Ivan Janiga

BASICS OF STATISTICAL ANALYSIS

STU
SLOVENSKÁ TECHNICKÁ
UNIVERZITA V BRATISLAVE

Ivan Janiga

BASICS OF STATISTICAL ANALYSIS

SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
2014

Recenzenti: PhDr. Jozef Galata, CSc.
doc. RNDr. Karol Pastor, CSc.
prof. Ing. Ladislav Starek, PhD.

Schválila Vedecká rada Strojníckej fakulty STU v Bratislave.

ISBN 978-80-227-4161-3

Contents

FOREWORD

Dear Readers,

We have written this textbook for the subject Basics of Applied Statistics, which is taught in the second year of the Bachelor program. It contains the following chapters: Probability, Random variables, Multivariate random variables, Creation of random sample and descriptive statistics, Point estimation, Statistical intervals and sample sizes at a given point estimation accuracy, Tests of hypotheses for a single sample and Statistical inference for two samples.

When writing the text, we placed emphasis on keeping the text as close as possible to the Engineer´s way of thinking. We avoided the exact mathematical formulations of the definitions. We tried to define new terms so as to be easier for engineers to understand and yet not lose their "exactness". The concepts are therefore explained using examples and figures.

Although the textbook was written primarily for the Bachelor program, it will also prove useful for students in higher engineering and doctoral studies. For researchers and workers in technical fields, it will also be helpful in the processing and evaluation of experimental data.

The textbook contains a lot of example problems and their solutions, in which the basic terms are clearly set out. At the end of the textbook, most necessary statistical tables are listed.

I wish to thank my reviewers, prof. Ing. Ladislav Starek, CSc., doc. RNDr. Karol Pastor, PhD. and PhDr. Jozef Galata, CSc., for their comments and reviewing the manuscript. Finally, I would like to express my great appreciation and gratitude to RNDr. Daniela Richtáriková, PhD., for her editing work, Mgr. Jana Gabková, PhD., for his valuable methodological comments, and Mgr. Milada Omachelová, PhD., for her beautiful pictures.

author

1 PROBABILITY

1.1 Random experiment, sample space and event

Learning goals

- Explain the terms *random experiment*, *sample space* and *event*.
- Define the sample space and event of a random experiment.
- Define a new joint event from existing events by using set operations.
- Assess if events are mutually exclusive and/or exhaustive.
- Explain the difference between discrete and continuous random variables.

Random experiment

When different results are obtained in repeated trials, the experiment is called a random experiment. Some sources of variability in the results are controllable and some are uncontrollable in the random experiment.
For example, when testing the life length of light bulbs, the sources of variability include:

- material,
- manufacturing process,
- production environment (temperature, humidity, etc.).,
- measuring instrument,
- drift of current,
- observer.

Sample space *Ω*

The sample space is the set of all possible results of the random experiment. We define two types of sample spaces.

1. **Discrete sample space**: consists of a finite (or countably infinite) number of outcomes. For example, a coin toss: $\Omega = \{\text{head}, \text{tail}\}$.
2. **Continuous sample space**: consists of infinite and innumerable outcomes. For example, life length of light bulbs: $\Omega = \{x\colon x \geq 0\}$.

Event *E*

An event is a subset of the sample space belonging to the random experiment.

Set Operations

To determine a new composite (joint) event from existing events we will use three set operations:

1. **union** ($E_1 \cup E_2$): combines all outcomes of E_1 and E_2,
2. **intersection** ($E_1 \cap E_2$): includes outcomes that are common to E_1 and E_2,
3. **complement** (E' or $\bar{E}$): contains outcomes that are not in E. Note that $(E')' = E$, while $E \cup E' = \Omega$.

Laws for set operations

The following laws are used in set operations:

1. **commutative law**

$$E_1 \cap E_2 = E_2 \cap E_1, \qquad E_1 \cup E_2 = E_2 \cup E_1,$$

2. **distributive law**

$$\left(E_1 \cap E_2\right) \cup E_3 = \left(E_1 \cup E_3\right) \cap \left(E_2 \cup E_3\right),$$

$$\left(E_1 \cup E_2\right) \cap E_3 = \left(E_1 \cap E_3\right) \cup \left(E_2 \cap E_3\right),$$

3. **deMorgan´s law**

$$\left(E_1 \cap E_2\right)' = E_1' \cup E_2', \ \left(E_1 \cup E_2\right)' = E_1' \cap E_2'.$$

Mutually exclusive events and complete system

A collection of events $E_1, E_2, \ldots, E_k$ is said to be **mutually exclusive** (**disjoint**), if the events do not have any outcomes in common, i.e.:

$$E_i \cap E_j = \varnothing \quad \text{for all pairs } (i, j):\ i \neq j.$$

The set of events $E_1, E_2, \ldots, E_k$ are said to be **exhaustive** (form a **complete system**) if their union is equal to Ω, that is

$$E_1 \cup E_2 \cup \ldots \cup E_k = \Omega.$$

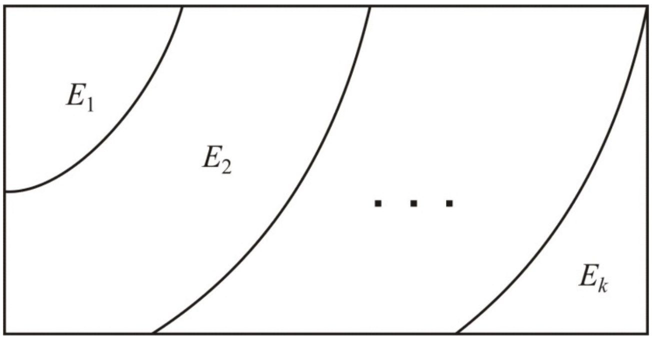


Figure 1.1 Mutually exclusive and exhaustive events

Example 1.1

The rise time (unit: min) of a reactor for two batches are measured in an experiment.

1. Define the sample space of the experiment.

   $\Omega = \{x:\ x > 0\}$, where *x* represents a rise time of the reactor for a certain batch.

2. Define an event A where the reactor rise time of the first batch is <u>less</u> than 55 minutes and B where the reactor rise time of the second batch is <u>greater</u> than 70 minutes.

   $A = \{x: 0 < x < 55\}$

   $B = \{x:\ x > 70\}$

3. Find $A \cup B$, $A \cap B$ and B'.

   $A \cup B = \{x: 0 < x < 55 \vee x > 70\}$ – the reactor rise time is less than 55 min or greater than 70 min.

   $A \cap B = \varnothing$ – the reactor rise time is less than 55 min and greater than 70 minutes; it is impossible.

   $A' = \{x: x \geq 55\}$ – the reactor rise time is not less than 55 minutes.

4. Are *A* and *B* mutually exclusive?

   Yes, because $A \cap B = \varnothing$.

5. Are *A* and *B* exhaustive?

   No, because $A \cup B \neq \Omega$.

Diagrams

Diagrams are often used to display a sample space and events in an experiment:

1. **Venn diagram**: A rectangle represents the sample space and circles indicate individual events, as illustrated in Fig. 1.2.

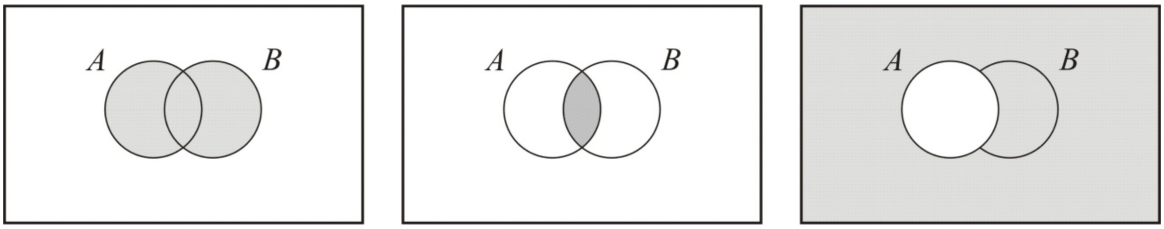


Figure 1.2 Venn diagrams: union, intersection and complement

2. **Tree diagram**: Branches represent possible outcomes, as shown in the following figure. The tree diagram method is useful when the sample space is established through several steps or stages.

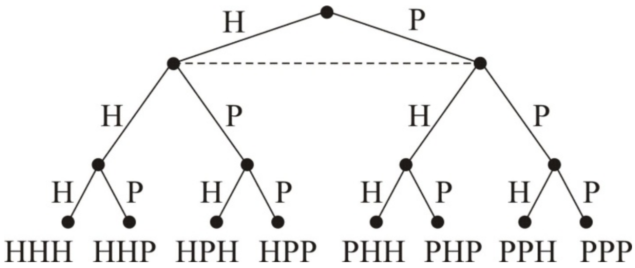


Figure 1.3 Tree diagram for outcomes of tossing three coins at the same time

1.2 Interpretations of probability

Learning goals

- Explain the term *probability*.
- Define the probability of an event.

Probability

The probability of an event means the likelihood of the event occurring in a random experiment. If Ω denotes the sample space and $A, A_1, A_2, A_3, \ldots$ denote events, the following conditions should be met:

1. $P(\Omega) = 1$
2. $0 \le P(A) \le 1$
3. $P(A_1 \cup A_2 \cup A_3 \cup \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots,$ where the events are mutually exclusive.

Classical definition of probability

If the sample space consists of n outcomes that are equally likely, the probability of each outcome is $1/n$. Then the probability of an event A consisting of k equally likely outcomes is

$$P(A) = \frac{k}{n}$$

where n is the number of possible outcomes in Ω and k is the number of equally likely elements in A.

Note. For any event A, $P(A') = 1 - P(A)$.

Statistical definition of probability

When we conduct n independent trials in the random experiment and monitored event A occurs k times, then the relative frequency of the occurrence of events A is $h_n(A) = \frac{k}{n}$; if for $n \to \infty$ the relative frequencies vary increasingly close within about a specific number, we can assume that this number is the probability of event A, i.e. $P(A)$. We estimate the value of $P(A)$ with a relative frequency

$$P(A) \approx h_n(A) = \frac{k}{n}.$$

Note. There is a significant difference between classical definition of probability and statistical definition of probability.

1.2.1 Probability of joint events

Learning goals

- Find the probability of a joint event by using probabilities of individual events.

Probability of joint events

The probability of a joint event can often be calculated by using the probabilities of the individual events involved. The following rules can be used to determine the probability of a joint event when the probabilities of existing events are known:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad \text{applies generally;}$$

$$P(A \cup B) = P(A) + P(B) \qquad \text{if } A \cap B = \varnothing;$$

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$

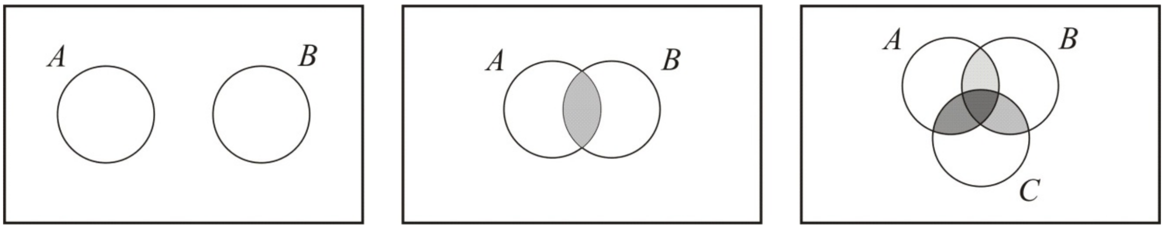


Figure 1.4 Venn diagrams for the probability of joint events

Example 1.2

A teacher of statistics tells students that the probabilities of obtaining grades of A, B, C, and D or below are 1/5, 2/5, 3/10 and 1/10, respectively. Find the probabilities of obtain signs:

1. A or B;
2. B or below.

<u>Solution</u>

Let E_1, E_2, E_3, E_4 denote the events of earning an A, B, C, and D or below, respectively. These individual events are mutually exclusive and exhaustive because

$$P(E_1 \cup E_2 \cup E_3 \cup E_4) = P(E_1) + P(E_2) + P(E_3) + P(E_4) =$$
$$= \frac{1}{5} + \frac{2}{5} + \frac{3}{10} + \frac{1}{10} = 1.$$

1. *The event of earning an* A *or* B is $E_1 \cup E_2$. Therefore,

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = \frac{1}{5} + \frac{2}{5} - 0 = \frac{3}{5}.$$

2. *The event of earning a* B *or below* is $E_2 \cup E_3 \cup E_4$, which is equal to E_1'. Therefore,

$$P(E_2 \cup E_3 \cup E_4) = P(E_1') = 1 - P(E_1) = 1 - \frac{1}{5} = \frac{4}{5}.$$

Example 1.3

Test results of scratch resistance and shock resistance for 100 disks of polycarbonate plastic are as follows:

Scratch resistance	Shock resistance	
	High	Low
High	80	9
Low	6	5

Let A denote the event that a disk has high scratch resistance and A' denote the event that a disk has low scratch resistance. Let B denote the event that a disk has high shock resistance and B' denote the event that a disk has low shock resistance (see below).

Scratch resistance	Shock resistance		$\sum$
	High (B)	Low (B')	
High (A)	80	9	89
Low (A')	6	5	11
$\sum$	86	14	**100**

1. When a disk is selected at random, find the probability that both the scratch and shock resistances of the disk are high.

$$P(A \cap B) = \frac{80}{100} = 0{,}8 = 80\,\%.$$

2. When a disk is selected at random, find the probability that the scratch or shock resistance of the disk is high.

   We know that $P(A) = \frac{89}{100}$, $P(B) = \frac{86}{100}$ a $P(A \cap B) = \frac{80}{100}$

   Therefore

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{89}{100} + \frac{86}{100} - \frac{80}{100} = \frac{95}{100} = 95\%.$$

3. Consider the event that a disk has high scratch resistance and the event that a disk has high shock resistance. Are these two events mutually exclusive?

   Because $P(A \cap B) = \frac{80}{100} \neq 0,$ the events A and B are not mutually exclusive.

1.3 Conditional probability

Learning goals

- Explain the term *conditional probability* of events.
- Calculate the conditional probability of events.

Conditional probability

The conditional probability $P(B|A)$ is the probability of an event B, given an event A. The following formula is used to calculate the conditional probability:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \text{ where } P(A) > 0.$$

Example 1.4

A new method of monitoring carpal tunnel syndrome at the workplace is tested with two groups of people: 50 workers having CTS and 50 healthy workers without CTS (see table below).

Group	Test result	
	Negative	Positive
CTS	10	40
Healthy	45	5

Let A denotes the event that a worker has CTS and A' denotes the event that a worker does not have CTS. Let B denotes the event that a CTS test is positive and B' denotes the event that a CTS test is negative. The summary of CTS test results is as follows:

Group	Test result		$\sum$
	Negative (B')	Positive (B)	
CTS (A)	10	40	50
Healthy (A')	45	5	50
$\sum$	55	45	100

1. Find the probability that a CTS test is positive (B) when a worker has CTS (A).

   We know that $P(A) = \frac{50}{100}$, $P(B) = \frac{45}{100}$ and $P(A \cap B) = \frac{40}{100}$, then it is valid:

   $$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{40/100}{50/100} = \frac{4}{5} = 80\%.$$

2. Find the probability that a worker has CTS (A), when a CTS test is positive (B).

   $$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{40/100}{45/100} = \frac{40}{45} = 88{,}89\%.$$

1.4 Multiplication and total probability rules

Learning goals

- Explain the multiplication rule.
- Explain the total probability rule.
- Apply the total probability rule to find the probability of an event when the event is partitioned into several mutually exclusive and exhaustive subsets.

Multiplication rule

From the definition of conditional probability

$$P(A\cap B)=P(B|A)P(A)=P(A|B)P(B)=P(B\cap A).$$

Total probability rules

1. When event B is partitioned into two mutually exclusive events $B\cap A$ and $B\cap A'$, then it is valid:

$$P(B)=P(B\cap A)+P(B\cap A')=$$
$$=P(B|A)P(A)+P(B|A')P(A').$$

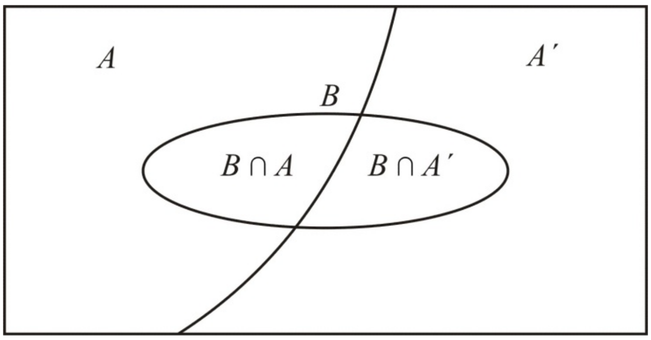


Figure 1.5 Partitioning event *B* into two mutually exclusive events

2. Let $A_1, A_2, ..., A_k$ be <u>mutually exclusive and exhaustive</u> events, then it is valid:

$$P(B)=P(B\cap A_1)+P(B\cap A_2)+\cdots+P(B\cap A_k)=...$$
$$=P(B|A_1)P(A_1)+P(B|A_2)P(A_2)+\cdots+P(B|A_k)P(A_k).$$

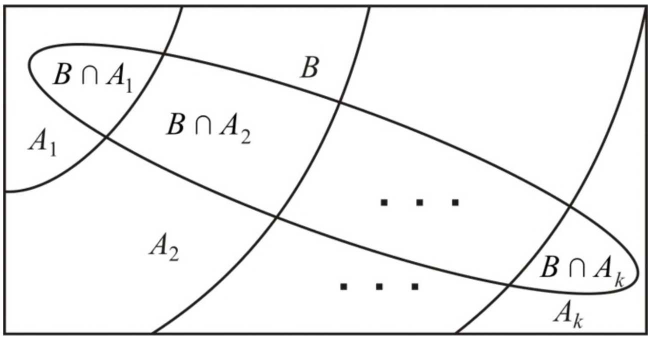


Figure 1.6 Partitioning event *B* into k mutually exclusive events

Example 1.5

In Example 1.4, the CTS screening method experiment indicates that the probability of screening a worker having CTS (A) as positive (B) is 0,8 and the probability of screening a worker without CTS (A') as positive (B) is 0,1. Then it is valid

$$P(B|A) = 0{,}8 \quad \text{a} \quad P(B|A') = 0{,}1 .$$

Suppose that the appearance of CTS in industry has probability $P(A) = 0{,}0017 = 0{,}17\%$. We will find probability that a randomly selected worker has positive CTS test (B) at the workplace.

We know that $P(A) = 0{,}0017$, then $P(A') = 1 - P(A) = 1 - 0{,}0017 = 0{,}9983$.

By using a total probability rule we will get:

$$P(B) = P(B|A) \times P(A) + P(B|A') \times P(A') =$$
$$= 0{,}8 \times 0{,}0017 + 0{,}1 \times 0{,}9983 = 0{,}101.$$

Example 1.6

Customer reviews are used to evaluate preliminary product design. In the past, 95% of very successful products, 60% of moderately successful products and 10% of poor products received good ratings. In addition, 40% of the product designs were very successful, 35% were moderately successful and 25% of the product designs were poor. We find the probability that the product will get good ratings.

Let A_1, A_2 and A_3 represent events – “very successful product,” “moderately successful product,” and “poor product.” Let us denote G the event of getting good rating from customers. Then

$$P(G|A_1) = 0{,}95 ; \quad P(G|A_2) = 0{,}60 ; \quad P(G|A_3) = 0{,}10 ;$$

$$P(A_1) = 0{,}40 ; \quad P(A_2) = 0{,}35 \quad \text{and} \quad P(A_3) = 0{,}25 .$$

The events A_1, A_2 and A_3 are mutually exclusive and exhaustive because:

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) =$$
$$= 0,40 + 0,35 + 0,25 = 1 = P(\Omega).$$

When we use the total probability rule, we get:

$$P(G) = P(G|A_1) \times P(A_1) + P(G|A_2) \times P(A_2) + P(G|A_3) \times P(A_3) =$$
$$= 0,95 \times 0,40 + 0,60 \times 0,35 + 0,10 \times 0,25 = 0,62 = 62\,\%.$$

1.5 Independence of two events

Learning goals

- Explain the term *independence* between events.
- Assess the independence of two events.

Independence of events

Two events A and B are stochastically independent if the occurrence of A does not affect the probability of B and vice versa. In other words, two events A and B are independent if and only if applies one of the following relations:

1. $P(A|B) = P(A)$
2. $P(B|A) = P(B)$
3. $P(A \cap B) = P(A)P(B)$

Derivation of the relationship $P(A \cap B) = P(A)P(B)$:

When events A and B are independent, then it is valid:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = P(B) \quad \Rightarrow \quad P(A \cap B) = P(A)P(B)\,.$$

Example 1.7

For the CTS test results in Example 1.4, the following probabilities have been calculated:

$$P(B) = \frac{45}{100} \quad \text{and} \quad P(B|A) = \frac{4}{5}\,.$$

We will find out if events A and B are independent. Because $P(B|A) = \frac{4}{5} \neq P(B) = \frac{45}{100}$, events A and B are not independent. This means that the information from the CTS test is useful for monitoring workers having CTS at the workplace.

1.6 Bayes´ theorem

Learning goals

- Apply Bayes´ theorem to find the conditional probability of an event when the event is partitioned into several mutually exclusive and exhaustive subsets.

Bayes´ theorem

From the definition of conditional probability we get:

$$P(A|B)=\frac{P(A\cap B)}{P(B)}=\frac{P(B|A)P(A)}{P(B)}=\frac{P(B|A)P(A)}{P(B\cap A)+P(B\cap A')}=$$
$$=\frac{P(B|A)P(A)}{P(B|A)\,P(A)+P(B|A')\,P(A')}.$$

The multiplication rule for a collection of k mutually exclusive and exhaustive events $A_1, A_2, ..., A_k$ and any event B is

$$P(B)=P(B|A_1)P(A_1)+P(B|A_2)P(A_2)+\cdots+P(B|A_k)P(A_k)$$

From the two expressions above, the following general result (known as Bayes´ theorem) is derived:

$$P(A_i|B)=\frac{P(B|A_i)P(A_i)}{P(B)}=\frac{P(B|A_i)P(A_i)}{P(B\cap A_1)+P(B\cap A_2)+\cdots+P(B\cap A_k)}=$$
$$=\frac{P(B|A_i)P(A_i)}{P(B|A_1)\,P(A_1)+P(B|A_2)\,P(A_2)+\cdots+P(B|A_k)\,P(A_K)}.$$

Example 1.8

In Example 1.4 and Example 1.5, the following probabilities have been calculated:

$$P(B|A)=0{,}8\,; \quad P(B|A')=0{,}1\,; \quad P(A)=0{,}0017 \quad \text{and} \quad P(B)=0{,}101\,.$$

We will find the probability that a worker has CTS (A) when the test is positive (B).

Using Bayes´ theorem we get

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}=\frac{P(B|A)P(A)}{P(B|A)\,P(A)+P(B|A')\,P(A')}=$$
$$=\frac{0{,}8\times 0{,}0017}{0{,}101}=0{,}013=1{,}3\%.$$

Since the occurrence of CTS in industry is low (0,17%), the probability that a worker has CTS is quite small (1,3 %) even if the test is positive.

We show the calculation using the following table.

Events A_i	Prior probabilities $P(A_i)$	Conditional probabilities $P(B\|A_i)$	Joint probabilities $P(A_i \cap B)$	Posterior probabilities $P(A_i\|B)$
A	0,0017	0,8	0,00136	0,01344
$A´$	0,9983	0,1	0,09983	0,98656
	1,0000		$P(B)$ = 0,10119	1,0000

2 RANDOM VARIABLES

Learning goals

- Explain the terms *random variable X* and *range of X*.
- Distinguish between discrete and continuous random variables.

Random variable

A random variable, denoted by an uppercase (capital letters) such as *X*, associates real numbers individual outcomes of a random experiment. Note that a measured value of *X* is denoted by a lowercase such as $x = 70$.

The set of possible numbers of *X* is referred to as the **range** of *X*. Depending on the type of the range, two categories of random variables are defined:

1. **Discrete random variable**: has a finite (or countably infinite) range.

   E.g. tossing a coin: $X = 0$ for head and $X = 1$ for tail.

2. **Continuous random variable**: has an interval of real numbers for its infinite range.
   E.g. the life length of an Infinity light bulb: $X \geq 0$.

2.1 Discrete random variables

2.1.1 Probability distributions and probability mass functions

Learning goals

- Distinguish between probability mass function and cumulative distribution function.
- Determine the probability mass function of a discrete random variable.

Probability distribution

A probability distribution indicates how probabilities are distributed over possible values of *X*.

Two types of functions are used to express the probability distribution of a discrete random variable *X*:

1. **probability mass function (p.m.f.):** describes the probability of a value of *X*, i.e., $P(X = x_i)$,
2. **cumulative distribution function (c.d.f.):** describes the sum of the probabilities of values of *X* that are less than or equal to a specified value, i.e., $P(X \leq x_i)$.

Probability mass function (p.m.f.)

The probability mass function of a discrete random variable *X*, denoted as $f(x)$, is

$$f(x_i) = P(X = x_i),\ x_i = x_1, x_2, \ldots, x_n,$$

which can be expressed by the table

x	x_1	x_2	x_3	$\cdots$	x_n
$f(x)$	$f(x_1)$	$f(x_2)$	$f(x_3)$	$\cdots$	$f(x_n)$

Probability mass function satisfies the following properties:

1. $f(x_i) \geq 0$ for all x_i
2. $\sum_{i=1}^{n} f(x_i) = 1$

Then its graph is as follows:

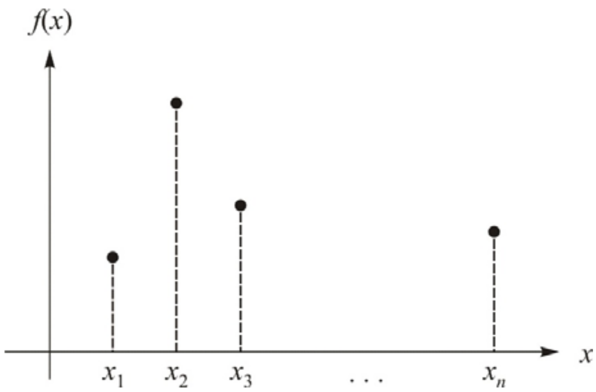


Figure 2.1

Example 2.1

The grades of $n = 50$ students in a statistics class are summarized as follows:

Marks	A	B	C	D	E	FX
Number of students	5	8	10	12	10	5

We determine the probability mass function of X and plot $f(x)$.

Solution:

Let random variable X (grade for the course) take its values $x = 1, 2, 3, 4, 5, 6$ representing marks A, B, C, D, E and FX.

x	1	2	3	4	5	6
Number of students	5	8	10	12	10	5

At first we calculate all the values of probability mass function:

$$f(x_1) = P(X=1) = \frac{5}{50} = 0{,}1 \qquad f(x_2) = P(X=2) = \frac{8}{50} = 0{,}16 \qquad f(x_3) = P(X=3) = \frac{10}{50} = 0{,}2$$

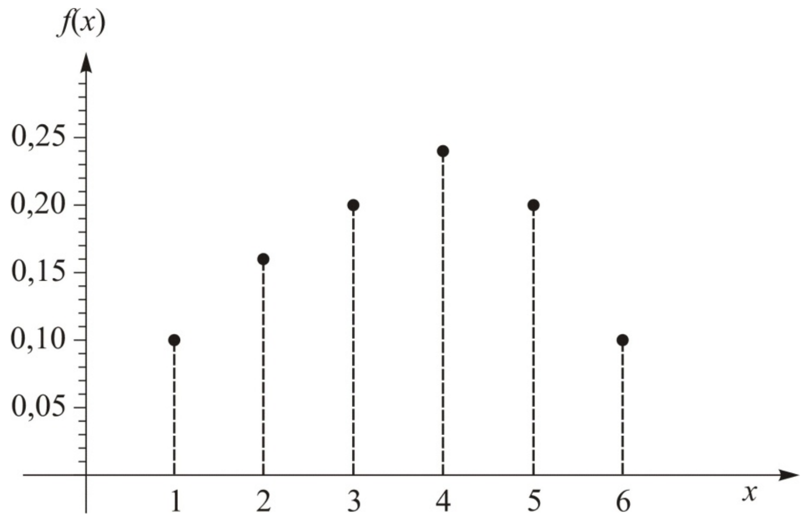
$$f(x_4) = P(X=4) = \frac{12}{50} = 0{,}24 \qquad f(x_5) = P(X=5) = \frac{10}{50} = 0{,}2 \qquad f(x_6) = P(X=6) = \frac{5}{50} = 0{,}1$$

Then

$f(x)$ given by table:

x	$f(x)$
1	0,10
2	0,16
3	0,20
4	0,24
5	0,20
6	0,10
$\sum$	1

$f(x)$ given by graph:



2.1.2 Cumulative distribution function

Learning goals

- Explain the term cumulative distribution function of a discrete random variable X, denoted as $F(x)$.
- Determine the cumulative distribution function of the discrete random variable.

Cumulative distribution function (c.d.f.)

The cumulative distribution function of a discrete random variable *X*, denoted as $F(x)$, is

$$F(x) = P(X \le x) = \sum_{x_i \le x} f(x_i) = \sum_{x_i \le x} P(X = x_i)$$

which can be expressed as follows

$$F(x) = \begin{cases} 0; & x < x_1 \\ f(x_1); & x_1 \le x < x_2 \\ f(x_1) + f(x_2); & x_2 \le x < x_3 \\ \quad \ldots & \quad \ldots \\ f(x_1) + f(x_2) + \ldots + f(x_{i-1}); & x_{n-1} \le x < x_n \\ 1; & x_n \le x \end{cases}$$

Cumulative distribution function has the following properties:

1. $0 \le F(x) \le 1$ for any real *x*
2. $F(x_1) \le F(x_2)$ for $x_1 < x_2$
3. $f(x_i) = F(x_i) - F(x_{i-1})$

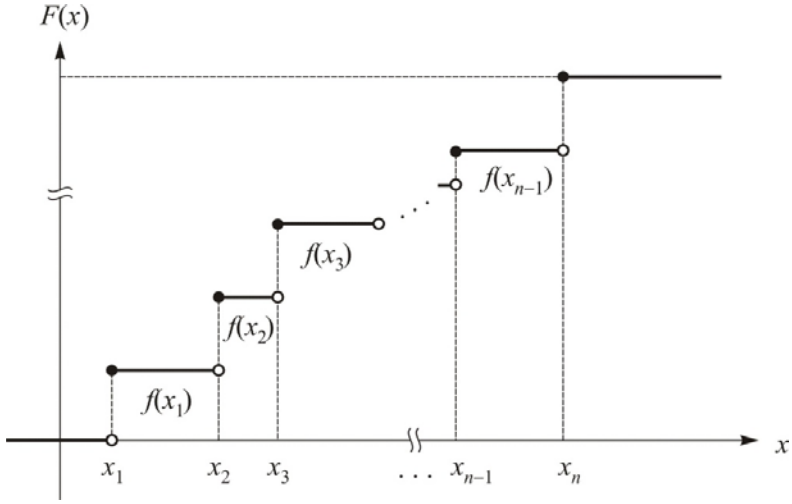


Figure 2.2 Distribution function given by graph

Example 2.2

In the previous example, we calculated the following probabilities:

$$P(X = 1) = 0,10; \quad P(X = 2) = 0,16; \quad P(X = 3) = 0,20;$$
$$P(X = 4) = 0,24; \quad P(X = 5) = 0,20; \quad P(X = 6) = 0,10$$

We determine the cumulative distribution function of the variable *X* and draw its graph**.**

By using the probability mass functions of *X* we get values of c.d.f. at individual points:

$F(1) = P(X \le 1) = P(X = 1) = 0{,}1$

$F(2) = P(X \le 2) = P(X = 1) + P(X = 2) = 0{,}1 + 0{,}16 = 0{,}26$

$F(3) = P(X \le 3) = 0{,}10 + 0{,}16 + 0{,}20 = 0{,}46$

$F(4) = P(X \le 4) = 0{,}10 + 0{,}16 + 0{,}20 + 0{,}24 = 0{,}70$

$F(5) = P(X \le 5) = 0{,}10 + 0{,}16 + 0{,}20 + 0{,}24 + 0{,}20 = 0{,}90$

$F(6) = P(X \le 6) = 0{,}10 + 0{,}16 + 0{,}20 + 0{,}24 + 0{,}20 + 0{,}10 = 1{,}00$

$P(X = 2) = P(X \le 2) - P(X \le 1) = 0{,}26 - 0{,}10 = 0{,}16$

Functional notation of c.d.f.:

$$F(x) = \begin{cases} 0 & x < 1 \\ 0{,}10 & 1 \le x < 2 \\ 0{,}26 & 2 \le x < 3 \\ 0{,}46 & 3 \le x < 4 \\ 0{,}70 & 4 \le x < 5 \\ 0{,}90 & 5 \le x < 6 \\ 1 & 6 \le x \end{cases}$$

Graph of c.d.f.:

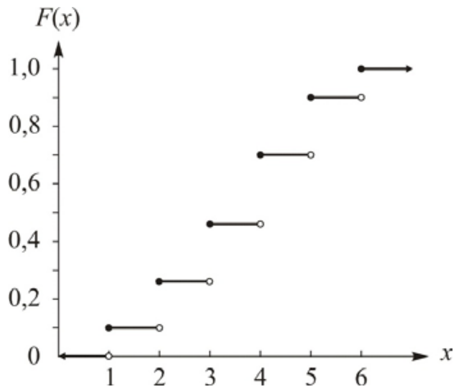


Figure 2.3 Cumulative distribution function *X*

2.1.3 Mean and variance of a discrete random variable

Learning goals

- Calculate the mean (expected value), variance and standard deviation of a discrete random variable.

Mean of *X*

The mean of *X*, denoted as μ or $E(X)$, means that the expected value of *X* and is defined by the relationship

$$\mu = E(X) = \sum_x x f(x)$$

Variance of *X*

The variance of *X*, denoted as σ^2 or $D(X)$, indicates the <u>dispersion</u> of *X* about μ and is defined by the relationship

$$\sigma^2 = D(X) = \sum_x (x-\mu)^2 f(x) = \sum_x x^2 f(x) - \mu^2$$

Standard deviation of *X*

The standard deviation of *X*, denoted as σ , is defined by the relationship

$$\sigma = \sqrt{\sigma^2} = \sqrt{D(X)} = \sqrt{\sum_x (x-\mu)^2 f(x)} = \sqrt{\sum_x x^2 f(x) - \mu^2}$$

Example 2.3

We determine the mean, variance and standard deviation of *X* (Example 2.1).
The probabilities of the values of *X*, we have calculated (see Example 2.1), are in the following table.

x	1	2	3	4	5	6
$f(x)$	0,10	0,16	0,20	0,24	0,20	0,10

Then

$$\mu = \sum_x xf(x) = 1\times 0,10 + 2\times 0,16 + 3\times 0,20 + 4\times 0,24 + 5\times 0,20 + 6\times 0,10 = 3,58$$

$$\sigma^2 = \left(\sum_x x^2 f(x)\right) - \mu^2 =$$

$$= \left(1^2\times 0,10 + 2^2\times 0,16 + 3^2\times 0,20 + 4^2\times 0,24 + 5^2\times 0,20 + 6^2\times 0,10\right) - 3,58^2 =$$

$$= 14,98 - 12,8164 =$$

$$= 2,1636$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{2,1636} = 1,47092$$

2.1.4 Discrete uniform distribution

Learning goals

- Describe the probability distribution of a discrete uniform random variable.
- Determine the probability function, mean, variance and standard deviation of a discrete uniform random variable.

Probability mass function of a discrete uniform distribution

A discrete uniform random variable X has an equal probability for each value in its range $\{a, a+1, a+2, \ldots, .b\}$, $a < b$ (see Figure 2.4). Thus, the probability mass function of X has the form

$$f(x) = \frac{1}{b-a+1}, \text{ where } \ x = a, a+1, \ldots, b$$

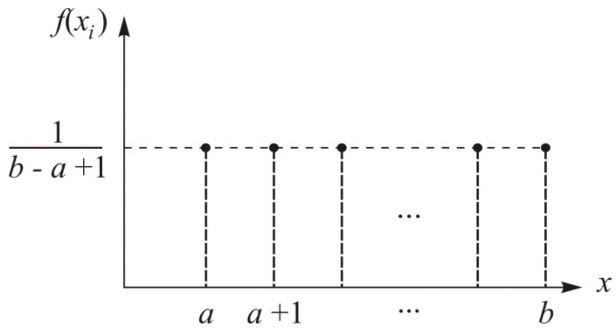


Figure 2.4 A discrete uniform distribution

The mean and **variance** of X are given by relations

$$\mu = \frac{a+b}{2} \quad \text{and} \quad \sigma^2 = \frac{(b-a+1)^2 - 1}{12}$$

Example 2.4

Suppose that six outcomes are equally likely in the experiment of casting a single die.

1. *Probability mass function of the discrete uniform distribution*

Determine the probability mass function of the number (X) of the die.

We know that X takes the values $x = 1, 2, \ldots, 6$, $a = 1$ and $b = 6$. Thus, the probability function of X is

$$f(x) = \frac{1}{b-a+1} = \frac{1}{6-1+1} = \frac{1}{6}, \ x = 1, 2, \ldots, 6$$

2. *Probability*

We find the probability that the number of points in the roll of the dice X in the experiment is <u>greater</u> than two.

$$P(X>2)=1-P(X\leq 2)=1-\sum_{i=1}^{2}\frac{1}{6}=1-\frac{2}{6}=\frac{2}{3}$$

3. *Mean, variance and standard deviation*

We know that $a=1$ and $b=6$, then the mean, variance and standard deviation of X is as follows:

$$\mu=\frac{a+b}{2}=\frac{1+6}{2}=3,5$$

$$\sigma^2=\frac{(b-a+1)^2-1}{12}=\frac{(6-1+1)^2-1}{12}=2,917=1,708^2$$

$$\sigma\ =\sqrt{2,917}=1,708$$

2.1.5 Binomial distribution

Learning goals

- Explain the terms *Bernoulli trial* and *binomial experiment*.
- Describe the probability distribution of a binomial random variable.
- Determine the probability mass function, mean and variance of a binomial random variable.

Binomial experiment

Binomial experiment refers to a random experiment consisting of n repeated trials which satisfy the following conditions:

1. The trials are independent, i.e. the outcome of a trial does not affect the outcomes of other trials.
2. Each trial has only two outcomes, labeled as “success“ and “failure“.
3. The probability of a „success „in each trial is constant and equals *p*.

In other words, a binomial experiment consists of a series of n independent Bernoulli trials (see the definition of Bernoulli trial below) with a constant probability of success (*p*) in each trial.

Bernoulli trial

A Bernoulli refers to a trial that has only two possible outcomes.

E.g. Bernoulli trials

1. flipping a coin: $\Omega = \{\text{head}, \text{tail}\}$
2. truth of an answer: $\Omega = \{\text{right}, \text{wrong}\}$
3. status of a machine: $\Omega = \{\text{working}, \text{broken}\}$
4. quality of a product: $\Omega = \{\text{good}, \text{defective}\}$
5. outcome of a task: $\Omega = \{\text{úspešný}, \text{neúspešný}\}$

Probability mass function of Bernoulli random variable X is

$$f(x) = \begin{cases} 1-p, & x = 0 \\ p, & x = 1 \end{cases}$$

The **mean, variance** and **standard deviation** of a Bernoulli random variable X are

$$\mu = p, \qquad \sigma^2 = p(1-p) \quad \text{and} \quad \sigma = \sqrt{p(1-p)}$$

Derivation of the relations for μ, σ^2 and σ of a Bernoulli random variable

$$\mu = \sum_x x f(x) = 0 \times (1-p) + 1 \times p = p$$

$$\sigma^2 = \left(\sum_x x^2 f(x)\right) - \mu^2 = \left(0^2 \times (1-p) + 1^2 \times p\right) - p^2 = p(1-p).$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{p(1-p)}$$

Binomial random variable

A binomial random variable X represents the number of trials whose outcome is a „success“ out of n trials in a binomial experiment with a probability of „success“ p (see Table 2.1).
General notation of a binomial distribution is $X \sim Bi(n, p)$.
The **probability mass function** of X is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, ..., n$$

General notation of a binomial distribution is $X \sim Bi(n, p)$.

Note. The number of combinations of x from n is equal to $C_x^n = \binom{n}{x} = \frac{n!}{x!(n-x)!}$.

The **mean**, **variance** and **standard deviation** of a binomial random variable X are

$$\mu = np, \quad \sigma^2 = np(1-p) \quad \text{and} \quad \sigma = \sqrt{np(1-p)}$$

Table. 2.1 Properties of binomial distribution

Distribution	Population*	Parameters		
		Probability of „success“	Number of trials	Frequency of „success“
binomial	infinite	p = constant °	n = constant	variable X

* If an item selected from a population is replaced before the next trial, the size of the population is considered <u>infinite</u> even if it may be finite.

° If the probability of success p is constant, the trials are considered <u>independent</u>; otherwise, the trials are dependent.

Example 2.5

A test has 50 multi-choice questions. Each question has four choices but only one answer is right. Suppose that a student gives his/her answers by simple guess.

1. *Probability mass function of a binomial distribution*

Determine the probability mass function of the number of right answers (X) that the student gives in the test.

