Poisson(ϑ) distributed.

- From the list of BME female students, we randomly and independently select n students, and let X_i denote the height of the i-th student. Then $X_1, \dots, X_n$ are (approximately) independent and (approximately) $\mathrm{N}(\mu; \sigma^2)$ distributed, where both parameters are unknown.

[2] Similarly, a randomly selected male student's height can be modeled by a normal distribution with a different mean. However, the height of a randomly selected BME student (without conditioning on gender) cannot, since male and female averages differ, leading to a density with two local maxima — which is therefore not approximately normal.

The two main branches of mathematical statistics are *estimation theory* and *hypothesis testing*. Estimation theory aims to determine, as precisely as possible, the value of the unknown parameter based on the sample. (Of course, "as precisely as possible" is not a mathematically precise expression; we shall clarify its meaning later for specific estimation methods.) Hypothesis testing, on the other hand, aims to verify or reject a given hypothesis using the sample. Examples include:

- The coin is fair.
- The average number of accidents per month is 2.
- The average height of female BME students is at most 166 cm.
- The standard deviation of female BME students' height is 3 cm.

1.1 Sample and Realization

We define the concept of a sample without referring to the unknown parameter ϑ. This will be useful because later we will encounter methods applicable not only to parametric families but also to completely unknown distributions.

Definition 1.1.1. *Let $X_1, \dots, X_n$ be independent, identically distributed random variables with possibly unknown marginal distributions. Then the random vector*

$$\mathbf{X} = (X_1, \dots, X_n)$$

*is called an **independent and identically distributed sample of size** n (abbreviated as an **i.i.d. sample of size** n).*

Now we introduce some useful notation for the parametric case:

Notation 1.1.1. Let $\mathbf{X} = (X_1, \dots, X_n)$ be an i.i.d. sample of size n, where the common distribution depends on a parameter $\vartheta \in \theta \subseteq \mathbb{R}^d$ for some $d \geq 1$. Then, for a given $\vartheta \in \theta$:

1. The distribution function of X_1 corresponding to parameter ϑ is denoted by F_ϑ.
2. If X_1 has a density function under this parameter, it is denoted by f_ϑ, i.e.

$$f_\vartheta(x) = \begin{cases} F'_\vartheta(x), & \text{if } F_\vartheta \text{ is differentiable at } x, \\ 0, & \text{otherwise.} \end{cases}$$

3. If X_1 is discrete under this parameter, its probability mass function is denoted by p_ϑ, meaning that for $x \in \mathbb{R}$, $p_\vartheta(x)$ denotes the probability that $X_1 = x$, given that ϑ is the true parameter.

Example 1.1.1.

- For the coin found on the street: $p_\vartheta(0) = 1 - \vartheta$ and $p_\vartheta(1) = \vartheta$, $\vartheta \in [0, 1]$.
- For the number of railway crossing accidents:

$$p_\vartheta(k) = \frac{\vartheta^k}{k!} e^{-\vartheta}, \qquad \vartheta > 0,\ k = 0, 1, \ldots.$$

- For the height of female BME students, the unknown parameter is (μ, σ^2), hence

$$f_{(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad x \in \mathbb{R}, \quad \mu \in \mathbb{R},\ \sigma^2 > 0.$$

We leave the corresponding distribution functions to the reader.

Sometimes even the range of possible values of the sample elements depends on ϑ, as shown by the following example (unlike the previous ones).

Example 1.1.2. A friend generated 5 random numbers distributed as $\mathrm{U}(0; 1)$, then multiplied each by the same unknown constant $\vartheta \in (1, 2)$, chosen secretly. The resulting vector $\mathbf{x} = (x_1, \ldots, x_5) \in \mathbb{R}^5$ (i.e., the scaled numbers) was given to us. This vector is a realization of an i.i.d. sample $\mathbf{X} = (X_1, \ldots, X_5)$, whose elements are independent and $\mathrm{U}(0; \vartheta)$ distributed (is this clear? If not, try deriving the density!).
Given ϑ, the possible values of $X_1, \ldots, X_n$ (where the density is positive) lie in the interval $(0, \vartheta)$, thus depending on ϑ.

To define the term *realization* precisely, we first introduce the concept of the set of *essential values* of a random variable (depending on ϑ). In the discrete case, this definition coincides with the one encountered in regression theory (see Definition 11.1.5 in Szabolcs Mészáros's notes), except that there the parameter was not unknown.

Definition 1.1.2 (and notation)**.** *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an i.i.d. sample of size n, where the distribution of the sample elements depends on a parameter $\vartheta \in \theta \subseteq \mathbb{R}^d$. For $\vartheta \in \theta$ and $i = 1, \ldots, n$, if the density function f_ϑ exists, define*

$$S_{X_i}^{(\vartheta)} = \{x \in \mathbb{R} \mid f_\vartheta(x) > 0\} \subseteq \mathbb{R}.$$

If instead the probability mass function p_ϑ exists, define

$$S_{X_i}^{(\vartheta)} = \{x \in \mathbb{R} \mid p_\vartheta(x) > 0\} \subseteq \mathbb{R}.$$

In both cases, $S_{X_i}^{(\vartheta)}$ is called the ***set of essential values of*** *X_i for the parameter ϑ.*[3]

[3] The set of essential values of X_i (for a given ϑ) is almost the same as the image of X_i as a function $\Omega \to \mathbb{R}$. The image may be slightly larger, but all values outside $S_{X_i}^{(\vartheta)}$ occur with probability zero.

Example 1.1.3. For the i.i.d. sample $(X_1, \ldots, X_n)$ obtained from tossing a coin n times with unknown head probability ϑ, where X_i is an indicator with parameter ϑ, we have $S_{X_i}^{(\vartheta)} = \{0, 1\}$ for all $\vartheta \in (0, 1)$, $S_{X_i}^{(0)} = \{0\}$ and $S_{X_i}^{(1)} = \{1\}$. Furthermore, $p_\vartheta(1) = \vartheta$, $p_\vartheta(0) = 1 - \vartheta$, and $p_\vartheta(x) = 0$ for all $x \in \mathbb{R} \setminus \{0, 1\}$.

In the uniform example (1.1.2), where $X_i \sim \mathrm{U}(0, \vartheta)$ and $n = 5$,

$$S_{X_i}^{(\vartheta)} = (0, \vartheta) \quad \text{and} \quad f_\vartheta(x) = \begin{cases} \frac{1}{\vartheta}, & \text{if } x \in (0, \vartheta), \\ 0, & \text{otherwise.} \end{cases}$$

for all $\vartheta \in \theta = (1, 2)$.

Definition 1.1.3. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an i.i.d. sample of size n, where the sample distribution depends on a parameter $\vartheta \in \theta \subseteq \mathbb{R}^d$. A vector $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is called a (possible) **realization** of $\mathbf{X} = (X_1, \ldots, X_n)$ for parameter $\vartheta \in \theta$, if $x_i \in S_{X_i}^{(\vartheta)}$ holds for all $i \in \{1, \ldots, n\}$.*

Example 1.1.4. Let us illustrate the notion of realization using Example 1.1.3.
For the coin found on the street, the sequence $(x_1, \ldots, x_7) = (1, 0, 0, 0, 1, 1, 0)$ is a realization of the i.i.d. sample $\mathbf{X} = (X_1, \ldots, X_7)$ for all parameters $0 < \vartheta < 1$ (but not for $\vartheta = 0$ or $\vartheta = 1$).
For the i.i.d. sample $\mathbf{X} = (X_1, \ldots, X_5)$ uniformly distributed on $(0, \vartheta)$,

$$(x_1, \ldots, x_5) = (0.14, 0.79, 1.13, 1, 1.2)$$

is a possible realization if $1.2 < \vartheta < 2$, but not if $1 < \vartheta \leq 1.2$.

1.2 Basic sample statistics

For an i.i.d. sample of size n, a *statistic* is any function of the sample elements which is symmetric, that is, it "depends on all sample elements in the same way." The following definition formalizes this property.

Definition 1.2.1. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an i.i.d. sample of size n. If $T\colon \mathbb{R}^n \to \mathbb{R}$ is a symmetric function, that is,*

$$T(x_1, \ldots, x_n) = T(x_{\pi(1)}, \ldots, x_{\pi(n)})$$

*for all $x_1, \ldots, x_n \in \mathbb{R}$ and for every permutation $\pi\colon \{1, \ldots, n\} \to \{1, \ldots, n\}$ in the combinatorics chapter for an equivalent definition, then the random variable $T(\mathbf{X}) = T(X_1, \ldots, X_n)$ is called a **statistic** of $X_1, \ldots, X_n$.*

We now introduce some classical statistics, partly already encountered earlier in the course. The first one is the sample mean, familiar already from high school.

Example 1.2.1. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an i.i.d. sample of size n. Then the quantity

$$\overline{X_n} = \frac{X_1 + \ldots + X_n}{n},$$

introduced in Section 9.2, is called the **sample mean** of $\mathbf{X}$, and it is a statistic of the sample. If $\mathbf{x} = (x_1, \ldots, x_n)$ is a realization of $\mathbf{X} = (X_1, \ldots, X_n)$, then we denote the mean of the realization by $\overline{x_n}$:

$$\overline{x_n} = \frac{x_1 + \ldots + x_n}{n}.$$

Based on what we have seen so far, it is not surprising that the sample mean is a kind of estimator of the expected value based on the sample. Clearly, if $\mathbb{E}(X_i)$ exists (that is, if $\mathbb{E}[|X_i|] < \infty$), then the expected value of the sample mean coincides with the expected value of each sample element:

$$\mathbb{E}\big[\overline{X_n}\big] = \frac{1}{n}\big(\mathbb{E}(X_1) + \ldots + \mathbb{E}(X_n)\big) = \mathbb{E}(X_1).$$

Further properties of this estimator will be analyzed later, in the language of estimation theory.