Since a "right answer" is a success, the probability of a success for each question is $p = \frac{1}{4} = 0,25$. Thus, the probability mass function of X is given by the relationship

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{50}{x} \times 0,25^x \times 0,75^{50-x}, \quad x = 0,1,2,\ldots,50$$

2. *Probability*

Find the probability that the student answers <u>at least</u> 30 questions correctly.

$$P(X \geq 30) = 1 - P(X < 30) = 1 - \sum_{x=0}^{29} \binom{50}{x} \times 0,25^x \times 0,75^{50-x} = 1,6 \times 10^{-7}$$

3. *Mean, variance and standard deviation of the correct answers*
The calculation is as follows:

$$\mu = n\,p = 50 \times 0{,}25 = 12{,}5$$

$$\sigma^2 = n\,p(1-p) = 50 \times 0{,}25 \times 0{,}25 = 9{,}375 \quad \text{and} \quad \sigma = \sqrt{9{,}375} = 3{,}0619$$

2.1.6 Hypergeometric distribution

Learning goals

- Describe the probability distribution of a hypergeometric random variable.
- Compare the hypergeometric distribution with the binomial distribution.
- Determine the probability mass function, mean, variance and standard deviation of a hypergeometric random variable.

Hypergeometric random variable

A hypergeometric random variable X represents the number of successes in a sample of size n that is selected at random without replacement from a finite population of size N consisting of M successes and $(N-M)$ failures. Since each item selected from the population is not replaced, the outcome of a trial depends on the outcome(s) of the previous trial(s). Therefore, the probability of success p at each trial is not constant.

The **probability mass function** of X is given by the relationship

$$f(x) = \frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}}; \quad x = \max\{0, n+M-N\}, \ldots, \min\{M, n\}$$

where N, M, n are natural numbers that meet inequalities: $1 \le n < N, \quad 1 \le M < N$.
General notation of a hypergeometric distribution is $X \sim H(N, M, n)$.

The **mean, variance** and **standard deviation** of X are given by the relationships

$$\mu = np, \quad \sigma^2 = np(1-p)\frac{N-n}{N-1} \quad \text{and} \quad \sigma = \sqrt{np(1-p)\frac{N-n}{N-1}} \quad \text{where} \quad p = \frac{M}{N}$$

Note. The variance of a hypergeometric random variable is different from the variance of a binomial random variable by $(N-n)/(N-1)$, which is called **finite population correction factor**.

Hypergeometric versus binomial distribution

In the hypergeometric distribution the population is finite and probability of success is changing, whereas in the binomial distribution the population is infinite and the probability of success is constant (see Table 2.3).

Table 2.2 The characteristics of binomial and hypergeometric distributions

Distribution	Population*	Parameters		
		Probability of „success“	Number of trials	Number of „successes“
Binomial	infinite	p = constant °	n = constant	variable X
Hypergeometric	finite	p is changing °	n = constant	variable X

* If an item selected from a population is replaced before the next trial, the size of the population is considered infinite even if it may be finite.

° If the probability of success p is constant, the trials are considered independent; otherwise, the trials are dependent.

Example 2.6

Physical education tutor has prepared interview for a sample of ten randomly selected students from the class. The class consists of 30 students, 20 of which are football players and 10 are basketball players.

1. *Probability mass function of the hypergeometric distribution*

Determine the probability mass function of the number of basketball players *X* in the sample.

We know that the population includes $N = 30$ students. The number of selected students is $n = 10$. Since the number of basketball players is 10, $M = 10$.

To determine the range of the number of basketball players (*X*) in the sample, calculate the following:

$$\max\{0, n-(N-M)\} = \max\{0, 10-(30-10)\} = \max\{0, -10\} = 0$$

$$\min\{M, n\} = \min\{10, 10\} = 10$$

Therefore, the probability mass function of X is given by the relationship:

$$f(x) = \frac{\binom{M}{x}\cdot\binom{N-M}{n-x}}{\binom{N}{n}} = \frac{\binom{10}{x}\cdot\binom{30-10}{10-x}}{\binom{30}{10}} = \frac{\binom{10}{x}\cdot\binom{20}{10-x}}{\binom{30}{10}};\ x = 0, 1, 2, \ldots, 10$$

2. *Probability*

Find the probability that at least one basketball player is in the sample.

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{\binom{10}{0} \cdot \binom{20}{10-0}}{\binom{30}{10}} = 1 - 0{,}006 = 0{,}994$$

3. *Mean, variance and standard deviation of the number of basketball players in the sample*

The calculation is as follows

$$p = \frac{M}{N} = \frac{10}{30} = \frac{1}{3}$$

$$\mu = n\,p = 10 \times \frac{1}{3} = 3{,}333$$

$$\sigma^2 = np(1-p)\frac{N-n}{N-1} = 10 \times \frac{1}{3} \times \frac{2}{3} \times \frac{30-10}{30-1} = 1{,}532567 = 1{,}2379^2$$

2.1.7 Poisson distribution

Learning goals

- Explain the term *Poisson process*
- Describe the probability distribution of a Poisson random variable.
- Determine the probability mass function, mean and variance of a Poisson random variable.
- Compare the Poisson distribution with the binomial distribution.

Poisson process

Suppose that the occurrence of an event over an interval (of time, length, area, space, etc.) is countable and the interval can be partitioned into <u>subinterval</u>. Then a random experiment is defined as a Poisson process (see Figure 2.5) if it is valid:

1. The probability of more than one occurrence in a subinterval is infinitesimal (approximately zero).
2. The occurrences of the event in non-overlapping subintervals are stochastically indecent-dent.
3. The probability of one occurrence of the event in a subinterval is the <u>same</u> throughout all subintervals and <u>proportional</u> to the length of the subinterval.

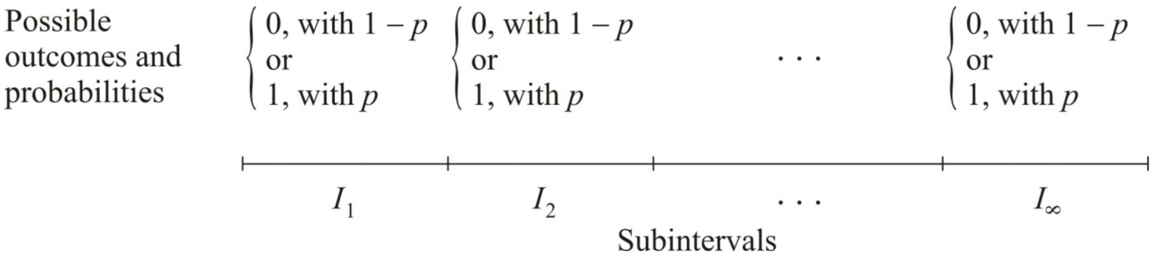


Figure 2.5. Poisson process

In other words, the Poisson process is a binomial experiment with infinite n trials. For example: the number of defects of product; the number of customers in a store; the number of automobile accidents; the number of e-mails received.

Poisson random variable

A Poisson random variable X represents the number of occurrences of an event of interest in a unit interval (of time, space, etc.) specified.

The **probability mass function** of X is given by the following relationship:

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \quad x = 0, 1, 2, \ldots$$

The **mean** and **variance** of X are

$$\mu = \lambda \quad \text{and} \quad \sigma^2 = \lambda$$

Note. Use consistent units to define a Poisson random variable X and the corresponding parameter λ. For example, the following pairs of X and λ are equivalent to each other:

X **counts/unit interval**	λ **average no. of counts/unit interval**
No. of flaws a disk	1
No. of flaws every 10 disks	10
No. of flaws every 100 disks	100

Poisson versus binomial distributions

In the Poisson distribution, the number of trials is infinite, whereas in the binomial distribution the number of trials is finite (see Table 2.3). In other words, the Poisson distribution with $E(X) = \lambda$ is the limiting form of binomial distribution with $E(X) = np$:

$$\lim_{x\to\infty} Bi(n,p) = \lim_{x\to\infty}\binom{n}{x}p^x(1-p)^{n-x} = \lim_{x\to\infty}\binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} = \frac{e^{-\lambda}\lambda^x}{x!}$$

Proof. Poisson versus binomial distributions

Suppose that X is a binomial random variable with parameters n and p , and let $\lambda = n\,p$.

Then

$$\lim_{x\to\infty}\binom{n}{x}p^x(1-p)^{n-x} = \lim_{x\to\infty}\frac{n!}{x!(n-x)!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} =$$

$$= \frac{\lambda^x}{x!}\lim_{x\to\infty}\frac{n(n-1)\cdots(n-x+1)}{n^x}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}$$

Table 2.3 The characteristics of binomial and Poisson distributions

Distribution	Population*	Parameters		
		Probability of „success“	No. of trials	No. of „successes“
Binomial	infinite	p = constant °	n = constant	variable X
Poisson	infinite	$p = \lambda / n$ = constant	n – infinite	variable X

* If an item selected from a population is replaced before the next trial, the size of the population is considered <u>infinite</u> even if it may be finite.

° If the probability of success p is constant, the trials are considered <u>independent</u>; otherwise, the trials are dependent.

As n rises above all limits, then the following applies:

$$\lim_{x\to\infty}\frac{n(n-1)\cdots(n-x+1)}{n^x} = 1$$

$$\lim_{x\to\infty}\left(1-\frac{\lambda}{n}\right)^n = \lim_{x\to\infty}\left[\left(1+\frac{1}{(-n/\lambda)}\right)^{-n/\lambda}\right]^{-\lambda} = e^{-\lambda}$$

$$\lim_{x\to\infty}\left(1-\frac{\lambda}{n}\right)^{-x} = 1$$

Therefore

$$\lim_{x\to\infty}\binom{n}{x}p^x(1-p)^{n-x} = \frac{e^{-\lambda}\lambda^x}{x!}$$

Example 2.7

The number of customers who buy at a local store has a Poisson distribution with mean 5 customers every 10 minutes.

1. *Probability mass function of the Poisson distribution*

Determine the probability mass function of number of customers X per <u>hour</u> coming to the local store.

The mean of X is

$\lambda = E(X) = 5$ customers/10 min. × 60 min. = 30 customers /hour

Therefore, the probability mass function of X is given by the next relationship:

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-30}30^x}{x!}, \quad x = 0,1,2,\ldots$$

2. *Probability*

Find the probability that 40 customers come to the local store in an hour.

$$f(40) = P(X = 40) = \frac{e^{-30}30^{40}}{40!} = 0{,}014 = 1{,}4\,\%$$

3. *Mean, variance and standard deviation of the number of customers per hour*

We calculate the mean and variance of X.

$\mu = \lambda = 30$ customers per hour

$\sigma^2 = \lambda = 30$ and $\sigma = \sqrt{30} = 5{,}478$

2.2 Continuous random variables

2.2.1 Probability distribution and probability density function

Learning goals

- Explain the term *probability density function* of *X*.
- Determine probability distribution of a continuous random variable by using the corresponding probability density function.

Probability distribution

Probability distribution of a continues random variable *X* is unambiguously defined by the **probability density function** $f(x)$ or **cumulative distribution function** $F(x)$.

Probability density function

Probability density function $f(x)$ of a continues random variable X satisfies the following properties:

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)\,dx$ pre arbitrary x_1 a x_2
4. $P(X = x) = 0$

From properties of density it follows that

$$P(x_1 \leq X \leq x_2) = P(x_1 \leq X < x_2) = P(x_1 < X \leq x_2) = P(x_1 < X < x_2)$$

Example 2.8

Suppose that X has the probability density function

$$f(x) = \begin{cases} e^{-x}, & x > 0 \\ 0, & \text{inde} \end{cases}$$

Calculate the following probabilities: $P(X < 2)$, $P(2 \leq X < 4)$ a $P(X \geq 4)$.

a) $P(X < 2) = \int_0^2 f(x)dx = \int_0^2 e^{-x}dx = \left[-e^{-x}\right]_0^2 = 0{,}86$

b) $P(2 \leq X < 4) = \int_2^4 f(x)dx = \int_2^4 e^{-x}dx = \left[-e^{-x}\right]_2^4 = -e^{-4} + e^{-2} = 0{,}12$

c) $P(X \geq 4) = \int_0^{\infty} f(x)dx = \int_0^{\infty} e^{-x}dx = \lim_{x\to\infty}(-e^{-x}) + e^{-4} = 0{,}02$

Note. $P(X < 2) + P(2 \leq X < 4) + P(X \geq 4) = 1$

2.2.2 Cumulative distribution functions

Learning goals

- Explain the term *cumulative distribution function of* X.
- Determine the cumulative distribution function of a continuous random variable.

Cumulative distribution function (c.d.f.)

The cumulative distribution function of a continuous random variable X is

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(u)du .$$

and satisfies the following properties:

1. $0 \le F(x) \le 1$
2. $F(x_1) \le F(x_2)$ if $x_1 < x_2$
3. $f(x) = \dfrac{dF(x)}{dx}$ for all x for which the derivative exists
4. $F(-\infty) = \lim_{x \to -\infty} F(x) = 0$ and $F(\infty) = \lim_{x \to \infty} F(x) = 1$

Other features of the probability density function and cumulative distribution function

1. $P(X \le x_0) = F(x_0) = \int_{-\infty}^{x_0} f(x)dx$
2. $P(x_1 \le X \le x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)dx$
3. $P(X \ge x_3) = 1 - F(x_3) = \int_{x_3}^{\infty} f(x)dx$

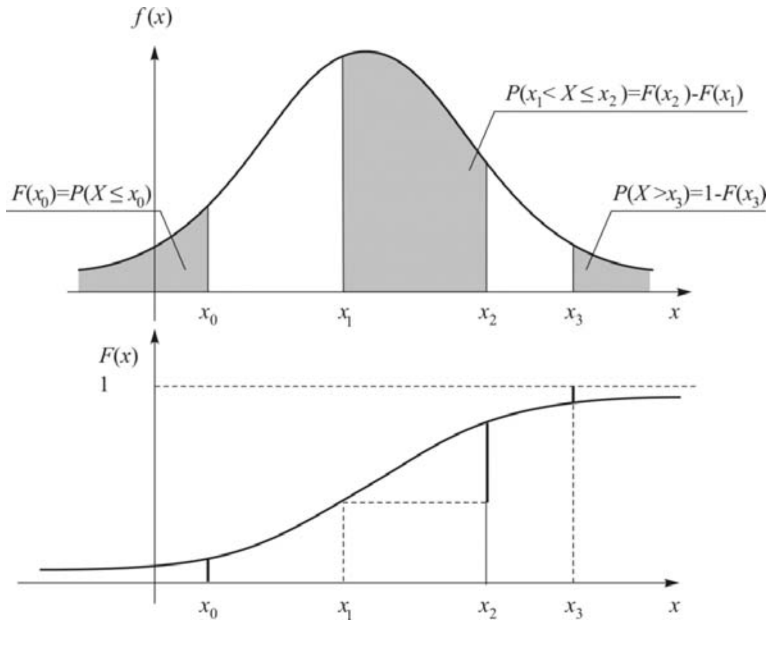


Figure 2.6 Properties of continuous distribution

Example 2.9

Let the probability density function of *X* (Example 2.8) is

$$f(x) = \begin{cases} e^{-x}, & x > 0 \\ 0, & \text{elsewhere} \end{cases}$$

Determine the cumulative distribution function of *X*. In the calculation we use the probability density function of *X*. Then it holds

$$F(x) = P(X \le x) = \int_0^x f(u)du = \int_0^x e^{-u}du = \left[-e^{-u}\right]_0^x = -e^{-x} + e^{-0} = 1 - e^{-x}$$

Therefore,

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-x}, & 0 \le x \end{cases}$$

2.2.3 Numerical characteristics of a continuous random variable

Learning goals

- Calculate the mean, variance and standard deviation of a continuous random variable.

Mean of X

The mean (expected value) of X is given by the relationship:

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

Variance of X

The variance of X is

$$\sigma^2 = D(X) = \int_{-\sigma\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2$$

Standard deviation of X

The standard deviation of X is given by the formula:

$$\sigma = \sqrt{D(X)}$$

Example 2.10

The probability density function of X is defined (Example 2.8) as

$$f(x)=\begin{cases} e^{-x}, & x>0 \\ 0, & \text{inde} \end{cases}$$

Determine the mean, variance and standard deviation of X.

1. The mean of X is

$$\mu=\int_0^\infty xf(x)dx=\int_0^\infty xe^{-x}dx$$

We use the method of <u>integration by parts</u> and get:

$$\mu=\int_0^\infty xe^{-x}dx=\lim_{x\to\infty}\left(-xe^{-x}\right)+\int_0^\infty e^{-x}dx=\lim_{x\to\infty}(e^{-x})+e^0=1$$

2. The variance of X is

$$\sigma^2=\int_0^\infty x^2f(x)dx-\mu^2=\int_0^\infty x^2e^{-x}dx-\mu^2=\int_0^\infty x^2e^{-x}dx-1$$

The integral is computed using the method per-partes:

$$\int_0^\infty x^2e^{-x}dx=\lim_{x\to\infty}(-x^2e^{-x})+0^2e^{-0}+\int_0^\infty 2xe^{-x}dx=2\int_0^\infty xe^{-x}dx=2\times1=2$$

Therefore,

$$\sigma^2=\int_0^\infty x^2e^{-x}dx-1=2-1=1$$

3. The standard deviation of X is $\sigma=1$.

2.2.4 Continuous uniform distribution

Learning goals

- Describe the probability distribution of a continuous uniform random variable.
- Determine the probability density function, cumulative distribution function, mean, variance and standard deviation of a continuous uniform random variable.

Probability density function

A continuous uniform random variable X has a <u>constant</u> **probability density function** over the range of X (see Figure 2.7):

$$f(x)=\begin{cases}\dfrac{1}{b-a} & \text{pre } a\le x\le b\\ 0 & \text{inde}\end{cases}$$

Cumulative distribution function

A continuous uniform random variable X has a **cumulative distribution function**

$$F(x)=\begin{cases}0 & x<a\\ \dfrac{x-a}{b-a} & a\le x<b\\ 1 & b\le x\end{cases}$$

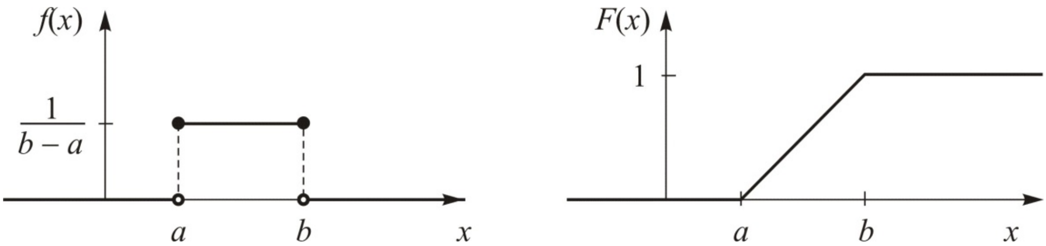


Figure 2.7 Continuous uniform distribution

The **mean** and **variance** of X are given by the following formulas:

$$\mu=\frac{a+b}{2} \quad \text{and} \quad \sigma^2=\frac{(b-a)^2}{12}, \text{ where } a\le x\le b$$

Example 2.11

Suppose that a random number generator produces real numbers that are uniformly distributed between numbers 0 and 100. Determine the probability density function, cumulative distribution function, probability, mean and variance σ^2 the random variable generated.

1. *Probability density function*

We know that $a=0$ and $b=100$, then applies:

$$f(x)=\frac{1}{b-a}=\frac{1}{100-0}=\frac{1}{100},\quad 0\le x\le 100$$

2. *Cumulative distribution function*

For $0 < x < 100$ is $F(x) = \int_a^x \frac{1}{100-0} du = \frac{x}{100-0} - \frac{0}{100-0} = \frac{x}{100}$, then

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{100} & 0 \le x < 100 \\ 1 & 100 \le x \end{cases}$$

3. *Probability*

Find the probability that a random variable (*X*) generated is between 10 and 90.

$$P(10 \le X \le 90) = \int_{10}^{90} f(x)dx = \int_{10}^{90} \frac{1}{100} dx = \frac{1}{100} \times [x]_{10}^{90} = \frac{1}{100} \times (90-10) = \frac{4}{5}$$

4. *Mean and variance*

Calculate the mean and variance of *X*.

$$\mu = \frac{a+b}{2} = \frac{0+100}{2} = 50$$

$$\sigma^2 = \frac{(b-a)^2}{12} = \frac{(100-0)^2}{12} = 833{,}333 = 28{,}8675^2$$

2.2.5 Normal and standard normal distributions

Learning goals

- Describe the properties of a normal distribution.
- Standardize a normal random variable.
- Using statistical tables to calculate probabilities.

Probability density function

A normal random variable with mean μ and variance σ^2 has the **probability density function**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty \le x \le \infty$$

A normal (Laplace – Gauss) distribution with **mean** μ and **variance** σ^2, denoted as $N(\mu, \sigma^2)$, is symmetric about μ and bell-shaped (see Figure 2.8). The symmetry of a normal curve implies

$$P(X < \mu) = P(X > \mu) = 0,5$$

The parameters μ and σ^2 determine the center (location) and shape of the normal (Gauss) curve, respectively. As illustrated in Figure 2.8, the larger the value of μ, the more to the right the center of the Gauss curve is located; the smaller the value of σ^2, the sharper the Gauss curve.

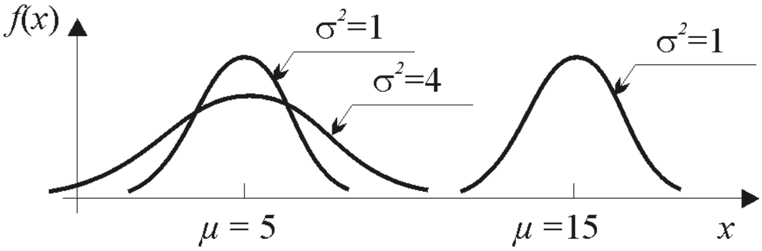


Figure 2.8 Gauss curves for selected parameter values μ a σ^2

Probabilities of normal distribution

Selected probabilities of a normal distribution are displayed in Figure 2.9. The area under the normal curve beyond $\pm 3\sigma$ is quite small (less than 0,003). Since 99,73 % of possible values of X are within the interval $(\mu - 3\sigma, \mu + 3\sigma)$, the range of 6σ is considered as **width of a normal distribution**.

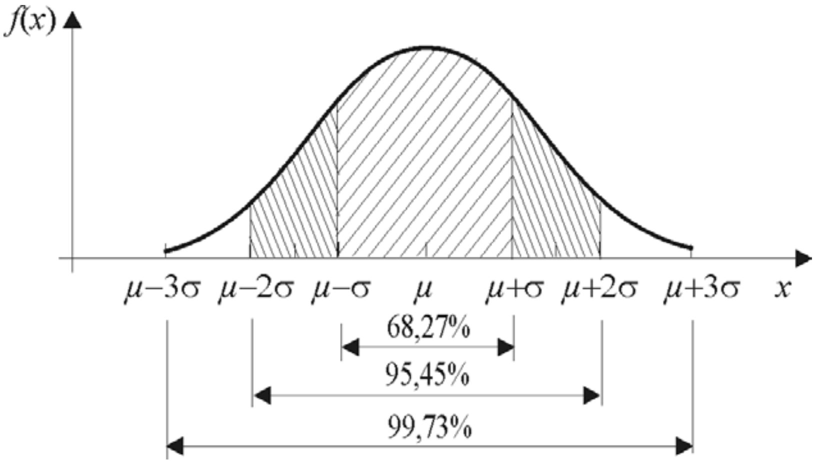


Figure 2.9 Probabilities of a normal distribution

Standard normal random variable

Any normal random variable of X with the parameters μ and σ^2 can be transformed to a standard normal random variable of Z with the parameters $\mu = 1$ and $\sigma^2 = 1$, denoted as $Z \sim N(0,1)$.

Table 2.4 Relationship between cumulative distribution functions

Transformation		
$N(\mu,\sigma^2)$	$\to$	$N(0,1)$
X	$\to$	$Z=\dfrac{X-\mu}{\sigma}$
x	$\to$	$z=\dfrac{x-\mu}{\sigma}$
$F(x)=P(X\le x)$	$\to$	$\Phi(z)=P(Z\le z)$
$F(x)=\Phi\left(\dfrac{x-\mu}{\sigma}\right)$		

Note. The value of $z=\dfrac{x-\mu}{\sigma}$ is called the ***z*-score**.

$$\Phi(-z)=1-\Phi(z) \quad \Leftrightarrow \quad \Phi(-z)+\Phi(z)=1$$

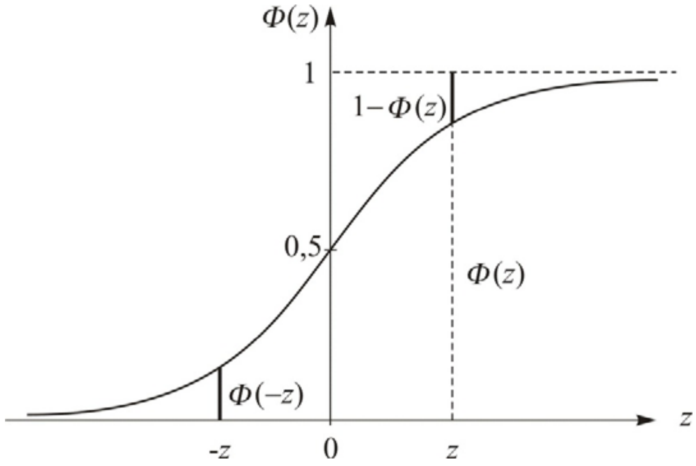
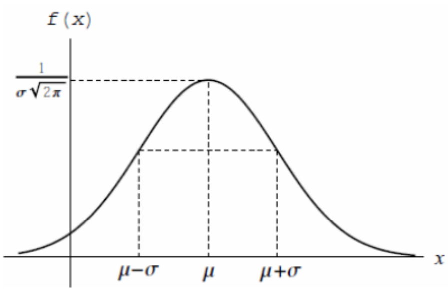
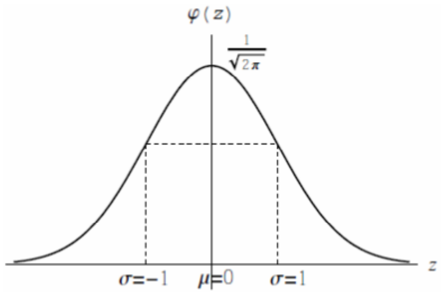
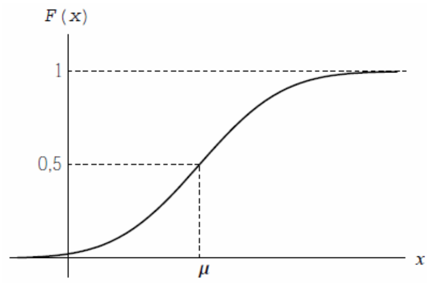
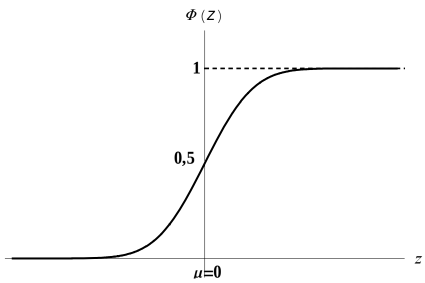


Figure 2.10

To calculate the probabilities we use **statistical tables**, that contain values of the function $\Phi(z)=\dfrac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{t^2}{2}}\,\mathrm{d}t$ for given z.

Table 2.5 A brief overview of the comparison of normal and standard normal distribution

Distribution	normal	standard normal
	$N(\mu,\sigma^2)$	$N(0,1)$
Random variable	X	$Z=\dfrac{X-\mu}{\sigma}$
Value of random variable	x	$z=\dfrac{x-\mu}{\sigma}$
Mean	$\mu=E(X)\in R$	$\mu=0$
Variance	$\sigma^2=D(X)$	$\sigma^2=1$
Probability density function	$f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ for $-\infty<x<\infty$ 	$\varphi(z)=\dfrac{1}{\sqrt{2\pi}}e^{\frac{-z^2}{2}}$ for $-\infty<z<\infty$ 
Cumulative distribution function	$F(x)=\dfrac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}\mathrm{d}t$ for $-\infty<x<\infty$ 	$\Phi(z)=\dfrac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-\frac{t^2}{2}}\mathrm{d}t$ for $-\infty<z<\infty$ 
Properties	– symmetrical about μ – has a bell shape $P(\mu-\sigma<X<\mu+\sigma)=68,27\%$ $P(\mu-2\sigma<X<\mu+2\sigma)=95,45\%$ $P(\mu-3\sigma<X<\mu+3\sigma)=99,73\%$	– symmetrical about μ – has a bell shape $P(-1<Z<1)=68,27\%$ $P(-2<Z<2)=95,45\%$ $P(-3<X<3)=99,73\%$

Example 2.12

Compute the following probabilities:

1. $P(Z > 2,11) = 1 - P(Z \leq 2,11) = 1 - 0,98257 = 0,01743$
2. $P(Z < -0,41) = P(Z > 0,41) = 1 - P(Z \leq 0,41) = 1 - 0,65910 = 0,3409$
3. $P(Z > -2.91) = P(Z < 2,91) = 0,99819$
4. $P(-1,09 < Z < 2,37) = \Phi(2,37) - \Phi(-1,09) =$
$$= \Phi(2,37) - \big(1 - \Phi(1,09)\big) =$$
$$= 0,99111 - (1 - 0,86214) =$$
$$= 0,99111 - 0,13786 = 0,85325$$
5. $P(Z \leq -4,6)$ – is not in statistical tables, can be calculated using statistical software Statgraphics, Minitab, Statistica, SPSS, SPlus and the other.
6. We want to find the value z such that $P(Z > z) = 0,02$. This probability expression can be written as $1 - P(Z \leq z) = 0,02 \Leftrightarrow P(Z \leq z) = 0,98$. Now statistical table is used in reverse. We search through the probabilities to find the value that corresponds to 0,98. Because we do not find 0,98 exactly, we take the nearest value of the probability that is 0,98030, corresponding to $z = 2,06$.
7. We are finding the value of z such that $P(-z < Z < z) = 0,98$. Because of symmetry of the standard normal distribution, if the area of the shaded region in Figure 2.9 is to equal 0,98, the area in each tail of distribution must equal $(1\text{- }0,98)/2 = 0,01$. Therefore, the value for z corresponds to a probability of 0,99. The nearest probability in Table is 0,99010, when $z = 2,33$.

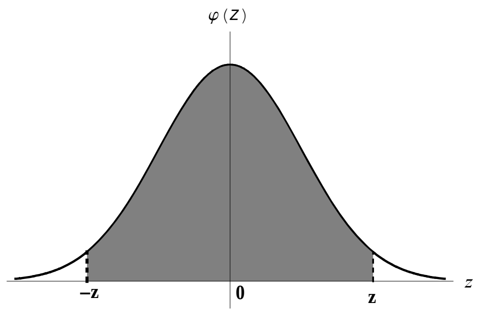


Figure 2.11

Example 2.13

The line width of for semiconductor manufacturing is assumed to be normally distributed with mean of 0,5 micrometer (μm) and the standard deviation of 0,05 micrometer (μm).

1. We calculate the probability that a line width is greater than 0,62 μm.

When the width of the line is marked as X its distribution is then $X \sim N(0,5;0,05^2)$. We calculate the following probability:

$$P(X > 0,62) = 1 - P(X \le 0,62) = 1 - F(0,62) =$$
$$= 1 - P\left(Z \le \frac{0,62 - 0,5}{0,05}\right) = 1 - P(Z \le 2,4) = 1 - \Phi(2,4) =$$
$$= 1 - 0,9918 = 0,0082 = 0,82\%$$

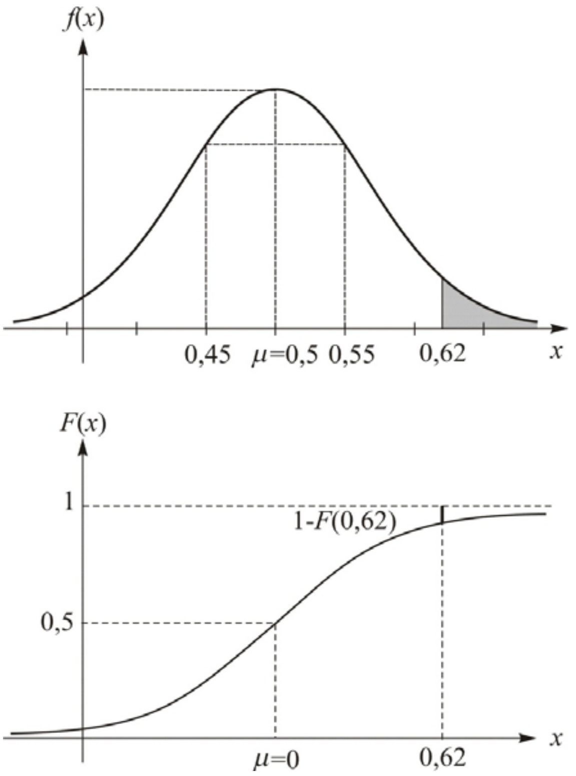


Figure 2.12a

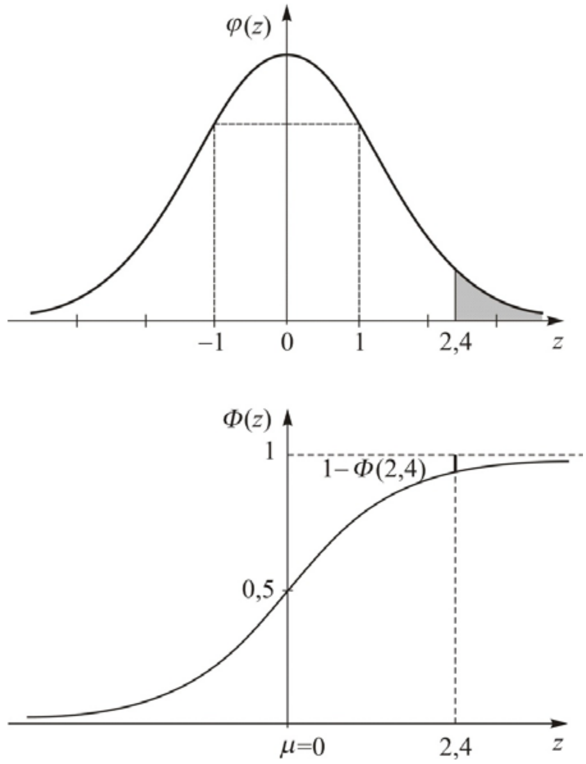


Figure 2.12b

2. We want to calculate the probability that a line width is between 0,47 µm and 0,63 µm. We calculate the following probability:

$$P(0,47 \le X \le 0,63) = F(0,63) - F(0,47) =$$
$$= P\left(\frac{0,47-0,5}{0,05} \le Z \le \frac{0,63-0,5}{0,05}\right) = P(-0,6 \le Z \le 2,6) =$$
$$= \Phi(2,6) - (1 - \Phi(0,6) = 0,9953 - 1 + 0,7257 = 0,721 = 72,1\%$$

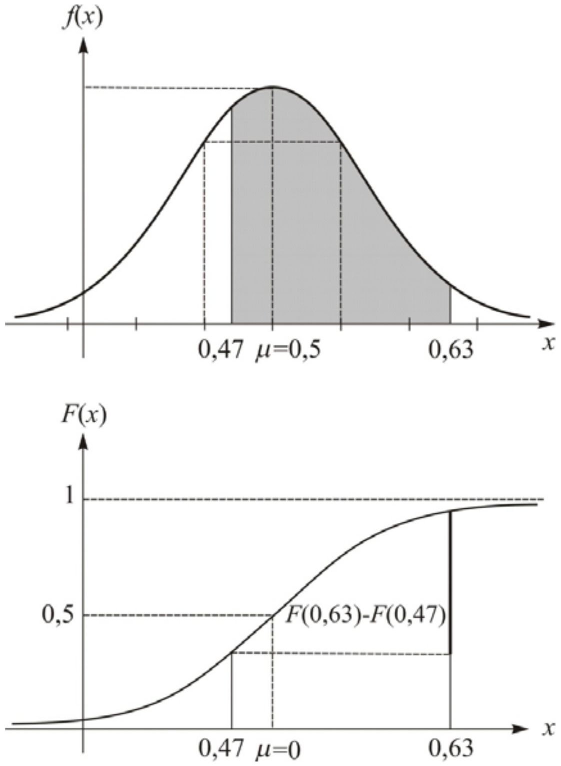


Figure 2.13a

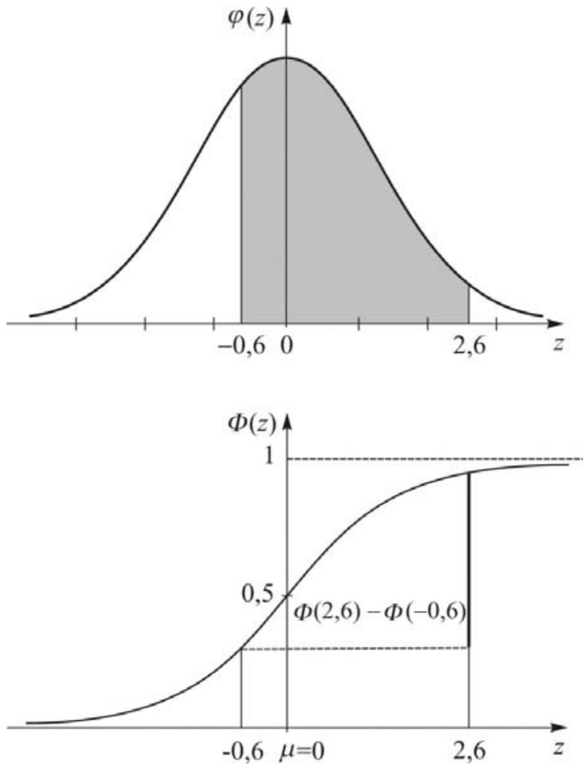


Figure 2.13b

3. We want to find the value of *x* below which is the 90% of values of the sample.
We are looking for the value of *x* for which is valid:

$$P(X < x) = 0,90$$

$$P\left(Z < \frac{x - 0,5}{0,05}\right) = P(Z < z) = 0,90$$

The closest value to the value of the probability 0,90, found in statistical tables, is 0,8997. The corresponding value is $z = 1,28$.

Then from $\frac{x - 0,5}{0,05} = 1,28$, we get $x = 1,28 \times 0,05 + 0,5 = 0,064 + 0,5 = 0,564$.

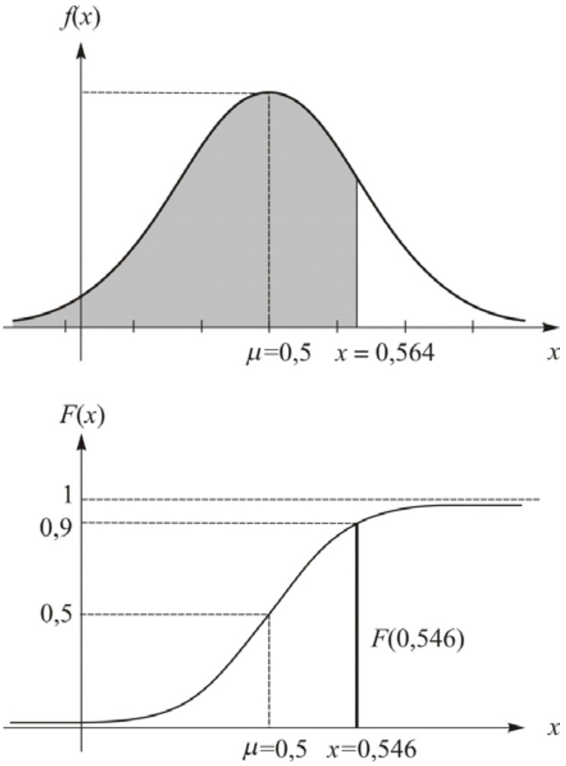


Figure 2.14a

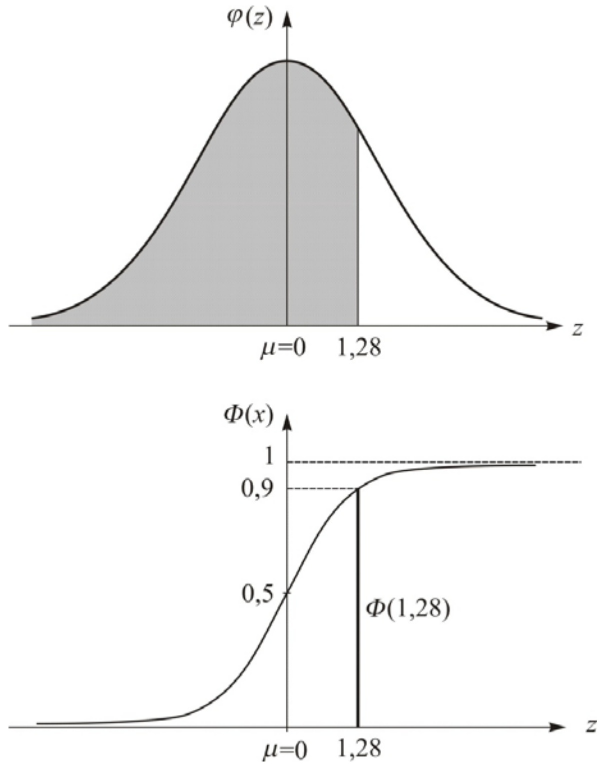


Figure 2.14b

3 MULTIVARIATE RANDOM VARIABLES

3.1 Two discrete random variables

Learning goals

- Determine joint, marginal and conditional probabilities for two discrete random variables X and Y by using corresponding probability distributions.
- Calculate the mean and variance of X and Y by using corresponding marginal probability distribution.
- Calculate the conditional mean and conditional variance of X given $Y = y$ (or Y given $X = x$) by using the corresponding conditional probability distribution.
- Assess the independence of X and Y.

Probability distribution of two random variables

Three kinds of probability distributions are used to describe the stochastic characteristics of two random variables *X* and *Y*:

1. joint probability distribution
2. marginal probability distribution
3. conditional probability distribution

Joint probability mass function

The joint probability mass function (p.m.f.) of two discrete random variables *X* and *Y*, denoted as $f_{XY}(x, y)$, satisfies the following conditions:

1. $f_{XY}(x, y) \geq 0$
2. $\sum_x \sum_y f_{XY}(x, y) = 1$
3. $f_{XY}(x, y) = P\left(X = x, Y = y\right)$

Marginal probability mass function

The marginal p.m.f.´s of X and Y with the joint p.m.f. $f_{XY}(x, y)$ are

$$f_X(x) = P(X = x) = \sum_{R_x} f_{XY}(x, y)$$

$$f_Y(y) = P(Y = y) = \sum_{R_y} f_{XY}(x, y)$$

where R_x and R_y denote the set of all points in the range of (X, Y) for which $X = x$ and $Y = y$, respectively.

The marginal p.m.f. $f_X(x)$ satisfies the following properties:

1. $f_X(x) = P(X = x) \geq 0$
2. $\sum_x f_X(x) = 1$

Similar relationships can be applied to $f_Y(y)$.

The **mean** and **variance** of X are

$$E(X) = \mu_X = \sum_x x f_X(x)$$

$$D(X) = \sigma_X^2 = \sum_x (x - \mu_X)^2 f_X(x) = \sum_x x^2 f_X(x) - \mu_X^2$$

Conditional probability mass function

Recall that $P(B|A) = P(A \cap B)/P(A)$ (see Section 1.3. Conditional probability). In parallel, the conditional p.m.f. of X given $Y = y$, denoted as $f_{X|y}(x)$, with the joint p.m.f. $f_{XY}(x, y)$ is

$$f_{X|y}(x) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f_{XY}(x, y)}{f_Y(y)}$$

The conditional p.m.f. $f_{X|y}(x)$ satisfies the following:

1. $f_{X|y}(x) \geq 0$
2. $\sum_x f_{X|y}(x) = 1$

Similar relationships apply to $f_{Y|x}(y)$.

The **conditional mean** and **conditional variance** of X given $Y = y$ are

$$E(X|y) = \mu_{X|y} = \sum_x x f_{X|y}(x)$$

$$D(X|y) = \sigma^2_{X|y} = \sum_x (x - \mu_{X|y})^2 f_{X|y}(x) = \sum_x x^2 f_{X|y}(x) - \mu^2_{X|y}$$

Similar relationships apply to $E(Y|x), D(Y|x)$.

Independence

Two random variables X and Y are independent if knowledge of the values of X does not affect <u>any</u> probabilities of the values of Y and vice versa. Thus, two independent random variables X a Y satisfy any of the following:

1. $f_{X|y}(x) = f_X(x)$
2. $f_{Y|x}(y) = f_Y(y)$
3. $f_{XY}(x, y) = f_X(x) f_Y(y)$

Example 3.1

The number of defects on the front side (X) of a wooden panel and the number of defects on the rear side (Y) of the panel are under study.

1. Suppose that the joint p.m.f. of X and Y is modeled as

$$f_{XY}(x, y) = c(x + y),\ x = 1,2,3 \text{ and } y = 1,2,3$$

Determine the value of c.

The joint p.m.f. of $f_{XY}(x, y)$ must satisfy any of the following:

a) $f_{XY}(x, y) = c(x + y) \geq 0$

b) $\sum_x \sum_y f_{XY}(x, y) = 1$

For the first condition, $c \geq 0$ because $x > 0$ and $y > 0$.

Next, for the second condition:

$$\sum_{x=1}^{3} \sum_{y=1}^{3} f(x, y) = \sum_{x=1}^{3} \sum_{y=1}^{3} c(x + y) = 36c = 1 \Rightarrow c = 1/36$$

Therefore, the joint p.m.f. of X and Y is

$$f_{XY}(x,y)=\frac{1}{36}(x+y)\ ,\ x=1,2,3 \text{ and } \ y=1,2,3$$

2. We determine the marginal p.m.f. of *X*. We find the mean and variance of *X*.

The marginal probabilities of *X* are:

$$f_X(1)=\sum_{y=1}^{3} f_{XY}(1,y)=\sum_{y=1}^{3}\frac{1}{36}(1+y)=\frac{1}{36}\cdot 9=\frac{1}{4}$$

$$f_X(2)=\sum_{y=1}^{3} f_{XY}(2,y)=\sum_{y=1}^{3}\frac{1}{36}(2+y)=\frac{1}{36}\cdot 12=\frac{1}{3}$$

$$f_X(3)=\sum_{y=1}^{3} f_{XY}(3,y)=\sum_{y=1}^{3}\frac{1}{36}(3+y)=\frac{1}{36}\cdot 15=\frac{5}{12}$$

Note that $\sum_{x} f_X(x)=f_X(1)+f_X(2)+f_X(3)=1$.

The mean and variance of *X* are

$$E(X)=\mu_X=\sum_{x=1}^{3} xf(x)=1\cdot f_X(1)+2\cdot f_X(2)+3\cdot f_X(3)=$$

$$=1\times\frac{1}{4}+2\times\frac{1}{3}+3\times\frac{5}{12}=2{,}17$$

$$D(X)=\sigma_X^2=\left(\sum_{x=1}^{3} x^2 f_X(x)\right)-\mu_X^2=$$

$$=\left(1^2\cdot f_X(1)+2^2\cdot f_X(2)+3^2\cdot f_X(3)\right)-2{,}17^2=$$

$$=\left(1^2\times\frac{1}{4}+2^2\times\frac{1}{3}+3^2\times\frac{5}{12}\right)-2{,}17^2=0{,}80^2$$

3. We determine the conditional p.m.f. of *Y* given $X=2$. We find the conditional mean and conditional variance of *Y* given $X=2$.

The conditional marginal probabilities of *Y* given $X=2$ are:

$$f_{Y|2}(1)=\frac{f_{XY}(2,1)}{f_X(2)}=\frac{(1/36)(2+1)}{1/3}=\frac{1}{4}$$

$$f_{Y|2}(2)=\frac{f_{XY}(2,2)}{f_X(2)}=\frac{(1/36)(2+2)}{1/3}=\frac{1}{3}$$

$$f_{Y|2}(3)=\frac{f_{XY}(2,3)}{f_X(2)}=\frac{(1/36)(2+3)}{1/3}=\frac{5}{12}$$

Note that $\sum_{y} f_{Y|2}(y)=f_{Y|2}(1)+f_{Y|2}(2)+f_{Y|2}(3)=1$.

The conditional mean and conditional variance of *Y* given $X = 2$ are:

$$E(Y|2) = \mu_{Y|2} = \sum_{y=1}^{3} y f_{Y|2}(y) = 1 \cdot f_{Y|2}(1) + 2 \cdot f_{Y|2}(2) + 3 \cdot f_{Y|2}(3) =$$

$$= 1 \times \frac{1}{4} + 2 \times \frac{1}{3} + 3 \times \frac{5}{12} = 2{,}17$$

$$D(Y|2) = \sigma^2_{Y|2} = \left(\sum_{y=1}^{3} y^2 f_{Y|2}(y) \right) - \mu^2_{Y|2} =$$

$$= \left(1^2 \cdot f_{Y|2}(1) + 2^2 \cdot f_{Y|2}(2) + 3^2 \cdot f_{Y|2}(3) \right) - 2{,}17^2 =$$

$$= \left(1^2 \times \frac{1}{4} + 2^2 \times \frac{1}{3} + 3^2 \times \frac{5}{12} \right) - 2{,}17^2 = 0{,}80^2$$

4. We verify that the number of defects on the front side (X) of the wood panel and the number of defects on the rear side (Y) of the panel are independent.
Check if

$$f_{XY}(1,1) = f_X(1) \cdot f_Y(1)$$

We know that $f_{XY}(1,1) = \frac{1}{36}(1+1) = \frac{1}{18}$, $f_X(1) = \frac{1}{4}$ a $f_Y(1) = \frac{1}{4}$, then

$$\left(f_{XY}(1,1) = \frac{1}{18} \right) \neq \left(f_X(1) f_Y(1) = \frac{1}{4} \times \frac{1}{4} = \frac{1}{16} \right)$$

Thus, the number of defects on the front side (X) of a wooden panel and the number of defects on the rear side (Y) of the panel are not independent.

3.2 Multiple discrete random variables

3.2.1 Joint probability distributions

Learning goals

- Explain the joint, marginal and conditional probability distribution of multi discrete random variables.
- Explain the independence of multi discrete random variables.

Joint probability mass function

A joint p.m.f. of discrete random variables $X_1, X_2, ..., X_n$ is given by the relationship

$$f_{X_1 X_2 \ldots X_n}(x_1, x_2, ..., x_n) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$$

and defined for all $x_1, x_2, ..., x_n$ in the range of $X_1, X_2, ..., X_n$.

Marginal probability mass function

Let the discrete random variables $X_1, X_2, ..., X_n$ have joint p.m.f. $f_{X_1X_2...X_n}(x_1, x_2, ..., x_n)$, then the marginal p.m.f. of X_i is

$$f_{X_i}(x_i) = P(X_i = x_i) = \sum_{R(x_i)} f_{X_1X_2...X_n}(x_1, x_2, ..., x_n) \ ,$$

where $R(x_i)$ denotes the set of all points in the range of $X_1, X_2, ..., X_n$ for which $X_i = x_i$.

Independence

The discrete random variables $X_1, X_2, ..., X_n$ are independent if and only if

$$f_{X_1X_2...X_n}(x_1, x_2, ..., x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \ldots f_{X_n}(x_n) \text{ for all } x_1, x_2, ..., x_n .$$

3.3 Two continuous random variables

Learning goals

- Determine joint, marginal and conditional probabilities for two continuous random variables X and Y by using corresponding probability distributions.
- Calculate the mean and variance of X and Y a continuous random variable by using the corresponding marginal probability distribution.
- Calculate the conditional mean and conditional variance of X given $Y = y$ and of Y given $X = x$ by using the corresponding conditional probability distributions.
- Assess the independence of X and Y.

Joint probability density function

The joint probability density function (p.d.f.) of two continuous random variables X and Y satisfies the following:

1. $f_{XY}(x, y) \geq 0$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) dx dy = 1$

Marginal probability density function

The marginal p.d.f.´s of *X* and *Y* with the joint p.d.f. $f_{XY}(x,y)$ are

$$f_X(x) = \int_y f_{XY}(x,y)dy$$

$$f_Y(y) = \int_x f_{XY}(x,y)dx$$

The marginal p.d.f. of *X* satisfies the following:

1. $f_X(x) \geq 0$
2. $\int_x f_X(x)dx = 1$

The **mean** and **variance** of *X* are

$$E(X) = \mu_X = \int_x x f_X(x)dx$$

$$D(X) = \sigma_X^2 = \int_x (x-\mu_X)^2 f_X(x)dx = \int_x x^2 f_X(x)dx - \mu_X^2$$

Conditional probability distribution

The conditional p.d.f. of *X* given $Y = y$ with joint p.d.f. $f_{XY}(x,y)$ is

$$f_{X|y}(x) = \frac{f_{XY}(x,y)}{f_Y(y)}$$

The conditional p.d.f. $f_{X|y}(x)$ satisfies the following:

1. $f_{X|y}(x) \geq 0$
2. $\int_x f_{X|y}(x)dx = 1$

The **conditional mean** and **conditional variance** of *X* given $Y = y$ are

$$E(X|y) = \mu_{X|y} = \int_x x f_{X|y}(x)dx$$

$$D(X|y) = \sigma_{X|y}^2 = \int_x (x-\mu_{X|y})^2 f_{X|y}(y)dy = \int_{R_x} y^2 f_{X|y}(y)dy - \mu_{X|y}^2$$

Similar relationships apply to $f_Y(y), E(Y), D(Y), f(Y|x), E(Y|x), D(Y|x)$.

Independence

Two continuous random variables *X* and *Y* are independent if any of the following is true:

1. $f_{X|y}(x) = f_X(x)$
2. $f_{Y|x}(y) = f_Y(y)$
3. $f_{XY}(x, y) = f_X(x) f_Y(y)$

Example 3.2

Two measurement methods are used to evaluate the surface smoothness of a paper product. Let *X* and *Y* denote the measurements of each of the two methods.

1. Suppose that the join p.d.f. of *X* and *Y* is modeled by

$$f_{XY}(x, y) = c, \ \ 0 < x < 4, \ \ x - 1 < y < x + 1$$

Determine the value of *c*.

The ranges of *X* and *Y* are given in the following picture. Note that the range of integration for *X* is divided into two parts:

I. $0 < x \le 1$, $0 < y < x + 1$ and II. $1 < x < 4$, $x - 1 < y < x + 1$

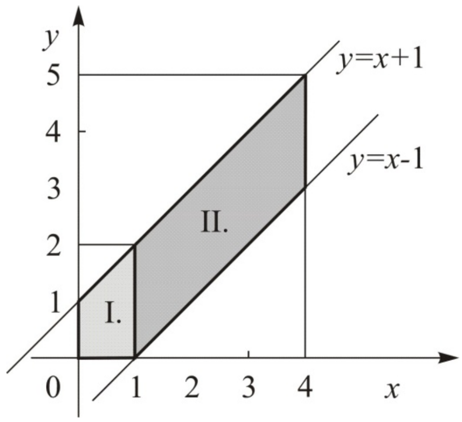


Figure 3.1

The joint p.d.f. of *X* and *Y* must satisfy:

a) $f_{XY}(x, y) = c \ge 0$

b) $\int_y \int_x f_{XY}(x, y) dx dy = 1$

For the first condition, $c \ge 0$ because $x > 0$ and $y > 0$.

According to the second conditions we calculate:

$$\int_y \int_x f_{XY}(x,y)dxdy = \int_0^1 \int_0^{x+1} cdydx + \int_1^4 \int_{x-1}^{x+1} cdydx = c\int_0^1 (x+1)dx + c\int_1^4 2dx =$$

$$= c\left(\left[\frac{1}{2}x^2\right]_0^1 + \left[x\right]_0^1\right) + c\left(\left[2x\right]_1^4\right) = \frac{3}{2}c + 6c = 7{,}5c = 1$$

$c = 1/7{,}5 = 2/15$

Then the joint p.d.f. of X and Y is

$$f_{XY}(x,y) = \frac{2}{15},\ \ 0 < x < 4,\ \ x-1 < y < x+1$$

2. We find the marginal p.d.f., the mean and variance of X.

The marginal p.d.f. of X is

$$f_X(x) = \int_y f_{XY}(x,y)dy = \begin{cases} \int_0^{x+1} \frac{2}{15}dy = \frac{2}{15}(x+1), & 0 < x \le 1 \\ \int_{x-1}^{x+1} \frac{2}{15}dy = \frac{2}{15}\times 2 = \frac{4}{15}, & 1 < x < 4 \end{cases}$$

The mean and variance of X is

$$E(X) = \mu_X = \int_x xf(x)dx = \int_0^1 \frac{2}{15}x(x+1)dx + \int_1^4 \frac{4}{15}xdx =$$

$$= \frac{2}{15}\left(\left[\frac{1}{3}x^3\right]_0^1 + \left[\frac{1}{2}x^2\right]_0^1\right) + \frac{2}{15}\left(\left[x^2\right]_1^4\right) = \frac{2}{15}\cdot\frac{5}{6} + \frac{2}{15}\cdot 15 = \frac{19}{9} = 2{,}11$$

$$D(X) = \sigma_X^2 = \left(\int_x x^2 f(x)dx\right) - \mu_X^2 = \left(\int_0^1 \frac{2}{15}x^2(x+1)dx + \int_1^4 \frac{4}{15}x^2dx\right) - \left(\frac{19}{9}\right)^2 =$$

$$= \frac{2}{15}\left(\left[\frac{1}{4}x^4\right]_0^1 + \left[\frac{1}{3}x^3\right]_0^1\right) + \frac{4}{15}\left(\left[\frac{1}{3}x^3\right]_1^4\right) - \left(\frac{19}{9}\right)^2 =$$

$$= \frac{2}{15}\cdot\frac{7}{12} + \frac{4}{15}\cdot\frac{63}{3} - \left(\frac{19}{9}\right)^2 = \frac{989}{810} = 1{,}221 = 1{,}11^2$$

3. We find the conditional p.m.f., conditional mean and conditional variance of Y given $X = 2$.

The conditional p.m.f. of Y given $X = 2$ is

$$f_{Y|2}(y) = \frac{f_{XY}(2,y)}{f_X(2)} = \frac{2/15}{4/15} = \frac{1}{2},\ \ 1 < y < 3$$

The conditional mean and conditional variance of Y given $X = 2$ are

$$E(Y|2) = \mu_{Y|2} = \int_1^3 y f_{Y|2}(y)dy = \frac{1}{2}\int_1^3 y dy = \frac{1}{2}\left(\left[\frac{1}{2}y^2\right]_1^3\right) = \frac{1}{4}\cdot(3^2-1) = 2$$

$$D(Y|2) = \sigma_{Y|2}^2 = \int_1^3 y^2 f_{Y|2}(y)dy - \mu_{Y|2}^2 = \frac{1}{2}\int_1^3 y^2 dy - 2^2 =$$

$$= \frac{1}{2}\left(\left[\frac{1}{3}y^3\right]_1^3\right) - 2^2 = \frac{1}{6}\cdot(3^3-1) - 2^2 = 0,33 = 0,58^2$$

4. Independence

Check if $f_{Y|2}(y) = f_Y(y)$:

$$f_{Y|2}(y) = \frac{f_{XY}(2,y)}{f_X(2)} = \frac{1}{2}, \quad 1 < y < 3$$

Note that range of X for $1 < y < 3$ is $y-1 < x < y+1$.

Then the marginal p.d.f. of Y for $1 < y < 3$ is

$$f_Y(y) = \int_x f_{XY}(x,y)dx = \int_{y-1}^{y+1} \frac{2}{15}dx = \frac{2}{15}\left([x]_{y-1}^{y+1}\right) = \frac{4}{15}, \quad 1 < y < 3$$

Since $\left(f_{Y|2}(y) = \frac{1}{2}\right) \neq \left(f_Y(y) = \frac{4}{15}\right)$, the measurement of the two methods X and Y are not independent.

3.4 Multiple continuous random variables

Learning goals

- Explain the joint, marginal and conditional probabilities of multiple continuous random variables.
- Explain the independence of multiple continuous random variables.

Joint probability density function

The joint p.d.f. of multiple continuous random variables $X_1, X_2, ..., X_n$ satisfies the following:

1. $f_{X_1 X_2 \dots X_n}(x_1, x_2, ..., x_n) \geq 0$
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_{X_1 X_2 \dots X_n}(x_1, x_2, ..., x_n)dx_1 dx_2 \dots dx_n = 1$

Marginal probability density function

When $f_{X_1X_2\ldots X_n}(x_1, x_2, ..., x_n)$ is the joint p.d.f. of the continuous random variables X_1, X_2, ... X_n, then the marginal p.d.f. of X_i is

$$f_{X_i}(x_i) = \iint\limits_{R(x_i)} \cdots \int f_{X_1X_2\ldots X_n}(x_1, x_2, ..., x_n) dx_1 dx_2 \ldots dx_{i-1} dx_{i+1} \ldots dx_n$$

where $R(x_i)$ is the set of all points in the range of $X_1, X_2, ..., X_n$ for which $X = x_i$.

Mean and variance

The **mean** of X_i is given by

$$E(X_i) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_i f_{X_1X_2\ldots X_n}(x_1, x_2, ..., x_n) dx_1 dx_2 \ldots dx_n$$

The **variance** of X_i is given by

$$D(X_i) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (x_i - \mu_{X_i})^2 f_{X_1X_2\ldots X_n}(x_1, x_2, ..., x_n) dx_1 dx_2 \ldots dx_n$$

Independence of random variables

The continuous random variables $X_1, X_2, ..., X_n$ are independent if and only if satisfies:

$$f_{X_1X_2\ldots X_n}(x_1, x_2, ..., x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \ldots f_{X_n}(x_n)$$

for all $x_1, x_2, ..., x_n$.

3.5 Covariance and correlation

Learning goals

- Explain the terms *covariance* and *correlation* between two random variables *X* and *Y*.
- Calculate the covariance and correlation coefficient of the random variables *X* and *Y*.

Covariance

The covariance between two random variables *X* and *Y* (denoted as $\text{cov}(X,Y)$ or σ_{XY}) indicates the linear relationship between *X* and *Y*:

$$\sigma_{XY} = E\left[(X - \mu_X)(Y - \mu_Y)\right] = E(XY) - \mu_X \mu_Y, \quad -\infty < \sigma_{XY} < \infty$$

Derivation of the relationship

$$E\left[(X-\mu_X)(Y-\mu_Y)\right]=E\left[XY-\mu_X Y-\mu_Y X+\mu_X\mu_Y\right]=$$
$$=E(XY)-\mu_X E(Y)-\mu_Y E(X)+\mu_X\mu_Y=$$
$$=E(XY)-\mu_X\mu_Y-\mu_Y\mu_X+\mu_X\mu_Y=$$
$$=E(XY)-\mu_X\mu_Y$$

Covariance properties:

1. $\mathrm{cov}(X,Y)=\mathrm{cov}(Y,X)$
2. $\mathrm{cov}(X,Y)\le D(X)\cdot D(Y)$
3. σ_{XY} depends on the variances of X and Y

Correlation coefficient

The correlation between two random variables X and Y represents the normalized linear relationship between X and Y (σ_{XY} normalized by σ_X and σ_Y):

$$\rho_{XY}=\frac{\mathrm{cov}\left(X,\ Y\right)}{\sqrt{D(X)D(Y)}}=\frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

Properties of the correlation coefficient:

1. $\rho_{XY}\in[-1;1]$,
2. $\rho_{XY}=1$ – direct linear dependence; with increasing values of X the values of Y increase,
3. $\rho_{XY}=-1$ – indirect linear dependence; with increasing values of X the values of Y decrease,
4. ρ_{XY} is a dimensionless.

Independence of *X* an *Y*

When X and Y are independent, then

$$\sigma_{XY}=\rho_{XY}=0$$

This is only necessary (not sufficient) condition for the independence of X and Y. In other words, even if $\sigma_{XY}=\rho_{XY}=0$ we cannot say that X and Y are independent.

Covariance matrix

The covariance of random variables X a Y was defined above. Let us have a random vector $\left(X_1,X_2,\dots,X_n\right)^{\mathrm{T}}=\boldsymbol{X}^{\mathrm{T}}$, then we can define the covariance matrix.

The covariance matrix of the random vector $(X_1, X_2, ..., X_n)^{\mathrm{T}} = \boldsymbol{X}^{\mathrm{T}}$ is

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}$$

where

$\sigma_{ij} = \mathrm{cov}(X_i, X_j),\ i, j = 1, 2, ..., n,\ i \neq j$ are the covariance between the components of a random vector;

$\sigma_{ii} = \mathrm{cov}(X_i, X_i) = \sigma_i^2,\ i = 1, 2, ..., n$ are the variances of the individual components of a random vector.

The covariance matrix is a square and symmetric matrix. When any two elements of a random vector are independent or at least uncorrelated, then the covariance matrix is diagonal. This means that the elements out of diagonal are equal to zero. In addition, if the variances of all the variables X_i ($i = 1, 2, ..., n$) are the same, $D(x_i) = \sigma^2$ ($i = 1, 2, ..., n$), then the covariance matrix of the random vector $(X_1, X_2, ..., X_n)^{\mathrm{T}} = \boldsymbol{X}^{\mathrm{T}}$ has the form $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{E}$, where $\boldsymbol{E}$ is identity matrix which means that the diagonal elements are equal to one and the others are zero.

Correlation matrix

Correlation matrix of a random vector $(X_1, X_2, ..., X_n)^{\mathrm{T}} = \boldsymbol{X}^{\mathrm{T}}$ is

$$\boldsymbol{P} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{pmatrix}$$

where

$\rho_{ij} = \dfrac{\mathrm{cov}(X_i, X_j)}{\sigma_i \sigma_j} = \dfrac{\sigma_{ij}}{\sigma_i \sigma_j},\ i, j = 1, 2, ..., n,\ i \neq j$ the correlation coefficient between the *i*-th and *j*-th component of the random vector;

$\rho_{ii} = \dfrac{\mathrm{cov}(X_i, X_i)}{\sigma_i \sigma_j} = \dfrac{\sigma_{ii}}{\sigma_i \sigma_i} = 1,\ i = 1, 2, ..., n$ is the correlation coefficient between the i-th and *i*-th component of the random vector.

The correlation matrix is square and symmetric matrix. When any two elements of a random vector are independent or at least uncorrelated, then the correlation matrix is identity matrix.

Example 3.3

The number of defects on the front side (X) of a wooden panel and the number of defects on the rear side (Y) of the panel are under study (Example 3.1). Calculate the value of the covariance and correlation coefficient.

We known that joint p.m.f. of X and Y is $f_{XY}(x,y)=\frac{1}{36}(x+y)$, $x=1,2,3$ and $y=1,2,3$, then applies:

$$E(XY)=\sum_{y=1}^{3}\sum_{x=1}^{3}xyf_{XY}(x,y)=\frac{1}{36}\sum_{y=1}^{3}\sum_{x=1}^{3}xy(x+y)=$$

$$=\frac{1}{36}\left[1\cdot 1\cdot(1+1)+2\cdot 1\cdot(2+1)+3\cdot 1\cdot(3+1)+\cdots\right.$$

$$\left.+1\cdot 3\cdot(1+3)+2\cdot 3\cdot(2+3)+3\cdot 3\cdot(3+3)\right]=\frac{1}{36}\times 168=4{,}67$$

From the symmetry of $f_{XY}(x,y)$ is known that $\mu_X=\mu_Y$ and $\sigma_X=\sigma_Y$. In Example 3.1 was calculated $\mu_X=2{,}17$ and $\sigma_X=0{,}80$. Therefore

$$\sigma_{XY}=E(XY)-\mu_X\mu_Y=4{,}67-2{,}17\times 2{,}17=-0{,}04$$

$$\rho_{XY}=\frac{\sigma_{XY}}{\sigma_X\sigma_Y}=\frac{-0{,}04}{0{,}80\times 0{,}80}=-0{,}06$$

Thus, there is a weak negative correlation between the number of defects on the front side (X) of a wooden panel and the number of defects at the rear side (Y) of the wood panel.

Example 3.4

Consider two measuring methods (Example 3.2).

a) Calculate the value of the covariance and correlation coefficient.
b) Determine the covariance and correlation matrix.

We know that measurements of X and Y have joint p.m.f. $f_{XY}(x,y)=\frac{2}{15}$, $0<x<4$, $x-1<y<x+1$, then:

$$E(XY)=\int_0^4\int_{x-1}^{x+1}xy\frac{2}{15}dydx=\frac{2}{15}\int_0^4 x\left[\frac{y^2}{2}\right]_{x-1}^{x+1}dx=\frac{4}{15}\int_0^4 x^2dx=$$

$$= \frac{4}{15}\left[\frac{x^3}{3}\right]_0^4 = \frac{4}{15}\cdot\frac{64}{3} = \frac{256}{45} = 5{,}68888$$

The mean and variance of X were calculated in Example 3.3: $\mu_X = 2{,}11$ and $\sigma_X = 1{,}11$. To be able to calculate the value of the covariance and correlation coefficient, it remains to calculate μ_Y and σ_Y.

The region of integration for variable Y is divided into three parts:

I. $0 < y \le 1$, $0 < x < y+1$

II. $1 < y \le 3$, $y-1 < x < y+1$

III. $3 < y < 5$, $y-1 < x < 4$

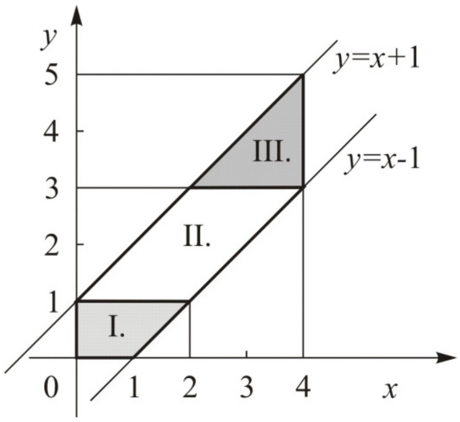


Figure 3.2

The marginal p.d.f. of Y is

$$f_Y(y) = \int_x f_{XY}(x,y)dx = \begin{cases} \int_0^{y+1} \frac{2}{15}dx = \frac{2}{15}(y+1), & 0 < y \le 1 \\ \int_{y-1}^{y+1} \frac{2}{15}dx = \frac{2}{15}\times 2 = \frac{4}{15}, & 1 < y \le 3 \\ \int_{y-1}^{4} \frac{2}{15}dx = \frac{2}{15}(5-y), & 3 < y < 5 \end{cases}$$

The mean and variance of Y are

$$E(Y) = \mu_Y = \int_y yf(y)dy = \int_0^1 \frac{2}{15}y(y+1)dy + \int_1^3 \frac{4}{15}y\,dy + \int_3^5 \frac{2}{15}y(5-y)dy =$$

$$= \frac{2}{15}\left(\left[\frac{1}{3}y^3\right]_0^1 + \left[\frac{1}{2}y^2\right]_0^1\right) + \frac{2}{15}\left(\left[y^2\right]_1^3\right) + \frac{2}{15}\left(\left[\frac{5}{2}y^2\right]_3^5 - \left[\frac{1}{3}y^3\right]_3^5\right) =$$

$$= \frac{1}{9} + \frac{16}{15} + \frac{44}{45} = \frac{97}{45} = 2{,}1556$$

$$D(Y) = \sigma_Y^2 = \left(\int_y y^2 f(y)\, dy \right) - \mu_y^2 = \int_0^1 \frac{2}{15} y^2 (y+1) dy + \int_1^3 \frac{4}{15} y^2\, dy + \int_3^5 \frac{2}{15} y^2 (5-y) dy - \left(\frac{97}{45} \right)^2 =$$

$$= \int_0^1 \frac{2}{15} y^2 (y+1) dy + \int_1^3 \frac{4}{15} y^2\, dy + \int_3^5 \frac{2}{15} y^2 (5-y) dy - \left(\frac{97}{45} \right)^2 =$$

$$= \frac{2}{15} \left(\left[\frac{1}{4} y^4 \right]_0^1 + \left[\frac{1}{3} y^3 \right]_0^1 \right) + \frac{4}{15} \left(\left[\frac{1}{3} y^3 \right]_1^3 \right) + \frac{2}{15} \left(\left[\frac{5}{3} y^3 \right]_3^5 - \left[\frac{1}{4} y^4 \right]_3^5 \right) - \left(\frac{97}{45} \right)^2 =$$

$$= \left(\frac{7}{90} + \frac{104}{45} + \frac{164}{45} \right) - \left(\frac{97}{45} \right)^2 = \frac{5617}{4050} = 1{,}38691 = 1{,}17767^2$$

a) Then the value of covariance is

$$\sigma_{XY} = E(XY) - \mu_X \mu_Y = \frac{256}{45} - \frac{190}{90} \cdot \frac{97}{45} = \frac{461}{405} = 1{,}13827$$

And the value of correlation coefficient is

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\frac{461}{405}}{\sqrt{\frac{989}{810}} \times \sqrt{\frac{5647}{4050}}} = 0{,}872386$$

b) For the random vector $(X,\, Y)^T = \boldsymbol{X}^T$ from the calculated values can be determined <u>correlation matrix</u>

$$\boldsymbol{P} = \begin{pmatrix} 1 & \rho_{XY} \\ \rho_{YX} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0{,}87 \\ 0{,}87 & 1 \end{pmatrix}$$

and <u>covariance matrix</u>

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{XX} & \sigma_{XY} \\ \sigma_{YX} & \sigma_{YY} \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{YX} & \sigma_Y^2 \end{pmatrix} = \begin{pmatrix} 1{,}11^2 & 1{,}14 \\ 1{,}14 & 1{,}18^2 \end{pmatrix}$$

3.6 Bivariate normal distribution

Learning goals

- Explain the joint probability density function of bivariate normal random variables.
- Determine the joint, marginal and conditional probabilities of bivariate normal random variables.

Bivariate normal random variables

The **joint probability density function** of two normal random variables X and Y with means μ_X and μ_Y, variances σ_X^2 and σ_Y^2 and correlation coefficient ρ_{XY} $(-1 < \rho_{XY} < 1)$ is

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}} \exp\left\{-\frac{1}{2(1-\rho_{XY}^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho_{XY}(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\}, -\infty < x, y < \infty$$

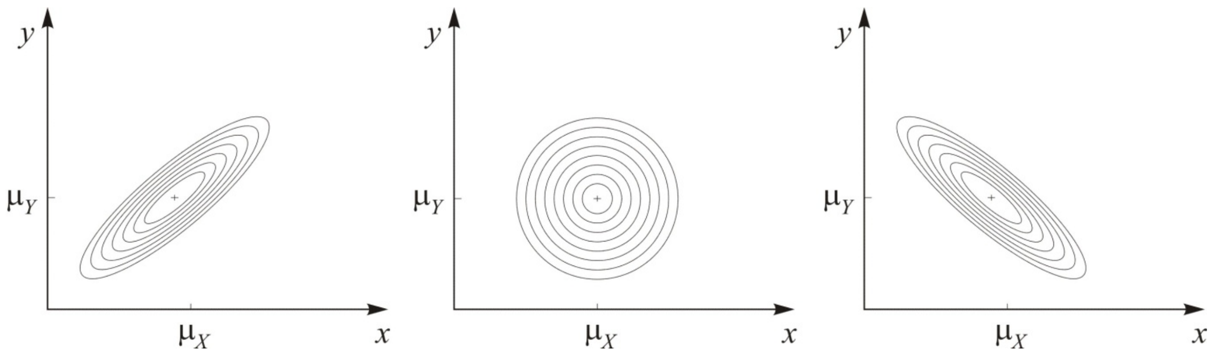


Figure 3.3 Bivariate normal distribution with different values of ρ_{XY}

Marginal probability distribution

The **marginal probability distributions** of X and Y are normal with means μ_X and μ_Y and variances σ_X^2 and σ_Y^2, respectively.

Conditional probability distribution

The **conditional probability distribution** of Y given $X = x$ is normal with mean

$$E(Y|x) = \mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$$

and variance

$$D(Y|x) = \sigma_Y^2(1-\rho_{XY}^2)$$

Example 3.5

Let *X* and *Y* represent two dimensions of an injection molded part. Suppose that *X* and *Y* have a bivariate normal distribution with means $\mu_X = 3,00$ and $\mu_Y = 7,70$, and variances $\sigma_X^2 = 0,04^2$ and $\sigma_Y^2 = 0,08^2$. Assume that *X* and *Y* are independent, i.e., $\rho_{XY} = 0$. Determine probability that $2,95 < X < 3,05$ and $7,60 < Y < 7,80$.

Because *X* and *Y* are independent,

$$P(2,95 < X < 3,05; 7,60 < Y < 7,80) = P(2,95 < X < 3,05)P(7,60 < Y < 7,80)$$

By standardizing both *X* and *Y* we have:

$$P(2,95 < X < 3,05) = P\left(\frac{2,95-3,00}{0,04} < \frac{X-\mu_X}{\sigma_X} < \frac{3,05-3,00}{0,04}\right) =$$

$$= P(-1,25 < Z < 1,25) = P(Z < 1,25) - P(Z < -1,25) =$$

$$= 0,894 - 0,106 = 0,789$$

$$P(7,60 < Y < 7,80) = P\left(\frac{7,60-7,70}{0,08} < \frac{Y-\mu_Y}{\sigma_Y} < \frac{7,80-7,70}{0,08}\right) =$$

$$= P(-1,25 < Z < 1,25) = 0,789$$

Thus,

$$P(2,95 < X < 3,05; 7,60 < Y < 7,80) = 0,789 \times 0,789 = 0,623$$

3.7 Linear combinations of random variables

Learning goals

- Explain the term *linear combination* of random variables.
- Determine the mean and variance of linear combination of random variables.

Linear combination

A random variable *Y* is sometimes defined by a linear combination of several random variables $X_1, X_2, \ldots, X_n$:

$$Y = k_1X_1 + k_2X_2 + \ldots + k_nX_n, \text{ where } k_i\text{'s are constants}$$

Rules for linear combination

The following rules are useful to determine the mean and variance of a linear combination of X and Y.

1. Rules for means

a) $E(b) = b$

b) $E(aX) = aE(X)$

c) $E(aX + b) = aE(X) + b$

d) $E(aX \pm bY) = aE(X) \pm bE(Y)$

e) $E\left((aX)^k\right) = a^k E(X^k)$

where a, b, k are constants.

2. Rules for variances

a) $D(b) = 0$

b) $D(aX) = a^2 D(X)$

c) $D(aX + b) = a^2 D(X) + D(b) = a^2 D(X)$

d) $D(aX \pm bY) = a^2 D(X) + b^2 D(Y) \pm 2ab\,\mathrm{cov}(X,Y)$

$D(aX \pm bY) = a^2 D(X) + b^2 D(Y)$, if X and Y are independent.

Note. Notice that $\mathrm{cov}(X,Y) = \sigma_{XY} = \rho_{XY}\sigma_X\sigma_Y$

The rules 1d) and 2d) can be extended for $Y = k_1X_1 + k_2X_2 + \ldots + k_nX_n$ as follows:

$$E(Y) = k_1E(X_1) + k_2E(X_2) + \ldots + k_nE(X_n) = \sum_{i=1}^{n} k_iE(X_i)$$

$$D(Y) = k_1^2E(X_1) + k_2^2E(X_2) + \ldots + k_n^2E(X_n) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_ik_j\,\mathrm{cov}(X_i, X_j) =$$

$$= k_i^2\sum_{i=1}^{n} E(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_ik_j\,\mathrm{cov}(X_i, X_j)$$

Example 3.6

The width of a casing X and the width of a door Y (both variables in meters) are normally distributed with means $\mu_X = 0{,}61\,\mathrm{m}$ and $\mu_Y = 0{,}51\,\mathrm{m}$, and standard deviations $\sigma_X = 0{,}0030\,\mathrm{m}$ and $\sigma_Y = 0{,}0015\,\mathrm{m}$, respectively. Assume that the width of the casing X and the width of the door Y are independent. Determine the mean and standard deviation of the difference between the width of the casing *X* and the width of the door *Y*.

$X \sim N(0{,}61;0{,}0030^2)$, $Y \sim N(0{,}51;0{,}0015^2)$ and $\operatorname{cov}(X,Y)=0,$ because *X* and *Y* are independent.

Therefore,

$$E(X-Y)=E(X)-E(Y)=0{,}61-0{,}51=0{,}10\,\text{m}$$

$$D(X-Y)=D(X)+D(Y)-2\operatorname{cov}(X,Y)=0{,}0030^2+0{,}0015^2-0=3{,}15\cdot 10^{-6}$$

3.8 Moment generating functions

Learning goals

- Explain the term *moment generating function*.
- Determine the moment generating function a *k*th moments of a random variable *X*.
- Find the mean and variance of *X* by using the first and second moments of *X*.

Moment generating function

The moment generating function of a random variable *X* (denoted as $M_X(t)$) is the expected value of e^{tX}, i.e.,

$$M_X(t)=E(e^{tX})=\begin{cases}\sum_i e^{tx_i} f(x_i), & \text{if } X \text{ is a discrete random variable}\\ \int_{-\infty}^{\infty} e^{tx} f(x)dx, & \text{if } X \text{ is a continuous random variable}\end{cases}$$

The moment generating function of *X* is <u>unique</u> if it exist and completely determines the probability distribution of *X*. Thus, if two random variables have the same moment generating function, they have the same probability distribution.

Moment

If $M_X^{(k)}(t)$ denotes the *k*th derivative of $M_X(t)$, the *k*th moment of *X* about the origin ($t=0$) is

$$E(X^k)=M_X^{(k)}(0)=\begin{cases}\sum_i x_i^k f(x_i), & \text{if } X \text{ is a diskrete random variable}\\ \int_{-\infty}^{\infty} x^k f(x)dx, & \text{if } X \text{ is a continuous random variable}\end{cases}$$

Derivation of the relationship

We know that the kth derivative of $M_X(t)$ is

$$M_X^{(k)}(t) = \frac{d^k M_X(t)}{dt^k} = \begin{cases} \sum_x x^k e^{tx} f(x), & \text{if } X \text{ is discrete random variable} \\ \int_{-\infty}^{\infty} x^k e^{tx} f(x)dx, & \text{if } X \text{ is continuous random variable} \end{cases}$$

Application of moments

The mean and variance of X can be determined by using the first and second moments of X:

$$\mu_X = E(X) = M'_X(0)$$

$$\sigma_X^2 = E(X^2) - [E(X)]^2 = M''_X(0) - [M'_X(0)]^2$$

Example 3.7

The geometric random variable X has probability distribution

$$f(x) = p(1-p)^{x-1}, \ \ x = 1, 2, \ldots, n$$

1. *We find the moment generating function of X.*

From the definition of moment generating function we get

$$M_X(t) = \sum_x e^{tx} f(x) = \sum_{x=1}^{\infty} e^{tx} p(1-p)^{x-1} = \sum_{x=1}^{\infty} pe^t \left[(1-p)e^t\right]^{x-1}$$

Note that the sum of the infinite geometric sequence $(a, ar, ar^2, \ldots)$ is

$$S = \sum_{n=1}^{\infty} ar^{n-1} = \frac{a}{1-r}, \text{ where } |r| < 1$$

Thus

$$M_X(t) = \frac{pe^t}{1-(1-p)e^t}$$

2. *We determine the mean and variance of X* by using the first and second moments of X *about the origin.*

The first moment of X about the origin ($t = 0$) is

$$M'_X(0) = \left[\frac{dM_X(t)}{dt}\right]_{t=0} = \left[\frac{d\left\{pe^t\left[1-(1-p)e^t\right]^{-1}\right\}}{dt}\right]_{t=0} =$$

$$=\left[\frac{pe^t}{1-(1-p)e^t}+\frac{p(1-p)e^{2t}}{\left[1-(1-p)e^t\right]^2}\right]_{t=0}=\frac{p}{1-(1-p)}+\frac{p(1-p)}{\left[1-(1-p)\right]^2}=$$

$$=\frac{p}{p}+\frac{p(1-p)}{p^2}=\frac{1}{p}$$

The second moment of X about the origin ($t=0$) is

$$M''_X(0)=\left[\frac{d^2M_X(t)}{dt^2}\right]_{t=0}=\left[\frac{dM'_X(t)}{dt}\right]_{t=0}=$$

$$=\left[\frac{d\left\{pe^t\left[1-(1-p)e^t\right]^{-1}\right\}}{dt}\right]_{t=0}+\left[\frac{d\left\{p(1-p)e^{2t}\left[1-(1-p)e^t\right]^2\right\}}{dt}\right]_{t=0}=$$

$$=\frac{1}{p}+\left[\frac{2p(1-p)e^{2t}}{\left[1-(1-p)e^t\right]^2}+\frac{2p(1-p)^2e^{2t}}{\left[1-(1-p)e^t\right]^3}\right]_{t=0}=$$

$$=\frac{1}{p}+\frac{2p(1-p)}{\left[1-(1-p)\right]^2}+\frac{2p(1-p)^2}{\left[1-(1-p)\right]^3}=\frac{1}{p}+\frac{2p(1-p)}{p^2}+\frac{2p(1-p)^2}{p^3}=$$

$$=\frac{1}{p}+\frac{2(1-p)}{p}+\frac{2(1-p)^2}{p^2}=\frac{2-p}{p^2}$$

Therefore the mean and variance of X are

$$\mu_X=E(X)=M'_X(0)=\frac{1}{p}$$

$$\sigma_X^2=E(X^2)-\left[E(X)\right]^2=M''_X(0)-\left[M'_X(0)\right]^2=$$

$$=\frac{2-p}{p^2}-\left(\frac{1}{p}\right)^2=\frac{1-p}{p^2}$$

3.9 Chebyshev´s inequality

Learning goals

- Explain the use of Chebyshev´s inequality rule.
- Bound the probability of a random variable *X* by using Chebyshev´s inequality rule and compare the bound probability with the corresponding actual probability.