The next classical statistic is the *corrected empirical variance*, which approximates the variance using the sample.

Definition 1.2.2. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an i.i.d. sample of size n. Then*

$$(S_n^*)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2 \tag{1}$$

*is called the **corrected empirical variance** of $\mathbf{X}$,*[4] *and*

$$S_n^* = \sqrt{(S_n^*)^2}$$

*is called the **corrected empirical standard deviation** of $\mathbf{X}$.*

Clearly, $(S_n^*)^2$ is a statistic of the sample $\mathbf{X}$. One may ask why we divide by $n-1$ (and not, say, by n) in formula 1. The reason is that in this way we obtain a statistic whose expected value coincides with the variance of the sample elements:

Statement 1.2.1. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an i.i.d. sample of size n, and assume that $\mathbb{E}(X_i^2) < \infty$. Then*

$$\mathbb{E}\big((S_n^*)^2\big) = \mathbb{D}^2(X_1).$$

[4]The word "empirical" means "based on observations." The term "corrected empirical variance" is also used.

Proof. The computations in the proof are somewhat lengthy, but apart from the linearity of expectation, they only use the fact that the expectation of the product of independent random variables equals the product of their expectations (Claim 6.1.4).
First,

$$\begin{aligned}\mathbb{E}[(S_n^*)^2] &= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\big(X_i - \overline{X_n}\big)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\Big(\mathbb{E}(X_i^2) - 2\mathbb{E}(X_i\overline{X_n}) + \mathbb{E}(\overline{X_n}^2)\Big) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}(X_i^2) - \frac{2}{n-1}\sum_{i=1}^{n}\mathbb{E}(X_i\overline{X_n}) + \frac{n}{n-1}\mathbb{E}(\overline{X_n}^2).\end{aligned}$$

We compute the three terms on the right-hand side separately. For the first term,

$$\frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}(X_i^2) = \frac{n}{n-1}\mathbb{E}(X_1^2),$$

since $X_1, \ldots, X_n$ are identically distributed.
For the second term, we compute its negative:

$$\begin{aligned}\frac{2}{n-1}\sum_{i=1}^{n}\mathbb{E}(X_i\overline{X_n}) &= \frac{2}{n-1}\sum_{i=1}^{n}\mathbb{E}\Big(X_i\cdot\frac{1}{n}\sum_{j=1}^{n}X_j\Big) \\ &= \frac{2}{n-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i^2) + \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}(X_iX_j)\Big) \\ &= \frac{2}{n-1}\Big(\mathbb{E}(X_1^2) + (n-1)\mathbb{E}(X_1)^2\Big) \\ &= \frac{2}{n-1}\mathbb{E}(X_1^2) + 2\mathbb{E}(X_1)^2,\end{aligned}$$

since $X_1, \ldots, X_n$ are identically distributed and independent.
For the third term,

$$\begin{aligned}\frac{n}{n-1}\mathbb{E}(\overline{X_n}^2) &= \frac{n}{(n-1)n^2}\mathbb{E}\Big[\Big(\sum_{i=1}^{n}X_i\Big)^2\Big] \\ &= \frac{1}{n(n-1)}\mathbb{E}\Big(\sum_{i=1}^{n}X_i^2 + 2\sum_{1\leq i<j\leq n}X_iX_j\Big) \\ &= \frac{1}{n(n-1)}\Big(\sum_{i=1}^{n}\big(\mathbb{E}(X_1^2) + (n-1)\mathbb{E}(X_1)^2\big)\Big) \\ &= \frac{1}{n-1}\mathbb{E}(X_1^2) + \mathbb{E}(X_1)^2,\end{aligned}$$

again using independence and identical distribution.
Collecting all terms, we obtain

$$\mathbb{E}[(S_n^*)^2] = \mathbb{E}(X_1^2)\Big(\frac{n}{n-1} - \frac{2}{n-1} + \frac{1}{n-1}\Big) + \mathbb{E}(X_1)^2(-2+1) = \mathbb{E}(X_1^2) - \mathbb{E}(X_1)^2 = \mathbb{D}^2(X_1).$$

□

Remark 1.2.1. We see that if we divided by $\frac{1}{n}$ instead of $\frac{1}{n-1}$ in formula (1), then (by linearity of expectation) the expected value of the resulting statistic would not be $\mathbb{D}^2(X_1)$, but $\frac{n-1}{n}\mathbb{D}^2(X_1)$. The statistic

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

is called the *empirical variance* (and its square root the *empirical standard deviation*), while $(S_n^*)^2$ is called *corrected* because its expected value equals $\mathbb{D}^2(X_1)$. This is the standard terminology in the Hungarian literature. In foreign-language texts, $(S_n^*)^2$ may itself be called the empirical variance, but we shall not adopt this convention here.

Another basic statistic is the **mode** of the sample, that is, the most frequent value in the sample. If there are several such values, each of them is considered a mode.
To define further basic statistics, we first introduce the notion of an ordered sample.

Definition 1.2.3. *Let* $\mathbf{X} = (X_1, \ldots, X_n)$ *be an i.i.d. sample of size* n. *Let* $(X_1^*, X_2^*, \ldots, X_n^*)$ *be a listing of the sample elements* $X_1, \ldots, X_n$ *such that*

$$X_1^* \le X_2^* \le \ldots \le X_n^*.$$

Then $(X_1^*, X_2^*, \ldots, X_n^*)$ *is called the* ***ordered sample****.*

Example 1.2.2. Let $\mathbf{x} = (x_1, \ldots, x_8) = (1, 2, 1, 3, 4, 6, 5, 2)$ be a realization of an i.i.d. sample $\mathbf{X} = (X_1, \ldots, X_8)$ from eight tosses of a fair die. Then the ordered sample $(X_1^*, \ldots, X_8^*)$ has realization $(x_1^*, \ldots, x_8^*) = (1, 1, 2, 2, 3, 4, 5, 6)$. Thus the realization of the ordered sample is uniquely determined by the realization of the original (unordered) sample, even if there are repeated values.

It is important to emphasize that $X_1^*, \ldots, X_n^*$ are not independent (we do not prove this formally, but one can feel that, for example, the distribution of the smallest sample element affects the distributions of all the other ordered elements, since they must be at least as large as the smallest one), and they are typically not identically distributed either.
Using the ordered sample, we can define the empirical median, which for odd sample size is the middle element of the ordered sample, and for even sample size is the arithmetic mean of the two middle elements:

Definition 1.2.4. *For an i.i.d. sample* $\mathbf{X} = (X_1, \ldots, X_n)$ *of size* n*, the* ***empirical median*** *is defined as* $m_{n,\mathbf{X}} = X^*_{k+1}$ *when* $n = 2k + 1$*, and as*

$$m_{n,\mathbf{X}} = \frac{X^*_k + X^*_{k+1}}{2}$$

when $n = 2k$*.*

Example 1.2.3. If a sample of size $n = 4$ has realization $(x_1, \ldots, x_4) = (\sqrt{2}, \pi, -1, \mathrm{e})$, then the ordered sample has realization $(x^*_1, x^*_2, x^*_3, x^*_4) = (-1, \sqrt{2}, \mathrm{e}, \pi)$, and the empirical median is $\frac{\sqrt{2}+\mathrm{e}}{2}$. If a sample of size $n = 3$ has realization $(x_1, x_2, x_3) = (-1, 3, 2)$, then the ordered sample has realization $(x^*_1, x^*_2, x^*_3) = (-1, 2, 3)$, and the empirical median is 2.

1.3 Empirical distribution function

In this subsection we assume that the random variables $X_1, \ldots, X_n$ are independent and identically distributed with some *unknown* distribution, whose distribution function we denote by $x \mapsto F(x) = \mathbb{P}(X_i < x)$. (The parameter ϑ will not appear here either.) We have no prior information about the distribution of the X_i's: for instance, we do not know whether it is discrete, continuous, or neither of the two.[5]
Given a realization $\mathbf{x} = (x_1, \ldots, x_n)$ of the sample $\mathbf{X} = (X_1, \ldots, X_n)$, how can we "estimate" the distribution function F so that, as $n \to \infty$, we obtain something converging to the true distribution function? This is exactly what the empirical distribution function does.

Definition 1.3.1. *Let* $n \in \mathbb{N}$ *and let* $X_1, \ldots, X_n$ *be i.i.d. random variables. The function* $\mathbb{R} \to \mathbb{R}$ *defined by*

$$x \mapsto F^*_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i < x\}} = \frac{1}{n} \big| \{ i \in \{1, \ldots, n\} \mid X_i < x \} \big| \tag{2}$$

is called the ***empirical distribution function*** *associated with the i.i.d. sample* $\underline{X} = (X_1, \ldots, X_n)$ *of size* n*.*

It is important to emphasize from the outset that F^*_n is a *random* function: its values depend on the random variables $X_1, \ldots, X_n$.
Using the ordered sample $(X^*_1, \ldots, X^*_n)$ (see Definition 1.2.3), the definition in (2) can be rewritten in the following simple explicit form:

$$F^*_n(x) = \begin{cases} 0, & \text{if } x \le X^*_1, \\ \frac{k}{n}, & \text{if } X^*_k < x \le X^*_{k+1}, \quad k = 1, \ldots, n-1, \\ 1, & \text{if } x > X^*_n. \end{cases} \tag{3}$$

[5]For the latter, consider the random variable X defined as follows: toss a coin, and if it comes up heads, toss a die and let X be the outcome; if it comes up tails, let X be a $\mathrm{U}(0; 1)$ distributed random number, independent of the die toss.

This reformulation also shows that, for any fixed realization $\mathbf{x} = (x_1, \ldots, x_n)$ of $\mathbf{X} = (X_1, \ldots, X_n)$, we indeed obtain a distribution function: the limit at $-\infty$ is 0, the limit at $+\infty$ is 1, and the function is nondecreasing and (thanks to the strict inequalities in the definition (2)) left-continuous.
The following statement shows that the empirical distribution function is a "good estimator" of the true distribution function: its expected value equals the true distribution function at every point, and its variance decreases as n grows. Moreover, the empirical distribution function converges (with probability 1) to the true distribution function at every point as $n \to \infty$. In other words, if the sample size is large enough, then for any fixed $x \in \mathbb{R}$, the value $F_n^*(x)$ of the empirical distribution function can (with probability 1) be made arbitrarily close to the value $F(x)$ of the true distribution function.

Statement 1.3.1. *Let $n \in \mathbb{N}$ and let $X_1, \ldots, X_n$ be independent, identically distributed random variables with distribution function F. Then for every $x \in \mathbb{R}$ we have:*

1. $\mathbb{E}(F_n^*(x)) = F(x)$,
2. $\mathbb{D}^2(F_n^*(x)) = \dfrac{F(x)(1-F(x))}{n}$, *and*
3. $\mathbb{P}(\lim_{n\to\infty} F_n^*(x) = F(x)) = 1$.

Proof. Fix $x \in \mathbb{R}$. From the definition (2) we see that $n \cdot F_n^*(x)$ is the sum of the i.i.d. indicator variables $\mathbb{1}_{\{X_i<x\}}$, each having parameter (and hence expected value) $\mathbb{P}(X_i < x) = F(x)$. Therefore $nF_n^*(x) \sim B(n; F(x))$. Using the well-known properties of the binomial distribution and of expectation and variance, we obtain

$$\mathbb{E}\big(F_n^*(x)\big) = \frac{1}{n}\mathbb{E}\big(nF_n^*(x)\big) = \frac{1}{n} \cdot nF(x) = F(x)$$

and

$$\mathbb{D}^2\big(F_n^*(x)\big) = \frac{1}{n^2}\mathbb{D}^2\big(nF_n^*(x)\big) = \frac{1}{n^2} \cdot nF(x)(1-F(x)) = \frac{F(x)(1-F(x))}{n},$$

as claimed in (1) and (2).
To prove (3), we apply the strong law of large numbers (Theorem 9.2.1) to the i.i.d. indicator variables $\mathbb{1}_{\{X_i<x\}}$. Their expected value is

$$\mathbb{E}\big(\mathbb{1}_{\{X_1<x\}}\big) = \mathbb{P}(X_1 < x) = F(x),$$

and they have finite variance, in fact $\mathbb{D}^2\big(\mathbb{1}_{\{X_1<x\}}\big) = F(x)(1-F(x))$. Hence, by Theorem 9.2.1, their average converges to their expected value with probability 1:

$$\mathbb{P}\Big(\lim_{n\to\infty} F_n^*(x) = F(x)\Big) = \mathbb{P}\Big(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i<x\}} = \mathbb{E}\big(\mathbb{1}_{\{X_1<x\}}\big)\Big) = 1.$$

This proves (3) as well. □

Statement 1.3.1 (3) can be strengthened further: the empirical distribution function not only converges to the true distribution function at every point (with probability 1), but it converges to it *uniformly on the entire real line* (with probability 1), that is, the distance between the two functions tends to zero. This is expressed by the following theorem:

Theorem 1.3.1 (Glivenko–Cantelli theorem)**.** *Let $n \in \mathbb{N}$ and let $X_1, \dots, X_n$ be independent, identically distributed random variables with distribution function F. Then the function $x \mapsto F_n^*(x)$ converges uniformly to $x \mapsto F(x)$ with probability 1, that is,*

$$\mathbb{P}\Big(\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \big|F_n^*(x) - F(x)\big| = 0 \Big) = 1.$$

Why is this stronger than part (3) of Statement 1.3.1? In Statement 1.3.1, for each fixed x the event where the convergence in (3) fails has probability 0, but in principle this null set may depend on x. The Glivenko–Cantelli theorem implies that even if we take the union of all these (uncountably many) null sets over all $x \in \mathbb{R}$, we still obtain an event of probability 0. We will not prove the Glivenko–Cantelli theorem in this course; a proof can be found in essentially any textbook on mathematical statistics.

2 Methods in estimation theory

In this chapter we again consider the situation where the distribution of the sample elements $X_1, \ldots, X_n$ depends on a parameter $\vartheta \in \theta$, and we wish to construct, based on some statistic of an i.i.d. sample of size n, as good an estimate as possible for the unknown parameter ϑ or for some function $\psi(\vartheta)$ of it.[6] Due to time constraints, the estimation methods discussed here will not be exhaustive; interested readers are referred to the course *Mathematical Statistics* in the MSc programmes in Computer Engineering and Business Informatics.

Definition 2.0.1. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an i.i.d. sample of size n, where the common distribution of the sample elements depends on an unknown parameter $\vartheta \in \theta \subseteq \mathbb{R}^d$, and let $\psi\colon \mathbb{R}^d \to \mathbb{R}$ be a function. Let $T(\mathbf{X}) = T(X_1, \ldots, X_n)$ be a statistic of the sample. We say that the statistic $T(\mathbf{X})$ is*

1. *an **unbiased estimator** of the parameter function $\psi(\vartheta)$ if for every $\vartheta \in \theta$,*

$$\mathbb{E}_\vartheta(T(X_1, \ldots, X_n)) = \psi(\vartheta),$$

   where $\mathbb{E}_\vartheta$ denotes expectation with respect to $\mathbb{P}_\vartheta$;

2. *an **asymptotically unbiased estimator** of the parameter function $\psi(\vartheta)$ if for every $\vartheta \in \theta$,*

$$\lim_{n\to\infty} \mathbb{E}_\vartheta(T(X_1, \ldots, X_n)) = \psi(\vartheta);$$

3. *a **strongly consistent estimator** of the parameter function $\psi(\vartheta)$[7] if for every $\vartheta \in \theta$,*

$$\mathbb{P}_\vartheta\big(\lim_{n\to\infty} T(X_1, \ldots, X_n) = \psi(\vartheta)\big) = 1; \tag{4}$$

4. *if both $T(\mathbf{X})$ and another statistic $T'(\mathbf{X}) = T'(X_1, \ldots, X_n)$ of the sample are unbiased estimators of $\psi(\vartheta)$, then we say that $T(\mathbf{X})$ is **at least as efficient as** $T'(\mathbf{X})$ if*

$$\mathbb{D}^2_\vartheta(T(\mathbf{X})) \le \mathbb{D}^2_\vartheta(T'(\mathbf{X})),$$

   *where $\mathbb{D}^2_\vartheta$ denotes the variance with respect to $\mathbb{P}_\vartheta$.[8] If $T(\mathbf{X})$ is at least as efficient as any unbiased estimator of $\psi(\vartheta)$, then we say that $T(\mathbf{X})$ is an **efficient estimator** of the parameter function $\psi(\vartheta)$.*

[6] ψ is also a Greek letter, called "psi."

[7] There are other notions of consistency for estimators, for example *weak consistency*, which expresses the same idea as strong consistency but with convergence in probability instead of almost sure convergence in (4). When we simply say that an estimator is *consistent*, we mean that it is weakly consistent. As we have already seen in the topic of laws of large numbers, almost sure convergence implies convergence in probability, so every strongly consistent estimator is consistent. There are further notions of consistency (which we do not detail here), for example *consistency in mean square*; see the course *Mathematical Statistics* in the MSc programmes in Computer Engineering and Business Informatics.

[8] In an analogous and straightforward way one can define the relations "(strictly) more efficient than", "at most as efficient as" and "(strictly) less efficient than" between two unbiased estimators.

Clearly, if a statistic is an unbiased estimator of a given parameter function, then it is also asymptotically unbiased. Let us now see some examples of unbiased, asymptotically unbiased and strongly consistent estimators.

Example 2.0.1. In the setting of Definition 2.0.1, one possible choice of the function ψ is

$$\theta \ni \vartheta \mapsto \psi(\vartheta) = \mathbb{E}_\vartheta(X_1),$$

provided that $\mathbb{E}_\vartheta(X_1)$ exists for every $\vartheta \in \theta$ (that is, $\mathbb{E}_\vartheta(|X_1|) < \infty$ for all $\vartheta \in \theta$). In this case

$$T(X_1, \ldots, X_n) = \overline{X}_n = \frac{X_1 + \ldots + X_n}{n}$$

(the sample mean) is an unbiased estimator of the parameter function $\psi(\vartheta) = \mathbb{E}_\vartheta(X_1)$, since for every $\vartheta \in \theta$,

$$\mathbb{E}_\vartheta(\overline{X}_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\vartheta(X_i) = \mathbb{E}_\vartheta(X_1)$$

(see Example 1.2.1). Furthermore, by the strong law of large numbers, if we assume that ϑ is the true parameter, then the sample mean converges almost surely to $\psi(\vartheta) = \mathbb{E}_\vartheta(X_1)$. Formally, for every $\vartheta \in \theta$,

$$\mathbb{P}_\vartheta\Big(\lim_{n\to\infty} \overline{X_n} = \mathbb{E}_\vartheta(X_1) \Big) = 1.$$

Thus $T(X_1, \ldots, X_n) = \overline{X_n}$ is also a strongly consistent estimator of the parameter $\psi(\vartheta) = \mathbb{E}_\vartheta(X_1)$.