Chebyshev´s inequality

A relationship between the mean and variance of a random variable *X* having a certain probability distribution is formulated by Chebyshev as follows:

$$P\left(\left|X-\mu\right|\geq c\sigma\right)\leq\frac{1}{c^2},\ c>0$$

By using Chebyshev´s inequality rule, a bound probability of any random variable can be determined. In Tab. 3.1 are presented bound probabilities of a normal random variable *X* with parameters μ and σ^2 and corresponding actual probabilities.

Tab. 3.1 *Bound probabilities and actual probabilities of a normal random variable X*

c	Probability condition $P\left(\left\|X-\mu\right\|\geq c\sigma\right)$	Bound probability ($1/c^2$)	Actual probability
1,5	$P\left(\left\|X-\mu\right\|\geq 1{,}5\sigma\right)$	0,444	0,134
2	$P\left(\left\|X-\mu\right\|\geq 2\sigma\right)$	0,250	0,046
3	$P\left(\left\|X-\mu\right\|\geq 3\sigma\right)$	0,111	0,003
4	$P\left(\left\|X-\mu\right\|\geq 4\sigma\right)$	0,063	< 0,001

The actual probability is calculated by adjusting the probabilistic relationship as follows:

$$P\left(\left|X-\mu\right|\geq c\sigma\right)=P\left(\frac{\left|X-\mu\right|}{\sigma}\geq c\right)=P\left(\left|Z\right|\geq c\right)=1-P\left(\left|Z\right|<c\right)=$$

$$=1-P(-c<Z<c)$$

When the last relationship gradually substituted for c values 1,5; 2; 3; 4 and do the calculation, we obtain the value of the actual probabilities (Tab. 3.1).

Example 3.8

Suppose that the photoresist thickness X in semiconductor manufacturing has a continuous uniform distribution with a mean of 10 μm and a standard deviation of 2,31 μm over the range $6 < x < 14$ μm. We bound the probability that the photoresist thickness is less than 7 or greater than 13 μm. Then we compare the bounded probability with the actual probability.

The probability density function of the uniform random variable X is

$$f(x) = \frac{1}{b-a} = \frac{1}{14-6} = \frac{1}{8}, \quad 6 < x < 14$$

Using Chebyshew´s inequality we get:

$$\begin{aligned} P(X<7)+P(X>13) &= P(X-10<7-10)+P(X-10>13-10) = \\ &= P(X-10<-3)+P(X-10>3) = \\ &= P\left(|X-10|>3\right) = P\left(|X-10|>c\sigma\right) < 1/c^2 \end{aligned}$$

Therefore $3 = c\cdot\sigma = c\cdot 2,31 \quad \Rightarrow \quad c = \dfrac{3}{2,31} \doteq 1,3$

Then the bound probability is

$$P\left(|X-10|>3\right) < \left(\frac{1}{c^2} = \frac{1}{1,30^2} = 0,59\right)$$

and the actual probability equals

$$P(X<7)+P(X>13) = 1-P(7<X<13) = 1-\int_7^{13}\frac{1}{8}dx = 1-\frac{1}{8}\left[x\right]_7^{13} = 1-\frac{6}{8} = 0,25$$

The actual probability is less than the bound probability, which supports the Chebyshev´s inequality.

4 CREATION OF RANDOM SAMPLE AND DESCRIPTIVE STATISTICS

Learning goals

- Recognize the difference between population and random sample.
- Describe the terms *random sample* and *statistic*.
- Distinguish between the terms *statistic* and *value of the statistic*.
- Distinguish between the terms *ordered random sample*, *order statistic* and *value of order statistic*.
- Explain why picking a *representative random sample* is important in research.

Population

Probability distribution is often used as a **model** for a population. The population that is normally distributed with parameters μ and σ^2, is called **normal population** or **population with normal distribution**. For example a design engineer may consider as normal population all values of the internal diameter of the piston ring automobile engine.

Random sample

Random sample is taken from the population under study, which is created by a certain random mechanism, to avoid bias (over- or underestimation).

Consider a random variable X with distribution function $F(x)$ and experiment, the results of which can be regarded as the value of this random variable. When we make *n* trials in a given experiment independently and under the same conditions, we get *n* observations $x_1, x_2, \ldots, x_n$, which represent the values of random variables $X_1, X_2, \ldots, X_n$.

Exactly is a random sample defined as follows:

The random variables $X_1, X_2, \ldots, X_n$ make **random sample** of size n if:

1. are independent of each other,
2. every X_i has the same probability distribution $f(x)$.

Thus, the joint probability density function (mass function) of $X_1, X_2, \ldots, X_n$ is

$$f_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n) = f(x_1) f(x_2) \cdots f(x_n)$$

Statistic

A **statistic** is a function of random variables $X_1, X_2, \dots, X_n$ from a random sample, which does not depend on the parameters of the probability distribution of the random variable X. The obvious is that statistic is a multivariate random variable, denoted generally by $T(X_1, X_2, ..., X_n)$.

Value of statistic

The value that can acquire statistic $T(X_1, X_2, ..., X_n)$ in one random sample realization $x_1, x_2, \dots, x_n$ is called a **value of statistic** and denoted $T(x_1, x_2, ..., x_n)$.

Ordered random sample and its realization

Let us have observations $x_1, x_2, \dots, x_n$, which are realizations of a random sample $X_1, X_2, \dots, X_n$. When we arrange observations by size in ascending order we get $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$, what are realization of an ordered random sample $X_{(1)}, X_{(2)}, \dots, X_{(n)}$.

Order statistic and its value

A random variable $X_{(i)}$ from an ordered random sample represents the *i*th order statistic and a value $x_{(i)}$ is the value of *i*th order statistic.

4.1 The numeric methods of descriptive statistics

Learning goals

- Describe the basic statistical characteristics and their values used in descriptive statistics.

The most commonly used statistics and their values

- **Sample mean** $\overline{X}$ — **Value of sample mean** $\overline{x}$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

A sample mean characterizes the central location of data.

- **Sample variance** S^2 **Value of sample variance** s^2

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = \frac{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n}{n-1}$$

A sample variance characterizes the variability of the data.

- **Sample standard deviation** S **Value of sample standard deviation** s

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \qquad s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$$

A sample standard deviation charakterizes the variability of the data.

- **Sample standard error** **Value of sample standard error**

$$\frac{S}{\sqrt{n}} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \qquad \frac{s}{\sqrt{n}} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$$

A sample standard error charakterizes the variability of sample mean of the data.

Other basic statistical characteristics used

- **Value of sample median** $x_{\text{med}} = \tilde{x}$

  divides the ordered data set $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ into two equal parts. In this data set the sample median represents the percentile $x_{0,50}$ and also the second quartile q_2, that is $x_{\text{med}} = \tilde{x} = x_{0,50} = q_2$.

- **Percentile** x_p

  The procedures to find a value of p–percentile x_p from n ascendingly ordered observations $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are as follows:

  Step 1: we calculate the position number r by using n and p

$$r = \begin{cases} np, & \text{if } n \text{ is odd} \\ (n+1)p & \text{if } n \text{ is even} \end{cases}$$

Step 2: we determine x_p based on the position number r

$$x_p = \begin{cases} x_{(r)}, & \text{if } r \text{ is an integer} \\ x_{(\lfloor r \rfloor)} + (x_{(\lceil r \rceil)} - x_{(\lfloor r \rfloor)})(r - \lfloor r \rfloor), & \text{if } r \text{ is not an integer,} \end{cases}$$

where the symbols $\lceil r \rceil$ means „rounding up“ and $\lfloor r \rfloor$ „rounding down“.

A value of percentile x_p for $p = \{0,25; 0,50; 0,75\}$ is called a **value of quartile** of the data set. The following three values of guartiles divide a set of data into four equal parts in ascending order:

– first (lower) quartile : $q_1 = x_{0,25}$

– second (middle) quartile (**median**) : $q_2 = x_{0,50}$

– third (upper) quartile : $q_3 = x_{0,75}$

- **Value of sample mode** $x_{\text{mod}} = \hat{x}$

  is the most frequently occurring value in the data. Data set can have no mode, one mode (unimodal), or more modes (bimodal, trimodal, etc.).

- **Minimum** $x_{\min}$

  is the minimum value of a realization of a sample, that is a set of data $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.

- **Maximum** $x_{\max}$

  is the maximum value of a realization of a sample, that is a set of data $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.

- **Sample range**

  is the difference between the maximum and minimum value of a data set $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$:

$$R = x_{\max} - x_{\min} = x_{(n)} - x_{(1)}$$

- **Value of lower (first) quartile** $q_1 = x_{0,25}$

  is the value dividing the data set $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ into two parts such that 25 % of the values is not greater than this value and 75 % of the values is not less than this value.

- **Value of upper (third) quartile** $q_3 = x_{0,75}$

  is the value dividing the data set $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ into two parts such that 75 % of the values is not greater than this value and 25 % of the values is not less than this value.

- **Intequartile range (IQR)**

  is the difference between the upper quartile and lower quartile of a data set:

$$\text{IQR} = q_3 - q_1 = x_{0,75} - x_{0,25}$$

- **Sample skewness** expresses the asymmetry of a frequency distribution of the data set $x_1, x_2, \ldots, x_n$. The measure of this asymmetry is **sample skewness coefficient**:

$$\frac{n\sum_{i=1}^{n}(x_i - \overline{x})^3}{(n-1)(n-2)s^3}, \text{ pre } n \geq 3 \text{ a } s \neq 0$$

- **Standardized sample skewness coefficient** is normally distributed $N(0,1)$ for $n > 150$:

$$\frac{n\sum_{i=1}^{n}(x_i - \overline{x})^3}{(n-1)(n-2)s^3} \Bigg/ \sqrt{\frac{6}{n}}$$

  Note. For symmetrical frequency distribution, this coefficient is equal to zero. For distribution skewed to the left, this coefficient is negative. For distribution skewed to the right this coefficient is a positive.

- **Sample kurtosis** expresses the taperness of a frequency distribution of the data set $x_1, x_2, \ldots, x_n$. The measure of this taperness is **sample kurtosis coefficient**:

$$\frac{n(n+1)\sum_{i=1}^{n}(x_i - \overline{x})^4}{(n-1)(n-2)(n-3)s^4} - \frac{3(n-1)^2}{(n-2)(n-3)} \text{ pre } n \geq 4 \text{ a } s \neq 0$$

- **Standardized sample kurtosis coefficient** is given by:

$$\left(\frac{n(n+1)\sum_{i=1}^{n}(x_i - \overline{x})^4}{(n-1)(n-2)(n-3)s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}\right) \Bigg/ \sqrt{\frac{24}{n}}$$

  Note. For values from the normal distribution the coefficient is approximately equal to zero. In comparison with normal distribution, a positive value of the coefficient means the distribution is more acute, while a negative value indicates a flatter distribution.

- **Sample variation coefficient** (v %) measures the magnitude of the standard deviation value as a percentage of the sample mean value according to:

$$\frac{s}{\overline{x}} \times 100 = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}}{\overline{x}} \times 100$$

4.2 Graphical methods of descriptive statistics

Learning goals

- Explain the use of *stem-and-leaf diagram*.
- Construct a stem-and-leaf diagram to visualize a set of data.
- Explain the terms *frequency, relative frequency, cumulative frequency a cumulative relative frequency*.
- Explain construction of a *frequency table.*
- Explain construction of a *histogram* and *polygon* from the frequency table.
- Explain the use of a *box plot*.
- Explain the use of a *normal probability plot*.

Stem-and-leaf diagram

A stem-and-leaf diagram is a good tool to graph a data set $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, where each number contains at least two digits. The following steps are applied to construct a stem-and-leaf diagram:

Step 1: Determine **stems** and **leaves.**

Divide each number $x_{(i)}$ into two parts: 1. stem for the "significant" digits (one or two digits in most cases) and 2. leaf for the "less significant" digit (last digit usually). The analyst should exercise his/her own discretion to determine which digits are most significant with consideration of the range of data.

Step 2: **Arrange** the stems and leaves.

The stems are arranged in ascending order. Then, beside each stem, corresponding leaves are listed next to each other in a row. The leaves of each stem are arranged in ascending order.

Step 3: Summarize the **frequency** of leaves for each stem.

Example 4.1

From the following data we construct diagram stem-and-leaf, which represents the frequency distribution diagram.

i	x_i	i	x_i	i	x_i	i	x_i
1	24,0	5	22,3	9	21,8	13	23,2
2	22,4	6	22,6	10	32,2	14	23,9
3	22,4	7	25,2	11	23,9	15	23,8
4	24,3	8	24,1	12	23,5	16	21,7

Solution

First, the data is arranged in ascending order:

i	$x_{(i)}$	i	$x_{(i)}$	i	$x_{(i)}$	i	$x_{(i)}$
1	21,7	5	22,4	9	23,5	13	24,0
2	21,8	6	22,6	10	23,8	14	24,1
3	22,3	7	23,2	11	23,9	15	24,3
4	22,4	8	23,2	12	23,9	16	25,2

Let the first two digits of the data form the stems. We list the stem values in ascending order into the column. Then we make a vertical line and write the leaf of each observation into the horizontal line of the corresponding stem. In this way we obtain the resulton on Figure 4.1.

Stem	Leaf	Frequency
21.	7 8	2
22.	3 4 4 6	4
23.	2 2 5 8 9 9	6
24.	0 1 3	3
25.	2	1

Figure 4.1 A stem-and-leaf diagram

Leaf					
			9		
			9		
		6	8		
		4	5	3	
	8	4	2	1	
	7	3	2	0	2
Stem	21.	22.	23.	24.	25.

Figure 4.2 Shape of the distribution

When we turn the diagram on the left hand side and look at columns of numbers above the line, we see the shape of the distribution.

Note. If necessary, stem-and-leaf can be:

a) compressed by merging of neighboring rows in one line (common class), see Figure 4.3.

50	0 1
51	4
52	5 6
53	3 6 8
54	2 4 5 7
55	3 4 9 9
56	0 1 2 7
57	3 5 8
58	1 2 6 9
59	1 7

→

50 – 51	0 1 ∗ 4
52 – 53	5 6 ∗ 3 6 8
54 – 55	2 4 5 7 ∗ 3 4 9 9
56 – 57	0 1 2 7 ∗ 3 5 8
58 – 59	1 2 6 9 ∗ 1 7

↑ 581 ↑ 591

Figure 4.3 Compression of stem-and-leaf diagram

b) splitting by dividing each of rows into two rows (classes), for example, see Figure 4.4.

51	6 8 9 9
52	0 3 4 7
53	3 7 8 8

→

51∗	
51∘	6 8 9 9
52∗	0 3 4
52∘	7
53∗	3
53∘	7 8 8

Figure 4.4 Split of stem-and-leaf diagram

The first class of the digits 0-4 we mark "∗" and the other with the digits 5-9 we mark "∘".

Table of frequencies

Creating a table of frequencies is a good method to describe a large set of data. Original data are classified into classes (categories). Then frequencies of each class are found and frequency distribution is created.The procedure of constructing the table of frequencies consists of the following steps:

1. We determine the number of classes which will contain the table of frequencies. If we can not determine the number of classes, one of the following formulas will help us

$$k = 1 + 3{,}322 \cdot \log(n) \qquad \text{or} \qquad k = \sqrt{n}$$

2. Determine the maximum $x_{\max}$ and minimum $x_{\min}$ value of a set of data.
3. Class width can be determined as follows:

$$h \geq \frac{x_{\max} - x_{\min}}{k} = \frac{R}{k}$$

   The result should be rounded up to an integer so that each value of the data is contained in the table of frequencies.
4. Create a class.
   Lower limit of the first class t_0 we choose either the smallest value in a data set or a value slightly smaller than the smallest value. The upper limit of the last k-th class t_k is chosen either the greatest value in a set of data or a value slightly greater than the maximum value. In the table is $t_0 \leq x_{\min}$, $t_k \geq x_{\max}$.

Each class represents the interval $[t_i, t_{i+1})$ of length h, specifically $t_{i+1} = t_i + h$, where. $i = 0,1,...,k-1$. For example, the first class: $[t_0, t_1)$, the second class: $[t_1, t_2) = [t_0 + h, t_1 + h)$, etc.

5. Class representative $\bar{t}_i$ is a middle of the ith interval (ith class), that is $\bar{t}_i = \frac{t_{i-1} + t_i}{2}$.
6. Each value of the data set is recorded in the row of the relevant class.
7. Count the number of the data set in each class. We get the frequency of classes n_i that we record in the appropriate column.
8. Calculate the relative frequencies, cumulative frequencies, cumulative relative frequencies, and write them in the next three columns.
   Thus, we created the entire table of frequencies (Table 4.1).

Table 4.1 Table of frequencies

Class number	Lower limit	Upper limit	Class representative	Absolute frequency	Relative frequency	Cumulative relative frequency
1.	t_0	t_1	$\bar{t}_1$	n_1	$f_1 = n_1 / n$	f_1
2.	t_1	t_2	$\bar{t}_2$	n_2	$f_2 = n_2 / n$	$f_1 + f_2$
. . .	. . .	. . .	. . .	. . .	. . .	. . .
k.	t_{k-1}	t_k	$\bar{t}_k$	n_k	$f_k = n_k / n$	$\sum_i f_i = 1$
				$\sum_i n_i = n$	$\sum_i f_i = 1$	

Note. When rounding relative frequency values it must be ensured that the sum of the rounded figures is equal to one.

Using the table of frequencies we can construct a histogram and polygon of frequencies, relative frequencies, cumulative frequency and cumulative relative frequencies. All these graphs show us a way of frequency distribution of the measured data. All these graphs show us a way of frequency distribution of the measured data.

Histogram

A histogram is a bar graph in which the length (on the vertical axis) and width (on the horizontal axis) of each bar are proportional to the frequency n_i (or relative frequency f_i) and size h of corresponding class interval $[t_i, t_{i+1})$, respectively. The shape of a histogram of a small set of data may vary significantly as the number of class intervals and corresponding class intervals width change. As the size of a data set becomes large (say, 75 or above), the shape of the histogram becomes stable.

Polygon of frequencies

A polygon of frequencies consists of line of segments that passes through the points $[\bar{t}_i; n_i]$ or $[\bar{t}_i; f_i]$, where $\bar{t}_i$ are midpoints of the intervals $[t_i, t_{i+1})$ and n_i or f_i are the absolute or relative frequencies of selected classes.

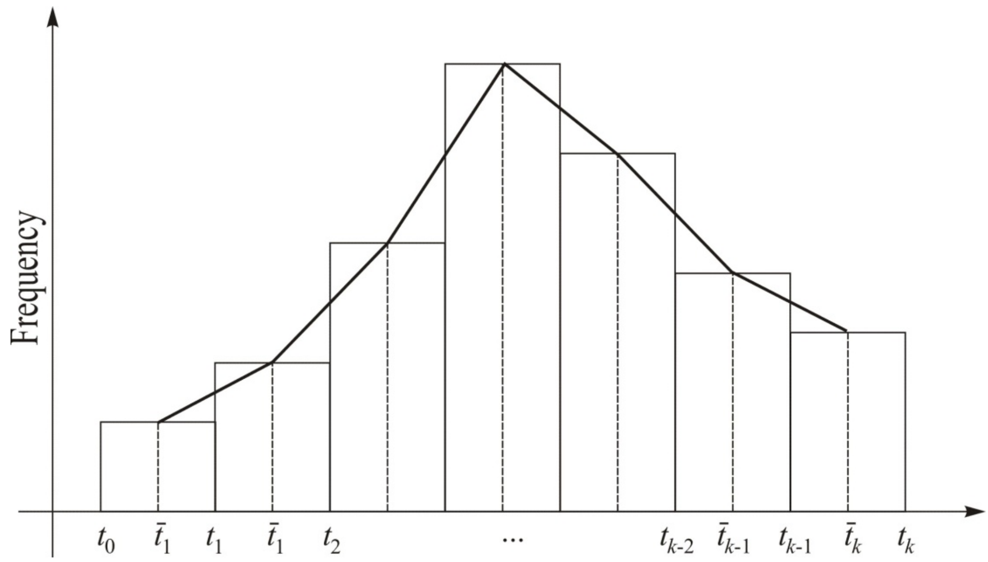


Figure 4.4

The stem-and-leaf diagram and histogram provide a general visual view of a data set, while numerical quantities such as $\bar{x}$ or s provide information about only one feature (characteristic properties) of the data.

Box plots

The box plot is a graphical display that simultaneously describes several important features of a data set, such as center, spread, departure from symmetry and identification of unusual observations or outliers.

A box plot displays (see Figure 4.5):

– the three quartiles, the minimum and maximum of the data on a rectangular box, aligned either horizontally or vertically;
– the box encloses the interquartile range (IQR) with the left (or lower) edge at the first (or lower) quartile $q_1 = x_{0,25}$ and the right (or upper) edge at the third (or upper) quartile;
– in the rectangle is a line segment parallel to the lower and upper limit, which is the second quartile (which is the 50th percentile or the median) $q_2 = x_{0,5} = x_{\text{med}} = \tilde{x}$;
– a line or whisker extends from each end of the box:
  a) *the lower whisker* is a line segment from the first quartile to the smallest data point within $1{,}5 \times \text{IQR}$ from the first (or lower) quartile,
  b) *the upper whisker* is a line segment from the third quartile to the largest data point within $1{,}5 \times \text{IQR}$ from the third (or upper) quartile;
– data farther from the box than the whiskers are plotted as individual points; a point beyond a whisker, but less than $3 \times \text{IQR}$ from the box edge, is called an **outlier** and is lying in intervals $\left(x_{0,25} - 3 \times \text{IQR}, x_{0,25} - 1{,}5 \times \text{IQR}\right)$ or $\left(x_{0,75} + 1{,}5 \times \text{IQR}, x_{0,75} + 3 \times \text{IQR}\right)$;
– a point more than $3 \times \text{IQR}$ from the box edge is called an **extreme outlier** and is lying in intervals $\left(-\infty, x_{0,25} - 3 \times \text{IQR}\right)$ or $\left(x_{0,75} + 3 \times \text{IQR}, \infty\right)$.

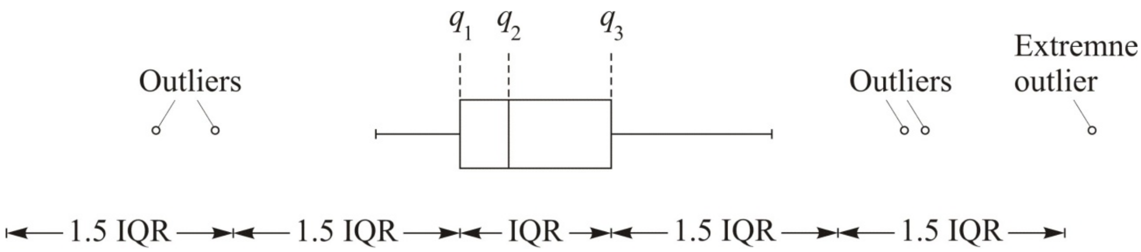


Figure 4.5 Description of a box plot

Note. Outliers and extreme outliers are values relatively very small or very large in relation to other data. Usually arise from three causes:

a) value is measured, recorded or inserted into the computer incorrectly,
b) measured value belongs to a different population,
c) value is measured and recorded correctly, but represents a rare event that may occur.

Given the above reasons it is necessary to consider whether these values in the random sampling of data leave or retire.

Normal probability plots

This chart is a special case of a probabilistic graph, which makes it possible to visually assess whether the data come from a normal distribution.

Its construction lies in the fact that the horizontal axis is plotted by the arranged values $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ and the vertical axis by values $\frac{j-0,375}{n+0,25}\times 100$ (in percentage). The graph is then the set of points $X_j = \left[x_{(j)}\,;\, \frac{j-0,375}{n+0,25}\times 100 \right]$, $j = 1,2,\ldots,n$, which is approximated in terms of the least-squares method by linear function – line. The fewer the points deviate from the straight line, the more likely it is that the measured data come from a normal distribution.

If we want to be sure that the measured data actually come from population with a normal distribution, it is necessary to test normality of the measured data, for example by using Shapiro-Wilk test (see chapter 7.6.2).

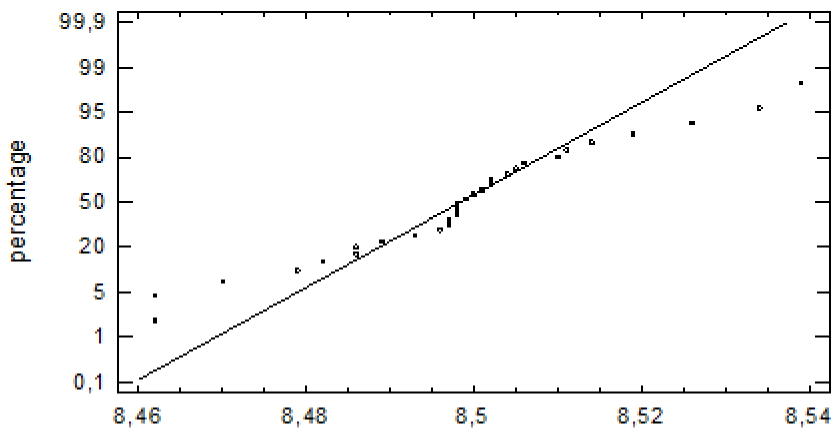


Figure 4.6

4.3 Presentation of numerical and graphical methods of a descriptive statistics on data from a random sample

Example 4.2

Alfa Machine carves a component used in the special security locks of the company Mul T Lock on the required width of 8,500 mm. Randomly select 30 parts and check them by one measuring tool for prescribed width.

On the measured data (Table 4.2) we present numerical and graphical methods of a descriptive statistics.

Table 4.2

i	x_i	i	x_i	i	x_i	i	x_i	i	x_i	i	x_i
1	8,462	6	8,505	11	8,482	16	8,493	21	8,511	26	8,497
2	8,489	7	8,519	12	8,499	17	8,510	22	8,496	27	8,506
3	8,500	8	8,486	13	8,498	18	8,502	23	8,470	28	8,501
4	8,486	9	8,502	14	8,462	19	8,526	24	8,539	29	8,49
5	8,504	10	8,498	15	8,534	20	8,498	25	8,514	30	8,479

Solution

Before access to the basic data processing, the measured values of the part width must be arranged in ascending order. The sorted data values represent the value of order statistic (Table 4.3).

Table 4.3

i	$x_{(i)}$	i	$x_{(i)}$	i	$x_{(i)}$	i	$x_{(i)}$	i	$x_{(i)}$	i	$x_{(i)}$
1	8,462	6	8,486	11	8,497	16	8,499	21	8,504	26	8,514
2	8,462	7	8,486	12	8,497	17	8,500	22	8,505	27	8,519
3	8,470	8	8,489	13	8,498	18	8,501	23	8,506	28	8,526
4	8,479	9	8,493	14	8,498	19	8,502	24	8,510	29	8,534
5	8,482	10	8,496	15	8,498	20	8,502	25	8,511	30	8,539

1. *The numeric methods of descriptive statistics*

The calculation of basic statistical (or sample) characteristics can be made by the current statistical software.

Value of sample mean:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{30}\sum_{i=1}^{30} x_i = \frac{(8,462 + 8,462 + 8,467 ... + 8,539)}{30} = 8,49883$$

Value of sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 =$$

$$= \frac{(8,462 - 8,49883)^2 + (8,462 - 8,49883)^2 + ... + (8,539 - 8,49883)^2}{29} = 0,00322006$$

Value of sample standard deviation:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \sqrt{0,00322006} = 0,0179445$$

Value of sample standard error:

$$\frac{s}{\sqrt{n}} = \sqrt{\frac{1}{n}\cdot\frac{1}{(n-1)}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \sqrt{\frac{0,00322006}{30}} = 0,00327621$$

Value of sample median (second quartile) $x_{\text{med}} = \tilde{x}\left(= q_2 = x_{0,50}\right)$:

For $n = 30$ calculated $r = (n+1)p = (30+1)\times 0,5 = 15,5$ is not an integer.

$$x_{0,5} = x_{(\lfloor r \rfloor)} + (x_{(\lceil r \rceil)} - x_{(\lfloor r \rfloor)})(r - \lfloor r \rfloor) =$$

$$= x_{(\lfloor 15,5 \rfloor)} + (x_{(\lceil 15,5 \rceil)} - x_{(\lfloor 15,5 \rfloor)})(15,5 - \lfloor 15,5 \rfloor) =$$

$$= x_{(15)} + (x_{(16)} - x_{(15)})(15,5 - 15) = 8,498 + (8,499 - 8,498)\times 0,5 = 8,4985$$

$$x_{\text{med}} = x_{0,50} = \tilde{x} = 8,4985$$

Value of sample mode $x_{\text{mod}} = \hat{x}$:

Most frequent value in the data (3 times) is only one value, namely 8.498. The data set thus have one mode – it is unimodal.

$$x_{\text{mod}} = \hat{x} = 8,498$$

Minimum: $x_{\min} = x_{(1)} = 8,462$

Maximum: $x_{\max} = x_{(30)} = 8,539$

Sample range**:** $R = x_{\max} - x_{\min} = 8,539 - 8,462 = 0,077$

Value of lower (first) quartile $q_1 = x_{0,25}$:

For $n = 30$ calculated $r = (n+1)p = (30+1)\times 0,25 = 7,75$ is not an integer.

$$q_1 = x_{0,25} = x_{(\lfloor r \rfloor)} + (x_{(\lceil r \rceil)} - x_{(\lfloor r \rfloor)})(r - \lfloor r \rfloor) =$$

$$= x_{(\lfloor 7,75 \rfloor)} + (x_{(\lceil 7,75 \rceil)} - x_{(\lfloor 7,75 \rfloor)})(7,75 - \lfloor 7,75 \rfloor) =$$

$$= x_{(7)} + (x_{(8)} - x_{(7)})(7,75 - 7) = 8,486 + (8,489 - 8,486)\times 0,75 = 8,48825$$

$$q_1 = x_{0,25} = 8,48825 \approx 8,489$$

Value of upper (third) quartile $q_3 = x_{0,75}$:

For $n = 30$ calculated $r = (n+1)p = (30+1)\times 0,75 = 23,25$ is not an integer.

$$q_3 = x_{0,75} = x_{(\lfloor r \rfloor)} + (x_{(\lceil r \rceil)} - x_{(\lfloor r \rfloor)})(r - \lfloor r \rfloor) =$$

$$= x_{(\lfloor 23,25 \rfloor)} + (x_{(\lceil 23,25 \rceil)} - x_{(\lfloor 23,25 \rfloor)})(23,25 - \lfloor 23,25 \rfloor) =$$

$$= x_{(23)} + (x_{(24)} - x_{(23)})(23,25 - 23) = 8,506 + (8,510 - 8,506) \times 0,25 = 8,5061$$

$$q_3 = x_{0,75} = 8,5061 \approx 8,506$$

Interquartile range: $\mathrm{IQR} = q_3 - q_1 = x_{0,75} - x_{0,25} = 8,506 - 8,489 = 0,017$

Standardized sample skewness coefficient: $n = 30 \geq 3$, $s = 0,0179445 \neq 0$

$$\frac{n\sum_{i=1}^{n}(x_i - \overline{x})^3}{(n-1)(n-2)s^3} \Bigg/ \sqrt{\frac{6}{n}} =$$

$$= \frac{30\left((8,462 - 8,49883)^3 + (8,462 - 8,49883)^3 + \ldots + (8,539 - 8,49883)^3\right)}{29 \cdot 28 \cdot 0,0179445^3} \Bigg/ \sqrt{\frac{6}{30}} =$$

$$= 0,0306566 > 0$$

The value of standardized sample skewness coefficient is positive and very small, it means that our data distribution is approximately symmetric distribution that is slightly skewed to the right.

Standardized sample kurtosis coefficient: $n = 30 \geq 4$, $s = 0,0179445 \neq 0$

$$\left(\frac{n(n+1)\sum_{i=1}^{n}(x_i - \overline{x})^4}{(n-1)(n-2)(n-3)s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}\right) \Bigg/ \sqrt{\frac{24}{n}} =$$

$$= \left(\frac{30 \cdot 31\left((8,462 - 8,49883)^4 + (8,462 - 8,49883)^4 + \ldots + (8,539 - 8,49883)^4\right)}{29 \cdot 28 \cdot 27 \cdot 0,0179445^4} - \frac{3 \cdot 29^2}{28 \cdot 27}\right) \Bigg/ \sqrt{\frac{24}{30}} =$$

$$= 0,66439 > 0$$

The value of the standardized sample kurtosis coefficient is positive, it means that our data distribution is compared with the normal distribution more sharp.

Sample variation coefficient (v %):

$$\frac{s}{\overline{x}} \times 100 = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}}{\overline{x}} \times 100 =$$

$$= \frac{\sqrt{\frac{(8,462 - 8,49883)^2 + (8,462 - 8,49883)^2 + \ldots + (8,539 - 8,49883)^2}{29}}}{8,49883} \times 100 =$$

$$= 0,21114$$

Table 4.4 Review of basic statistical (or sample) characteristics calculated using the statistical software Statgraphics Centurion XV

		Explanation
Count	30	sample size
Average	8,49883	value of sample mean
Median	8,4985	value of sample median
Mode	8,498	value of sample mode
Variance	0,000322006	value of sample variance
Standard deviation	0,0179445	value of sample standard deviation
Coeff. of variation	0,211141%	value of sample coefficient of variation
Standard error	0,00327621	value of sample standard error
Minimum	8,462	value of sample minimum
Maximum	8,539	value of sample maximum
Range	0,077	value of sample range
Lower quartile	8,489	value of sample lower quartile
Upper quartile	8,506	value of sample upper quartile
Interquartile range	0,017	value of sample interquartile range (IQR)
Stnd. skewness	0,0306566	value of standardized sample skewness coefficient
Stnd. kurtosis	0,66439	value of standardized sample kurtosis coefficient

2. *Graphical methods of descriptive statistics*

All of the above graphs, including tables of frequencies, we also present in the Statgraphics Centurion XV.

Stem-and Leaf Diagram

Let the first three digits of data are stems.

Stem	Leaf
8,46	22
8,47	09
8,48	2669
8,49	36778889
8,50	0122456
8,51	0149
8,52	6
8,53	49

Figure 4.7 Stem-and-leaf diagram

Stem-and-Leaf Display for Šírka: unit = 0,001 1|2 represents 0,012

```
         LO|8,462 8,462

  2    846|
  4    847|09
  8    848|2669
 (8)   849|36778889
 14    850|0122456
  7    851|0149
  3    852|6

         HI|8,534 8,539
```

Figure 4.8 Stem-and-leaf diagram in the Statgraphics Centurion XV

Note. In this type of graph Statgraphics Centurion XV marked out outliers:

- low: LO|8,462 8,462 t.j. $x_{(1)} = x_{(2)} = 8,462$
- high: HI|8,534 8,539 t.j. $x_{(29)} = 8,534$ a $x_{(30)} = 8,539$

We will check them even on a box plot and determine whether these values are only outliers or extreme outliers.

Table of frequencies

a) We determine the number of classes *k*:

b) $k = 1 + 3,322\log(n) = 1 + 3,322\log(30) = 5,907 \approx 6$ or $k = \sqrt{n} = \sqrt{30} = 5,477 \approx 6$.

c) We calculate the width of the class: $h = \dfrac{x_{\max} - x_{\min}}{k} = \dfrac{0,077}{6} = 0,012833 \approx 0,015$.

d) We choose: $t_0 = 8,45 < x_{\min} = 8,462$, $t_6 = 8,54 > x_{\max} = 8,539$.

e) We construct a table of frequencies (Table 4.5).

Table 4.5 Table of frequencies

Class number	Lower limit	Upper limit	Class representative	Absolute frequency	Relative frequency	Cumulative relative frequency
1.	8,450	8,465	8,4575	2	0,0667	0,0667
2.	8,465	8,480	8,4725	2	0,0667	0,1334
3.	8,480	8,495	8,4875	5	0,1667	0,3001
4.	8,495	8,510	8,5025	14	0,4666	0,7667
5.	8,510	8,525	8,5175	4	0,1333	0,9000
6.	8,525	8,540	8,5325	3	0,1000	1,0000
				$\sum_i n_i = 30$	$\sum_i f_i = 1$	

Note. In the rounding of values of relative frequency it must be ensured that the sum of the rounded figures is equal to one.

Table 4.6 Table of frequencies in Statgraphics Centurion XV

Frequency Tabulation for Šírka súčiastky

	Lower	*Upper*			*Relative*	*Cumulative*	*Cum. Rel.*
Class	*Limit*	*Limit*	*Midpoint*	*Frequency*	*Frequency*	*Frequency*	*Frequency*
	at or below	8,45		0	0,0000	0	0,0000
1	8,45	8,46667	8,45833	2	0,0667	2	0,0667
2	8,46667	8,48333	8,475	3	0,1000	5	0,1667
3	8,48333	8,5	8,49167	12	0,4000	17	0,5667
4	8,5	8,51667	8,50833	9	0,3000	26	0,8667
5	8,51667	8,53333	8,525	2	0,0667	28	0,9333
6	8,53333	8,55	8,54167	2	0,0667	30	1,0000
	above	8,55		0	0,0000	30	1,0000

Mean = 8,49883 Standard deviation = 0,0179445

<u>Histogram and polygon of frequencies</u>

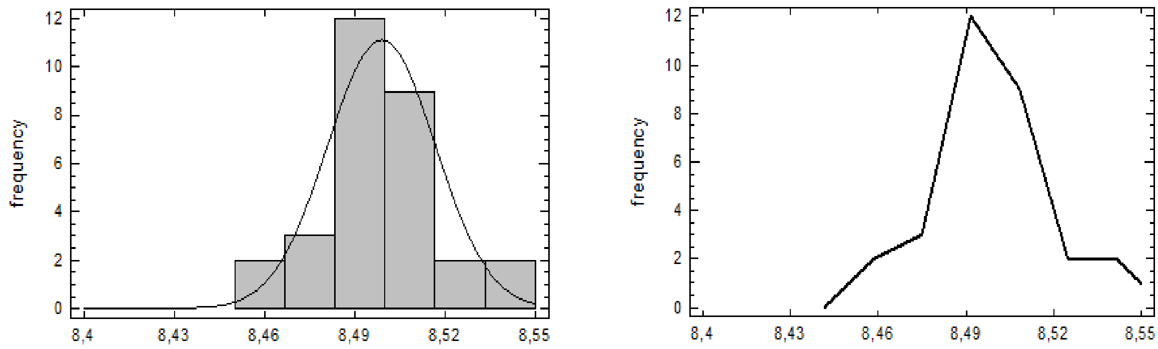


Figure 4.9 Histogram with normal distribution curve (Gaussian curve) and polygon frequencies in Statgraphics Centurion XV

Both chart show the character frequency distribution - measured data can come from a normal distribution.