Example 2.0.2. Another possible choice for ψ is

$$\theta \ni \vartheta \mapsto \psi(\vartheta) = \mathbb{D}^2_\vartheta(X_1),$$

provided that $\mathbb{E}_\vartheta(X_1^2)$ is finite for every $\vartheta \in \theta$. In this case

$$T(X_1, \ldots, X_n) = (S_n^*)^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X_n})^2$$

(the corrected empirical variance) is an unbiased estimator of the parameter function

$$\psi(\vartheta) = \mathbb{D}^2_\vartheta(X_1).$$

Indeed, for every $\vartheta \in \theta$ we have

$$\mathbb{E}_\vartheta((S_n^*)^2) = \mathbb{D}^2_\vartheta(X_1).$$

We can also see that the statistic

$$T'(X_1, \ldots, X_n) = S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X_n})^2 \tag{5}$$

is *not* an unbiased estimator of $\mathbb{D}_\vartheta^2(X_1)$, since (by linearity of expectation, see also Remark 1.2.1)

$$\mathbb{E}_\vartheta[S_n^2] = \frac{n-1}{n} \mathbb{E}_\vartheta[(S_n^*)^2] = \frac{n-1}{n} \mathbb{D}_\vartheta^2(X_1).$$

On the other hand, $T'(X_1, \ldots, X_n) = S_n^2$ is an asymptotically unbiased estimator of $\mathbb{D}_\vartheta^2(X_1)$, because

$$\lim_{n\to\infty} \mathbb{E}_\vartheta[S_n^2] = \lim_{n\to\infty} \frac{n-1}{n} \mathbb{D}_\vartheta^2(X_1) = \mathbb{D}_\vartheta^2(X_1).$$

Example 2.0.3. Unbiasedness alone does not guarantee that an estimator is good, for instance in terms of efficiency; there may exist another unbiased estimator that is strictly more efficient.
Let (X_1, X_2, X_3) be an i.i.d. sample from a $\mathrm{U}(\vartheta; 1+\vartheta)$ distribution with unknown parameter $\vartheta \geq 0$, and consider the ordered sample element $T(X_1, X_2, X_3) = X_2^*$ as a statistic of (X_1, X_2, X_3). In an earlier exercise we proved that, for $\vartheta = 0$, the density of X_2^* is

$$f_{X_2^*}(x) = \begin{cases} 6x - 6x^2, & \text{if } 0 < x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

We also computed that $\mathbb{E}(X_2^*) = 1/2$, and from the density it is easy to compute that $\mathbb{D}^2(X_2^*) = 0.05$.
For a general $\vartheta > 0$, we can obtain X_1, X_2, X_3 by taking i.i.d. random variables Y_1, Y_2, Y_3 uniformly distributed on $(0, 1)$ (with Y_2^* denoting the second order statistic) and then adding ϑ to each of them. Hence for general ϑ,

$$\mathbb{E}(X_2^*) = \mathbb{E}(Y_2^* + \vartheta) = 1/2 + \vartheta = \mathbb{E}_\vartheta(X_1)$$

and

$$\mathbb{D}^2(X_2^*) = \mathbb{D}^2(Y_2^* + \vartheta) = \mathbb{D}^2(Y_2^*) = 0.05.$$

Thus the statistic $T(X_1, X_2, X_3) = X_2^*$ is an unbiased estimator of the parameter function

$$\psi(\vartheta) = \mathbb{E}_\vartheta(X_1) = \vartheta + \frac{1}{2}.$$

However, this estimator is not efficient: for example,

$$S(X_1, X_2, X_3) = \overline{X_3}$$

(the sample mean of three observations) is also an unbiased estimator of $\psi(\vartheta)$, but its variance is

$$\mathbb{D}^2(\overline{X_3}) = \mathbb{D}^2\Big(\frac{X_1 + X_2 + X_3}{3}\Big) = \frac{1}{3}\mathbb{D}^2(X_1) = \frac{1}{12 \cdot 3} = \frac{1}{36} < 0.05,$$

that is, smaller than the variance of $T(X_1, X_2, X_3)$. Therefore $S(X_1, X_2, X_3)$ is more efficient than $T(X_1, X_2, X_3)$.

3 Confidence intervals

3.1 Basic concepts: confidence interval, quantiles of continuous distributions

The maximum likelihood and method-of-moments estimators introduced so far belong to the category of *point estimators*, since our estimate for the unknown parameter ϑ (or for some function of it) is a "point" that depends on the sample, that is, a real number or vector. Although a point estimate gives, in a certain sense, our "best guess" for the parameter given the observed sample, in practice it rarely happens that the true parameter is actually close to the obtained estimate.
Therefore, in this chapter we introduce confidence intervals, which belong to the class of *interval estimators*. Here the result of the estimation is an interval, depending on the realization, which contains the true parameter value ϑ with a pre-specified probability $1-\varepsilon$. By increasing the length of the interval, this probability can be made arbitrarily close to 1; for instance, it is often chosen to be 0.99 (the error probability is then $\varepsilon = 0.01$) or 0.95 (with $\varepsilon = 0.05$). The following definition formalizes this.

Definition 3.1.1. *Let $\mathbf{X} = (X_1, \dots, X_n)$ be an i.i.d. sample of size n, where the common distribution of the sample elements depends on a parameter $\vartheta \in \theta \subseteq \mathbb{R}^d$, let $\psi\colon \mathbb{R}^d \to \mathbb{R}$ be a function, and let $0 < \varepsilon < 1$.*
We say that the interval $[T_1(\mathbf{X}), T_2(\mathbf{X})]$ is a (precise) ***confidence interval of level*** *$1-\varepsilon$ for the parameter function $\psi(\vartheta)$ if*

$$\mathbb{P}_\vartheta\big(T_1(\mathbf{X}) \le \psi(\vartheta) \le T_2(\mathbf{X})\big) = 1 - \varepsilon$$

holds for all $\vartheta \in \theta$.

Remark 3.1.1.

1. The endpoints $T_1(\mathbf{X})$ and $T_2(\mathbf{X})$ in the definition are statistics of the sample $\mathbf{X}$, so they depend on the random sample, but once the realization is observed, they are no longer random.

2. A confidence interval of *exact* level $1-\varepsilon$ can typically be constructed only when the marginal distribution of $X_1, \dots, X_n$ is continuous. In the discrete case, one usually speaks of *at least $1-\varepsilon$ level confidence intervals*, where "$= 1-\varepsilon$" in the definition is replaced by "$\ge 1-\varepsilon$". This distinction will not play a major role in this course.

Before we can present examples of confidence intervals, we also need the notion of a *quantile*. We introduce this only for continuous random variables.
Let X be a continuous random variable. By Definition 4.2.1 this means that X has a density function f_X. In this case we know that the distribution function $x \mapsto F_X(x)$ is continuous. For simplicity, assume that the density is nonzero exactly on a single open interval I (that is, $f_X(x) > 0$ for all $x \in I$ and $f_X(x) = 0$ for all $x \in \mathbb{R} \setminus I$). Examples:

1. the uniform distribution, where $I = (a, b)$, with $-\infty < a < b < \infty$,

2. the exponential distribution, where $I = (0, \infty)$,

3. and the normal distribution, where $I = (-\infty, \infty) = \mathbb{R}$.

Then on the interval I the distribution function F_X is strictly increasing. Furthermore, if I is of the form (a, b) or (a, ∞), then $F_X(x) = 0$ holds for all $x \leq a$, and if I is of the form (a, b) or $(-\infty, b)$, then $F_X(x) = 1$ holds for all $x \geq b$. It follows that the inverse function

$$F_X^{-1} \colon (0, 1) \to I, \quad y \mapsto F_X^{-1}(y)$$

exists (for $x \in I$ and $0 < y < 1$, the equality $x = F_X^{-1}(y)$ holds if and only if $y = F_X(x)$). This inverse function is also continuous and strictly increasing. It is important to emphasize, however, that the range of F_X^{-1} is generally not all of $\mathbb{R}$, but only I. Indeed, the examples above show that the value 0 is taken by F_X either at infinitely many points (uniform and exponential distributions) or at no point at all (normal distribution), and similarly for the value 1.

Definition 3.1.2. *Let X be a continuous random variable with distribution function F_X and density f_X. Assume that there exists an open interval I such that for all $x \in \mathbb{R}$ we have $f_X(x) \neq 0$ if and only if $x \in I$.*
*Let $0 < y < 1$. Then the point $F_X^{-1}(y) \in I$ is called the y-**quantile** of X. The $\frac{1}{2}$-quantile (that is, the case $y = 1/2$) is called the **median**.*

If x is the y-quantile of X, then

$$F_X(x) = \mathbb{P}(X < x) = y,$$

that is, X takes values smaller than x with probability exactly y. In the case of the median, $y = 1/2$, so the probability that $X < x$ equals the probability that $X \geq x$ (and, due to continuity, also the probability that $X > x$). In this sense the median is indeed analogous to the empirical median, i.e. to the "middle element of the ordered sample" (or the average of the two middle elements in case of an even sample size).
For discrete random variables it may well happen that the value $0 < y < 1$ is skipped by the distribution function. For example, in the case of a single fair die roll, the distribution function never takes the value $1/12$, since it jumps from 0 directly to $1/6$ (at $x = 1$). For this reason, the above definition of the y-quantile is only correct when the distribution function is continuous; however, with some technical work, the definition can be generalized to discrete random variables as well.