<u>Normal probability plot</u>

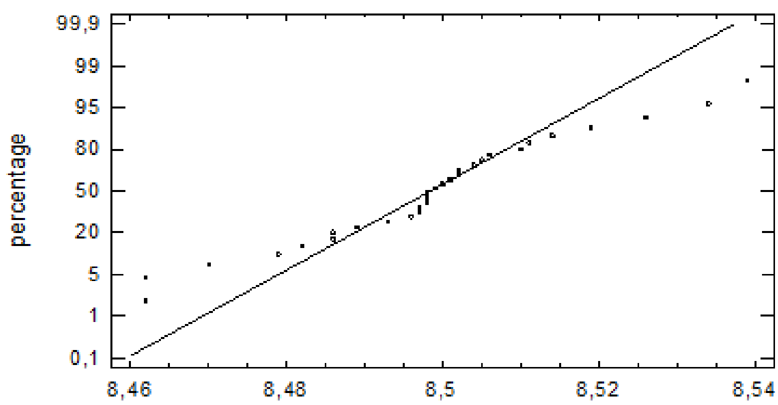


Figure 4.10 Normal probability plot in Statgraphics Centurion XV

The presented points (Figure 4.10) deviate only minimally from the approximation line.
Conclusion: From the last three graphs we conclude that the measured data can come from a normal distribution. More exactly, we check the normality of data using a normality test (e.g. Shapiro–Wilk test, see subchapter 7.6.2**).**

Box-and-whisker plot

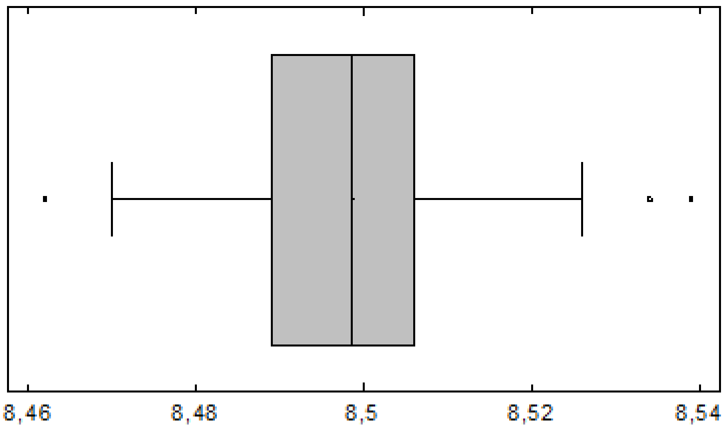


Figure 4.11 Box-and-whisker diagram in Statgraphics Centurion XV

The point in rectangle represents the mean of $\overline{x} = 8{,}49883$.
Even in this type of the chart the outliers are marked:

- lower: $x_{(1)} = x_{(2)} = 8{,}462$,
- upper: $x_{(29)} = 8{,}534$ a $x_{(30)} = 8{,}539$.

We find whether the values are only outliers or extreme outliers. Calculate the intervals:

- interval for lower outliers:

$$\left(x_{0,25} - 3 \times \text{IQR} \; ; \; x_{0,25} - 1{,}5 \times \text{IQR}\right) = \left(8{,}489 - 3 \times 0{,}017; 8{,}489 - 1{,}5 \times 0{,}017\right) = \\ = (8{,}435; 8{,}4635)$$

- interval for lower extreme outliers:

$$\left(-\infty, x_{0,25} - 3 \times \text{IQR}\right) = \left(-\infty; 8{,}489 - 3 \times 0{,}017\right) = (-\infty; 8{,}435)$$

- interval for upper outliers:

$$\left(x_{0,75} + 1{,}5 \times \text{IQR}; x_{0,75} + 3 \times \text{IQR}\right) = \left(8{,}506 + 1{,}5 \times 0{,}017; 8{,}506 + 3 \times 0{,}017\right) = \\ = (8{,}5315; 8{,}557)$$

- interval for upper extreme outliers:

$$\left(x_{0,75} + 3 \times \text{IQR}; \infty\right) = \left(8{,}506 + 3 \times 0{,}017; \infty\right) = (8{,}557; \infty)$$

Conclusion: The values $x_{(1)} = x_{(2)} = 8{,}462$ are from the interval $(8{,}435;\ 8{,}4635)$, therefore they are lower outliers. The values $x_{(29)} = 8{,}534$ and $x_{(30)} = 8{,}539$ are from the interval $(8{,}5315;\ 8{,}557)$, therefore they are upper outliers. Based on the causes of occurrence of these values it is necessary to consider whether these values will stay in or wil be discarded from a random sample. Since the lower outliers are very close to the upper limit of the interval $(8{,}435;\ 8{,}4635)$ and the upper outliers are in very close proximity to the lower limit of the interval $(8{,}5315;\ 8{,}557)$, we decided to retain them in the random sample.

5 POINT ESTIMATION

Learning goals

- Describe the terms *parameter*, *point estimator* and *point estimate*.
- Identify two major areas of statistical inference.
- Distinguish between point estimation and interval estimation.
- Determine the point estimate of a parameter.

Parameter (θ)

A parameter θ represents a characteristic of the population under study. It is constant but unknown in most cases e.g. mean (μ), variance (σ^2), proportion (p), correlation coefficient (ρ) and regression coefficient (β).

Statistical inference

Statistical inference refers to making decisions or drawing conclusions about a population by analyzing a random sample from the population. Two major areas of statistical inference can be defined:

1. **Parameter estimation**: Estimates the value of θ. E.g. $\mu = 150$
2. **Hypothesis testing**: Tests an assertion of θ. E.g. H_0: $\mu = 150$

Parameter estimation

Parameter estimation is further divided into two areas:

1. **Point estimation**: Estimates the exact location of θ. E.g. $\mu = 150$
2. **Interval estimation**: Establishes an interval that includes the true value of θ with a designated probability ($1-\alpha$, where α usually equals 0,1; 0,05; 0,01). E.g. $P(145 < \mu < 155) = 0,95$

Point estimator ($\hat{\Theta}$)

A point estimator **($\hat{\Theta}$)** is a statistic (function of random sampling) used to estimate θ. It is a random variable because a statistic is a random variable. E.g. Point estimator of μ is sample mean $\overline{X} = \sum_{i=1}^{n} X_i / n$.

A <u>single numerical value</u> of $\hat{\Theta}$ determined by a particular random sample is called a **point estimate** of θ (denoted by $\hat{\theta}$).

Example 5.1.

Suppose that the life length of an INFINITY light bulb X has a normal distribution with mean μ and variance σ^2. A random sample of size $n = 25$ light bulbs was examined and the sum of the life lengths was 14 900 hours. Estimate the mean life length (μ) of an INFINITY light bulb.

$$\hat{\mu} = \overline{x} = \frac{\sum_i x_i}{n} = \frac{14900}{25} = 745 \text{ hrs}$$

5.1 General concepts of point estimation

Learning goals

- Explain the terms *unbiased estimator*, *minimum variance unbiased estimator* (MVUE), *standard error* and *mean square error* (MSE) *of an estimator*.
- Select the appropriate point estimator of θ in terms of unbiasedness, minimum variance and minimum mean square error each.

Unbiased estimator

An unbiased estimator is a point estimator ($\hat{\Theta}$) whose expected value is equal to the true value of θ, i.e.

$$E(\hat{\Theta}) = \theta$$

Note that several unbiased estimators can be defined for a single parameter θ.

Bias of point estimator ($\hat{\Theta}$)

The bias of point estimator $\hat{\Theta}$ is the difference between the expected value of $\hat{\Theta}$ and the true value of θ:

$$B(\hat{\Theta}) = E(\hat{\Theta}) - \theta$$

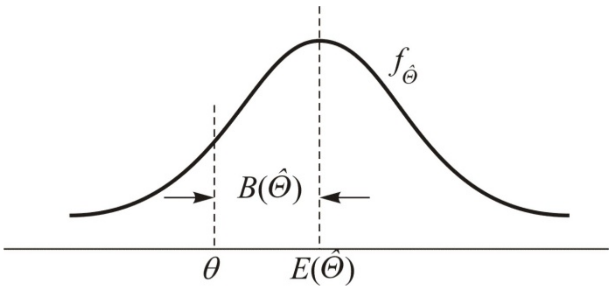


Figure 5.1 Bias of a point estimator $\hat{\Theta}$

Example 5.2

Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a probability distribution with $E(X) = \mu$ and $D(X) = \sigma^2$. Show if the sample mean $\overline{X} = \sum_{i=1}^{n} X_i / n$ is an unbiased estimator of the population mean μ:

$$E(\overline{X}) = E\left(\sum_{i=1}^{n} X_i / n\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu$$

Since $E(\overline{X}) = \mu$, $\overline{X}$ is an unbiased estimator of μ.

Minimum variance unbiased estimator (MVUE)

A minimum variance unbiased estimator of θ is the unbiased $\hat{\Theta}$ with the smallest variance. By using the MVUE of θ, the unknown parameter θ can be estimated accurately and precisely.

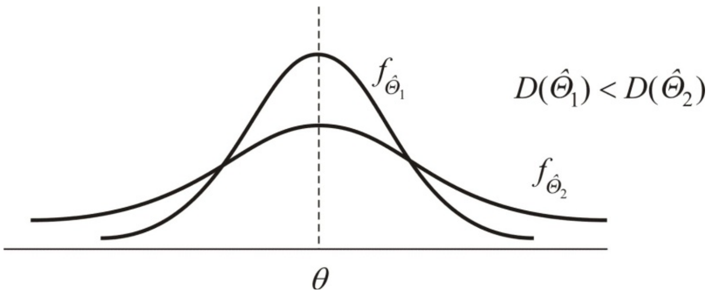


Figure 5.2 Variances of unbiased estimators of θ

Example 5.3

Let $X_1, X_2, \ldots, X_n$ denote a random sample of size $n > 1$ from a population with $E(X) = \mu$ and $D(X) = \sigma^2$. Both X_1 and $\overline{X}$ are unbiased estimators of μ because their expected values are equal to μ. Of the two estimators, which is preferred to estimate μ and why?

The variances of X_1 and $\bar{X}$ are:

$$D(X_1) = \sigma^2$$

$$D(\bar{X}) = D\left(\sum_i X_i / n\right) = \frac{1}{n^2}\sum_i D(X_i) = \frac{1}{n^2}\sum_i \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Since $\left(D(X_1) = \sigma^2\right) > \left(D(\bar{X}) = \frac{\sigma^2}{n}\right)$ for $n > 1$, $\bar{X}$ is preferred to estimate μ with higher accuracy.

Standard error ($\sigma_{\hat{\Theta}}$)

The standard error of a point estimator $\sigma_{\hat{\Theta}}$ is the standard deviation of a point estimator $\hat{\Theta}$. It can be used as a measure to indicate the **precision of** parameter estimation.
If $\sigma_{\hat{\Theta}}$ includes unknown parameters that can be estimated, use of the estimates of the parameters in calculating $\sigma_{\hat{\Theta}}$ produces an **estimated standard error** $s_{\hat{\Theta}}$.

Example 5.4

The random variable X (Example 5.1) has a normal distribution with mean μ and variance σ^2. The random sample of size $n = 25$ was examined.

1. *Standard error*

Assuming $\sigma^2 = 40^2$, determine the standard error of the sample mean ($\bar{X}$). Note that $\bar{X} = \sum_{i=1}^{n} X_i / n \sim N(\mu, \sigma^2 / n)$.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{25}} = 8\,\text{hrs}$$

2. *Estimated standard error*

Suppose that σ^2 is unknown and the sample variance $s_X^2 = 35^2$. Calculate the estimated standard error of the sample mean ($\bar{X}$).

$$s_{\bar{X}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s_X}{\sqrt{n}} = \frac{35}{\sqrt{25}} = 7\,\text{h}$$

Mean square error (MSE) of estimator

The mean square error (*MSE*) of a point estimator $\hat{\Theta}$ is the expected squared difference between $\hat{\Theta}$ and θ:

$$MSE(\hat{\Theta}) = E\left((\hat{\Theta}-\theta)^2\right) = E\left(\hat{\Theta}-E(\hat{\Theta})\right)^2 + \left(\theta - E(\hat{\Theta})\right)^2 =$$
$$= D(\hat{\Theta}) + B^2(\hat{\Theta})$$
$$= D(\hat{\Theta}) \quad \text{for unbiased } \hat{\Theta} \text{ because } B(\hat{\Theta}) = 0$$

<u>The derivation of the formula:</u>

$$MSE(\hat{\Theta}) = E\left((\hat{\Theta}-\theta)^2\right) = E\left(\left(\left(\hat{\Theta}-E(\hat{\Theta})\right)-\left(\theta-E(\hat{\Theta})\right)\right)^2\right) =$$
$$= E\left(\left(\hat{\Theta}-E(\hat{\Theta})\right)^2 - 2\left(\hat{\Theta}-E(\hat{\Theta})\right)\left(\theta-E(\hat{\Theta})\right)+\left(\theta-E(\hat{\Theta})\right)^2\right) =$$
$$= E\left(\left(\hat{\Theta}-E(\hat{\Theta})\right)^2\right) - 2E\left(\left(\hat{\Theta}-E(\hat{\Theta})\right)\left(\hat{\Theta}-E(\hat{\Theta})\right)\right) + E\left(\left(\theta-E(\hat{\Theta})\right)^2\right) =$$
$$= E\left(\left(\hat{\Theta}-E(\hat{\Theta})\right)^2\right) + E\left(\left(\theta-E(\hat{\Theta})\right)^2\right) =$$
$$= D(\hat{\Theta}) + B^2(\hat{\Theta})$$

Note. Biased estimate of the parameter with the smallest error MSE (the most accurate) is sometimes used instead of an unbiased estimate with less accuracy.

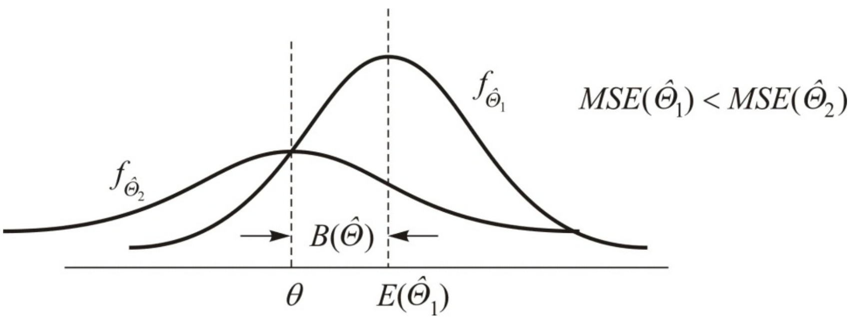


Figure 5.3 A biased estimator $\hat{\Theta}_1$ with a smaller mean square error than that of the unbiased estimator $\hat{\Theta}_2$

Example 5.5

Suppose that the means and variances of $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are $E(\hat{\Theta}_1) = \theta$, $E(\hat{\Theta}_2) = 0{,}9\theta$, $D(\hat{\Theta}_1) = 5$ and $D(\hat{\Theta}_2) = 4$. Which estimator is preferred to estimate θ and why?

For $\hat{\Theta}_1$

$$B(\hat{\Theta}_1) = E(\hat{\Theta}_1) - \theta = \theta - \theta = 0$$
$$MSE(\hat{\Theta}_1) = D(\hat{\Theta}_1) + B^2(\hat{\Theta}_1) = 5 + 0 = 5$$

For $\hat{\Theta}_2$

$$B(\hat{\Theta}_2) = E(\hat{\Theta}_2) - \theta = 0{,}9\theta - \theta = -0{,}1\theta$$

$$MSE(\hat{\Theta}_2) = D(\hat{\Theta}_2) + B^2(\hat{\Theta}_2) = 4 + 0{,}01\theta^2$$

By subtracting $MSE(\hat{\Theta}_2)$ from $MSE(\hat{\Theta}_1)$ we get:

$$MSE(\hat{\Theta}_1) - MSE(\hat{\Theta}_2) = 5 - (4 + 0{,}01\theta^2) = 1 - 0{,}01\theta^2$$

The preferred estimator of θ for precise estimation depends on the range of θ as follows:

a) $\hat{\Theta}_1$, if $\theta \geq 10$ because $MSE(\hat{\Theta}_1) \leq MSE(\hat{\Theta}_2)$ and $\hat{\Theta}_1$ is unbiased,

b) $\hat{\Theta}_2$, if $\theta < 10$ because $MSE(\hat{\Theta}_1) > MSE(\hat{\Theta}_2)$.

5.2 Methods of point estimation

Learning goals

- Explain the utility of the *maximum likelihood method*.
- Find a point estimator of θ by using the maximum likelihood method.

Maximum likelihood method

The method of maximum likelihood is used to derive a point estimator of θ. This method finds a **maximum likelihood estimator** of θ which maximizes the likelihood function of a random sample $X_1, X_2, \ldots, X_n$:

$$L(\theta) = f(x_1;\theta) f(x_2;\theta) \cdots f(x_n;\theta)$$

where $X_1, X_2, \ldots, X_n$ are independent random variables with the same probability density function $f(x;\theta)$.

Example 5.6

Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from an exponential distribution with the parameter λ. Find the *maximum likelihood estimator* of λ.

The probability density function of an exponential distribution is

$$f(x;\lambda) = \lambda e^{-\lambda x}$$

Thus the likelihood function of $X_1, X_2, \ldots, X_n$ is

$$L(\lambda) = f(x_1;\lambda) f(x_2;\lambda) \cdots f(x_n;\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

Then the log likelihood function ($L(\lambda) > 0$) is

$$\ln L(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

The derivative of $\ln L(\lambda)$ is

$$\frac{d \ln L(\lambda)}{d\lambda} = \frac{d}{d\lambda}\left(n \ln \lambda - \lambda \sum_{i=1}^{n} x_i \right) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

By equating this derivative of $\ln L(\lambda)$ to zero, the point estimator of λ which maximizes $L(\lambda)$ is

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\overline{X}}$$

5.3 Sampling distributions of means

Learning goals

- Explain the term sampling distribution.
- Explain the central limit theorem (CLT).
- Determine the distribution of a sample mean by applying the central limit theorem.

Sampling distribution

A sampling distribution is the probability distribution of a statistic (a function of random variables such as sample mean and sample variance). The sampling distribution of a statistic depends on the following:

- The distribution of the population
- The size of the sample
- The method of sample selection

Sampling distribution of $\overline{X}$

Suppose that a random sample of size n is taken from a normal distribution with mean μ and variance σ^2.

Then the sampling distribution of the sample mean is

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

<u>The derivation of the relationship</u>
Since $X_1, X_2, \ldots, X_n$ are independent and normally distributed with the same $E(X) = \mu$ and $D(X) = \sigma^2$, the distribution of $\overline{X}$ is normal with mean and variance

$$E(\overline{X}) = E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n}\left[E(X_1) + E(X_2) + \cdots E(X_n)\right] =$$

$$= \frac{1}{n}(\mu + \mu + \cdots + \mu) = \frac{1}{n} \times n\mu = \mu$$

$$D(\overline{X}) = D\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n^2}\left[D(X_1) + D(X_2) + \cdots + D(X_n)\right] =$$

$$= \frac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots + \sigma^2) = \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n}$$

Central limit theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n taken from a population (X) with mean μ and variance σ^2. Then the limiting form of the distribution of the sample mean $\overline{X}$ is

$$\overline{X} = \sum_{i=1}^{n} X_i / n \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \Rightarrow \quad Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

as n approaches infinity ($n \to \infty$). This normal approximation of $\overline{X}$ is called the *central limit theorem* (CLT).

As displayed in Figure 5.4, the distributions of the sample means from uniform, binomial and exponential distributions become normal distributions as their sample sizes n become sufficiently large. In most cases, if $n \geq 30$, normal approximation of $\overline{X}$ will be satisfactory <u>regardless of</u> the distribution of X. In case $n < 30$, if the distribution of X is **close to the normal**, a normal approximation of $\overline{X}$ will be acceptable.

Based on the central limit theorem, the sampling distribution of $\overline{X}$ where X is normal or non-normal is as follows:

1. **Normal population**, $X \sim N(\mu, \sigma^2)$, than

$$\overline{X} = \sum_{i=1}^{n} X_i / n \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

where $X_1, \ldots, X_n$ are independent random variables normally distributed $N(\mu, \sigma^2)$.

2. **Non-normal population** with parameters μ and σ^2

   a) **Normal approximation applicable**, then

   $\overline{X} = \sum_{i=1}^{n} X_i / n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, if $n \geq 30$ or the distribution of X is close to the normal

   b) **Normal approximation inapplicable**, then

   it would be difficult to find the distribution of $\overline{X}$ if $n < 30$ and the distribution of X is significantly deviated from the normal. In this case, use non-parametric statistics for statistical inference.

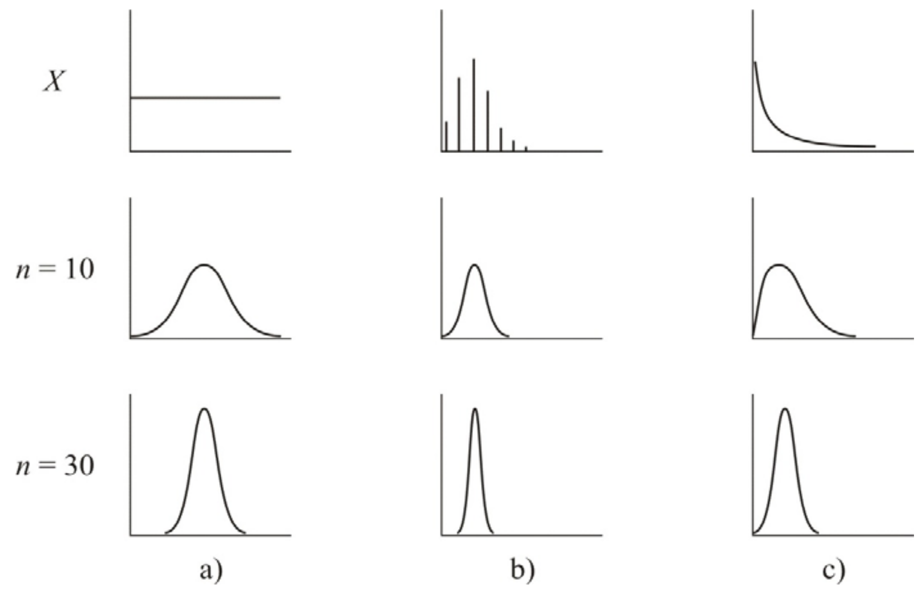


Figure 5.4 Sampling distribution of $\overline{X}$: a) $X \sim R(0, b)$; b) $X \sim Bi(n, 0{,}2)$; c) $X \sim E(\lambda = 1)$

Example 5.7

Suppose that the waiting time (X; unit: min.) of a customer to pick up his/her prescription at a drug store follows an exponential distribution with $E(X) = 20$ min and $D(X) = 400\,\text{min}^2$. A random sample of size $n = 40$ customers is observed. What is the distribution of the sample mean?

Since $n \geq 30$, normal approximation is applicable to $\overline{X}$ even if X is exponentially distributed. Thus, the sampling distribution of $\overline{X}$ is

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N(20, \frac{400}{40}) = N(20,10)$$

6 STATISTICAL INTERVALS AND SAMPLE SIZE AT A GIVEN POINT ESTIMATE ACCURACY

Learning goals

- Distinguish between confidence, prediction and statistical tolerance intervals.
- Interpret a $100(1-\alpha)$ % confidence interval (CI).
- Explain the relationship between the length of a CI and precision of estimation.
- Identify the error (E) when estimating the actual parameter.

Statistical intervals

While point estimation estimates the exact location of a parameter (θ), interval estimation establishes bounds of plausible values for θ. Three types if statistical intervals are defined:

1. **Confidence interval (CI):** Bounds a parameter of the population distribution.
   E.g. when $X \sim N(\mu,\sigma^2)$, then 90 % CI on μ indicates that the CI contains the value of μ with 90 % confidence.
2. **Prediction interval (PI):** Bounds a future observation.
   E.g. when $X \sim N(\mu,\sigma^2)$, then 90 % PI on a new observation indicates that the PI contains a new observation with 90 % confidence.
3. **Statistical tolerance interval (TI):** Bounds a selected proportion of the population distribution.
   E.g. a 95 % TI on X with 90 % confidence indicates that the TI contains 95 % of X values with 90 % confidence.

Confidence interval

A $100(1-\alpha)\%$ confidence interval on a parameter θ has both lower and upper bounds ($l \le \theta \le u$), or lower bound ($l \le \theta$) or upper bound ($\theta \le u$), which are computed by using a sample from a population. Since different samples will produce different values of l and u, the lower– and upper–confidence limits are considered values of random variables L and U which satisfy the following:

$$P(L \le \theta \le U) = 1-\alpha = 100(1-\alpha), \quad 0 \le \alpha \le 1.$$

Since L and U are random variables, a CI is a random interval. A $100(1-\alpha)$ % CI indicates that, if CIs are established from an infinite number of random samples, $100(1-\alpha)$ % of the CIs will contain the true value of θ.

There are two types of confidence intervals:

1. **Two–sided CI**: Specifies both the lower– and upper–confidence limits of θ, such as $l \le \theta \le u$.
   E.g. $730 \le \mu_X \le 750$ hrs, where X is life length of an INFINITY light bulb.
2. **One–sided CI**: Defines either the lower– or upper–confidence limits of θ, such as $l \le \theta$ or $\theta \le u$.
   E.g. $730 \le \mu_X$ or $\mu_X \le 750$ hrs, where X is life length of an INFINITY light bulb.

Length of CI and precision of estimation

The **length of a CI** refers to the distance between the upper and lower limits ($u-l$). The wider the CI, the more confident we are that the interval actually contains the true value of θ (see Figure 6.1), but less informed we are about the true value of θ.

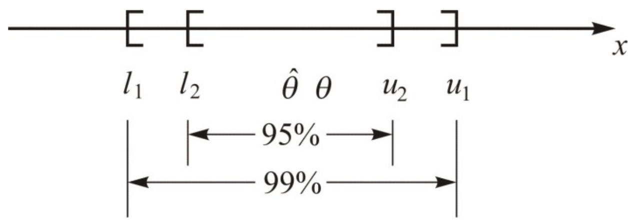


Figure 6.1 Confidence intervals on the mean life length ($\theta = \mu_X$) of an INFINITY light bulb with selected levels of confidence

The length of a CI on θ is inversely related to the **precision** of estimation on θ: the wider the CI, the less precise the estimation of θ.

Error *E* in estimation of parameter θ

$$E = \left|\hat{\theta} - \theta\right|,$$

where $\hat{\theta}$ is a point estimate of the true value of θ.

6.1 Confidence interval on the mean of a normal distribution with variance known

Learning goals

- Determine the point estimator of μ when σ^2 is known and the sampling distribution of the point estimator.
- Establish a $100(1-\alpha)\%$ CI on μ where σ^2 is known.
- Determine a sample size to satisfy a predetermined level of error (E) in estimating μ.
- Find the critical value of the normal distribution in tables in Appendix.

Inference context

- **Parameter** of interest: μ
- **Point estimator** of μ: $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$, σ^2 known
- **Statistic:** $Z = \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, where $N(0,1)$ denotes the standard normal distribution

Confidence interval formula

A $100(1-\alpha)\%$ CI on μ when σ^2 is known is

$$\overline{X} - k_\alpha \frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + k_\alpha \frac{\sigma}{\sqrt{n}} \qquad \text{for two−sided CI,}$$

$$\overline{X} - k_{2\alpha} \frac{\sigma}{\sqrt{n}} \le \mu \qquad \text{for one−sided IS with the lower bound,}$$

$$\mu \le \overline{X} + k_{2\alpha} \frac{\sigma}{\sqrt{n}} \qquad \text{for one−sided IS with the upper bound,}$$

where k_α and $k_{2\alpha}$ are critical values of $N(0,1)$ (see Appendix).

The derivation of the formula for a two−sided CI

By using the statistic $Z = \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, we get

$$P\left(-k_\alpha \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le k_\alpha\right) = 1-\alpha$$

$$P\left(\bar{X}-k_{\alpha}\frac{\sigma}{\sqrt{n}}\leq\mu\leq\bar{X}+k_{\alpha}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

Therefore

$$L=\bar{X}-k_{\alpha}\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad U=\bar{X}+k_{\alpha}\frac{\sigma}{\sqrt{n}}$$

Note that, for a one–sided CI, use $k_{2\alpha}$ instead of k_{α} to derive the corresponding limit.

Determination of sample size

To establish a $100(1-\alpha)\%$ CI on μ which does not exceed a predefined level of $E=|\bar{x}-\mu|$, the sample size is determined by the formula

$$n=\left(\frac{k_{\alpha}\sigma}{E}\right)^{2}$$

Note. In case n is not an integer, round up the value.

Example 6.1

Suppose that the life length of an INFINITY light bulb (X; unit: hour) follows the normal distribution with a mean μ and the variance $\sigma^2=40^2$, e.g. $X\sim N(\mu,40^2)$. A random sample of 30 bulbs is tested as shown below, and the sample mean is found to be $\bar{x}=780$ hours.

No.	Life length	No.	Life length	No.	Life length
1	727	11	831	21	725
2	755	12	742	22	735
3	714	13	784	23	770
4	840	14	807	24	792
5	772	15	820	25	765
6	750	16	812	26	749
7	814	17	804	27	829
8	820	18	754	28	821
9	753	19	715	29	816
10	796	20	845	30	743

1. *Confidence interval on μ, σ^2 is known*

Construct a 95 % two–sided confidence interval on the mean life length (μ) of an INFINITY light bulb.

$$P\left(l\leq\mu\leq u\right)=1-\alpha=0{,}95 \;\Rightarrow\; \alpha=0{,}05$$

95 % two−sided CI on μ:

$$\overline{X} - k_\alpha \frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + k_\alpha \frac{\sigma}{\sqrt{n}}$$

$$780 - k_{0,05} \frac{40}{\sqrt{30}} \le \mu \le 780 + k_{0,05} \frac{40}{\sqrt{30}}$$

$$780 - 1,96 \times \frac{40}{\sqrt{30}} \le \mu \le 780 + 1,96 \times \frac{40}{\sqrt{30}}$$

$$765,686 \le \mu \le 794,314$$

2. *Sample size selection*

Find a *sample size* n to construct a two−sided 95 % confidence interval on μ with an error 20 hours.

$$n = \left(\frac{k_\alpha \sigma}{E}\right)^2 = \left(\frac{k_{0,05} \times 40}{20}\right)^2 = \left(\frac{1,96 \times 40}{20}\right)^2 = 15,3664 \approx 16$$

6.2 Confidence interval on the mean of a normal distribution with unknown variance

Learning goals

- Determine the point estimator of μ when σ^2 is unknown and the sampling distribution of the point estimator.
- Establish a $100(1-\alpha)\%$ CI on μ where σ^2 is unknown.
- Find the critical value of the t-distribution in tables in Appendix.

Inference context

- **Parameter** of interest**:** μ
- **Point estimator** of μ: $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$, σ^2 is unknown
- **Statistic:** $T = \frac{\overline{X} - \mu}{S / \sqrt{n}} \sim t(n-1)$, where *S* is an estimator of σ and $t(n-1)$ denotes the *t*–distribution (Student) with the degrees of freedom $n-1$ (Janiga, 2013, subchapter 3.8.2).

Confidence interval formula

A $100(1-\alpha)\%$ CI on μ when σ^2 is unknown is

$\overline{X} - t(n-1;\alpha)\dfrac{S}{\sqrt{n}} \le \mu \le \overline{X} + t(n-1;\alpha)\dfrac{S}{\sqrt{n}}$ for two−sided CI,

$\overline{X} - t(n-1;2\alpha)\dfrac{S}{\sqrt{n}} \le \mu$ for one−sided CI with the lower bound,

$\mu \le \overline{X} + t(n-1;2\alpha)\dfrac{S}{\sqrt{n}}$ for one−sided CI with the upper bound,

where $t(n-1;\alpha)$ and $t(n-1;2\alpha)$ are critical values of a *t*–distribution with the degrees of freedom $n-1$ (see Appendix).

The derivation of the formula for a two−sided CI

$$P\left(-t(n-1;\alpha) \le T \le t(n-1;\alpha)\right) = 1-\alpha$$

$$P\left(-t(n-1;\alpha) \le \frac{\overline{X}-\mu}{S/\sqrt{n}} \le t(n-1;\alpha)\right) = 1-\alpha$$

$$P\left(\overline{X} - t(n-1;\alpha)\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t(n-1;\alpha)\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

Therefore

$$L = \overline{X} - t(n-1;\alpha)\frac{S}{\sqrt{n}} \quad \text{and} \quad U = \overline{X} + t(n-1;\alpha)\frac{S}{\sqrt{n}}$$

Example 6.2

Suppose that the life length of an INFINITY light bulb (X; unit: hour) follows the normal distribution with unknown parameters μ and σ^2. The random sample of size $n = 30$ bulbs is tested (see Example 6.1). Construct a 95 % two−sided CI on the mean life length (μ) of an INFINITY light bulb.

The point estimates values of $\bar{x} = 780$ and $s = 40,0164$ needed for the construction of CI were obtained from data on life bulbs given in Example 6.1.

$P\left(l \le \mu \le u\right) = 0,95 = 1-\alpha \Rightarrow \alpha = 0,05$

95 % two−sided CI on μ:

$$\overline{X} - t(n-1;\alpha)\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t(n-1;\alpha)\frac{S}{\sqrt{n}}$$

$$780 - t(29;0{,}05)\frac{40{,}0164}{\sqrt{30}} \le \mu \le 780 + t(29;0{,}05)\frac{40{,}0164}{\sqrt{30}}$$

$$780 - 2{,}045 \times \frac{40{,}0164}{\sqrt{30}} \le \mu \le 780 + 2{,}045 \times \frac{40{,}0164}{\sqrt{30}}$$

$$765{,}059 \le \mu \le 794{,}941$$

Notice that the CI constructed by using the t−distributed sample data $765{,}059 \le \mu \le 794{,}941$ is wider than the corresponding CI constructed on the bases of the normal distributed sample data $765{,}686 \le \mu \le 794{,}314$.

6.3 Confidence interval on the variance of a normal distribution

Learning goals

- Determine the point estimator of σ^2 and the sampling distribution of the point estimator.
- Establish a $100(1-\alpha)\%$ CI on σ^2.
- Find the critical value of the χ^2−distribution in tables in Appendix.

Inference context

- **Parameter** of interest: σ^2
- **Point estimator** of σ^2: $S^2 = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$, where $X_1, X_2, \ldots, X_n$ is random sample taken from $N(\mu, \sigma^2)$
- **Statistic:** $X^2 = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$, where S^2 is an estimator of σ^2 and $\chi^2(n-1)$ denoted χ^2–distribution with the degrees of freedom $n-1$ (Janiga, 2013, subchapter 3.8.1).

Confidence interval formula

A $100(1-\alpha)\%$ CI on σ^2 is

$$\frac{(n-1)S^2}{\chi^2(n-1;\alpha/2)} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2(n-1;1-\alpha/2)}$$ for twoo−sided CI,

$$\frac{(n-1)S^2}{\chi^2(n-1;\alpha)} \le \sigma^2$$ for one−sided CI with the lower bound,

$$\sigma^2 \le \frac{(n-1)S^2}{\chi^2(n-1;1-\alpha)}$$ for one−sided CI with the upper bound,

where $\chi^2(n-1;\alpha/2)$, $\chi^2(n-1;1-\alpha/2)$, $\chi^2(n-1;\alpha)$ and $\chi^2(n-1;1-\alpha)$ are critical values of a $\chi^2(n-1)$ distribution (see Annex).

The derivation of the formula for a two−sided CI

$$P\left(\chi^2(n-1;1-\alpha/2) \le \chi^2 \le \chi^2(n-1;\alpha/2)\right) = 1-\alpha$$

$$P\left(\chi^2(n-1;1-\alpha/2) \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2(n-1;\alpha/2)\right) = 1-\alpha$$

$$P\left(\frac{(n-1)S^2}{\chi^2(n-1;\alpha/2)} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2(n-1;1-\alpha/2)}\right) = 1-\alpha$$

Therefore

$$L = \frac{(n-1)S^2}{\chi^2(n-1;\alpha/2)} \quad \text{and} \quad U = \frac{(n-1)S^2}{\chi^2(n-1;1-\alpha/2)}$$

Example 6.3

Let the life length of an INFINITY light bulb X has a normal distribution with mean μ and variance σ^2, which are unknown. A random sample of size $n = 30$ bulbs is tested and the sample variance is found to be $s^2 = 40{,}0164^2$. Construct a 95% two–sided confidence interval on the variance of the life length of an INFINITY light bulb σ^2.

$$P\left(l \le \sigma^2 \le u\right) = 0{,}95 = 1-\alpha \;\Rightarrow\; \alpha = 0{,}05$$

95 % two−sided CI on σ^2:

$$\frac{(n-1)S^2}{\chi^2(n-1;\alpha/2)} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2(n-1;1-\alpha/2)}$$

$$\frac{(30-1)\times 40{,}0164^2}{\chi^2(29;0{,}05/2)} \le \sigma^2 \le \frac{(30-1)\times 40{,}0164^2}{\chi^2(29;1-0{,}05/2)}$$

$$\frac{46438{,}1}{45{,}7} \le \sigma^2 \le \frac{46438{,}1}{16{,}0}$$

$$1016{,}70 \le \sigma^2 \le 2902{,}38$$

$$31{,}88^2 \le \sigma^2 \le 53{,}87^2$$

6.4 A large–sample confidence interval for a population proportion

Learning goals

- Determine the point estimator of p and the sampling distribution of the point estimator.
- Establish a $100(1-\alpha)\%$ CI on p.
- Determine a sample size to satisfy a predetermined level of error (E) in estimating p.

Inference context

- **Parameter** of interest: p
- **Point estimator** of p: $\hat{P} = \dfrac{X}{n}$, where $X \sim B(n,p)$
- **Statistic:** $Z = \dfrac{\hat{P}-p}{\sqrt{p(1-p)/n}} \sim N(0,1)$ if $np(1-p) > 9$; $\hat{P}$ is an estimator of p

Sampling distribution of the estimator $\hat{P}$

The mean and variance of a binomial random variable $X \sim B(n,p)$ are

$$E(X) = np \quad \text{and} \quad D(X) = np(1-p)$$

Thus

$$E(\hat{P}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}np = p$$

$$D(\hat{P}) = D\left(\frac{X}{n}\right) = \frac{1}{n^2}D(X) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

Conditions for approximation of a binomial distribution $B(n, p)$ by a normal distribution are complied, because p is neither close to zero nor close to one and n is relatively large so that $np(1-p) > 9$.

Therefore the approximate distribution of $\hat{P}$ is

$$\hat{P} \approx N\left(p, \frac{p(1-p)}{n}\right) \;\Rightarrow\; Z = \frac{\hat{P}-p}{\sqrt{p(1-p)/n}} \approx N(0,1)$$

Confidence interval formula

A $100(1-\alpha)\%$ CI on p is

$$\hat{P} - k_{\alpha}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \le p \le \hat{P} + k_{\alpha}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \quad \text{for two−sided CI,}$$

$$\hat{P} - k_{2\alpha}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \le p \quad \text{for one−sided CI with the lower bound,}$$

$$p \le \hat{P} + k_{2\alpha}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \quad \text{for one−sided CI with the upper bound.}$$

The derivation of the formula for a two−sided CI

$$P\left(-k_{\alpha} \le Z \le k_{\alpha}\right) = 1-\alpha$$

$$P\left(-k_{\alpha} \le \frac{\hat{P}-p}{\sqrt{p(1-p)/n}} \le k_{\alpha}\right) = 1-\alpha$$

$$P\left(\hat{P} - k_{\alpha}\sqrt{\frac{p(1-p)}{n}} \le p \le \hat{P} + k_{\alpha}\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$$

We use an estimator $\hat{P}$ of the unknown parameter p.

$$P\left(\hat{P} - k_{\alpha}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \le p \le \hat{P} + k_{\alpha}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right) = 1-\alpha$$

Therefore

$$L = \hat{P} - k_{\alpha}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \quad \text{a} \quad U = \hat{P} + k_{\alpha}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$$

Determination of sample size

In estimating p the following formulas are used to calculate a sample size n for a predefined level of error:

$$n = \left(\frac{k_\alpha}{E}\right)^2 p(1-p) \quad \text{if } p \text{ is known}$$

$$n = \left(\frac{k_\alpha}{E}\right)^2 \times 0,25 \quad \text{if } p \text{ is unknown}$$

Example 6.4

A sample of $n = 40$ bridges in a certain region is tested for metal corrosion. It was found $x = 28$ corroded bridges.

1. *Confidence interval on* p

Construct a 95 % two−sided confidence interval on the proportion of corroded bridges p in the region.

$$P\left(l \le p \le u\right) = 0,95 = 1-\alpha \Rightarrow \alpha = 0,05 \qquad \hat{p} = \frac{x}{n} = \frac{28}{40} = 0,7$$

Since both $n\hat{p} = 40 \times 0,7 = 28$ and $n(1-\hat{p}) = 40 \times 0,3 = 12$ are greater than five, the sampling distribution of $\hat{P}$ is approximately normal.

95 % two−sided CI on p is

$$\hat{p} - k_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + k_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0,7 - k_{0,05} \sqrt{\frac{0,7 \times (1-0,7)}{40}} \le p \le 0,7 + k_{0,05} \sqrt{\frac{0,7 \times (1-0,7)}{40}}$$

$$0,7 - 1,96 \times 0,07 \le p \le 0,7 + 1,96 \times 0,07$$

$$0,5628 \le p \le 0,8372$$

$$0,56 \le p \le 0,84$$

2. *Sample size selection*

Determine a sample size n to establish a 95 % two−sided CI on p with an error equal to 0,05 from the true proportion.

$$n = \left(\frac{k_\alpha}{E}\right)^2 \times 0,25 = \left(\frac{k_{0,05}}{0,05}\right)^2 \times 0,25 = \left(\frac{1,96}{0,05}\right)^2 \times 0,25 = 384,2 \approx 385$$

6.5 A prediction interval for a future observation

Learning goals

- Determine the distribution of the prediction error $X_{n+1} - \overline{X}$.
- Establish a $100(1-\alpha)\%$ prediction interval (PI) for a new observation.

Sampling distribution of prediction error $E = X_{n+1} - \overline{X}$

Let $X_1, X_2, \ldots, X_n$ is a random sample from a normal population with mean μ and variance σ^2. We wish to predict a new observation X_{n+1}. If $\overline{X}$ is used as a point estimator of X_{n+1}, then the distribution of corresponding prediction error E is

$$E = X_{n+1} - \overline{X} \sim N\left(0, \sigma^2\left[1 + \frac{1}{n}\right]\right),$$

because $X_{n+1} \sim N\left(\mu, \sigma^2\right)$, $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and the statistics X_{n+1} and $\overline{X}$ are independent.