3.2 Confidence interval for the mean of a normal distribution with known variance

We now present our first example of a confidence interval, which is perhaps also the most classical one in this topic:

Example 3.2.1. Let $X_1, \ldots, X_n \sim \mathrm{N}(\mu; \sigma^2)$ be i.i.d. random variables, where the mean $\mu \in \mathbb{R}$ of the normal distribution is unknown, but the variance $\sigma^2 > 0$ is known. For $\varepsilon \in (0, 1)$, construct a confidence interval of level $1 - \varepsilon$ for the mean μ of the normal distribution in such a way that the length of the interval is as small as possible.

Since the density

$$f_\mu(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

is strictly decreasing as a function of the distance of x from μ (for any fixed value of the parameter μ), the shortest confidence interval of level $1 - \varepsilon$ for a given ε is obtained if we choose it symmetrically around the sample mean $\overline{X_n}$, that is, in the form

$$[\overline{X_n} - r_\varepsilon, \overline{X_n} + r_\varepsilon]$$

such that

$$\mathbb{P}_\mu(\mu \in [\overline{X_n} - r_\varepsilon, \overline{X_n} + r_\varepsilon]) = 1 - \varepsilon \tag{6}$$

holds for all $\mu \in \mathbb{R}$. Our task is thus to determine the value of r_ε (depending on ε).
The distribution of the sample mean is, by a well-known property of the normal distribution,

$$\overline{X_n} = \frac{1}{n} \underbrace{\sum_{i=1}^{n} \underbrace{X_i}_{\sim \mathrm{N}(\mu;\sigma^2) \text{ i.i.d.}}}_{\sim \mathrm{N}(n\mu; n\sigma^2)} \sim \mathrm{N}(\mu; \sigma^2/n).$$

It follows that

$$\frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\overline{X_n} - \mu)}{\sigma} \sim \mathrm{N}(0; 1). \tag{7}$$

Let us now rewrite equation (6) in terms of the random variable $\frac{\overline{X_n} - \mu}{\sigma}\sqrt{n}$:

$$\begin{aligned} & \mathbb{P}_\mu(\mu \in [\overline{X_n} - r_\varepsilon, \overline{X_n} + r_\varepsilon]) = 1 - \varepsilon \\ \Leftrightarrow \quad & \mathbb{P}_\mu(-r_\varepsilon \le \overline{X_n} - \mu \le r_\varepsilon) = 1 - \varepsilon \\ \Leftrightarrow \quad & \mathbb{P}_\mu\Big(-\frac{r_\varepsilon}{\sigma}\sqrt{n} \le \frac{\overline{X_n} - \mu}{\sigma}\sqrt{n} \le \frac{r_\varepsilon}{\sigma}\sqrt{n} \Big) = 1 - \varepsilon. \end{aligned}$$

The last equality holds (denoting the distribution function of the standard normal distribution by Φ) if and only if

$$\Phi\Big(\frac{r_\varepsilon}{\sigma}\sqrt{n}\Big) - \Phi\Big(-\frac{r_\varepsilon}{\sigma}\sqrt{n}\Big) = 2\Phi\Big(\frac{r_\varepsilon}{\sigma}\sqrt{n}\Big) - 1 = 1 - \varepsilon,$$

which can be rearranged as

$$\Phi\Big(\frac{r_\varepsilon}{\sigma}\sqrt{n}\Big) = 1 - \frac{\varepsilon}{2}. \tag{8}$$

The distribution function $\Phi\colon \mathbb{R} \to (0,1)$ is strictly increasing on all of $\mathbb{R}$, so the inverse function $\Phi^{-1}\colon (0,1) \to \mathbb{R}$ exists and is also strictly increasing. Applying Φ^{-1} to both sides of (8) yields

$$\frac{r_\varepsilon}{\sigma}\sqrt{n} = \Phi^{-1}(1-\varepsilon/2),$$

that is,

$$r_\varepsilon = \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\varepsilon/2).$$

To state the final result, let us introduce, for $\delta \in (0,1)$, the notation

$$u_\delta = \Phi^{-1}(1-\delta). \tag{9}$$

Note that this is exactly the $(1-\delta)$-quantile of the standard normal distribution, since if $X \sim \mathrm{N}(0;1)$, then

$$\mathbb{P}(X < u_\delta) = \Phi(u_\delta) = \Phi(\Phi^{-1}(1-\delta)) = 1-\delta.$$

Thus

$$r_\varepsilon = \frac{\sigma}{\sqrt{n}}u_{\varepsilon/2},$$

and the desired confidence interval is

$$\big[\overline{X_n} - \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}}, \overline{X_n} + \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}}\big].$$

We summarize our result in the following statement.

Statement 3.2.1. *Let $X_1, \ldots, X_n$ be i.i.d. $\mathrm{N}(\mu;\sigma^2)$ distributed random variables, where the mean μ is unknown and the variance $\sigma^2 > 0$ is known, and let $\overline{X_n}$ denote the sample mean. Then for $\varepsilon \in (0,1)$ a confidence interval of level $1-\varepsilon$ for the mean μ is given by*

$$\big[T_1(\mathbf{X}), T_2(\mathbf{X})\big] = \big[\overline{X_n} - \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}}, \overline{X_n} + \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}}\big],$$

where $u_{\varepsilon/2}$ is the $(1-\varepsilon/2)$-quantile of the standard normal distribution.

Remark 3.2.1. Observe that:

1. the larger the sample size n, the shorter the required confidence interval (cf. the law of large numbers);
2. the larger the variance σ^2, the longer the required confidence interval (since larger variance means a "flatter", more spread-out density);
3. the smaller the error level ε, the longer the required confidence interval (since μ must lie in the confidence interval with probability $1-\varepsilon$ if μ is the true parameter).

A natural question arises: how do we construct a confidence interval for the variance of a normal distribution when the mean is known? Or how can we construct a confidence interval for the mean of a normal distribution when the variance is unknown? The remainder of this chapter is devoted to these questions. Informatics (or the corresponding course in any BSc programme in mathematics).
After that, we discuss the construction of confidence intervals for the variance of a normal distribution with known mean.

4 Hypothesis testing methods

4.1 Introduction

Besides estimation theory, the other major branch of mathematical statistics is *hypothesis testing*, where we aim to confirm or reject some prior hypothesis (assumption) based on measurements or data analysis. Examples of such hypotheses include:[9]

- The expected value of the sample elements is 2.
- The expected value of the sample elements is at most 2.
- The expected value of the sample elements is equal to the expected value of the elements of a given second sample.
- The sample elements are normally distributed.
- The realizations are independent of the outcomes of some other, related measurements.

In this chapter our goal is to study hypotheses belonging to the first three types. For further topics we again refer the interested reader to the course *Mathematical Statistics* in the MSc programmes in Computer Engineering and Business Informatics.
After sampling, we use calculations to examine whether the measurement results contradict our hypothesis, that is, whether the observed data are very unlikely under the assumption that the hypothesis holds. If so, we reject the hypothesis.
In general, a hypothesis test (or *test* for short) consists of the following steps:

1. Formalization of the *null hypothesis* H_0.
2. Construction of the *acceptance region* and the *critical region* for the sample.
3. Execution of the experiment, yielding a realization (data) $x_1, \ldots, x_n$.
4. Checking whether the data fall into the acceptance region. If yes, we accept H_0. Otherwise we reject H_0.

We now give the formal definitions of the above notions, and of some further related concepts.

Definition 4.1.1. *Let* $\mathbf{X} = (X_1, \ldots, X_n)$ *be an i.i.d. sample on the probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*, where the marginal distribution of the sample elements is unknown. Let* $\mathbf{x} = (x_1, \ldots, x_n)$ *be a realization of* $\mathbf{X} = (X_1, \ldots, X_n)$*.*

[9] In addition to the examples listed here, we may of course think of the example hypotheses mentioned at the beginning of Section 1.