Thus

$$Z = \frac{X_{n+1} - \overline{X}}{\sigma\sqrt{1 + \frac{1}{n}}} \sim N\left(0, 1^2\right), \text{ if } \sigma^2 \text{ is known}$$

$$T = \frac{X_{n+1} - \overline{X}}{S\sqrt{1 + \frac{1}{n}}} \sim t(n-1), \text{ if } \sigma^2 \text{ is unknown}$$

Two–sided prediction interval formula

$$\overline{X} - k_\alpha \sigma\sqrt{1 + \frac{1}{n}} \le X_{n+1} \le \overline{X} + k_\alpha \sigma\sqrt{1 + \frac{1}{n}}\,,\ \sigma^2 \text{ is known}$$

$$\overline{X} - t(n-1;\alpha)\, S\sqrt{1 + \frac{1}{n}} \le X_{n+1} \le \overline{X} + t(n-1;\alpha)\, S\sqrt{1 + \frac{1}{n}}\,,\ \sigma^2 \text{ is unknown}$$

Example 6.5 (Prediction interval on X_{n+1}, σ^2 unknown)

From the light bulb life length data in Example 6.2 the following quantities have been obtained: $n = 30$, $\overline{x} = 780$ a $s^2 = 40^2$. Construct a 95 % two–sided prediction interval on the life length of the next light bulb tested (X_{31}**).**

$$P(l \le p \le u) = 0,95 = 1-\alpha \;\Rightarrow\; \alpha = 0,05$$

95 % two−sided PI: (σ^2 unknown) on X_{31}

$$\overline{X} - t(n-1;\alpha)S\sqrt{1+\frac{1}{n}} \le X_{n+1} \le \overline{X} + t(n-1;\alpha)S\sqrt{1+\frac{1}{n}}$$

$$780 - t(29;0,05)\times 40,0164\times\sqrt{1+\frac{1}{30}} \le X_{31} \le 780 + t(29;0,05)\times 40,0164\times\sqrt{1+\frac{1}{30}}$$

$$780 - 2,045\times 40,677873 \le X_{31} \le 780 + 2,045\times 40,677873$$

$$780 - 83,18625 \le X_{31} \le 780 + 83,18625$$

$$696,81375 \le X_{31} \le 863,18625$$

Note that this $t-$based PI $[696,81375;\ 863,18625]$ is <u>wider</u> than the corresponding $t-$based CI on μ $[765,065;\ 794,935]$ in Example 6.2).

6.6 Statistical tolerance intervals for a normal distribution with unknown parameters

Learning goals

- Establish a p tolerance interval with $100(1-\alpha)\%$ confidence for a normal population with unknown parameters μ and σ^2.

Statistical tolerance interval

Suppose that the life length X of an INFINITY light bulb follows a normal distribution with mean $\mu = 780$ and variance $\sigma^2 = 40^2$. Then the interval

$$(\mu - 1,96\sigma;\, \mu + 1,96\sigma) = (780 - 1,96\times 40;\, 780 + 1,96\times 40)$$

includes 95 % of the light bulb population in terms of life length. The interval $(\mu - 1,96\sigma;\, \mu + 1,96\sigma)$ is called **statistical tolerance interval**.

When μ and σ^2 are unknown, it may be used data $x_1, x_2, \ldots, x_n$ from the sample of size n to compute the values of sample mean $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$. Then it is possible to establish the interval $(\overline{x} - 1,96s;\, \overline{x} + 1,96s)$.

However, due to sampling variability in $\bar{x}$ and s^2, the estimated statistical tolerance interval includes less than 95% of the population values. The solution to this problem lies in replacing the value of 1.96 by some other value that will create the interval containing 95% of the values of the population with some level of confidence.

Definition of the two–sided and one–sided statistical tolerance interval

The $100(1-\alpha)$ % **two–sided** statistical tolerance interval (Garaj, I., Janiga, I., 2002) is the interval

$$(\bar{x}-ks;\bar{x}+ks)$$

for which the following equation is valid

$$P\left[P(\bar{x}-ks < X < \bar{x}+ks) \geq p\right] = 1-\alpha,$$

where $k = k(n, p, 1-\alpha)$ is tolerance factor (see Appendix), $1-\alpha$ is a confidence and p is the proportion of values from distribution $N(\mu, \sigma^2)$.

The $100(1-\alpha)$ % **one–sided** statistical tolerance interval (Garaj, I., Janiga, I., 2005) is the interval

$$(-\infty, \bar{x}+ks) \text{ or } (\bar{x}-ks, \infty)$$

for which the following is valid

$$P\left[P(X < \bar{x}+ks) \geq p\right] = 1-\alpha \quad \text{or} \quad P\left[P(\bar{x}-ks < X) \geq p\right] = 1-\alpha$$

where $k = k(n, p, 1-\alpha)$ is tolerance factor (see Appendix), $1-\alpha$ is a confidence and p is the proportion of values from distribution $N(\mu, \sigma^2)$.

Example 6.6

From the data on the light bulbs life length, which come from a normal distribution, we obtain values: $n = 30,$ $\bar{x} = 780$ a $s = 40{,}0164.$ Construct a two−sided statistical tolerance interval with 90 % confidence that covers at least 95 % of the life length measurement of light bulbs.

For $n = 30$, $p = 0{,}95$ a $1-\alpha = 0{,}90$ can be found in the appropriate table (see Appendix) the value of $k = 2{,}4166$. The values are substituted into the relationship $(\bar{x}-ks, \bar{x}+ks)$ and we obtain:

$$(780 - 2{,}4166 \times 40{,}0164; 780 + 2{,}4166 \times 40{,}0164)$$

$$(780 - 96{,}7036; 780 + 96{,}7036)$$

$$(683{,}2964; 876{,}7036)$$

After rounding down the lower limit and rounding up the upper limit we obtain the interval

$$(683{,}29;\, 876{,}71)\,.$$

We want to construct a one–sided statistical tolerance interval with 90 % confidence that covers at least 95 % of the life length measurement of light bulbs.

For $n = 30$, $p = 0{,}95$ and $1-\alpha = 0{,}90$ the value $k = 2{,}0799$ can be found in the appropriate table (see Appendix). The values are substituted into the relationship $(\overline{x} - ks, \infty)$ and we obtain:

$$(780 - 2{,}0799 \times 40{,}0164; \infty)$$

$$(780 - 83{,}2301; \infty)$$

$$(696{,}7698896; \infty)$$

After rounding down the lower limit to three decimal places we obtain the interval

$$(696{,}769; \infty)\,.$$

7 TESTS OF HYPOTHESES FOR A SINGLE SAMPLE

7.1 Hypothesis testing

Learning goals

- Explain the terms *null hypothesis*, *alternative hypothesis*, *test statistic*, *acceptance region*, *rejection region*, *critical value*, *type I error probability* (α), *type II error probability* (β) and *power of a test* $(1-\beta)$.
- Establish the acceptance and rejection regions of hypothesis test at α.
- Determine the type II error probability and power of a test.
- Explain the relationships between α and β.
- Identify the procedure of hypothesis testing.

Hypothesis

A hypothesis is an assertion about the parameters (θ) of one or more populations under study. There are two kinds of hypotheses:

1. **Null hypothesis** $\left(H_0 : \theta = \theta_0\right)$

   States the presumed condition of θ (based on experience, theory, design specification, regulation or contractual obligation) that will be held unless there is a strong evidence against it. Note that H_0 should always specify an exact value of θ. E.g. $H_0 : \mu_X = 750$ hrs, where X is the life length of an INFINITY light bulb.

2. **Alternatívna hypotéza (** H_1 **):**

   States the condition of θ that would be concluded if H_0 is rejected. The following types of H_1 are defined:

   – **two–sided** $H_1 : \theta \neq \theta_0$**:** Indicates no directionality of θ. E.g. $H_1 : \mu_X \neq 750$ hrs,

   – **one–sided** $H_1 : \theta < \theta_0$ or $H_1 : \theta > \theta_0$**:** Indicates the directionality of θ. E.g. lower–side inequality $H_1 : \mu_X < 750$ hrs or upper–side inequality $H_1 : \mu_X > 750$ hrs.

Test statistic

A test statistic refers to a statistic used for statistical inference about θ. E.g. test statistic for inference on μ, where $X \sim N(\mu, \sigma^2)$ and σ^2 is known, is

$$\bar{X} = \sum_{i=1}^{n} X_i / n \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{or} \quad Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

Test regions

Two regions of a test statistic (see Figure 7.1) are established for testing H_0 against H_1**:**

- **Acceptance region ($\bar{K}$):** The region of a test statistic that will lead to failure to reject of H_0.
- **Rejection (critical) region (*K*):** The region of a test statistic that will lead to rejection of H_0.

The boundaries between the acceptance and rejection regions are called **critical values.** When we mark the critical values (k_1, k_2), then $\bar{K} = [k_1, k_2]$ and $K = (-\infty, k_1) \cup (k_2, \infty)$.

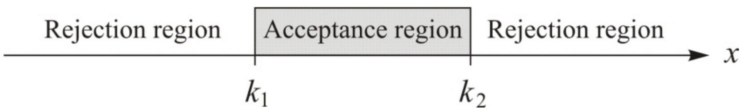


Figure 7.1 Acceptance and rejection regions for H_0

Test errors

The truth or falsity of a hypothesis can never be known with certainty unless the entire population is examined accurately and thoroughly. Therefore a hypothesis test based on a random sample may lead to one of the two types of wrong conclusions (see Table 7.1):

1. **Type I error:** Reject H_0 when H_0 is true.
2. **Type II error:** Fail to reject H_0 when H_0 is false.

Table 7.1 Decision matrix of hypothesis testing

	Fail to reject H_0	Reject H_0
H_0 is true	Correct decision	**Type I error**
H_0 is false	**Type II error**	Correct decision

The probability of type I error (denoted as α) and **the probability of type II error** (denoted as β) are conditional probabilities as follows:

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ is true})$$

$$\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 | H_0 \text{ is false})$$

The type I probability α is also called the **level of significance** of a test. The values of $\alpha = 0{,}05$ and $\alpha = 0{,}01$ are the most used.

Power of test

The **power** of a statistical test indicates the probability of rejecting H_0 when H_0 is false and indicates the ability (sensitivity) of the test to detect evidence against H_0.

$$\begin{aligned}\text{power of test} &= P(\text{reject } H_0 \mid H_0 \text{ is false}) = \\ &= 1 - P(\text{fail to reject } H_0 \mid \text{H}_0 \text{ is false}) = \\ &= 1 - \beta\end{aligned}$$

Test regions, hypotheses, α, β and power of test

The test regions, hypotheses, α, β and power of a test are related to each other. The acceptance and rejection regions of a test statistic $\hat{\Theta}$ are determined based on α and the hypothesized value of θ (denoted as θ_0) in H_0 (note that α and θ_0 are specified by the analyst). If L and U denote the lower and upper limits of an acceptance region, respectively, and we are testing $H_0 : \theta = \theta_0$, the acceptance region of $\hat{\Theta}$ is determined as follows:

$$1 - \alpha = P(\text{fail to reject } H_0 \mid H_0 \text{ is true}) = P(\text{fail to reject } H_0 \mid \theta = \theta_0) =$$

$$= \begin{cases} P(L \le \hat{\Theta} \le U \mid \theta = \theta_0), & \text{for two-sided } H_1 \\ P(L \le \hat{\Theta} \mid \theta = \theta_0), & \text{for lower-sided } H_1 \\ P(\hat{\Theta} \le U \mid \theta = \theta_0), & \text{for upper-sided } H_1 \end{cases}$$

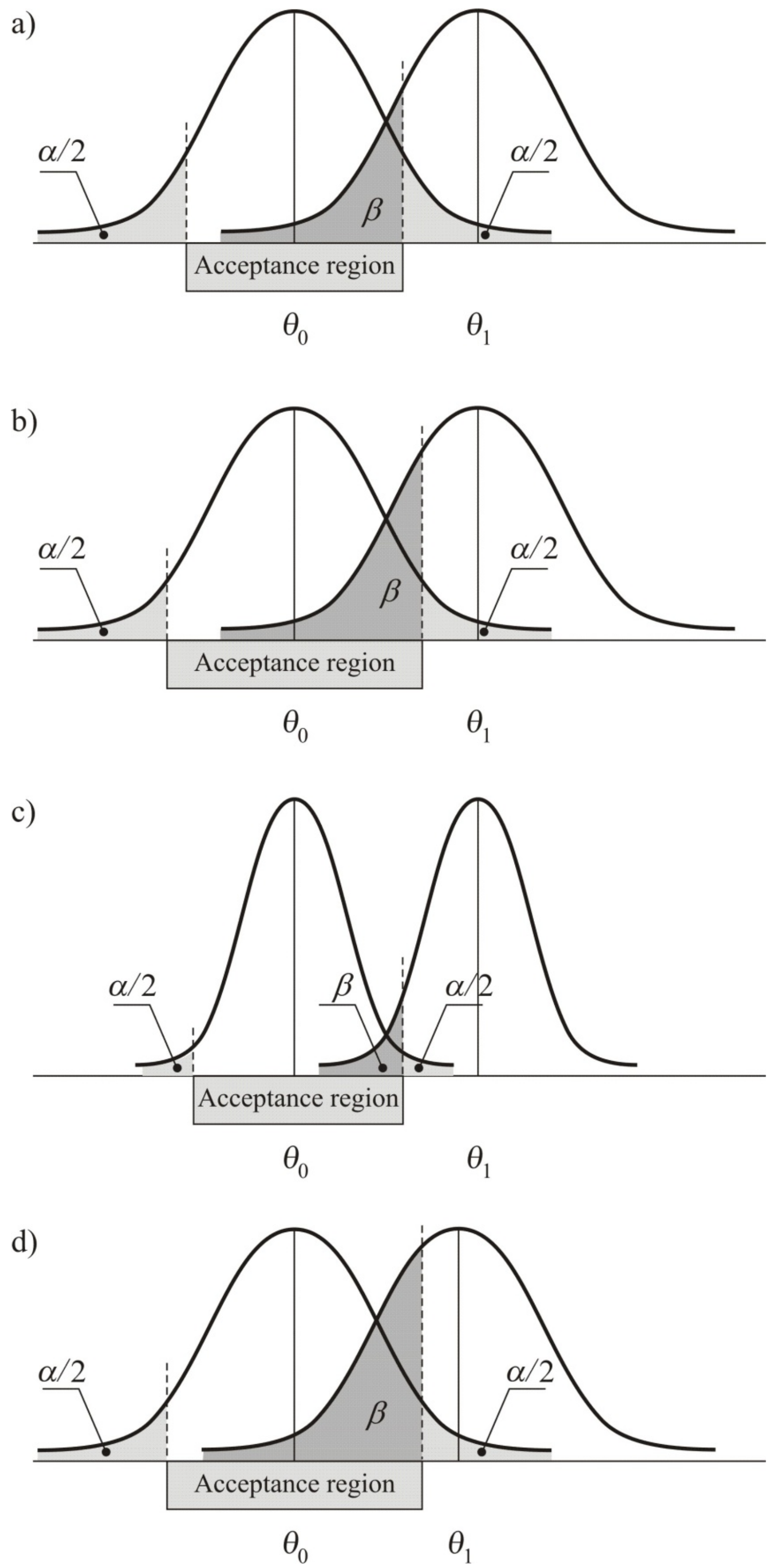
On the other hand, the β and power of a test $1 - \beta$ are determined based on the acceptance region of the test and the true value of θ as follows:

$$\beta = P(\text{fail to reject } H_0 \mid H_0 \text{ is false}) = P(\text{fail to reject } H_0 \mid \theta \ne \theta_0) =$$

$$= \begin{cases} P(L \le \hat{\Theta} \le U \mid \theta \ne \theta_0), & \text{for two-sided } H_1 \\ P(L \le \hat{\Theta} \mid \theta \ne \theta_0), & \text{for lower-sided } H_1 \\ P(\hat{\Theta} \le U \mid \theta \ne \theta_0), & \text{for upper-sided } H_1 \end{cases}$$

and power of the test $= 1 - \beta$.

Relationship between α and β



Hypothetical distributions of $\bar{X} \sim N(\mu, \sigma^2 / n)$ for $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$

For *n* fixed the folowing is valid:
As the acceptance region widens, α decreases but β increases.

For constant critical values the folowing is valid:
As n increases, both α and β decrease.

When $H_0 : \theta = \theta_0$ is false:
β increases as the true value of θ approaches to θ_1 and vice versa.

Figure 7.2 Relationships between α and β

Example 7.1

Suppose that the life length of an INFINITY light bulb (*X*; unit: hour) is normally distributed with $\sigma^2 = 40^2$. We wish to test H_0: $\mu_X = 750$ versus H_1: $\mu_X \neq 750$ with a random sample of size $n = 30$ light bulbs.

Solution

1. *Acceptance and rejection regions*

Construct the acceptance and rejection regions of the test on μ at $\alpha = 0{,}05$

The test statistic of μ is the sample mean with the following sampling distribution:

$$\bar{X} \sim N\left(\mu, \frac{40^2}{30}\right)$$

The acceptance region $l \le \bar{X} \le u$ for H_0: $\mu_X = 750$ versus H_1: $\mu_X \neq 750$ satisfies the following:

$$1-\alpha = 1-0{,}05 = 0{,}95 = P(\text{fail to reject } H_0 | \mu = 750)$$

$$0{,}95 = P(l \le \bar{X} \le u | \mu{=}750) =$$

$$= P\left(\frac{l-\mu}{40/\sqrt{30}} \le \frac{\bar{X}-\mu}{40/\sqrt{30}} \le \frac{u-\mu}{40/\sqrt{30}} | \mu{=}750\right) =$$

$$= P\left(\frac{l-750}{40/\sqrt{30}} \le Z \le \frac{u-750}{40/\sqrt{30}}\right) = P\left(-k_\alpha \le Z \le k_\alpha\right) =$$

$$= P(-1{,}96 \le Z \le 1{,}96)$$

Accordingly, the critical values are

$$\frac{l-750}{40/\sqrt{30}} = -1{,}96 \Rightarrow l = 750 - 1{,}96 \times \frac{40}{\sqrt{30}} = 735{,}686$$

$$\frac{u-750}{40/\sqrt{30}} = 1{,}96 \Rightarrow u = 750 + 1{,}96 \times \frac{40}{\sqrt{30}} = 764{,}314$$

Therefore

acceptance region of H_0: $735{,}686 \le \bar{x} \le 764{,}314$

rejection region of H_0: $\bar{x} < 735{,}686$ and $\bar{x} > 764{,}314$

2. *β and power of test*

Assume that the true value of $\mu_X = 730$ hrs. Find the β and power of the test if the acceptance region is $735{,}686 \le \bar{x} \le 764{,}314$.

$$\beta = P(\text{fail to reject } H_0 | H_0 \text{ is false}) = P(735{,}686 \le \overline{X} \le 764{,}314 | \mu = 730) =$$

$$= P\left(\frac{735{,}686-\mu}{40/\sqrt{30}} \le \frac{\overline{X}-\mu}{40/\sqrt{30}} \le \frac{764{,}314-\mu}{40/\sqrt{30}} | \mu{=}730 \right) =$$

$$= P\left(\frac{735{,}686-730}{40/\sqrt{30}} \le Z \le \frac{764{,}314-730}{40/\sqrt{30}} \right) = P(0{,}77859 \le Z \le 4{,}69864) =$$

$$= P(Z \le 4{,}70) - P(Z \le 0{,}78) = 1 - 0{,}7823 = 0{,}2177$$

Power of the test $= 1 - \beta = 0{,}7823$.

Hypothesis testing procedure

1. We formulate the null hypothesis H_0 and alternative hypothesis H_1 (double-sided or one-sided).
2. Determine the **test statistic and its distribution**.
3. Calculate the **value of the test statistic**.
4. Select the **level of significance** α.
5. Determine a **critical value(s) for the** α.
6. Make a **conclusion**, such as *we reject or fail to reject H_0 at α*.

7.2 Tests on the mean of a normal distribution, variance known

Learning goals

- Test a hypothesis on μ when σ^2 is known (*z*-test).
- Calculate the *P*-value of a *z*-test.
- Compare the α–value approach with the *P*-value approach in evaluating hypothesis test results.
- Explain the relationship between confidence interval estimation and hypothesis testing.
- Determine the sample size of a *z*-test for statistical inference on μ by applying an appropriate sample size formula and operating charakteristic (OC) curve.
- Explain the effect of the sample size n on the statistical significance and power of the test.
- Distinguish between statistical significance and practical significance.

Inference context

Parameter: μ

Point estimator of μ: $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$; σ^2 is known

Test statistic of μ: $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$

Test procedure (*z*-test):

Step 1: State the **null hypothesis** H_0 **and alternative hypothesis** H_1.

$H_0: \mu = \mu_0$ $\quad$ $H_1: \mu \neq \mu_0$ for two–sided test

$\mu < \mu_0$ for lower–sided test

$\mu > \mu_0$ for upper–sided test

Step 2: Determine a **test statistic and its value.**

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}, \qquad z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Step 3: Determine a **critical value(s) for *α*.**

k_α for two–sided test

$k_{2\alpha}$ for one–sided test

Step 4: Make a **conclusion.** Reject H_0 if

$|z_0| > k_\alpha$ for two–sided test

$z_0 < -k_{2\alpha}$ for lower–sided test

$z_0 > k_{2\alpha}$ for upper–sided test

***P*-value in hypothesis testing**

The ***P*-value** is the smallest level of significance that would lead to rejection of the null hypothesis H_0 with the given data.

The *P*-value of a test statistic z_0 can be computed by using the following formulas:

$$P = \begin{cases} 2[1 - \Phi(|z_0|)] & \text{for two-sided test} \\ 1\text{-}\Phi(\text{z}_0) & \text{for upper-sided test} \\ \Phi(\text{z}_0) & \text{for lower-sided test} \end{cases}$$

α verzus *P*-value approach

Two approaches are available to use in reporting the result of a hypothesis test:

1. ***α*-value approach:** States the test result at the value of α preselected.
2. ***P*-value approach:** Specifies how far the test statistic is from the critical value(s). Once the *P*-value is known, the decision maker can draw a conclusion at any specified level of significance α as follows:
   - reject H_0 at α , if $P \le \alpha$
   - fail to reject H_0 at α , otherwise

Note that the *P*-value approach is more flexible and informative than the *α*-value approach.

Formulas for confidence intervals (CI)

For $100(1-\alpha)\%$ confidence interval on μ, when σ^2 is known, the following formulas are applied:

$$\overline{X} - k_\alpha \frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + k_\alpha \frac{\sigma}{\sqrt{n}} \quad \text{for two–sided CI}$$

$$\overline{X} - k_{2\alpha} \frac{\sigma}{\sqrt{n}} \le \mu \quad \text{for lower–sided CI}$$

$$\mu \le \overline{X} + k_{2\alpha} \frac{\sigma}{\sqrt{n}} \quad \text{for upper–sided CI}$$

Testing hypotheses using the confidence interval

Null hypothesis $H_0\colon \mu = \mu_0$ is rejected on the level of significance α if

$$\mu_0 \notin \left[\overline{X} - k_\alpha \frac{\sigma}{\sqrt{n}}, \overline{X} + k_\alpha \frac{\sigma}{\sqrt{n}} \right] \quad \text{for two–sided test}$$

$$\mu_0 > \overline{X} + k_{2\alpha} \frac{\sigma}{\sqrt{n}} \quad \text{for lower–sided test}$$

$$\mu_0 < \overline{X} - k_{2\alpha} \frac{\sigma}{\sqrt{n}} \quad \text{for upper–sided test}$$

For example, suppose that a 95 % CI on μ is $751 \le \mu \le 779$ and we are testing $H_0\colon \mu = 750$ vs. $H_1\colon \mu \ne 750$ at $\alpha = 0{,}05$. Since the CI of μ does not include the hypothesized value $\mu = 750$, we will reject H_0 at $\alpha = 0{,}05$.

Sample size formula

Formulas are used to determine the sample size of a particular test for particular levels of β (or power of test $=1-\beta$), α and $\delta\,(=|\mu-\mu_0|$, that is difference between the true value μ and its hypothesized value μ_0). For a *z*-test on single sample, the following formulas are applied:

$$n=\frac{(k_\alpha+k_{2\beta})^2\sigma^2}{\delta^2} \quad \text{for two−sided test}$$

$$n=\frac{(k_{2\alpha}+k_{2\beta})^2\sigma^2}{\delta^2} \quad \text{for one−sided test}$$

Note that the sample size requirement icreases as α, β and δ decrease and σ increases.

Operating characteristic (OC) curve

Operating characteristic (OC) curves for a *z*-test on μ are provided in Appendix. The OC curves plot β against *d* for various sample sizes n and two levels of significance $\alpha=0{,}01$ and $\alpha=0{,}05$, i.e.

$$\beta=f(n,d,\alpha)$$

Table 7.2 Operating characteristic charts for z-test − single sample

Test		α	OC curve*	OC parameter
z-test	Two−sided	0,05	OC–a	$d=\frac{\lvert\mu-\mu_0\rvert}{\sigma}=\frac{\lvert\delta\rvert}{\sigma}$
		0,01	OC–b	
	One−sided	0,05	OC–c	
		0,01	OC–d	

*See in Appendix.

Effect of sample size

As the sample size *n* increases, both the statistical significance (inverse to *P*-value) and power ($1-\beta$) of the test increase. For example, Table 7.3 presents *P*-values and power of testing on μ for the following conditions:

- $\bar{X}\sim N\left(\mu;\frac{2^2}{n}\right)$ and $\bar{x}=50{,}5$
- H_0: $\mu=50$ vs. H_1: $\mu\neq 50$

– $\alpha = 0,05$ and the true value of $\mu = 50,5$.

The P-value column indicates that, for the same value of $\bar{x} = 50,5$:

– H_0 is <u>rejected</u> at $\alpha = 0,05$ when $n = 100$ because $P \le \alpha$, while

– H_0 is <u>not rejected</u> at $\alpha = 0,05$ when $n \le 50$ because $P > \alpha$.

Table 7.3 The P-values and powers of testing on μ for selected sample sizes

Sample size (*n*)		**1 - *P***	***P*-value**	**Power of test** $(1-\beta)$	
.	10	.	0,43	.	0,124
.	25	.	0,21	.	0,240
.	50	↓	0,08	.	0,424
↓	100	increasing	0,01	↓	0,705
increasing	400	statistical	$5,73\times10^{-7}$	increasing	0,998
sample size	1000	significance	$2,57\times10^{-15}$	power of test	>0,999

Statistical versus practical significance

The statistical significance of a test does <u>not</u> necessarily indicate its practical significance. For example, when the sample size increases, then the power of the test increases. In this case, any small departure of μ from the hypothesized value μ_0 will be detected (in other words, H_0: $\mu = \mu_0$ will be rejected) for a large sample, even when the departure is of little practical significance. Therefore, the analyst should check if the statistical test result has also practical significance.

Example 7.2

For the light bulb life length data in Table 6.1, the following results have been obtained: $n = 30$, $\bar{x} = 780$, $\sigma^2 = 40^2$.

1. *Hypothesis test on* μ, σ^2 *known; two–sided test*

Test H_0: $\mu = 765$ hrs vs. H_1: $\mu \ne 765$ hrs at $\alpha = 0,05$.

<u>Procedure:</u>

Step 1: State H_0 and H_1.

$$H_0: \mu = 765 \qquad H_1: \mu \ne 765$$

Step 2: Determine a **test statistic and its value.**

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{780 - 765}{40/\sqrt{30}} = 2,05396$$

Step 3: Determine a **critical value(s) for *α*.**

$k_{\alpha} = k_{0,05} = 1{,}96$

Step 4: Make a **conclusion.**

Since $|z_0| = 2{,}05396 > k_{0.05} = 1{,}96$, H_0 reject at the level of significance $\alpha = 0{,}05$.

2. *P-value approach*

Find the *P*-value for this two–sided *z*-test.

$$P = 2[1 - \Phi(|z_0|)] = 2[1 - \Phi(|2{,}05396|)] = 2[1 - 0{,}98001] = 0{,}03998$$

Conclusion: Since $P = 0{,}03998 \le \alpha = 0{,}05$, reject H_0 at $\alpha = 0{,}05$.

3. *Relationship between CI and hypothesis test*

Test H_0: $\mu = 765$ hrs vs. H_1: $\mu \ne 765$ hrs at $\alpha = 0{,}05$ based on the 95 % two–sided CI on μ.

Conclusion: Since the 95 % two–sided CI on μ, $765{,}686 \le \mu \le 794{,}314$, does not include the hypothesized value 765 hrs, reject H_0 at $\alpha = 0{,}05$.

4. *Sample size determination*

Determine the sample size n required for this two–sided *z*-test to detect the true mean as high as 785 hours with power of test 0,9. Apply an appropriate sample size formula and OC curve.

a) Sample size formula

Power of test $= P(\text{reject} H_0 | H_0 \text{ is false}) = 1 - \beta = 0{,}9 \Rightarrow \beta = 0{,}1 \Rightarrow 2\beta = 0{,}2$

$\delta = \mu - \mu_0 = 785 - 765 = 20$

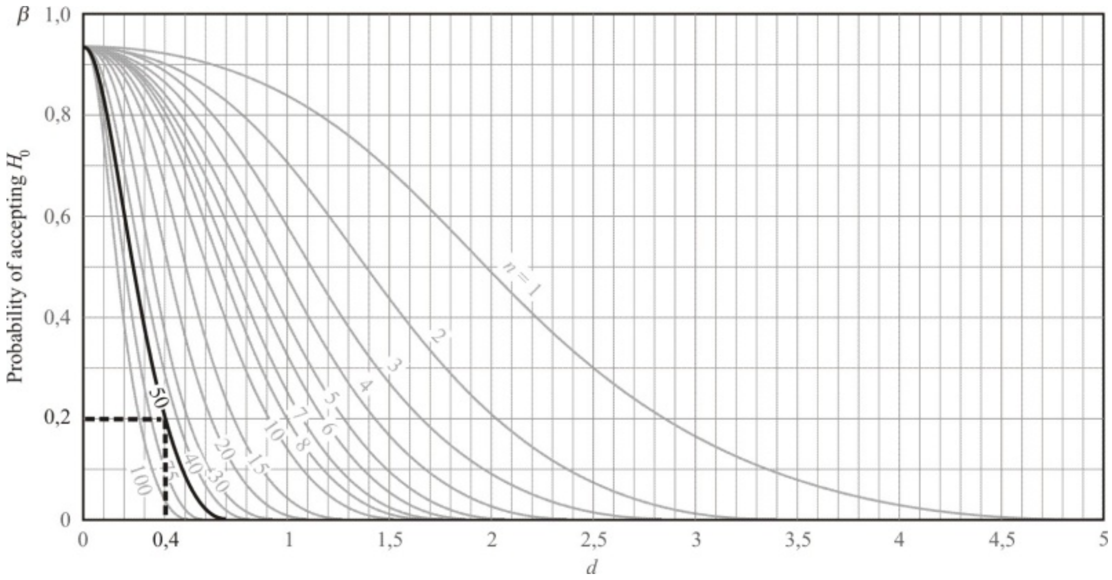
$$n = \frac{(k_{\alpha} + k_{2\beta})^2 \sigma^2}{\delta^2} = \frac{(k_{0,05} + k_{0,2})^2 40^2}{20^2} = \frac{(1{,}96 + 1{,}28)^2 40^2}{20^2} = 41{,}9904 \approx 42$$

b) OC curve

For two–sided *z*-test at $\alpha = 0{,}05$ and for a single sample, we calculate the value of the parameter d:

$$d = \frac{|\mu - \mu_0|}{\sigma} = \frac{|\delta|}{\sigma} = \frac{|20|}{40} = 0{,}5$$

For $d = 0{,}5$ and $\beta = 0{,}1$, the OC–a curve displayed below (see also in Appendix) provides the required sample size $n = 44$, which is close to the value $n = 42$ calculated by using the sample size formula.



OC-a curve for the two–sided normal test with different values of *n* and $\alpha = 0,05$.

7.3 Tests on the mean of a normal distribution, variance unknown

Learning goals

- Test a hypothesis on μ when σ^2 is known (*t*-test).
- Determine the sample size of a *t*-test for statistical inference on μ by using an appropriate operating charakteristic (OC) curve.

Inference context

Parameter: μ

Point estimator of μ: $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$; σ^2 is unknown

Test statistic of μ: $T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$

Test procedure (*t*-test):

Step 1: State the **null hypothesis** H_0 **and alternative hypothesis** H_1.

H_0: $\mu = \mu_0$ H_1: $\mu \neq \mu_0$ for two–sided test

$\mu < \mu_0$ for lower–sided test

$\mu > \mu_0$ for upper–sided test

Step 2: Determine a **test statistic and its value.**

$$T_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}, \qquad t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Step 3: Determine a **critical value(s) for *α*.**

$t(n-1;\alpha)$ for two–sided test

$t(n-1;2\alpha)$ for one–sided test

Step 4: Make a **conclusion.** Reject H_0 if

$|t_0| > t(n-1;\alpha)$ for two–sided test

$t_0 < -t(n-1;2\alpha)$ for lower–sided test

$t_0 > t(n-1;2\alpha)$ for upper–sided test

Formulas for confidence intervals (CI)

For $100(1-\alpha)\%$ confidence interval on μ, when σ^2 is unknown, the following formulas are applied:

$$\bar{X} - t(n-1,\ \alpha)\frac{S}{\sqrt{n}} \ \le\ \mu \ \le\ \bar{X} + t(n-1,\ \alpha)\frac{S}{\sqrt{n}} \quad \text{for two–sided CI}$$

$$\bar{X} - t(n-1,\ 2\alpha)\frac{S}{\sqrt{n}} \ \le\ \mu \quad \text{for lower–sided CI}$$

$$\mu \le \bar{X} + t(n-1,\ 2\alpha)\frac{S}{\sqrt{n}} \quad \text{for upper–sided CI}$$

Testing hypotheses using the confidence interval

Null hypothesis H_0: $\mu = \mu_0$ is rejected on the level of significance α if

$$\mu_0 \notin \left[\bar{X} - t(n-1,\ \alpha)\frac{S}{\sqrt{n}}\ ;\ \bar{X} + t(n-1,\ \alpha)\frac{S}{\sqrt{n}}\right] \quad \text{for two–sided test}$$

$$\mu_0 > \bar{X} + t(n-1,\ 2\alpha)\frac{S}{\sqrt{n}} \quad \text{for lower–sided test}$$

$$\mu_0 < \bar{X} - t(n-1,\ 2\alpha)\frac{S}{\sqrt{n}} \quad \text{for upper–sided test}$$

Operating characteristic (OC) curve

Operating characteristic (OC) curves for a *t*-test on μ are provided in Appendix. The OC curves plot β against d (for *t*-test) for various sample sizes n and two levels of significance $\alpha = 0,01$ and $\alpha = 0,05$, i.e.

$$\beta = f(n, d, \alpha)$$

Table 7.4 Operating characteristic charts for t-test − single sample

Test		α	OC°	OC parameter
t-test	Two−sided	0,05	OC–e	$d = \dfrac{\lvert\mu - \mu_0\rvert}{\hat{\sigma}} = \dfrac{\lvert\delta\rvert}{\hat{\sigma}}^{*}$
		0,01	OC–f	
	One−sided	0,05	OC–g	
		0,01	OC–h	

∗ As a $\hat{\sigma}$ use sample standard deviation. °See in Appendix.

Example 7.3

For the light bulb life length data in Table 6.1, the following results have been obtained: $n = 30$, $\bar{x} = 780$, $s^2 = 40,0164^2$.

1. *Hypothesis test on* μ, σ^2 *unknown; two–sided test*

Test H_0: $\mu = 765$ hrs vs. H_1: $\mu \neq 765$ hrs at $\alpha = 0,05$.

Procedure:

Step 1: State H_0 and H_1.

H_0: $\mu = 765$ $\quad$ H_1: $\mu \neq 765$

Step 2: Determine a **test statistic and its value.**

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{780 - 765}{40,0164/\sqrt{30}} = 2,05312$$

Step 3: Determine a **critical value(s) for *α*.**

$$t(n-1;\alpha) = t(30-1;0,05) = t(29;0,05) = 2,045$$

Step 4: Make a **conclusion.**

Since $|t_0| = 2,05312 > t(29;0,05) = 2,045$, reject H_0 at the level of significance $\alpha = 0,05$.

2. *Sample size determination*

Determine the sample size n required for this two–sided t-test to detect the true mean as high as 785 hours with power of test 0,9. Apply an appropriate OC curve.

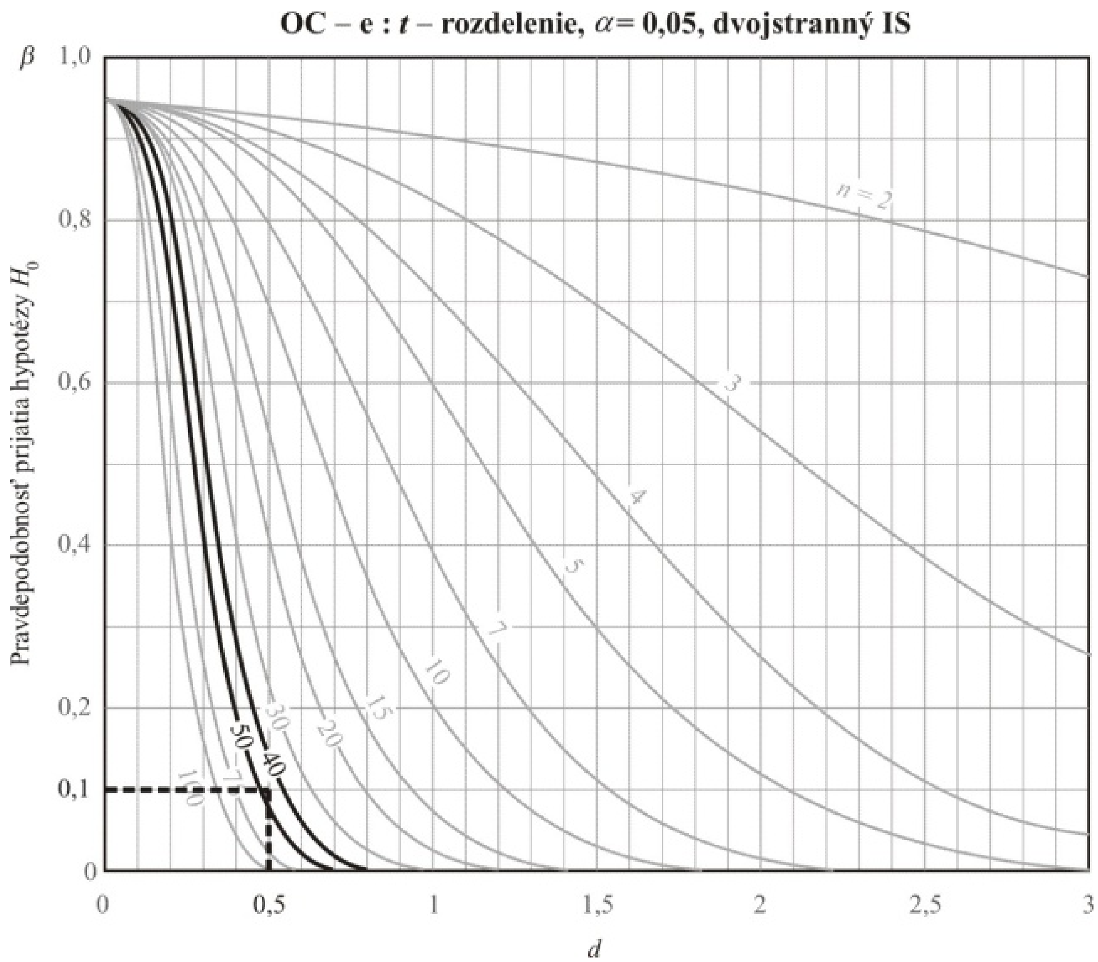
$$\text{Power of test } = P(\text{reject } H_0 | H_0 \text{ is false}) = 1 - \beta = 0{,}9 \Rightarrow \beta = 0{,}1$$

$$\delta = \mu - \mu_0 = 785 - 765 = 20$$

For two–sided t-test at $\alpha = 0{,}05$ and for a single sample, we calculate the value of the parameter d:

$$d = \frac{|\mu - \mu_0|}{\sigma} = \frac{|\delta|}{s} = \frac{|20|}{40{,}0164} = 0{,}499795$$

For $d = 0{,}5$ and $\beta = 0{,}1$, the OC–e curve displayed below (see also in Appendix) provides the required sample size $n = 45$.



OC−e curve for the two–sided normal test with different values of n and $\alpha = 0{,}05$.

7.4 Hypothesis tests on the variance of a normal population

Learning goals

- Test a hypothesis on σ^2 (χ^2-test).
- Determine the sample size of a χ^2-test for statistical inference on σ^2 by using an appropriate operating charakteristic (OC) curve.

Inference context

Parameter: σ^2

Point estimator of σ^2: $S^2 = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$

Test statistic of σ^2: $X^2 = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

Test procedure (χ^2-test):

Step 1: State the **null hypothesis** H_0 **and alternative hypothesis** H_1.