1. *A **null hypothesis** H_0 is a statement whose truth value depends only on the joint distribution of $X_1, \dots, X_n$. The **alternative hypothesis** H_1 is the negation of H_0: $H_1 = \neg H_0$. The **alternative** associated with the given test is*

$$H_0 \text{ vs. } H_1.$$

2. *An **acceptance region** $\mathfrak{X}_\mathrm{e}$ is a subset of $\mathbb{R}^n$. The **critical region** $\mathfrak{X}_\mathrm{k}$ associated with the acceptance region $\mathfrak{X}_\mathrm{e}$ is the complement of $\mathfrak{X}_\mathrm{e}$, that is, $\mathfrak{X}_\mathrm{k} = \overline{\mathfrak{X}_\mathrm{e}} = \mathbb{R}^n \setminus \mathfrak{X}_\mathrm{e}$.*

   *The **(statistical) test** defined by the critical region $\mathfrak{X}_\mathrm{k}$ is the procedure in which we accept H_0 if $\mathbf{X} \in \mathfrak{X}_\mathrm{e}$, and reject H_0 (i.e. accept H_1) if $\mathbf{X} \in \mathfrak{X}_\mathrm{k}$.*

3. *The **size** (or **significance level**) of the test defined by the critical region $\mathfrak{X}_\mathrm{k}$ is the number $0 < \alpha < 1$ for which*

$$\mathbb{P}\big((X_1, \dots, X_n) \in \mathfrak{X}_\mathrm{k} \big| H_0 \text{ is true}\big) = \alpha.$$

4. *A **type I error** occurs when, in the course of the test procedure, we reject H_0 even though H_0 is true. A **type II error** occurs when we accept H_0 even though H_1 is true (that is, H_0 is false).*

Remark 4.1.1. For a test of size α, the probability of a type I error is exactly the size:

$$\mathbb{P}(\text{we reject } H_0 \mid H_0 \text{ is true}) = \mathbb{P}\big((X_1, \dots, X_n) \in \mathfrak{X}_\mathrm{k} \big| H_0 \text{ is true}\big) = \alpha.$$

In contrast, there is in general no explicit method to compute the probability of a type II error, as concrete examples will illustrate.

Some general, mathematically imprecise considerations and examples[10] may help in understanding how specific tests work.

Remark 4.1.2. The null hypothesis H_0 is often formulated in such a way that, if it is true, then the distribution of the sample elements $X_1, \dots, X_n$ is fully specified. In this sense, for the coin found on the street, a good choice for H_0 is the statement

$$H_0\colon \text{ “the coin is fair”},$$

because if H_0 is true, then the sample elements are i.i.d. indicator variables with parameter $1/2$.

By contrast, a poor choice for the null hypothesis would be

$$H_0'\colon \text{ “the coin is not fair”},$$

because under H_0' we only know that the sample elements are i.i.d. indicator variables with some parameter different from $1/2$, which does not provide enough information to compute the size of the test.

[10] These originate from Marianna Bolla's 2012 lecture course *Mathematical Statistics*.

Remark 4.1.3. It is often useful to interpret the null hypothesis as expressing a kind of presumption of innocence. For instance, we typically examine the coin because we suspect that it is biased, yet we take the negation of this suspicion as H_0. In this interpretation:

- a type I error can be seen as analogous to convicting an innocent person;
- a type II error can be seen as analogous to acquitting a guilty person.

(We emphasize once more that these are mathematically imprecise, and not really formalizable, statements.)
In most standard tests we can control only the probability of a type I error, that is, the size of the test: we fix a value α in advance, and the construction of the test guarantees that its size is exactly α. However, it is intuitively clear — even before learning about specific tests — that if we choose the desired size very small, then the probability of a type II error will increase. ("If we want to be very sure we do not convict innocent people, then we will end up acquitting guilty people more often as well.")
In practice, the most commonly used significance levels are $\alpha = 0.05$ and $\alpha = 0.01$.

Example 4.1.1. Suppose we want to introduce a new drug to the market. Then the alternative

$$H_0\colon \text{ "the drug is ineffective or harmful"} \quad \text{vs.} \quad H_1\colon \text{ "the drug is effective"}$$

is a reasonable choice, because in this case a type I error corresponds to the scenario where, based on the sample, we conclude that a drug that is in fact ineffective or harmful is effective (and thus we would likely release it). Since this is a type I error, we can control its probability and keep it below a prescribed level $1 - \alpha$, at the cost of increasing the probability of a type II error.
Here, a type II error corresponds to the case where the drug is effective, but based on the sample we still deem it ineffective or harmful (and therefore we will not introduce it). Of course, a type II error is also undesirable (e.g. from an economic perspective), but it does not have such potentially fatal consequences as a type I error.
Therefore, if we are using a hypothesis testing method for drug testing in which only the probability of a type I error is controllable, then we should choose H_0 as above.

4.2 u-test

In the remainder of this chapter we describe several variants of the u-test and the t-test. These are tests concerning the mean of a normal distribution, in the cases of known and unknown variance, respectively. We will see that these tests are closely related to the topic of confidence intervals (which is already suggested by the notation u and t). It is worth clarifying right at the start that the u- and t-tests only work under the assumption of normally distributed samples. By the central limit theorem, many kinds of samples can be reasonably approximated as normal, but there are also numerous applications

where this is not the case. For example, in the course *Mathematical Statistics* in the MSc programmes in Computer Engineering and Business Informatics, so-called *nonparametric tests* are also covered; these do not assume normality and can be applied to a wide variety of distributions. The best-known examples of nonparametric tests include the χ^2-*test* and the *Kolmogorov–Smirnov test*.
We introduce the two-sided, one-sample version of the u-test via an example, and then summarize the procedure in a concise form.

Example 4.2.1 (Two-sided, one-sample u-test: buying bread)**.** We buy a "1 kg" loaf of bread every day at the corner bakery, and recently the breads look smaller than we were used to before, so we suspect that the baker has started selling loaves whose mean weight is now less than 1 kg. We assume that the weights of the loaves we buy (measured in kilograms) are independent and normally distributed with unknown mean μ and known standard deviation $\sigma = 0.02$ kilograms (so $\sigma^2 = 0.0004$ kg^2).
We therefore set up the alternative

$$H_0 \colon \mu = \mu_0 \qquad \text{vs.} \qquad H_1 \colon \mu \neq \mu_0,$$

where in our case $\mu_0 = 1$. We prescribe a size $\varepsilon = 0.05$, i.e. we want the probability of rejecting H_0 under H_0 to be at most 0.05.
We then take a sample: for $n = 25$ days we place each purchased loaf on the scale. We obtain a realization $\mathbf{x} = (x_1, \ldots, x_{25})$ with sample mean $\overline{x_n} = \overline{x_{25}} = 0.98$ kg. Based on this sample, we want to construct a test of size exactly $\varepsilon = 0.05$, that is, we want to choose the acceptance region $\mathfrak{X}_\mathrm{e}$ and the critical region $\mathfrak{X}_\mathrm{k}$ so that

$$\mathbb{P}(H_0 \text{ is rejected} \mid H_0) = \varepsilon = 0.05$$

holds.
We look for the acceptance region in the form

$$\mathfrak{X}_\mathrm{e} = [\overline{X_n} - h, \overline{X_n} + h],$$

where $h > 0$. We want to choose h so that, when $\mu = \mu_0 = 1$ (we denote the corresponding probability by $\mathbb{P}_{\mu_0}$), the true mean $\mu = \mu_0$ falls into the critical region with probability exactly ε, i.e. into the acceptance region with probability $1 - \varepsilon$:

$$\mathbb{P}_{\mu_0}\big(\mu_0 \in [\overline{X_n} - h, \overline{X_n} + h]\big) = 1 - \varepsilon.$$

In other words, $[\overline{X_n} - h, \overline{X_n} + h]$ must be a symmetric confidence interval of level $1 - \varepsilon$ around $\overline{X_n}$ for the parameter μ_0. From Subsection 3.2 we already know that such a confidence interval is

$$\big[\overline{X_n} - \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}}, \overline{X_n} + \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}}\big],$$

and thus

$$h = \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}}.$$

That is, we accept H_0 exactly when

$$\overline{X_n} - \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}} \le \mu_0 \le \overline{X_n} + \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}} \qquad \Leftrightarrow \qquad \mu_0 - \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}} \le \overline{X_n} \le \mu_0 + \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}}$$

$$\Leftrightarrow \qquad -u_{\varepsilon/2} \le \frac{\overline{X_n} - \mu_0}{\sigma}\sqrt{n} \le u_{\varepsilon/2}.$$

Hence the acceptance region is

$$\mathfrak{X}_{\mathrm{e}} = \left\{\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n \colon\ -u_{\varepsilon/2} \le \frac{\overline{x_n} - \mu_0}{\sigma}\sqrt{n} \le u_{\varepsilon/2}\right\}.$$

Thus, if the *test statistic*

$$u(\mathbf{X}) = \frac{\overline{X_n} - \mu_0}{\sigma}\sqrt{n}$$

takes a realized value

$$u(\mathbf{x}) = \frac{\overline{x_n} - \mu_0}{\sigma}\sqrt{n}$$

that lies between $-u_{\varepsilon/2}$ and $u_{\varepsilon/2}$, then we accept H_0; otherwise we reject it (and thus accept H_1).
In our concrete case $\varepsilon = 0.05$, $n = 25$, $\overline{x_n} = 0.98$, $\mu_0 = 1$ and $\sigma = 0.02$. Thus

$$u(\mathbf{x}) = \frac{0.98 - 1}{0.02}\sqrt{25} = -5.$$

Looking into the standard normal table we find that $u_{\varepsilon/2} = \Phi^{-1}(0.975) \approx 1.9600$. Hence we would accept H_0 if the value of the test statistic $u(\mathbf{X})$ lay between -1.96 and 1.96. Since this is not the case, we reject H_0, i.e. we have sufficient evidence that the breads do not have mean weight 1 kg.

We now formalize the general procedure of the one-sample u-test; some details depend on the subtype of the test (one-sided or two-sided):

One-sample u-test Assumptions: the data points $x_1, \ldots, x_n$ are realizations of i.i.d. normal random variables $X_1, \ldots, X_n$ with unknown mean μ and known standard deviation $\sigma > 0$. General procedure:

1. Formulate the null hypothesis H_0.
2. Choose the desired size $\varepsilon \in (0, 1)$.
3. Collect the data $x_1, \ldots, x_n$.
4. Determine the acceptance region. This depends on the subtype of the test, but always involves computing (or reading from a table) an appropriate quantile of the normal distribution.