$H_0: \sigma^2 = \sigma_0^2$ $\quad H_1: \sigma^2 \neq \sigma_0^2$ for two–sided test

$\sigma^2 < \sigma_0^2$ for lower–sided test

$\sigma^2 > \sigma_0^2$ for upper–sided test

Step 2: Determine a **test statistic and its value.**

$X_0^2 = \dfrac{(n-1)S^2}{\sigma_0^2}$ $\quad \chi_0^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$

Step 3: Determine a **critical value(s) for *α*.**

$\chi^2(n-1;1-\alpha/2)$ and $\chi^2(n-1;\alpha/2)$ for two–sided test

$\chi^2(n-1;1-\alpha)$ for lower–sided test

$\chi^2(n-1;\alpha)$ for upper–sided test

Step 4: Make a **conclusion.** Reject H_0 if

$\chi_0^2 < \chi^2(n-1;1-\alpha/2)$ or $\chi_0^2 > \chi^2(n-1;\alpha/2)$ for two–sided test

$\chi_0^2 < \chi^2(n-1;1-\alpha)$ for lower–sided test

$\chi_0^2 > \chi^2(n-1;\alpha)$ for upper–sided test

Formulas for confidence intervals (CI)

For $100(1-\alpha)\%$ confidence interval on σ^2 the following formulas are applied:

$$\frac{(n-1)S^2}{\chi^2(n-1;\alpha/2)} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2(n-1;1-\alpha/2)} \quad \text{for two–sided CI}$$

$$\frac{(n-1)S^2}{\chi^2(n-1;\alpha)} \le \sigma^2 \quad \text{for lower–sided CI}$$

$$\sigma^2 \le \frac{(n-1)S^2}{\chi^2(n-1;1-\alpha)} \quad \text{for upper–sided CI}$$

Testing hypotheses using the confidence interval

Null hypothesis H_0: $\sigma^2 = \sigma_0^2$ is rejected on the level of significance α if

$$\sigma_0^2 \notin \left[\frac{(n-1)S^2}{\chi^2(n-1;\alpha/2)}, \frac{(n-1)S^2}{\chi^2(n-1;1-\alpha/2)}\right] \quad \text{for two–sided test}$$

$$\sigma_0^2 > \frac{(n-1)S^2}{\chi^2(n-1;1-\alpha)} \quad \text{for lower–sided test}$$

$$\sigma_0^2 < \frac{(n-1)S^2}{\chi^2(n-1;\alpha)} \quad \text{for upper–sided test}$$

Operating characteristic (OC) curve

Operating characteristic (OC) curves for a χ^2-test on σ^2 are provided in Appendix. For the two–sided alternative hypothesis $H_1 : \sigma^2 \ne \sigma_0^2$, the OC–i and OC–j plot β against an abscissa parameter

$$\lambda = \frac{\sigma}{\sigma_0}$$

for various sample sizes n, where σ denotes the true value of the standard deviation, and two levels of significance $\alpha = 0,01$ and $\alpha = 0,05$. The OC–k and OC–l curves are for upper–sided alternative H_1: $\sigma^2 > \sigma_0^2$, while the OC–m and OC–n are for lower–sided alternative H_1: $\sigma^2 < \sigma_0^2$.

Table 7.5 Operating characteristic charts for χ^2-test − single sample

Test		α	OC curve*	OC parameter
χ^2-test	Two–sided	0,05	OC–i	$\lambda = \dfrac{\sigma}{\sigma_0}$
		0,01	OC–j	
	Upper–sided	0,05	OC–k	
		0,01	OC–l	
	Lower–sided	0,05	OC–m	
		0,01	OC–n	

* See in Appendix.

Example 7.4

For the light bulb life length data in Table 6.1 and Example 6.3, the following results have been obtained:

$n = 30$, $s^2 = 40{,}0164^2$; 95 % two–sided CI on σ^2: $31{,}88^2 \le \sigma^2 \le 53{,}87^2$.

1. *Hypothesis test on σ^2; two–sided test*

Test H_0: $\sigma^2 = 40^2$ vs. H_1: $\sigma^2 \ne 40^2$ at $\alpha = 0{,}05$.

<u>Procedure:</u>

Step 1: State H_0 and H_1.

H_0: $\sigma^2 = 40^2$ $\quad H_1$: $\sigma^2 \ne 40^2$

Step 2: Determine a **test statistic and its value.**

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(30-1)\cdot 40{,}0164^2}{40^2} = 29{,}0238$$

Step 3: Determine a **critical value(s) for *α*.**

$$\chi^2(n-1;\alpha/2) = \chi^2(29;0{,}025) = 45{,}7$$

$$\chi^2(n-1;1/\alpha/2) = \chi^2(29;0{,}975) = 16{,}0$$

Step 4: Make a **conclusion.**

Since $\chi_0^2 = 29{,}0238 \in (16{,}0;45{,}7)$, fail to reject H_0 at the level of significance $\alpha = 0{,}05$.

2. *Relationship between CI and hypothesis test*

Test H_0: $\sigma^2 = 40^2$ vs. H_1: $\sigma^2 \ne 40^2$ at $\alpha = 0{,}05$ based on the 95 % two–sided CI on σ^2.

<u>Conclusion:</u> Since the 95 % two–sided CI on σ^2, $31{,}88^2 \le \sigma^2 \le 53{,}87^2$, includes the hypothesized value 40^2, fail to reject H_0 at $\alpha = 0{,}05$.

3. *Sample size determination*

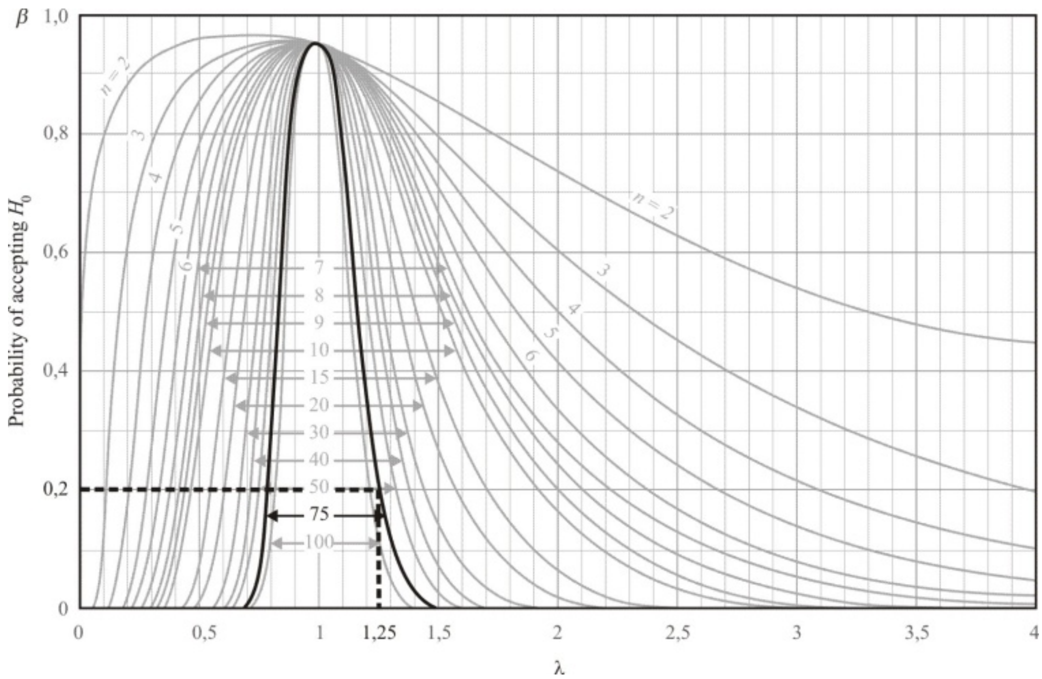
Determine the sample size n required for this two–sided χ^2-test to detect the true standard deviation as high as 50 hours with power of test 0,8. Apply an appropriate OC curve.

$$\text{Power of test} = P(\text{reject } H_0 | H_0 \text{ is false}) = 1 - \beta = 0,8 \Rightarrow \beta = 0,2$$

For two–sided χ^2-test at $\alpha = 0,05$ and for a single sample, we calculate the value of the parameter λ:

$$\lambda = \frac{\sigma}{\sigma_0} = \frac{50}{40} = 1,25$$

For $\lambda = 1,25$ and $\beta = 0,2$, the OC-i curve displayed below (see also in Appendix) provides the required sample size $n = 75$.



OC−i curve for the two–sided χ^2-test with different values of n and $\alpha = 0,05$.

7.5 Hypothesis tests on a population proportion

Learning goals

- Test a hypothesis on p (z-test) for a large sample.
- Determine the sample size for statistical inference on p by using an appropriate sample size formula.

Inference context

Parameter: p

Point estimator of p: $\hat{P} = \dfrac{X}{n}$, where $X \sim B(n,p)$

Test statistic of p: $Z = \dfrac{\hat{P} - p}{\sqrt{p(1-p)/n}} \sim N(0,1)$, $np(1-p) > 9$

Test procedure (z-test):

Step 1: State the **null hypothesis** H_0 **and alternative hypothesis** H_1.

H_0: $p = p_0$ H_1: $p \neq p_0$ for two–sided test

$p < p_0$ for lower–sided test

$p > p_0$ for upper–sided test

Step 2: Determine a **test statistic and its value.**

$$Z_0 = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \qquad z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Step 3: Determine a **critical value(s) for α.**

k_α for two–sided test

$k_{2\alpha}$ for one–sided test

Step 4: Make a **conclusion.** Reject H_0 if

$|z_0| > k_\alpha$ for two–sided test

$z_0 < -k_{2\alpha}$ for lower–sided test

$z_0 > k_{2\alpha}$ for upper–sided test

Formulas for confidence intervals (CI)

For $100(1-\alpha)\%$ confidence interval on p the following formulas are applied:

$$\hat{P} - k_\alpha\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \le p \le \hat{P} + k_\alpha\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \quad \text{for two–sided CI}$$

$$\hat{P} - k_{2\alpha}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \le p \quad \text{for lower–sided CI}$$

$$p \le \hat{P} + k_{2\alpha}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \qquad \text{for upper–sided CI}$$

Testing hypotheses using the confidence interval

Null hypothesis $H_0\colon p = p_0$ is rejected on the level of significance α if

$$p_0 \notin \left[\hat{P} - k_\alpha\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \;;\; \hat{P} + k_\alpha\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \right] \qquad \text{for two–sided test}$$

$$p_0 > \hat{P} + k_{2\alpha}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \qquad \text{for lower–sided test}$$

$$p_0 < \hat{P} - k_{2\alpha}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \qquad \text{for upper–sided test}$$

Sample size formula

For a hypothesis test on p, the following formulas are applied to determine the sample size:

$$n = \left(\frac{k_\alpha\sqrt{p_0(1-p_0)} + k_{2\beta}\sqrt{p(1-p)}}{p - p_0} \right)^2 \qquad \text{for two−sided test}$$

$$n = \left(\frac{k_{2\alpha}\sqrt{p_0(1-p_0)} + k_{2\beta}\sqrt{p(1-p)}}{p - p_0} \right)^2 \qquad \text{for one−sided test}$$

Example 7.5

For the corroded bridge data in Example 6.4, the following results have been obtained: $n = 40$ and $\hat{p} = \frac{x}{n} = \frac{28}{40} = 0{,}7$.

1. *Hypothesis test on* p *; two–sided test*

Test $H_0\colon p = 0{,}5$ vs. $H_1\colon p \ne 0{,}5$ at $\alpha = 0{,}05$.

Procedure:

Step 1: State H_0 and H_1.

$$H_0\colon p = 0{,}5 \qquad H_1\colon p \ne 0{,}5$$

Step 2: Determine a **test statistic and its value.**

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0,7-0,5}{\sqrt{\frac{0,5\times(1-0,5)}{40}}} = 2,5316$$

Step 3: Determine a **critical value(s) for *α*.**

$$k_\alpha = k_{0,05} = 1,96$$

Step 4: Make a **conclusion.**

Since $|z_0| = 2,5316 > k_{0,05} = 1,96$, reject H_0 at the level of significance $\alpha = 0,05$.

Sample size determination

Determine the sample size n required for this two–sided *z*-test to detect the true proportion p as high as 70 % with the power of test 0,9. Apply an appropriate sample size formula.

p = 70 % = 0,7

Power of test $= P(\text{reject}\, H_0 | H_0 \text{ is false}) = 1-\beta = 0,9 \Rightarrow \beta = 0,1 \Rightarrow 2\beta = 0,2$

$$n = \left(\frac{k_\alpha\sqrt{p_0(1-p_0)} + k_{2\beta}\sqrt{p(1-p)}}{p-p_0}\right)^2 = \left(\frac{k_{0,05}\sqrt{0,5\times(1-0,5)} + k_{0,2}\sqrt{0,7\times(1-0,7)}}{0,7-0,5}\right)^2 =$$

$$= \left(\frac{1,96\times 0,25 + 1,28\times 0,45826}{0,2}\right)^2 = 61,3535 \approx 62$$

7.6 Testing for goodness of fit

Statistical tests, which test a hypothesis about the type of distribution are called *goodness of fit tests*. This section lists three different tests.

7.6.1 Pearson χ^2-test

Learning goals

- Explain the term *categorical variable*.
- Distinguish between nominal and ordinal variables.
- Explain why the expected frequency of each class interval should be at least three in the goodness-of-fit test.
- Conduct a goodness-of-fit test on a hypothesized distribution.

Categorical variable

A categorical variable is used to represent a set of categories. Two types of categorical variables are defined depending on the significance of the order of the category listing.

1. **Nominal variable:** The order of listing of categories is not meaningful.

   E.g. gender (male and female) or hand dominance (left-handed, right-handed and ambidextrous).

2. **Ordinal variable:** The order of listing of categories is meaningful.

   E.g. education (less than 9 years, 9 – 12 years, more than 12 years), symptom severity (none, mild, moderate, severe).

Inference context

The underlying probability distribution of the population is unknown. Thus, we wish to test if a particular distribution fits the population.

E.g. $H_0 : X \sim P_0(\lambda)$ (Poisson – discrete distribution)

$H_0 : X \sim N(\mu, \sigma^2)$ (continuous distribution)

Test statistic

In these tests, the data from the random sample of size n are classified in the class intervals. As a measure of the difference between observed and expected frequencies of class is taken the test statistic

$$X^2 = \sum_{i=1}^{r} \frac{\left(n_i - np_i\right)^2}{np_i} \sim \chi^2(r - s - 1)$$

where r is number of class intervals,

n_i is observed frequency of i-th class interval,

np_i is expected frequency of i-th class interval, $p_i = P(t_{i-1} < X \le t_i \mid H_0 \text{ is true})$

s is number of parameters of the hypothesized distribution that are estimated by sample statistic.

Caution. *Minimum expected frequency*

One point to be noted in the application of this test procedure concerns the magnitude of the expected frequencies. If these expected frequencies are too small, the test statistic X_0^2 will not reflect the departure of observed from expected, but only the small magnitude of the expected frequencies. There is no general agreement regarding the minimum value of expected frequencies, but values of 3, 4 and 5 are widely used as minimal. To avoid this undesirable case, when an expected frequency is very small (say less then 3), the corresponding class interval should be combined with an adjacent class interval and the number of class intervals r is reduced by one.

Table 7.6 *Goodness of fit test table*

Class intervals i	Observed frequency n_i	Probability p_i	Expected frequency np_i	$n_i - np_i$	$\frac{(n_i - np_i)^2}{np_i}$
1					
2					
$\vdots$					
r					

A Pearson chi-square goodness of fit test (χ^2**-test**) is one of the most widely used tests, which allows you to test the type of continuous and discrete distributions.

Test procedure (χ^2-test)

Step 1: State H_0 and H_1.

H_0: X has a particular distribution vs. H_1: X has not a particular distribution

Step 2: Determine a **test statistic and its value.**

a) Estimate the parameter(s) of the hypothesized distribution if their values are not provided.

b) Define class intervals and summarize observed frequencies n_i accordingly.

c) Estimate the probabilities (p_i′s) of the class intervals.

d) Calculate the expected frequencies (np_i) of the class intervals. If an expected frequency of a class interval is too small (less than 3), combine it to an adjacent class interval. Then, repeat steps 2b) až 2d)

e) Calculate the value of test statistic

$$X_0^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}$$

Step 3: Determine a **critical value for *α*.**

$$\chi^2(r - s - 1, \alpha)$$

Step 4: Make a **conclusion.** Reject H_0 if

$$\chi_0^2 > \chi^2(r-s-1,\alpha)$$

Caution. Upper–sided critical region

Since the test statistic X_0^2 becomes smaller as the hypothesized distribution fits better, no lower limit is set as a critical value in the goodness of fit test.

Example 7.6 (Goodness-of-fit test; discrete distribution)

The number of e-mails per hour (*X*) coming to a certain firm´s e-mail account is assumed to follow a Poisson distribution. The following hourly e-mail arrival data are obtained during 100 hours:

No. e-mails/hour (X)	0	1	2	3
Frequency	60	28	7	5

Conduct a goodness of fit test at $\alpha = 0,05$ to confirm that the number of e-mails coming per hour is governed by Poisson distribution.

<u>Procedure</u>

Step 1: State H_0 and H_1.

$$H_0 : X \sim Po(\lambda) \qquad H_1 : X \not\sim Po(\lambda)$$

Step 2: Determine a **test statistic and its value.**

a) Estimate the <u>parameter</u> of the hypothesized distribution.

$$\hat{\lambda} = \widehat{E(X)} = \frac{0\times 60 + 1\times 28 + 2\times 7 + 3\times 5}{100} = 0,57$$

The number of parameters estimated is $s = 1$.

b) Define <u>class intervals</u> and summarize <u>observed frequencies</u> n_i accordingly.

c) Estimate the <u>probabilities</u> ($\hat{p}_i$) of the class intervals.

$$\hat{p}_1 = P(X=0) = \frac{e^{-0,57}(0,57)^0}{0!} = 0,57 \qquad \hat{p}_2 = P(X=1) = \frac{e^{-0,57}(0,57)^1}{1!} = 0,32$$

$$\hat{p}_3 = P(X=2) = \frac{e^{-0,57}(0,57)^2}{2!} = 0,09 \qquad \hat{p}_4 = P(X=3) = \frac{e^{-0,57}(0,57)^3}{3!} = 0,02$$

d) Calculate the <u>expected frequencies</u> ($n\hat{p}_i$) of the class intervals. If an expected frequency is too small (less than 3), adjust the class intervals.

e-mails intervals X	Observed frequency n_i	Probability $\hat{p}_i$	Expected frequency $n\hat{p}_i$	$n_i - n\hat{p}_i$	$\frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i}$
0	60	0,57	57		
1	28	0,32	32		
2	7	0,09	9		
3 or more	5	0,02	2		

Since the expected frequency of the last class interval in the above table is less than three, combine the last two cells as follows:

e-mails intervals X	Observed frequency n_i	Probability $\hat{p}_i$	Expected frequency $n\hat{p}_i$	$n_i - n\hat{p}_i$	$\frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i}$
0	60	0,57	57	3	0,15789
1	28	0,32	32	− 4	0,50000
2 or more	12	0,11	11	1	0,09000

e) Calculate the value of test statistic: $\chi_0^2 = \sum_{i=1}^{3} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i} = 0,7488$.

Step 3: Determine a **critical value for *α*.**

$$\chi^2(r-s-1,\alpha) = \chi^2(3-1-1;0,05) = \chi^2(1;0,05) = 3,84$$

Step 4: Make a **conclusion.**

Since $\chi_0^2 = 0,7488 < \chi^2(1,0,05) = 3,84$, fail to reject H_0 at $\alpha = 0,05$.

In other words, the number of e-mail arrivals per hour follows a Poisson distribution at the level of significance $\alpha = 0,05$.

Example 7.7 (Goodness-of-fit test; continuous distribution)

The final scores X of $n = 40$ students in a statistics class are summarized as follows:

Final scores (*X*)	$x < 60$	$60 \le x < 70$	$70 \le x < 80$	$80 \le x < 90$	$90 \le x$
Frequence	3	2	9	12	14

The mean and variance of the scores are 83 and $11,874^2$, respectively. Test if a normal distribution fits the test scores at $\alpha = 0,05$.

Procedure:

Step 1: State H_0 and H_1.

$$H_0: X \sim N(83;11,874^2) \qquad H_1: X \not\sim N(83;11,874^2)$$

Step 2: Determine a **test statistic and its value.**

a) Estimate the parameter of the hypothesized distribution.
   Since μ and σ^2 are known, skip this step and the number of parameters estimated is $s=0$.

b) Define class intervals and summarize observed frequencies n_i accordingly.

c) Estimate the probabilities ($\hat{p}_i$) of the class intervals.

$$\hat{p}_1 = P(X < 60) = P\left(Z \le \frac{60_i - 83}{11,874} \right) = P(Z < -1,937) = 1 - \Phi(1,937) = 0,0263725$$

$$\hat{p}_2 = P(60 \le X < 70) = P\left(\frac{60-83}{11,874} \le Z \le \frac{70-83}{11,874} \right) =$$
$$= P(-1,937 \le Z < -1,0948) = \Phi(-1,0948) - \Phi(-1,937) = 0,1104295$$

$$\hat{p}_3 = P(70 \le X < 80) = P(-1,0948 \le Z < -0,25265) = 0,263465$$

$$\hat{p}_4 = P(80 \le X < 90) = P(-0,25265 \le Z < 0,589523) = 0,321979$$

$$\hat{p}_5 = P(90 \le X) = P(0,589523 \le Z) = 0,277754$$

d) Calculate the expected frequencies ($n\hat{p}_i$) of the class intervals. If an expected frequency is too small (less than 3), adjust the class intervals.

X	Observed frequency n_i	Probability $\hat{p}_i$	Expected frequency $n\hat{p}_i$
$x < 60$	3	0,0263725	1
$60 \le x < 70$	2	0,1104295	4
$70 \le x < 80$	9	0,2634650	11
$80 \le x < 90$	12	0,321979	13
$90 \le x$	14	0,277754	11

Since the expected frequency of the first class interval in the previous table is less than three, combine the first two class intervals as follows:

X	Observed frequency n_i	Probability $\hat{p}_i$	Expected frequency $n\hat{p}_i$
$x < 70$	5	0,136802	5
$70 \le x < 80$	9	0,2634650	11
$80 \le x < 90$	12	0,321979	13
$90 \le x$	14	0,277754	11

e) Calculate the value of test statistic $\chi_0^2 = \sum_{i=1}^{4} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i} = 1,2586$

X	Observed frequency n_i	Expected frequency $n\hat{p}_i$	$n_i - n\hat{p}_i$	$\frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i}$
$x < 70$	5	5	0	0
$70 \le x < 80$	9	11	–2	0,3636
$80 \le x < 90$	12	13	–1	0,0769
$90 \le x$	14	11	3	0,8181

Step 3: Determine a **critical value for α.**

$$\chi^2(r-s-1,\alpha) = \chi^2(4-0-1;0,05) = \chi^2(3;0,05) = 7,81$$

Step 4: Make a **conclusion.**

Since $\chi_0^2 = 1,2586 < \chi^2(3;0,05) = 7,81$, fail to reject H_0 at the level of significance $\alpha = 0,05$.

7.6.2 Shapiro-Wilk normality test

The Shapiro-Wilk test can be used for a random sample of sizes $2 \le n \le 2000$ and for individual measured values (not for grouped data like in Pearson χ^2-test).

Let $x_1, x_2, ..., x_n$ are realizations of the random sample $X_1, X_2, ..., X_n$. When we arrange observations by size in ascending order we get $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$, what are realizations of an ordered random sample $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$.

Procedure for testing hypotheses

Step 1: State the **null hypothesis H_0 and alternative hypothesis H_1**.

$H_0 : X \sim N(\mu, \sigma^2)$ versus $H_1 : X \not\sim N(\mu, \sigma^2)$, where μ, σ^2 are unknown

Step 2: Determine a **test statistic.**

$$W = \frac{\left(\sum_{i=1}^{m} a_i(n)(X_{(n-i+1)} - X_{(i)})\right)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

where

- $a_i(n)$ are coefficients listed in the table (see Annex);
- $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$;
- $m = \begin{cases} \frac{n}{2} & \text{for } n \text{ even} \\ \frac{n\text{-}1}{2} & \text{for } n \text{ odd} \end{cases}$.

Step 3: Determine a **critical value for *α*.**

$W_\alpha(n)$ is a value listed in the table for given n and $\alpha = 0{,}01$ or $\alpha = 0{,}05$ (see Annex).

Step 4: Make a **conclusion.** Reject H_0 if

$$W \le W_\alpha(n)$$

Note. In the case of large sample size it is possible to determine the critical values $W_\alpha(n)$ by using statistical software such as using Statgraphics Centurion XV.

7.7 Contingency table tests

Learning goals

- Describe a contingency table.
- Conduct a contingency table test for independence/homogeneity of categorical variables.

Contingency table $r \times c$

Let us have a two-dimensional random vector $\boldsymbol{Z} = (X, Y)^{\mathrm{T}}$ of categorical variables. The X may take values $1, 2, \ldots, r$ and Y values $1, 2, \ldots, c$ ($r > 1$, $c > 1$). Denote:

$$p_{ij} = P(X = i, Y = j), \quad p_{i.} = P(X = i) = \sum_{j=1}^{c} p_{ij}, \quad p_{.j} = P(Y = j) = \sum_{i=1}^{r} p_{ij} .$$

Suppose that $p_{ij} > 0$ for all twosome (i, j).

Let the n elements of a sample from a population may be classified according to two different criteria. When denote n_{ij} the number of those cases in which $X = i$ and $Y = j$, the results can be written in the form of so-called contingency table:

Table 7.7 An $r \times c$ Contingency table

X	Y			
	1	2	$\cdots$	c
1	n_{11}	n_{12}	$\cdots$	n_{1c}
2	n_{21}	n_{21}	$\cdots$	n_{11}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
r	n_{r1}	n_{r2}	$\cdots$	n_{rc}

Inference context

We wish to test the association between two categorical variables X and Y by using an $r \times c$ contingency table for independence or homogeneity as follows:

- **Independence:** To examine if X and Y are independent, a representative sample is selected from a single population and then each element in the sample is classified into one of r categories in X and one of c categories in Y.
  E.g. classifying a sample of residents in Rohožník in terms of sex X and occupation Y.
- **Homogeneity:** To examine if r populations X_i, $1, 2, \ldots, r$ are homogeneous in terms of Y, representative samples are selected from the r populations and then the elements of each sample are classified into c categories in Y.
  E.g. classifying five samples of residents from different counties X in terms of occupation Y.

Recall that for two indepenent events A and B is valid:

$$P(A|B) = P(A),\ P(B|A) = P(B) \text{ and } P(A \cap B) = P(A)P(B).$$

Likewise, the relationship of two categorical variables is considered independent or homogeneous if

1. $P(X = x_i | Y) = P(X = x_i) = p_{i.}$,
2. $P(Y = y_j | X) = P(Y = y_j) = p_{.j}$,
3. $P(X = x_i,\ Y = y_j) = p_{ij} = P(X = x_i)P(Y = y_j) = p_{i.}p_{.j}$,

where: $P(X=x_i|Y)$ and $P(Y=y_j|X)$ are conditional probabilities,

$P(X=x_i)$ and $P(Y=y_j)$ are marginal probabilities, and

$P(X=x_i,\ Y=y_j)$ is the joint probability of X and Y.

Test statistic

$$X^2=\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{\left(n_{ij}-np_{ij}\right)^2}{np_{ij}}\sim\chi^2(\nu),\ \nu=(r-1)(c-1)$$

where

n_{ij} is observed frequency of cell ij,

$np_{ij}=n(p_{i.}\times p_{.j})=n\dfrac{n_{i.}\times n_{.j}}{n\times n}=\dfrac{n_{i.}\times n_{.j}}{n}$ is expected frequancy of cell ij.

Table 7.8 *Independence/Homogeneity Test table*

X	Y				Totals
	1	2	$\cdots$	c	
1	n_{11} np_{11}	n_{12} np_{12}	$\cdots$	n_{1c} np_{1c}	$n_{1.}$
2	n_{21} np_{21}	n_{21} np_{21}	$\cdots$	n_{2c} np_{2c}	$n_{2.}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
r	n_{r1} np_{r1}	n_{r2} np_{r2}	$\cdots$	n_{rc} np_{rc}	$n_{r.}$
Totals	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.c}$	n

Caution. *Minimum expected frequency*
Like the minimum expected frequency for the goodness of fit test (see Section 7.6.1), if an expected frequency is too small (say less then 3), X^2 can be improperly large for a small departure of the observed frequency from the expected one. Thus, any category whose expected frequency is small (less then 3) should be combined with an adjacent category.

Test procedure (χ^2-test)

Step 1: State the **null hypothesis** H_0 **and alternative hypothesis** H_1.

1. Testing for **idependence**

H_0: X and Y are independent

H_1: X and Y are not independent

2. Testing for **homogeneity.**

H_0: X_i $(1,2,\ldots,r)$ are homogenous in terms of Y

H_1: X_i $(1,2,\ldots,r)$ are not homogenous in terms of Y

Step 2: Determine a **test statistic and its value.**

$$X_0^2 = \sum_{j=1}^{c}\sum_{i=1}^{r} \frac{\left(n_{ij} - np_{ij}\right)^2}{np_{ij}} \sim \chi^2((r-1)(c-1)) \text{, where } np_{ij} = \frac{n_{i.}n_{.j}}{n}$$

Step 3: Determine a **critical value for *α*.**

$$\chi^2((r-1)(c-1),\alpha)$$

Step 4: Make a **conclusion.** Reject H_0 if

$$\chi_0^2 > \chi^2((r-1)(c-1),\alpha)$$

Caution. Upper-sided critical region

Since the test statistic X_0^2 becomes small as the null hypothesis of independence or homogeneity is true, no lower limit is set as a critical value in the independence or homogeneity test.

Example 7.8 (Contigency table test; Independence)

Grades in ergonomics X and grades in statistics Y of a hundred students are summarized as follows:

Ergonomics grade X	Statistics grade Y		
	A	B	Others
A	12	5	4
B	10	19	17
Others	4	8	21

Test if grades in ergonomics X and grades in statistics Y are independent at $\alpha = 0,05$.

Procedure:

Step 1: State H_0 and H_1.

H_0 Grades in ergonomics X and grades in statistics Y are independent

H_1 Grades in ergonomics X and grades in statistics Y are not independent

Step 2: Determine a **test statistic and its value.**

$$\chi_0^2 = \sum_{j=1}^{3}\sum_{i=1}^{3} \frac{\left(n_{ij} - np_{ij}\right)^2}{np_{ij}} =$$

$$= \frac{(12-5,46)^2}{5,46} + \frac{(5-6,72)^2}{6,72} + \frac{(4-8,82)^2}{8,82} + \frac{(10-11,96)^2}{11,96} + \ldots + \frac{(21-13,86)^2}{13,86} =$$

$$= 19,4958 \approx 19,5$$

Ergonomics grade X	Statistics grade Y			Totals
	A	B	Others	
A	12 / 5,46	5 / 6,72	4 / 8,82	21
B	10 / 11,96	19 / 14,72	17 / 19,32	46
Others	4 / 8,58	8 / 10,56	21 / 13,86	33
Totals	26	32	42	100

Step 3: Determine a **critical value for** $\boldsymbol{\alpha}$**.**

$$\chi^2((r-1)(c-1),\alpha) = \chi^2((3-1)(3-1),0,05) = \chi^2(4;0,05) = 9,49$$

Step 4: Make a **conclusion.**

Since $\chi_0^2 = 19,5 > \chi^2(4;0,05) = 9,49$, reject H_0 at $\alpha = 0,05$.

It is concluded that grades in ergonomics X and grades in statistics Y are not independent at $\alpha = 0,05$.

Example 7.9

A random sample of 300 adults with different hand sized X evaluates two mouse designs Y. The evaluation results are summarized as follows:

Hand size X	Mouse designs Y	
	Conventional	New
Small	35	65
Medium	20	80
Large	30	70

Test if users in different hand size groups X have homogeneous opinions on the mouse designs at $\alpha = 0,05$.

Procedure:

Step 1: State H_0 and H_1.

H_0: Users in different hand-size groups are homogeneous in terms of opinions on the mouse designs.

H_1: Users in different hand-size groups are not homogeneous in terms of opinions on the mouse designs.

Step 2: Determine a **test statistic and its value.**

Hand size X	Mouse designs Y		Totals
	Conventional	New	
Small	35 28,3	65 71,7	100
Medium	20 28,3	80 71,7	100
Large	30 28,3	70 71,7	100
Totals	85	215	300

$$\chi_0^2 = \sum_{j=1}^{2}\sum_{i=1}^{3}\frac{\left(n_{ij} - np_{ij}\right)^2}{np_{ij}} = 5,74981 \approx 5,75$$

Step 3: Detarmine a **critical value for *α***.

$$\chi^2((r-1)(c-1), \alpha) = \chi^2((3-1)(2-1); 0,05) = \chi^2(2; 0,05) = 5,99$$

Step 4: Make a **conclusion**.

Since $\chi_0^2 = 5,7 < \chi^2(2; 0,05) = 5,99$, fail to reject H_0 at $\alpha = 0,05$.

It is concluded that users in different hand size groups do not have significantly different opinions on the mouse designs at $\alpha = 0,05$.

8 STATISTICAL INFERENCE FOR TWO SAMPLES

8.1 Inference for a difference in means of two normal distributions, variances known

Learning goals

- Test a hypothesis on $\mu_1 - \mu_2$ when σ_1^2 and σ_2^2 are known (z-test).
- Determine the sample size of a z-test for statistical inference on $\mu_1 - \mu_2$ by using an appropriate sample size formula and operating characteristic (OC) curve.
- Establish a $100(1-\alpha)\%$ confidence interval (CI) on $\mu_1 - \mu_2$ when σ_1^2 and σ_2^2 are known.
- Determine the sample size of a z-test to satisfy a preselected level of error E in estimating $\mu_1 - \mu_2$.

Inference context

- **Parameter** of interest: $\mu_1 - \mu_2$
- **Point estimator** of $\mu_1 - \mu_2$: $\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$, where

  $\bar{X}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$ and $\bar{X}_2 \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$;

  σ_1^2 and σ_2^2 are <u>known</u>;

  $\bar{X}_1$ and $\bar{X}_2$ are <u>independent</u>.
- **Test statistic** of $\mu_1 - \mu_2$: $Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$

Sampling distribution of $\bar{X}_1 - \bar{X}_2$

The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

where $\bar{X}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$, $X_{11}, X_{12}, \ldots, X_{1n_1} \sim i.i.d.\ N(\mu_1, \sigma_1^2)$;

$\bar{X}_2 \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$, $X_{21}, X_{22}, \ldots, X_{2n_2} \sim i.i.d.\ N(\mu_2, \sigma_2^2)$;

$\bar{X}_1$ and $\bar{X}_2$ are independent.

Derivation of relationship $\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$

Since $\bar{X}_1$ and $\bar{X}_2$ are independent and normal with means and variances $E(\bar{X}_1)$, $E(\bar{X}_2)$, $D(\bar{X}_1) = \sigma_1^2 / n_1$ and $D(\bar{X}_2) = \sigma_2^2 / n_2$, respectively, $\bar{X}_1 - \bar{X}_2$ is normal with mean and variance

$$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

$$D(\bar{X}_1 - \bar{X}_2) = D(\bar{X}_1) + D(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Test procedure (*z*-test):

Step 1: State the **null hypothesis** H_0 **and alternative hypothesis** H_1.

H_0: $\mu_1 - \mu_2 = \delta_0$ $\quad$ H_1: $\mu_1 - \mu_2 \neq \delta_0$ for two–sided test

$\mu_1 - \mu_2 < \delta_0$ for lower–sided test

$\mu_1 - \mu_2 > \delta_0$ for upper–sided test

Step 2: Determine a **test statistic and its value.**

$$Z_0 = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \qquad z_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Step 3: Determine a **critical value(s) for** α

k_α for two–sided test

$k_{2\alpha}$ for one–sided test

Step 4: Make a **conclusion**. Reject H_0 if

$|z_0| > k_\alpha$ for two–sided test

$z_0 < -k_{2\alpha}$ for lower–sided test

$z_0 > k_{2\alpha}$ for upper–sided test

Sample size formula

It is possible to obtain formulas for calculating the sample sizes directly. Suppose that the null hypothesis $H_0 : \mu_1 - \mu_2 = \delta_0$ is false and that the true difference in means is $\mu_1 - \mu_2 = \delta$, where $\delta > \delta_0$. One may find formulas for the sample size required to obtain a specific value of the type II error probability β for a given difference in means δ and level of significance α .

For the two-sided alternative hypothesis with significance level α, the sample size $n_1 = n_2 = n$ required to detect a true difference in means of δ with power of the test at least $1-\beta$ is

$$n = \frac{(k_\alpha + k_{2\beta})^2(\sigma_1^2 + \sigma_2^2)}{(\delta - \delta_0)^2}$$

For the one-sided alternative hypothesis with significance level α, the sample size $n_1 = n_2 = n$ required to detect a true difference in means of $\delta(\neq \delta_0)$ with power of the test at least $1-\beta$ is

$$n = \frac{(k_{2\alpha} + k_{2\beta})^2(\sigma_1^2 + \sigma_2^2)}{(\delta - \delta_0)^2}$$

Operating characteristic (OC) curve

Operating characteristic (OC) curves for a *z*-test on $\mu_1 - \mu_2$ are provided in Appendix. The OC curves plot β against *d* for various sample sizes $n \; (= n_1 = n_2)$ and two levels of significance $\alpha = 0{,}01$ and $\alpha = 0{,}05$, i.e.

$$\beta = f(n, d, \alpha)$$

Table 8.1 Operating characteristic charts for z-test – two samples

Test		α	OC curve*	OC parameter
z-test	Two−sided	0,05	OC–a	$d = \frac{\lvert\delta - \delta_0\rvert}{\sqrt{\sigma_1^2 + \sigma_2^2}}$
		0,01	OC–b	
	One−sided	0,05	OC–c	
		0,01	OC–d	

*See in Appendix.

Confidence interval formula

A $100(1-\alpha)\%$ CI on $\mu_1-\mu_2$, when σ_1^2 and σ_2^2 are known, is

$$\left(\bar{X}_1-\bar{X}_2\right)-k_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_1^2}{n_1}}\le\mu_1-\mu_2\le\left(\bar{X}_1-\bar{X}_2\right)+k_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_1^2}{n_1}}\quad\text{for two–sided CI}$$

$$\left(\bar{X}_1-\bar{X}_2\right)-k_{2\alpha}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_1^2}{n_1}}\le\mu_1-\mu_2\quad\text{for lower–sided CI}$$

$$\mu_1-\mu_2\le\left(\bar{X}_1-\bar{X}_2\right)+k_{2\alpha}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_1^2}{n_1}}\quad\text{for upper–sided CI}$$

<u>Derivation of formula for two-sided CI</u>

By using the test statistic $Z=\dfrac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}\sim N(0,1)$, we get

$$P(-k_\alpha\le Z\le k_\alpha)=1-\alpha$$

$$P\left(-k_\alpha\le\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}\le k_\alpha\right)=1-\alpha$$

$$P\left(\left(\bar{X}_1-\bar{X}_2\right)-k_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\le\mu_1-\mu_2\le\left(\bar{X}_1-\bar{X}_2\right)+k_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)=1-\alpha$$

Therefore,

$$L=\left(\bar{X}_1-\bar{X}_2\right)-k_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\quad\text{and}\quad U=\left(\bar{X}_1-\bar{X}_2\right)+k_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

Testing hypotheses using the confidence interval

Null hypothesis $H_0\!:\ \mu_1-\mu_2=\delta_0$ is rejected on the level of significance α if

$$\delta_0\notin\left[\left(\bar{X}_1-\bar{X}_2\right)-k_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\left(\bar{X}_1-\bar{X}_2\right)+k_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]\quad\text{for two–sided test}$$

$$\delta_0>\left(\bar{X}_1-\bar{X}_2\right)+k_{2\alpha}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\quad\text{for lower–sided test}$$

$$\delta_0 < \left(\overline{X}_1 - \overline{X}_2\right) - k_{2\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad \text{for upper–sided test}$$

Sample size formula for predefined error

When determining the $100(1-\alpha)\%$ IS for $\mu_1 - \mu_2$ which shall not exceed a predefined error E, the sample size is determined by the formula

$$n = \left(\frac{k_\alpha}{E}\right)^2 \times (\sigma_1^2 + \sigma_2^2), \text{ where } n_1 = n_2 = n$$

Example 8.1

The life lengths of INFINITY (X_1; unit: hour) and FOREVER (X_2; unit: hour) light bulbs are under study. Suppose that X_1 and X_2 are normally distributed with $\sigma_1^2 = 40^2$ and $\sigma_2^2 = 30^2$, respectively. The random samples of INFINITY and FOREVER light bulbs are shown below:

i	Life lengths		i	Life lengths		i	Life lengths	
	X_1	X_2		X_1	X_2		X_1	X_2
1	727	789	11	831	755	21	725	837
2	755	835	12	742	813	22	735	798
3	714	765	13	784	828	23	770	837
4	840	796	14	807	771	24	792	841
5	772	797	15	820	829	25	765	766
6	750	776	16	812	756	26	749	
7	814	769	17	804	787	27	829	
8	820	836	18	754	788	28	821	
9	753	847	19	715	794	29	816	
10	796	769	20	845	822	30	743	

The two random samples are summarized as follows:

Brand of light bulb	Sample size	Value of sample mean	Variance
INFINITY (X_1)	$n_1 = 30$	$\overline{x}_1 = 780$ hrs	$\sigma_1^2 = 40^2$
FOREVER (X_2)	$n_2 = 25$	$\overline{x}_2 = 800,04$ hrs	$\sigma_2^2 = 30^2$

1. *Hypothesis Test on* $\mu_1 - \mu_2$, σ_1^2 *and* σ_2^2 *are known; two-sided test*

Test if the mean life length of an INFINITY light bulb is different from that of a FOREVER light bulb at $\alpha = 0{,}05$.

Step 1: State H_0 and H_1

$$H_0: \mu_1 - \mu_2 = 0 \qquad H_1: \mu_1 - \mu_2 \neq 0$$

Step 2: Determine a **test statistic and its value**.

$$z_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(780 - 800{,}04) - 0}{\sqrt{\frac{40^2}{30} + \frac{30^2}{25}}} = -2{,}12$$

Step 3: Determine a **critical value(s) for** α.

$$k_\alpha = k_{0,05} = 1{,}96$$

Step 4: Make a **conclusion**.

Since $|z_0| = 2{,}12 > k_{0.05} = 1{,}96$, reject H_0 at $\alpha = 0{,}05$.