5. Compute the value of the test statistic

$$u(\mathbf{x}) = \frac{\overline{x_n} - \mu_0}{\sigma}\sqrt{n}.$$

6. Decide whether the value of the test statistic falls into the acceptance region. If yes, we accept H_0. If not, we reject H_0 (and thus accept H_1).

Based on our example, in the two-sided, one-sample u-test, the procedure can be made more precise as follows:

- Null hypothesis: $H_0 \colon \mu = \mu_0$.
- Relevant quantile: $u_{\varepsilon/2} = \Phi^{-1}(1 - \varepsilon/2)$.
- Acceptance region: $\mathfrak{X}_{\mathrm{e}} = \{(x_1, \ldots, x_n) \colon |u(\mathbf{x})| \leq u_{\varepsilon/2}\}$, that is:
  - if $|u(\mathbf{x})| > u_{\varepsilon/2}$, we reject H_0,
  - otherwise we accept H_0.

We now return to our previous example and introduce the one-sided, one-sample u-test as well.

Example 4.2.2 (One-sided, one-sample u-test: buying bread)**.** In the previous example, one may ask why we chose as null hypothesis that the mean weight of the breads is exactly 1 kg. After all, no reasonable person would interpret a sample mean of 0.98 kg as evidence that the mean is *greater* than 1 kg; we only see it as evidence that it is smaller. Accordingly, we can set up a different alternative (again in the spirit of a "presumption of innocence"):

$$H_0 \colon \mu \geq \mu_0 \qquad \text{vs.} \qquad H_1 \colon \mu < \mu_0, \tag{10}$$

where μ_0 is still equal to 1.
Does this alternative satisfy the requirement that if it holds, then the distribution of the sample elements is known? Strictly speaking, no: if H_0 holds, the sample elements might be distributed as $N(\mu_0; \sigma^2)$ or as $N(\mu_0 + 113; \sigma^2)$, for example. However, under H_0, a sample with mean less than μ_0 is most likely when μ equals μ_0 (rather than being larger). Therefore, if for any event A we (somewhat abusively) introduce the notation

$$\mathbb{P}(A \mid H_0 \text{ is true}) := \mathbb{P}(A \mid \text{the mean of the sample elements is } \mu_0),$$

i.e. we treat H_0 being true as the mean being equal to μ_0, then we can proceed similarly as in the two-sided, one-sample u-test.
The realized test statistic $u(\mathbf{x})$ is the same as in the two-sided case; the only difference is that we now reject H_0 only when the value $u = \frac{\overline{x_n} - \mu_0}{\sigma}\sqrt{n}$ is very small. More precisely, we choose an acceptance region that is unbounded to the right, such that for $\mu = \mu_0$ the

value μ_0 lies in the acceptance region $\mathfrak{X}_\mathrm{e}$ with probability $1-\varepsilon$, and in the critical region $\mathfrak{X}_\mathrm{k}$ to its left with probability ε:

$$\mathfrak{X}_\mathrm{k} = \{\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n \colon u(\mathbf{x}) < -u_\varepsilon\},$$

where $u_\varepsilon = \Phi^{-1}(1-\varepsilon)$ is again the $(1-\varepsilon)$-quantile of the standard normal distribution. In our example we found $u(\mathbf{x}) = -5$. The $(1-\varepsilon) = 0.95$ quantile of the standard normal distribution is $\Phi^{-1}(1-\varepsilon) = \Phi^{-1}(0.95) \approx 1.6449$. Since $u(\mathbf{x}) = -5 < -u_\varepsilon$, we again reject H_0, and conclude that the mean weight of the breads is less than 1 kg.
Thus, in the one-sided, one-sample u-test with $H_0 \colon \mu \geq \mu_0$, we never reject H_0 when the sample mean exceeds μ_0; however, for sample means smaller than μ_0 the test is stricter than the two-sided one-sample test: in our example, if $u(\mathbf{x})$ lies between -1.9600 and -1.6449, then the one-sided test already rejects H_0, whereas the two-sided test still accepts it. This is the main advantage of the one-sided test.

If we instead want to construct a one-sided, one-sample u-test for the alternative

$$H_0 \colon \mu \leq \mu_0 \qquad \text{vs.} \qquad H_1 \colon \mu > \mu_0, \tag{11}$$

then by symmetry we obtain the critical region

$$\mathfrak{X}_\mathrm{k} = \{\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n \colon u(\mathbf{x}) > u_\varepsilon\}.$$

Now we reject H_0 only when the value of the test statistic is very large. Summarizing the properties of the one-sided, one-sample u-test for both possible null hypotheses:

$H_0 \colon \mu \geq \mu_0$ **version**

- Null hypothesis: $H_0 \colon \mu \geq \mu_0$.
- Relevant quantile: $-u_\varepsilon = -\Phi^{-1}(1-\varepsilon)$.
- Acceptance region: $\mathfrak{X}_\mathrm{e} = \{(x_1, \ldots, x_n) \colon u(\mathbf{x}) \geq -u_\varepsilon\}$, that is:
  - if $u(\mathbf{x}) < -u_\varepsilon$, we reject H_0,
  - otherwise we accept H_0.

$H_0 \colon \mu \leq \mu_0$ **version**

- Null hypothesis: $H_0 \colon \mu \leq \mu_0$.
- Relevant quantile: $u_\varepsilon = \Phi^{-1}(1-\varepsilon)$.
- Acceptance region: $\mathfrak{X}_\mathrm{e} = \{(x_1, \ldots, x_n) \colon u(\mathbf{x}) \leq u_\varepsilon\}$, that is:
  - if $u(\mathbf{x}) > u_\varepsilon$, we reject H_0,

– otherwise we accept H_0.

Remark 4.2.1 (Power of a test, optimality of the u-test)**.** We have proved (at least sketchily) that the size of the two-sided, one-sample u-test, i.e. the probability of a type I error, is at most the prescribed parameter ε. What can we say about the type II error? If H_1 holds, then we do not know the distribution of the sample elements, since the mean of their normal distribution is unknown. Take some parameter $\mu_1 \neq \mu_0$, where μ_0 is the mean under the null hypothesis. Thus if μ_1 is the true parameter, then H_1 is true.
If we knew that μ_1 is the true parameter, then the type II error probability would be

$$\mathbb{P}(H_0 \text{ is accepted} \mid \ \mu_1 \text{ is the true parameter}) = \mathbb{P}\big(u(\mathbf{X}) \in [-u_{\varepsilon/2}, u_{\varepsilon/2}] \big|\ X_1 \sim \mathrm{N}(\mu_1; \sigma^2)\big),$$

which is a concrete value that can be computed. In this case, the **power** of the test is defined as 1 minus the probability of a type II error. The power is thus defined only for a specific parameter value μ_1 in H_1. It is clear that the further μ_1 is from μ_0, the smaller the type II error probability will be, and hence the larger the power of the test.
Why do we use the (two-sided, one-sample) u-test in practice, and not some other test of the same size ε for the same alternative (assuming the sample elements are normal with known σ)? Because it can be shown that the u-test is the **uniformly most powerful** test among all such tests. This means that, for any choice of μ_1 in H_1, the type II error probability for $\mu = \mu_1$ is minimized when we use the u-test. We do not prove this statement here, nor do we formulate it in a precise way; we only note that it follows from the **Neyman–Pearson lemma**. The Neyman–Pearson lemma states the existence of a uniformly most powerful test and provides its construction in general, and is typically covered in mathematical statistics courses aimed at mathematics students (for instance at BME-TTK).
In this sense, the one-sided, one-sample and the two-sided, two-sample u-tests, as well as all the t-tests discussed in these notes, are uniformly most powerful for their respective alternatives.

We now give an example where the underlying distribution is not normal, but the CLT still allows us to apply the u-test.

Example 4.2.3. We toss the coin found on the street, which shows heads with an unknown probability $\vartheta \in (0, 1)$, $n = 100$ times; the result is 60 heads and 40 tails. Decide, using the one-sided, one-sample u-test, whether the coin can be considered fair if the size of the test is $\varepsilon = 0.05$ and $\varepsilon = 0.01$.
The sample elements $X_i = \mathbb{1}_{\{\text{the } i\text{-th toss is heads}\}}$ are not normally distributed. They are indicator variables with parameter ϑ, so $\mathbb{E}_\vartheta[X_i] = \vartheta$, and their variance is also unknown: $\mathbb{D}^2_\vartheta[X_i] = \vartheta(1 - \vartheta)$.
The sample size $n \geq 30$ is large enough to apply the central limit theorem. Consider the alternative

$$H_0 \colon \vartheta \leq \frac{1}{2} \qquad \text{vs.} \qquad H_1 \colon \vartheta > \frac{1}{2}.$$

Again, under H_0, a sample with mean larger than $n\vartheta = 50$ is most likely when $\vartheta = \frac{1}{2}$. Therefore, we may (again somewhat abusively) assume that under H_0 we have $\vartheta = \frac{1}{2}$. In this case the sample elements X_i have mean $\frac{1}{2}$ and variance $\frac{1}{4}$.

Hence, by the central limit theorem,

$$u(X_1, \ldots, X_n) = \frac{\sum_{i=1}^n X_i - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}}$$

is approximately $\mathrm{N}(0;1)$ distributed. In the spirit of the one-sided, one-sample u-test, if the value of this statistic lies in $(-\infty, u_\varepsilon]$, we accept H_0, and otherwise we reject H_0.
In our concrete example,

$$u(x_1, \ldots, x_n) = \frac{60 - 50}{10/2} = 2.$$

We compare this to $u_\varepsilon = \Phi^{-1}(1-\varepsilon) = \Phi^{-1}(0.95) = 1.6449$ for $\varepsilon = 0.05$, and to $u_\varepsilon = \Phi^{-1}(0.99) = 2.3263$ for $\varepsilon = 0.01$. Therefore, for $\varepsilon = 0.05$ we reject H_0 (and conclude that the coin is not fair), whereas for $\varepsilon = 0.01$ we accept H_0 (and conclude that the coin is fair).

Remark 4.2.2 (p-value)**.** We see that decreasing ε makes it more likely that we accept H_0; this is not surprising, since ε coincides with the size of the test, i.e. with the probability of a type I error. In the coin example above, the largest size for which we still accept H_0 given the observed test statistic $u(\mathbf{x})$ lies between $\varepsilon = 0.01$ and $\varepsilon = 0.05$. This value is called the **p-value** of the test. The smaller the p-value, the more justified the rejection of H_0.
In the coin example (one-sided, one-sample u-test) the p-value is exactly 1 minus the value of the standard normal distribution function at the test statistic:

$$u(\mathbf{x}) = u_\varepsilon \;\Leftrightarrow\; \mathbb{P}(u(\mathbf{X}) > u(\mathbf{x})) = \varepsilon \;\Leftrightarrow\; \varepsilon = 1 - \Phi(u(\mathbf{x})) = 1 - \Phi(2) \approx 1 - 0.9772 = 0.0228.$$

So the p-value is approximately 2.28%. For sizes smaller than or equal to this value we accept H_0, and for larger sizes we reject it.
In the two-sided u-test, for example in the bread example, we have

$$|u(\mathbf{x})| = u_{\varepsilon/2} \;\Leftrightarrow\; \varepsilon = 2(1 - \Phi(|u(\mathbf{x})|)) = 2(1 - \Phi(5)) \approx 5.733 \times 10^{-7}.$$

Thus the p-value of the test is extremely small; we accept H_0 only if $\varepsilon \leq 5.733 \times 10^{-7}$.
Using the p-value represents an alternative viewpoint to the preset size: instead of choosing the size ε before sampling, and then deciding on acceptance or rejection of H_0 based on this size and the sample, we determine (from the sample) the size ε that lies exactly on the borderline between acceptance and rejection of H_0, i.e. the p-value.
It is common terminology to say that a statement about a sample is *significant* if the p-value is at most 0.05, and *highly significant* if it is at most 0.01. This terminology can be used for almost any hypothesis testing method; it means that, if we reject the null hypothesis based on the sample, then the probability of a type I error (i.e. the size) is at most 0.05 or 0.01, respectively.

At the end of this subsection, we turn to the **two-sample** u-test, which can be used to test the equality of the means of two normal populations. We only present the two-sided version here; the one-sided version can be derived from it in the same way as in the one-sample case. We again start with the bread example, in a slightly extended form.

Example 4.2.4 (Two-sided, two-sample u-test: buying bread)**.** After learning that, with size $\varepsilon = 0.05$, we must reject the null hypothesis that the breads bought at our original baker have mean weight 1 kg, we look for an alternative source of bread. Near our home there is another bakery that produces "1 kg" loaves that look, taste and cost very similarly to those of the first baker. This baker advertises that, using precision equipment, they can produce loaves whose weights have standard deviation only 0.01 kg (i.e. $\sigma_1 = 0.01$ kg, so $\sigma_1^2 = 0.0001$ kg^2).
We now buy one loaf per day from this baker for $n_1 = 10$ days and measure their weights. We obtain a realization $\mathbf{x} = (x_1, \dots, x_{n_1})$ with sample mean $\overline{x_{n_1}} = 0.9825$ kg. We introduce a new notation (which will be useful later) for the weights of the breads bought at the old bakery: $\mathbf{y} = (y_1, \dots, y_{n_2})$, where the sample size is $n_2 = 25$, the sample mean is $\overline{y_{n_2}} = 0.98$, and the standard deviation is $\sigma_2 = 0.02$. (We assume that the weight of any bread produced by either baker is independent of the weight of any bread produced by the same or the other baker.)
We would like to decide whether we are better off buying from this new baker, i.e. whether the unknown mean μ_1 of the weights of the breads from the new baker is greater than the mean μ_2 of those from the old baker. Therefore, for the alternative

$$H_0\colon \mu_1 = \mu_2 \qquad \text{vs.} \qquad H_1\colon \mu_1 \neq \mu_2$$

we want to construct a test of size $\varepsilon = 0.05$. Assuming that the sample $(X_1, \dots, X_{n_1})$ is independent of the sample $(Y_1, \dots, Y_{n_2})$, the sample mean $\overline{X_{n_1}}$ is independent of the sample mean $\overline{Y_{n_2}}$. Therefore, linear combinations of $\overline{X_{n_1}}$ and $\overline{Y_{n_2}}$ are also normally distributed. It can be shown by elementary calculations (which we omit here) that

$$u(\mathbf{X}, \mathbf{Y}) = \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has a standard normal distribution. Knowing this, we can proceed similarly to the two-sided, one-sample u-test. If

$$u(\mathbf{X}, \mathbf{Y}) \in [-u_{\varepsilon/2}, u_{\varepsilon/2}] = \big[-\Phi^{-1}(1-\varepsilon/2), \Phi^{-1}(1-\varepsilon/2) \big],$$

then we accept H_0, and otherwise we reject it. That is, the critical region is

$$\mathfrak{X}_{\mathrm{k}} = \big\{ (\mathbf{x}, \mathbf{y}) = (x_1, \dots, x_{n_1}, y_1, \dots, y_{n_2}) \in \mathbb{R}^{n_1+n_2} \colon |u(\mathbf{x}, \mathbf{y})| > u_{\varepsilon/2} \big\}.$$

In our concrete example,

$$u(\mathbf{x}, \mathbf{y}) = \frac{0.9825 - 0.98}{\sqrt{\frac{0.0001}{10} + \frac{0.0004}{25}}} \approx 0.4903.$$

Since $|u(\mathbf{x}, \mathbf{y})| < u_{\varepsilon/2} = \Phi^{-1}(0.975) = 1.9600$, we accept H_0. In other words, the mean weights of the breads at the new and the old bakeries are the same, so switching bakers does not improve our situation...

We now summarize the two-sided, two-sample u-test:

Two-sided, two-sample u-test Assumptions: the data points $x_1, \dots, x_{n_1}$ are realizations of i.i.d. normal random variables $X_1, \dots, X_{n_1}$ with unknown mean μ_1 and known standard deviation $\sigma_1 > 0$. The data points $y_1, \dots, y_{n_2}$ are realizations of i.i.d. normal random variables $Y_1, \dots, Y_{n_2}$ that are jointly independent of $X_1, \dots, X_{n_1}$, with unknown mean μ_2 and known standard deviation $\sigma_2 > 0$. Procedure:

1. Null hypothesis: $H_0 \colon \mu_1 = \mu_2$ vs. $H_1 \colon \mu_1 \neq \mu_2$.
2. Choose the desired size $\varepsilon \in (0, 1)$.
3. Collect the data $x_1, \dots, x_{n_1}, y_1, \dots, y_{n_2}$.
4. Test statistic:
$$u(\mathbf{x}, \mathbf{y}) = \frac{\overline{x_{n_1}} - \overline{y_{n_2}}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$
5. Acceptance region:
$$\mathfrak{X}_{\mathrm{e}} = \big\{(\mathbf{x}, \mathbf{y}) = (x_1, \dots, x_{n_1}, y_1, \dots, y_{n_2}) \in \mathbb{R}^{n_1+n_2} \colon |u(\mathbf{x}, \mathbf{y})| \leq u_{\varepsilon/2}\big\},$$
where $u_{\varepsilon/2} = \Phi^{-1}(1 - \varepsilon/2)$.
6. Decision: if $|u(\mathbf{x}, \mathbf{y})| > u_{\varepsilon/2}$, we reject H_0, otherwise we accept it.

References

The material of this supplementary note (in addition to the lecture notes by Szabolcs Mészáros) is based on Noemi Kurt's textbook *Stochastik für Informatiker* (Springer, 2020) and on the lecture material of Marianna Bolla's (BME-TTK) *Mathematical Statistics* course from the spring semester of 2013.

Formulas for confidence intervals and hypothesis tests

$s_n^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 - \overline{x_n}^2, \quad (s_n^*)^2 = \frac{n}{n-1}s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x_n})^2, \quad s_n^* = \sqrt{(s_n^*)^2}.$

u-test

1. Two-sided, one-sample: $u = \frac{\overline{x_n}-\mu_0}{\sigma}\sqrt{n}, \quad u_{\varepsilon/2} = \Phi^{-1}(1-\varepsilon/2)$,
   confidence interval for μ:

$$\big[\overline{x_n} - \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}}, \overline{x_n} + \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}}\big].$$

2. One-sided, one-sample: $u = \frac{\overline{x_n}-\mu_0}{\sigma}\sqrt{n}, \quad u_\varepsilon = \Phi^{-1}(1-\varepsilon)$.

3. Two-sided, two-sample: $u = \frac{\overline{x_{n_1}}-\overline{y_{n_2}}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}, \quad u_{\varepsilon/2} = \Phi^{-1}(1-\varepsilon/2)$.