2. *Sample size determination for predefined power of test*

Determine the sample size $n \ (= n_1 = n_2)$ required for this two-sided *z*-test to detect the true difference in mean life length as high as 20 hours with 0,8 of power. Apply an appropriate sample size formula and OC curve.

a) Sample size formula

True difference: $\delta = \mu_1 - \mu_2 = 20$

Hypothetical difference: $\delta_0 = \mu_1 - \mu_2 = 0$

Power of test $= P(\text{reject}\, H_0 \,|\, H_0 \text{ is false}) = 1 - \beta = 0{,}8 \Rightarrow \beta = 0{,}2 \Rightarrow 2\beta = 0{,}4$

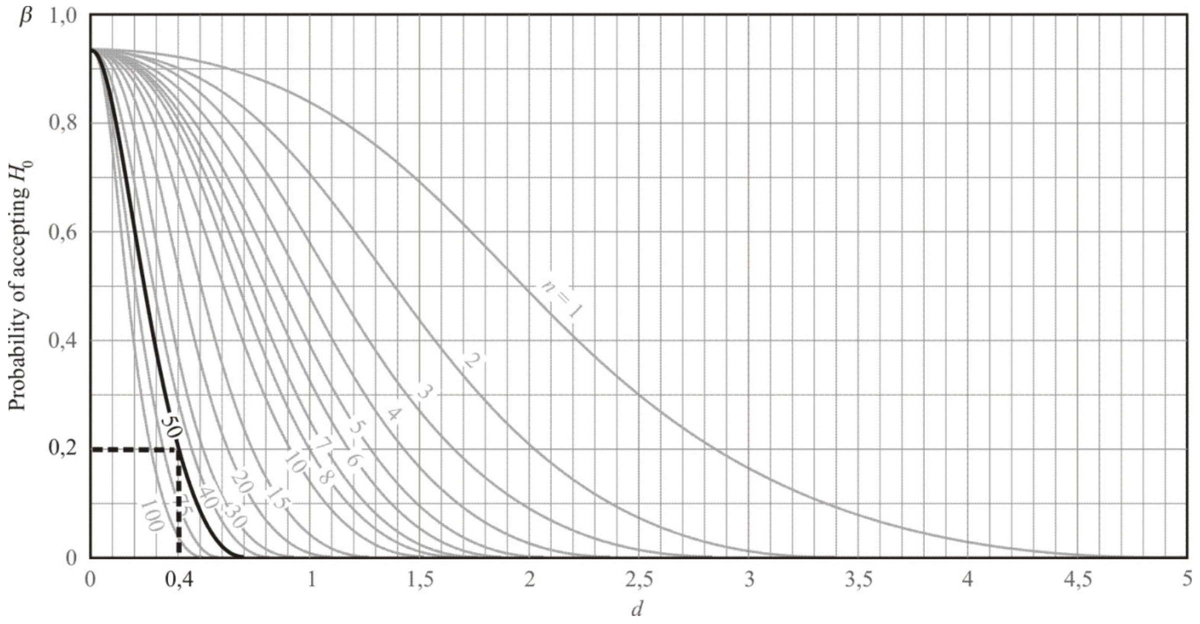
$$n = \frac{(k_\alpha + k_{2\beta})^2(\sigma_1^2 + \sigma_2^2)}{(\delta - \delta_0)^2} = \frac{(k_{0,05} + k_{0,4})^2(40^2 + 30^2)}{(20-0)^2} == \frac{(1{,}96 + 0{,}84)^2 \times (40^2 + 30^2)}{(20-0)^2} \approx 50$$

b) OC curve

For two–sided *z*-test at $\alpha = 0{,}05$ and for two samples, we calculate the value of the parameter d:

$$d = \frac{|\delta - \delta_0|}{\sqrt{\sigma_1^2 + \sigma_2^2}} = \frac{|20 - 0|}{\sqrt{40^2 + 30^2}} = 0{,}4$$

For $d = 0{,}4$ and $\beta = 0{,}2$, the OC–a curve displayed below (see also in Appendix) provides the required sample size $n = 50$ which is the same value as calculated by using the sample size formula.



OC-a curves for the two–sided normal test with different values of *n* and $\alpha = 0,05$.

3. *Two-sided confidence interval*

Construct a 95 % two-sided confidence interval on the mean difference in life length. Based on this 95 % two-sided CI on $\mu_1 - \mu_2$, test H_0: $\mu_1 - \mu_2 = 0$ vs. H_1: $\mu_1 - \mu_2 \neq 0$ at $\alpha = 0,05$.

$$P(l \leq \mu_1 - \mu_2 \leq u) = 0,95 = 1 - \alpha \quad \Rightarrow \quad \alpha = 0,05$$

95 % two-sided CI on $\mu_1 - \mu_2$:

$$\left(\bar{x}_1 - \bar{x}_2\right) - k_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \left(\bar{x}_1 - \bar{x}_2\right) + k_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\left(780 - 800,04\right) - k_{0,05} \sqrt{\frac{40^2}{30} + \frac{30^2}{25}} \leq \mu_1 - \mu_2 \leq \left(780 - 800,04\right) + k_{0,05} \sqrt{\frac{40^2}{30} + \frac{30^2}{25}}$$

$$-20,04 - 1,96 \times \sqrt{\frac{40^2}{30} + \frac{30^2}{25}} \leq \mu_1 - \mu_2 \leq -20,04 + 1,96 \times \sqrt{\frac{40^2}{30} + \frac{30^2}{25}}$$

$$-38,56 \leq \mu_1 - \mu_2 \leq -1,51$$

Since this 95 % two-sided CI on $\mu_1 - \mu_2$ does not include the hypothesized value of $\delta_0 = 0$, reject H_0 at $\alpha = 0,05$.

4. *Sample size determination for predefined error*

Find the sample size $n \left(= n_1 = n_2\right)$ to construct a two-sided confidence interval on $\mu_1 - \mu_2$ within 20 hours of error at $\alpha = 0,05$.

$\sigma_1^2 = 40^2$, $\sigma_2^2 = 30^2$, $1-\alpha = 0,95 \Rightarrow \alpha = 0,05$, $E = 20$

$$n = \left(\frac{k_\alpha}{E}\right)^2 \times (\sigma_1^2 + \sigma_2^2) = \left(\frac{k_{0,05}}{20}\right)^2 \times (40^2 + 30^2) = \left(\frac{1,96}{20}\right)^2 \times 2500 = 24,01 \approx 25$$

8.2 Inference for a difference in means of two normal distributions, variances unknown

Learning goals

- Test a hypothesis on $\mu_1 - \mu_2$ when σ_1^2 and σ_2^2 are unknown (*t*-test).
- Determine the sample size of a *t*-testu for statistical inference on $\mu_1 - \mu_2$ by using an appropriate operating characteristic (OC) curve.
- Establish a $100(1-\alpha)\%$ confidence interval (CI) on $\mu_1 - \mu_2$ when σ_1^2 and σ_2^2 are unknown.

Inference context

- **Parameter** of interest: $\mu_1 - \mu_2$
- **Point estimator** of $\mu_1 - \mu_2$: $\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$, where

  $\bar{X}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$, $\bar{X}_2 \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$;

  σ_1^2 and σ_2^2 are <u>unknown</u>;

  $\bar{X}_1$ and $\bar{X}_2$ are <u>independent</u>.
- **Test statistic** of $\mu_1 - \mu_2$: Different test statistics of $\mu_1 - \mu_2$ are used depending on the equality of σ_1^2 and σ_2^2 as follows:

  <u>Case 1: Equal variances</u> ($\sigma_1^2 = \sigma_2^2 = \sigma^2$)

  $$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(\nu),\ \nu = n_1 + n_2 - 2$$

  where $S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ (pooled estimator of σ^2)

Case 2: Unequal variances ($\sigma_1^2 \neq \sigma_2^2$)

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t(\nu), \quad \nu = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1+1} + \frac{(S_2^2/n_2)^2}{n_2+1}} - 2$$

Note. The equality of two variances ($\sigma_1^2 = \sigma_2^2$ or $\frac{\sigma_1^2}{\sigma_2^2} = 1$) can be checked by using an *F*-test.

Test procedure (*t*-test):

Step 1: State the **null hypothesis** H_0 **and alternative hypothesis** H_1.

H_0: $\mu_1 - \mu_2 = \delta_0$ $\quad$ H_1: $\mu_1 - \mu_2 \neq \delta_0$ for two–sided test

$\mu_1 - \mu_2 < \delta_0$ for lower–sided test

$\mu_1 - \mu_2 > \delta_0$ for upper–sided test

Step 2: Determine a **test statistic and its value**.

Case 1: Equal variances ($\sigma_1^2 = \sigma_2^2 = \sigma^2$)

$$T_0 = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(\nu), \; \nu = n_1 + n_2 - 2; \quad t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(\nu),$$

where $S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ (estimator of σ^2) and

$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ (estimate of σ^2)

Case 2: Unequal variances ($\sigma_1^2 \neq \sigma_2^2$)

$$T_0 = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t(\nu), \; \nu = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1+1} + \frac{(S_2^2/n_2)^2}{n_2+1}} - 2; \; t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(\nu)$$

Step 3: Determine a **critical value(s) for** α **.**

$t(\nu;\alpha)$ for two–sided test

$t(\nu;2\alpha)$ for one–sided test

Step 4: Make a **conclusion**. Reject H_0 if

$|t_0| > t(\nu;\alpha)$ for two–sided test

$t_0 < -t(\nu;2\alpha)$ for lower–sided test

$t_0 > t(\nu;2\alpha)$ for upper–sided test

Operating characteristic (OC) curve

Table 8.2 displays a list of OC charts and a formula of the OC parameter d for a *t*-test on $\mu_1 - \mu_2$ where $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and $n_1 = n_2 = n$. Note that OC curves are unavailable for a *t*-test when $\sigma_1^2 \neq \sigma_2^2$ because the corresponding *t*-distribution is unknown. In this case, we proceed as follows. By using the Table 8.2, the appropriate OC chart for a particular *t*-test is chosen. The sample size n^* obtained from an OC curve is used to determine the size $n(= n_1 = n_2)$ as follows:

$$n = \frac{n^* + 1}{2}, \text{ where } n^* \text{ is from an OC curve}$$

Table 8.2 Operating characteristic charts for t-test – two samples

Test		α	OC*	OC parameter
t-test	Two−sided	0,05 0,01	OC–e OC–f	$d = \frac{\lvert\delta - \delta_0\rvert}{2\hat{\sigma}}$°
	One−sided	0,05 0,01	OC–g OC–h	

°For $\hat{\sigma}$, use s_p (pooled estimate of common standard deviation) or subjective estimate.

* See in Appendix.

Confidence interval formula

A $100(1-\alpha)\%$ CI on $\mu_1 - \mu_2$ when σ_1^2 and σ_2^2 are unknown depends on the equality of σ_1^2 and σ_2^2 as follows:

Case 1: Equal variances ($\sigma_1^2 = \sigma_2^2 = \sigma^2$)

$$\left(\bar{X}_1 - \bar{X}_2\right) - t(\nu;\alpha)S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le \left(\bar{X}_1 - \bar{X}_2\right) + t(\nu;\alpha)S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{for two-sided CI}$$

$$\left(\bar{X}_1 - \bar{X}_2\right) - t(\nu;2\alpha)S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \quad \text{for lower–sided CI}$$

$$\mu_1 - \mu_2 \le \left(\bar{X}_1 - \bar{X}_2\right) + t(\nu;2\alpha)S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{for upper–sided CI}$$

Case 2: Unequal variances ($\sigma_1^2 \neq \sigma_2^2$)

$$\left(\bar{X}_1 - \bar{X}_2\right) - t(\nu;\alpha)\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \left(\bar{X}_1 - \bar{X}_2\right) + t(\nu;\alpha)\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \quad \text{for two-sided CI}$$

$$\left(\bar{X}_1 - \bar{X}_2\right) - t(\nu;2\alpha)\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq \mu_1 - \mu_2 \quad \text{for lower–sided CI}$$

$$\mu_1 - \mu_2 \leq \left(\bar{X}_1 - \bar{X}_2\right) + t(\nu;\alpha)\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \quad \text{for upper–sided CI}$$

Testing hypotheses using the confidence interval

Null hypothesis $H_0\!: \mu_1 - \mu_2 = \delta_0$ is rejected on the level of significance α if

Case 1: Equal variances ($\sigma_1^2 = \sigma_2^2 = \sigma^2$)

$$\delta_0 \notin \left[\left(\bar{X}_1 - \bar{X}_2\right) - t(\nu;\alpha)S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ \left(\bar{X}_1 - \bar{X}_2\right) + t(\nu;\alpha)S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right] \quad \text{for two-sided test}$$

$$\delta_0 > \left(\bar{X}_1 - \bar{X}_2\right) + t(\nu;2\alpha)S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{for lower–sided test}$$

$$\delta_0 < \left(\bar{X}_1 - \bar{X}_2\right) - t(\nu;2\alpha)S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{for upper–sided test}$$

Case 2: Unequal variances ($\sigma_1^2 \neq \sigma_2^2$)

$$\delta_0 \notin \left[\left(\bar{X}_1 - \bar{X}_2\right) - t(\nu;\alpha)\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},\ \left(\bar{X}_1 - \bar{X}_2\right) + t(\nu;\alpha)\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right] \quad \text{for two–sided test}$$

$$\delta_0 > \left(\bar{X}_1 - \bar{X}_2\right) + t(\nu;2\alpha)\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \quad \text{for lower–sided test}$$

$$\delta_0 < \left(\bar{X}_1 - \bar{X}_2\right) - t(\nu;2\alpha)\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \quad \text{for upper–sided test}$$

CI and hypothesis test for a large sample

If the sample sizes are large ($n_1 \geq 30$ and $n_2 \geq 30$), the *z*-based CI formulas and test procedure in Section 8.1 can be applied to inference on $\mu_1 - \mu_2$ regardless of whether the underlying populations are normal or non-normal according to the central limit theorem (described in Section 5.3.

Example 8.2

For the light bulb life length data in Example 8.1, the following results have been obtained:

Brand of light bulb	Sample size	Value of sample mean	Variance
INFINITY (X_1)	$n_1 = 30$	$\bar{x}_1 = 780$ hrs	$s_1^2 = 40{,}0164^2$
FOREVER (X_2)	$n_2 = 25$	$\bar{x}_2 = 800{,}04$ hrs	$s_2^2 = 30{,}0048^2$

Case 1: Equal variances **(** $\sigma_1^2 = \sigma_2^2 = \sigma^2$ **)**

1. *Hypothesis Test on* $\mu_1 - \mu_2$ *,* σ_1^2 *and* σ_2^2 *are unknown and equal; two-sided test*

Assuming $\sigma_1^2 = \sigma_2^2$, test if the mean life length of an INFINITY light bulb is different from that of a FOREVER light bulb at $\alpha = 0{,}05$.

Step 1: State H_0 and H_1

$H_0\colon \mu_1 - \mu_2 = 0 \qquad H_1\colon \mu_1 - \mu_2 \neq 0$

Step 2: Determine a **test statistic and its value**.

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(780 - 800{,}04) - 0}{35{,}83\sqrt{\frac{1}{30} + \frac{1}{25}}} = -2{,}065$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} - 2 = \frac{(30-1)\times 40{,}0164^2 + (25-1)\times 30{,}0048^2}{30 + 25 - 2} = 1283{,}87$$

$$\Rightarrow s_p = 35{,}83$$

Step 3: Determine a **critical value(s) for** α .

$t(\nu, \alpha) = t(53, 0, 05) = 2{,}006 \quad , \quad \nu = n_1 + n_2 - 2 = 30 + 25 - 2 = 53$

Step 4: Make a **conclusion**.

Since $|t_0| = 2{,}065 > t(53; 0{,}05) = 2{,}006$, reject H_0 at $\alpha = 0{,}05$.

2. *Sample size determination (* $\sigma_1^2 = \sigma_2^2$ *)*

Assuming $\sigma_1^2 = \sigma_2^2$, determine the sample size n $(= n_1 = n_2)$ required for this two-sided *t*-test to detect the true difference in mean life length as high as 20 hours with 0,8 of power. Apply an appropriate OC curve.

True difference: $\delta = \mu_1 - \mu_2 = 20$

Hypothetical difference: $\delta_0 = \mu_1 - \mu_2 = 0$

For a two–sided t-test at $\alpha = 0,05$ and for two samples, we calculate the value of the parameter d:

$$d = \frac{|\delta - \delta_0|}{2\hat{\sigma}} = \frac{|\delta - \delta_0|}{2s_P} = \frac{|20-0|}{2\times 35,83} \approx 0,28$$

where

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} =$$

$$= \frac{(30-1)\times 40,0164^2 + (25-1)\times 30,0048^2}{30+25-2} = 1283,0 = 35,83^2$$

For $d = 0,28$ and $\beta = 0,2$, the OC–e curve (see in Appendix) provides the required sample size $n = 100$.

3. *Confidence interval on* $\mu_1 - \mu_2$, σ_1^2 *and* σ_2^2 *unknown but equal; two-sided CI*

Assuming $\sigma_1^2 = \sigma_2^2$, construct a 95 % two-sided confidence interval on the difference in mean life length $\mu_1 - \mu_2$.

$$P(l \le \mu_1 - \mu_2 \le u) = 0,95 = 1-\alpha \quad \Rightarrow \quad \alpha = 0,05$$

$$\nu = n_1 + n_2 - 2 = 30 + 25 - 2 = 53; \quad s_P^2 = 35,83^2$$

95% two-sided CI on $\mu_1 - \mu_2$:

$$\left(\overline{x}_1 - \overline{x}_2\right) - t(\nu;\alpha)s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le \mu_1 - \mu_2 \le \left(\overline{x}_1 - \overline{x}_2\right) + t(\nu;\alpha)s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$\left(780-800,04\right) - t(53;0,05)\times 35,83\sqrt{\frac{1}{30}+\frac{1}{25}} \le \mu_1 - \mu_2 \le \left(780-800,04\right) + t(53;0,05)\times 35,83\sqrt{\frac{1}{30}+\frac{1}{25}}$$

$$-20,04 - 2,065\times 9,703 \le \mu_1 - \mu_2 \le -20,04 + 2,065\times 9,703$$

$$-39,502 \le \mu_1 - \mu_2 \le -0,577995$$

$$-39,5 \le \mu_1 - \mu_2 \le -0,600$$

Note that this t-based CI ($-39,5 \le \mu_1 - \mu_2 \le -0,600$) is <u>wider</u> than the corresponding z-based CI ($-38,56 \le \mu_1 - \mu_2 \le -1,51$) in Example 8.1.

<u>Case 2: Unequal variances</u> ($\sigma_1^2 \ne \sigma_2^2$)

1. *Hypothesis Test on* $\mu_1 - \mu_2$, σ_1^2 *and* σ_2^2 *are unknown and unequal; two-sided test*

Assuming $\sigma_1^2 \neq \sigma_2^2$, test if the mean life length of an INFINITY light bulb is different from that of a FOREVER light bulb at $\alpha = 0,05$.

Step 1: State H_0 and H_1

$H_0: \mu_1 - \mu_2 = 0 \qquad H_1: \mu_1 - \mu_2 \neq 0$

Step 2: Determine a **test statistic and its value**.

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(780 - 800,04) - 0}{\sqrt{\frac{40,0164^2}{30} + \frac{30,0084^2}{25}}} = -2,12$$

$$\nu = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1+1} + \frac{(S_2^2/n_2)^2}{n_2+1}} - 2 = \frac{(40,0164^2/30 + 30,0048^2/25)^2}{\frac{(40,0164^2/30)^2}{30+1} + \frac{(30,0048^2/25)^2}{25+1}} - 2 = 56,3552 \approx 57$$

Step 3: Determine a **critical value(s) for** α .

$t(\nu;\alpha) = t(57;0,05) = 2,00$

Step 4: Make a **conclusion**.

Since $|t_0| = 2,12 > t(57;0,05) = 2,00$, reject H_0 at $\alpha = 0,05$.

2. *Sample size determination (* $\sigma_1^2 \neq \sigma_2^2$ *)*

Assuming $\sigma_1^2 = \sigma_2^2$, determine the sample size $n (= n_1 = n_2)$ required for this two-sided *t*-test to detect the true difference in mean life length as high as 20 hours with 0,8 of power. Apply an appropriate OC curve.

True difference: $\delta = \mu_1 - \mu_2 = 20$

Hypothetical difference: $\delta_0 = \mu_1 - \mu_2 = 0$

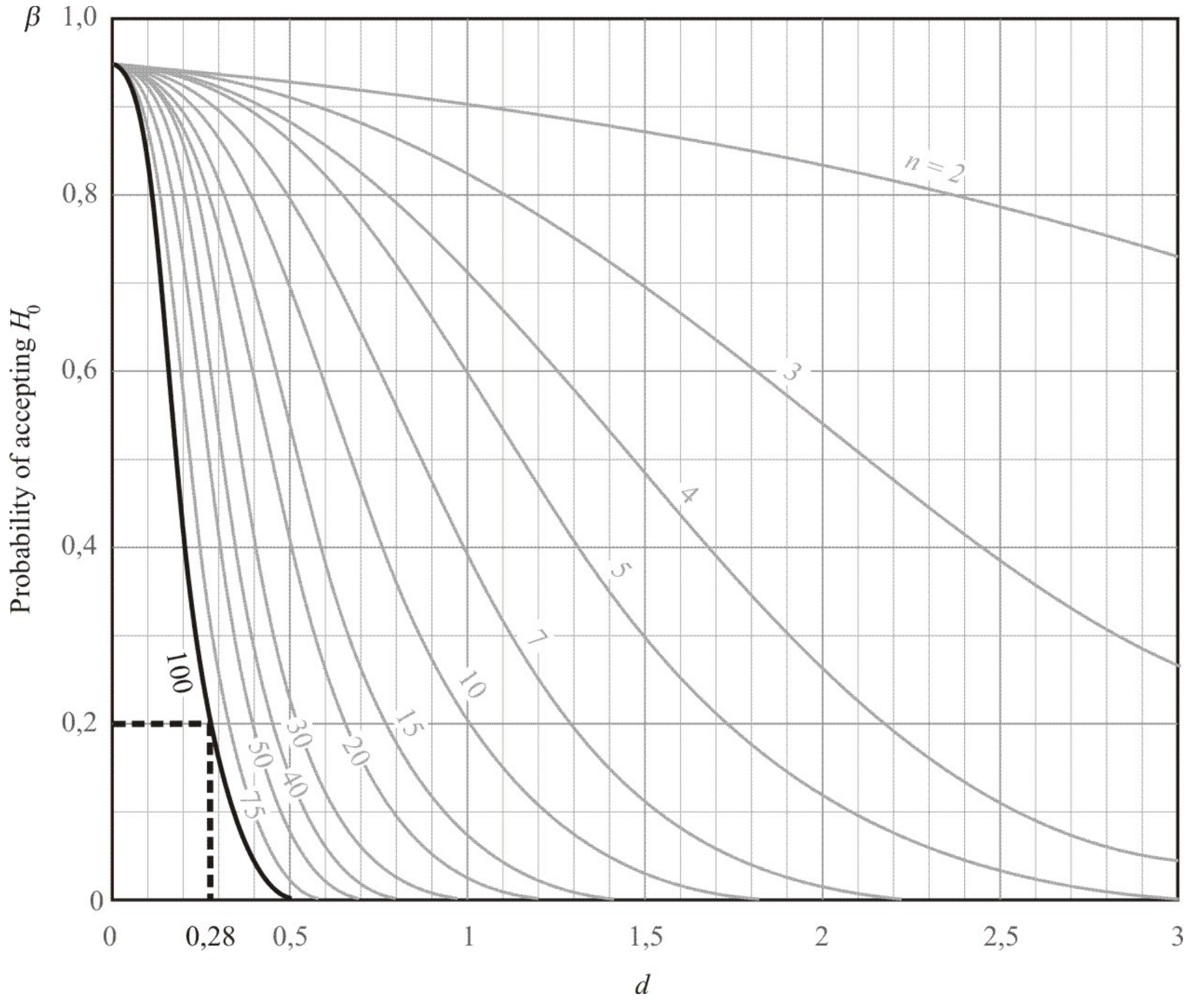
For two–sided *t*-test at $\alpha = 0,05$ and for two samples, we calculate the value of the parameter d :

$$d = \frac{|\delta - \delta_0|}{2\hat{\sigma}} = \frac{|\delta - \delta_0|}{2s_P} = \frac{|20-0|}{2 \times 35,8} \approx 0,28$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(30-1)\times 40,0164^2 + (25-1)\times 30,0048^2}{30+25-2} = 1283,0 = 35,8^2$$

For $d = 0,28$ and $\beta = 0,2$, the OC–e curve (see in Appendix) provides the value of $n^* = 100$ as displayed below.



OC−e curves for the two–sided *t*-test with different values of *n* and $\alpha = 0,05$.

Thus the required sample size *n* is

$$n = \frac{n^* + 1}{2} = \frac{100 + 1}{2} = 50,5 \approx 51$$

3. *Confidence interval on* $\mu_1 - \mu_2$, σ_1^2 *and* σ_2^2 *unknown and unequal; two-sided CI*

Assuming $\sigma_1^2 \neq \sigma_2^2$, construct a 95 % two-sided confidence interval on the difference in mean life length $\mu_1 - \mu_2$. Based on this 95% two-sided CI on $\mu_1 - \mu_2$, test H_0: $\mu_1 - \mu_2 = 0$ vs. H_1: $\mu_1 - \mu_2 \neq 0$ at $\alpha = 0,05$.

$$P(l \leq \mu_1 - \mu_2 \leq u) = 0,95 = 1 - \alpha \quad \Rightarrow \quad \alpha = 0,05$$

$$\nu = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1 + 1} + \frac{(S_2^2/n_2)^2}{n_2 + 1}} - 2 \approx 57$$

95 % two–sided CI on $\mu_1 - \mu_2$:

$$\left(\bar{x}_1 - \bar{x}_2\right) - t(\nu;\alpha)\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \le \mu_1 - \mu_2 \le \left(\bar{x}_1 - \bar{x}_2\right) + t(\nu;\alpha)\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

$$\left(\bar{x}_1 - \bar{x}_2\right) - t(57;0,05)\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \le \mu_1 - \mu_2 \le \left(\bar{x}_1 - \bar{x}_2\right) + t(57;0,05)\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

$$\left(780-800,04\right) - 2,00\sqrt{\frac{40,0164^2}{30}+\frac{30,0048^2}{25}} \le \mu_1 - \mu_2 \le \left(780-800,04\right) + 2,00\sqrt{\frac{40,0164^2}{30}+\frac{30,0048^2}{25}}$$

$$-20,04 - 2,00 \times 18,9091 \le \mu_1 - \mu_2 \le -20,04 + 2,00 \times 18,9091$$

$$-39,0079 \le \mu_1 - \mu_2 \le -1,07207$$

$$-39,0 \le \mu_1 - \mu_2 \le -1,1$$

Since this 95 % two–sided CI on $\mu_1 - \mu_2$ when σ_1^2 and σ_2^2 are unknown and unequal does not include the hypothesized value zero ($\delta_0 = 0$), reject H_0 at $\alpha = 0,05$.

8.3 Paired *t*-test

Learning goals

- Explain a paired experiment and its purpose.
- Test a hypothesis on μ_D for paired observations when σ_D^2 is unknown (paired *t*-test).
- Establish a $100(1-\alpha)\%$ confidence interval (CI) on μ_D paired observations when σ_D^2 is unknown.

Paired experiment

A paired experiment collects a pair of observations (X_1 and X_2) for each specimen (experimental unit) and analyzes their differences (instead of the original data). This paired experiment is used when **heterogeneity** exist**s** between specimens and this heterogeneity can significantly affect X_1 and X_2; in other words, X_1 and X_2 are <u>not</u> independent.

Inference context

- **Parameter** of interest: μ_D
- **Point estimator** of μ_D : $\bar{D} = \overline{X_1 - X_2} \sim N(\mu_D; \frac{\sigma_D^2}{n})$, μ_D unknown;

  X_1 and X_2 are <u>not</u> independent.
- **Test statistic** of μ_D : $T = \dfrac{\bar{D} - \mu_D}{S_D / \sqrt{n}} \sim t(n-1)$

Test procedure (paired *t*-test):

Step 1: State the **null hypothesis** H_0 **and alternative hypothesis** H_1.

$H_0: \mu_D = \delta_0$ $\quad H_1: \mu_D \neq \delta_0$ for two–sided test

$\mu_D < \delta_0$ for lower–sided test

$\mu_D > \delta_0$ for upper–sided test

Step 2: Determine a **test statistic and its value**.

$$T_0 = \frac{\bar{D} - \delta_0}{S_D / \sqrt{n}} \sim t(n-1); \qquad t_0 = \frac{\bar{d} - \delta_0}{s_D / \sqrt{n}}$$

Step 3: Determine a **critical value(s) for** α .

$t(n-1; \alpha)$ for two–sided test

$t(n-1; 2\alpha)$ for one–sided test

Step 4: Make a **conclusion**. Reject H_0 if

$|t_0| > t(n-1; \alpha)$ for two–sided test

$t_0 < -t(n-1; 2\alpha)$ for lower–sided test

$t_0 > t(n-1; 2\alpha)$ for upper–sided test

Confidence interval formula

A $100(1-\alpha)\%$ CI on μ_D, when σ_D^2 is unknown, is

$$\bar{D} - t(n-1; \alpha)\frac{S_D}{\sqrt{n}} \le \mu_D \le \bar{D} + t(n-1; \alpha)\frac{S_D}{\sqrt{n}} \quad \text{for two–sided CI}$$

$$\bar{D} - t(n-1; 2\alpha)\frac{S_D}{\sqrt{n}} \le \mu_D \quad \text{for lower–sided CI}$$

$$\mu_D \le \bar{D} + t(n-1; 2\alpha)\frac{S_D}{\sqrt{n}} \quad \text{for upper–sided CI}$$

CI and hypothesis test for large sample

If the sample size is large ($n \geq 30$), the *z*-based CI formulas and test procedure in Section 7.2 can be applied to inference on μ_D according to the central limit theorem.

Example 8.3

The weights (unit: kg) before and after a diet program for 30 participants are measured below.

i	**Before** (X_1)	**After** (X_2)	***i***	**Before** (X_1)	**After** (X_2)	***i***	**Before** (X_1)	**After** (X_2)
1	72,575	69,400	**11**	71,668	63,503	**21**	77,111	69,853
2	78,018	72,575	**12**	92,986	88,904	**22**	98,883	96,616
3	69,853	61,689	**13**	74,389	71,668	**23**	66,678	60,781
4	95,254	89,811	**14**	102,058	93,894	**24**	78,471	71,668
5	78,471	75,296	**15**	84,368	82,554	**25**	88,451	84,822
6	65,771	61,689	**16**	70,307	67,585	**26**	99,790	94,801
7	89,811	82,554	**17**	83,461	79,832	**27**	97,522	93,440
8	74,843	72,575	**18**	78,471	70,760	**28**	93,440	91,172
9	81,647	80,739	**19**	81,193	75,750	**29**	74,843	70,760
10	78,018	77,564	**20**	76,204	68,946	**30**	77,111	69,853

The summary of the weight data is as follows:

Sample size (no. participants)	Sample mean (weight loss)	Sample variance
$n = 30$	$\bar{d} = 4,687$	$s_D^2 = 5,297.$

where $d = "\text{Before} - \text{After}"$.

1. *Hypothesis test on* μ_D, σ_D^2 *unknown; two-sided test*

Test if there is a significant effect of the diet program on weight loss. Use $\alpha = 0,05$.

Step 1: State H_0 and H_1.

$$H_0: \mu_D = 0 \qquad H_1: \mu_D \neq 0$$

Step 2: Determine a **test statistic and its value**.

$$t_0 = \frac{\bar{d} - \delta_0}{s_D / \sqrt{n}} = \frac{4,687 - 0}{2,3015 / \sqrt{30}} = 11,15436$$

Step 3: Determine a **critical value(s) for** α.

$$t(n-1; \alpha) = t(30-1; 0,05) = t(29; 0,05) = 2,045$$

Step 4: Make a **conclusion**.

Since $|t_0| = 11{,}15436 > t(29;0{,}05) = 2{,}045$, reject H_0 at $\alpha = 0{,}05$.

2. *Confidence interval on* μ_D, σ_D^2 *unknown; two-sided CI*

Construct a 95% two-sided confidence interval on the mean weight loss μ_D due to diet program. Based on this 95% two-sided CI on μ_D, test H_0: $\mu_D = 0$ vs. H_1: $\mu_D \neq 0$ at $\alpha = 0{,}05$.

$$P(l \le \mu_D \le u) = 0{,}95 = 1-\alpha \quad \Rightarrow \quad \alpha = 0{,}05 \ ; \ n-1 = 30-1 = 29$$

95 % two-sided CI on μ_D:

$$\bar{d} - t(n-1;\alpha)\frac{s_D}{\sqrt{n}} \le \mu_D \le \bar{d} + t(n-1;\alpha)\frac{s_D}{\sqrt{n}}$$

$$4{,}687 - t(29;0{,}05)\frac{2{,}3015}{\sqrt{30}} \le \mu_D \le 4{,}687 + t(29;0{,}05)\frac{2{,}3015}{\sqrt{30}}$$

$$4{,}687 - 2{,}045 \times 0{,}4202 \le \mu_D \le 4{,}687 + 2{,}045 \times 0{,}4202$$

$$2{,}968382 \le \mu_D \le 6{,}405618$$

$$2{,}968 \le \mu_D \le 6{,}406$$

Since this 95 % two-sided CI on μ_D does not include the hypothesized value zero ($\delta_0 = 0$), reject H_0 at $\alpha = 0{,}05$.

8.4 Inference on the variances of two normal populations

Learning goals

- Test a hypothesis on the ratio of two variances σ_1^2 / σ_2^2 (F-test).
- Determine the sample size of an F-test for statistical inference on σ_1^2 / σ_2^2 by using an appropriate operating charakteristic (OC) curve.
- Establish a $100(1-\alpha)\%$ confidence interval (CI) for σ_1^2 / σ_2^2.

Inference context

- **Parameter** of interest: $\dfrac{\sigma_1^2}{\sigma_2^2}$
- **Point estimator** of $\dfrac{\sigma_1^2}{\sigma_2^2}$: $\dfrac{S_1^2}{S_2^2}$, where

$$S_1^2 = \frac{\sum_{i=1}^{n_1}(X_{1i} - \overline{X}_1)^2}{n_1 - 1} \text{ and } S_2^2 = \frac{\sum_{i=1}^{n_2}(X_{2i} - \overline{X}_2)^2}{n_2 - 1};$$

$X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$;

X_1 and X_2 are <u>independent</u>.

- **Test statistic** of $\dfrac{\sigma_1^2}{\sigma_2^2}$: $F_0 = \dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$

Test procedure (*F*-test):

Step 1: State the **null hypothesis** H_0 **and alternative hypothesis** H_1.

$H_0: \dfrac{\sigma_1^2}{\sigma_2^2} = \dfrac{\sigma_{1,0}^2}{\sigma_{2,0}^2}$ $\quad H_1: \dfrac{\sigma_1^2}{\sigma_2^2} \neq \dfrac{\sigma_{1,0}^2}{\sigma_{2,0}^2}$ for two–sided test

$\dfrac{\sigma_1^2}{\sigma_2^2} < \dfrac{\sigma_{1,0}^2}{\sigma_{2,0}^2}$ for lower–sided test

$\dfrac{\sigma_1^2}{\sigma_2^2} > \dfrac{\sigma_{1,0}^2}{\sigma_{2,0}^2}$ for upper–sided test,

Step 2: Determine a **test statistic and its value**.

$$F_0 = \frac{S_1^2/\sigma_{1,0}^2}{S_2^2/\sigma_{2,0}^2} = \frac{S_1^2}{S_2^2}\frac{\sigma_{2,0}^2}{\sigma_{1,0}^2} \sim F(n_1 - 1; n_2 - 1); \qquad f_0 = \frac{s_1^2/\sigma_{1,0}^2}{s_2^2/\sigma_{2,0}^2} = \frac{s_1^2}{s_2^2}\frac{\sigma_{2,0}^2}{\sigma_{1,0}^2}$$

Step 3: Determine a **critical value(s) for** α.

$f(n_1 - 1, n_2 - 1, 1 - \alpha/2)$ a $f(n_1 - 1, n_2 - 1, \alpha/2)$ for two–sided test

$f(n_1 - 1, n_2 - 1, 1 - \alpha)$ $\left(= \dfrac{1}{f(n_2 - 1, n_1 - 1, \alpha)}\right)$ for lower–sided test

$f(n_1 - 1, n_2 - 1, \alpha)$ for upper–sided test,

Step 4: Make a **conclusion**. Reject H_0 if

$f_0 < f(n_1 - 1, n_2 - 1, 1 - \alpha/2)$ or $f_0 > f(n_1 - 1, n_2 - 1, \alpha/2)$ for two–sided test

$f_0 < f(n_1 - 1, n_2 - 1, 1 - \alpha)$ for lower–sided test

$f_0 > f(n_1 - 1, n_2 - 1, \alpha)$ for upper–sided test

Operating characteristic (OC) curve

Table 8. displays a list of OC charts and a formula of the OC parameter λ for an *F*-test on σ_1^2/σ_2^2 where $n_1 = n_2 = n$. By using the Table 8., the appropriate OC chart for a particular *F*-test is chosen.

Table 8.3 Operating charakteristic for F-test (two random samples)

Test		α	OC curve	OC parameter
F-test	Two-sided	0,05 0,01	OC–o OC–p	$\lambda = \frac{\sigma_1}{\sigma_2}$
	One-sided	0,05 0,01	OC–q OC–r	

Confidence interval formula

A $100(1-\alpha)\%$ CI on $\frac{\sigma_1^2}{\sigma_2^2}$ is as follows:

two–sided CI

$$\frac{S_1^2}{S_2^2}\frac{1}{f(n_1-1,n_2-1,\alpha/2)} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}\frac{1}{f(n_1-1,n_2-1,1-\alpha/2)}$$

or

$$\frac{S_1^2}{S_2^2} f(n_2-1,n_1-1,1-\alpha/2) \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2} f(n_2-1,n_1-1,\alpha/2)$$

lower–sided CI

$$\frac{S_1^2}{S_2^2}\frac{1}{f(n_1-1,n_2-1,\alpha)} \le \frac{\sigma_1^2}{\sigma_2^2} \quad \text{or} \quad \frac{S_1^2}{S_2^2} f(n_2-1,n_1-1,1-\alpha) \le \frac{\sigma_1^2}{\sigma_2^2}$$

upper–sided CI

$$\frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}\frac{1}{f(n_1-1,n_2-1,1-\alpha)} \quad \text{or} \quad \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2} f(n_2-1,n_1-1,\alpha)$$

Derivation of formula for two-sided CI on σ_1^2/σ_2^2

By using the test statistic $F_0 = \frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} \sim F(n_2-1,n_1-1)$, we get

$$P\left(f(n_1-1,n_2-1;1-\alpha/2) \le F_0 \le f(n_1-1,n_2-1;\alpha/2)\right) = 1-\alpha$$

$$P\left(f(n_1-1,n_2-1;1-\alpha/2) \le \frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} \le f(n_1-1,n_2-1;\alpha/2) \right) = 1-\alpha$$

$$P\left(\frac{S_1^2}{S_2^2} f(n_1-1,n_2-1;1-\alpha/2) \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2} f(n_1-1,n_2-1;\alpha/2) \right) = 1-\alpha$$

or

$$P\left(\frac{S_1^2}{S_2^2} \frac{1}{f(n_1-1,n_2-1;\alpha/2)} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2} \frac{1}{f(n_1-1,n_2-1;1-\alpha/2)} \right) = 1-\alpha$$

Therefore,

$$L = \frac{S_1^2}{S_2^2} f(n_2-1,n_1-1;1-\alpha) = \frac{S_1^2}{S_2^2} \frac{1}{f(n_1-1,n_2-1;\alpha)}$$

$$U = \frac{S_1^2}{S_2^2} f(n_2-1,n_1-1;\alpha) = \frac{S_1^2}{S_2^2} \frac{1}{f(n_1-1,n_2-1;1-\alpha)}$$

Example 8.4

For the light bulb life length data in Example 8.1, the following results have been obtained:

Brand of light bulb	Sample size	Value of sample mean	Variance
INFINITY (X_1)	$n_1 = 30$	$\bar{x}_1 = 780$ hrs	$s_1^2 = 40{,}0164^2$
FOREVER (X_2)	$n_2 = 25$	$\bar{x}_2 = 800$ hrs	$s_2^2 = 30{,}0048^2$

1. *Hypothesis test on* σ_1^2/σ_2^2 *; two-sided test*

Test H_0: $\sigma_1^2/\sigma_2^2 = 1$ vs. H_1: $\sigma_1^2/\sigma_2^2 \ne 1$ at $\alpha = 0{,}05$.

Step 1: State the **null hypothesis** H_0 **and alternative hypothesis** H_1.

$$H_0\colon \frac{\sigma_1^2}{\sigma_2^2} = 1 \qquad\qquad H_1\colon \frac{\sigma_1^2}{\sigma_2^2} \ne 1$$

Step 2: Determine a **test statistic and its value**.

$$f_0 = \frac{s_1^2}{s_2^2} \times \frac{\sigma_{2,0}^2}{\sigma_{1,0}^2} = \frac{40{,}0164^2}{30{,}0048^2} \times 1 = 1{,}77867$$

Step 3: Determine a **critical value(s) for** α .

$$f(n_1-1,n_2-1,1-\alpha/2) = f(29,24,0{,}975) = \frac{1}{f(24,29,0{,}025)} = 0{,}46$$

$$f(n_1-1,n_2-1,\alpha/2) = f(29,24,0{,}025) = 2{,}22$$

Step 4: Make a **conclusion**.

Since $f_0 = 1,78 > f(29,24,0,975) = 0,46$ and $f_0 = 1,78 < f(29,24,0,025) = 2,22$, fail to reject H_0 at $\alpha = 0,05$.

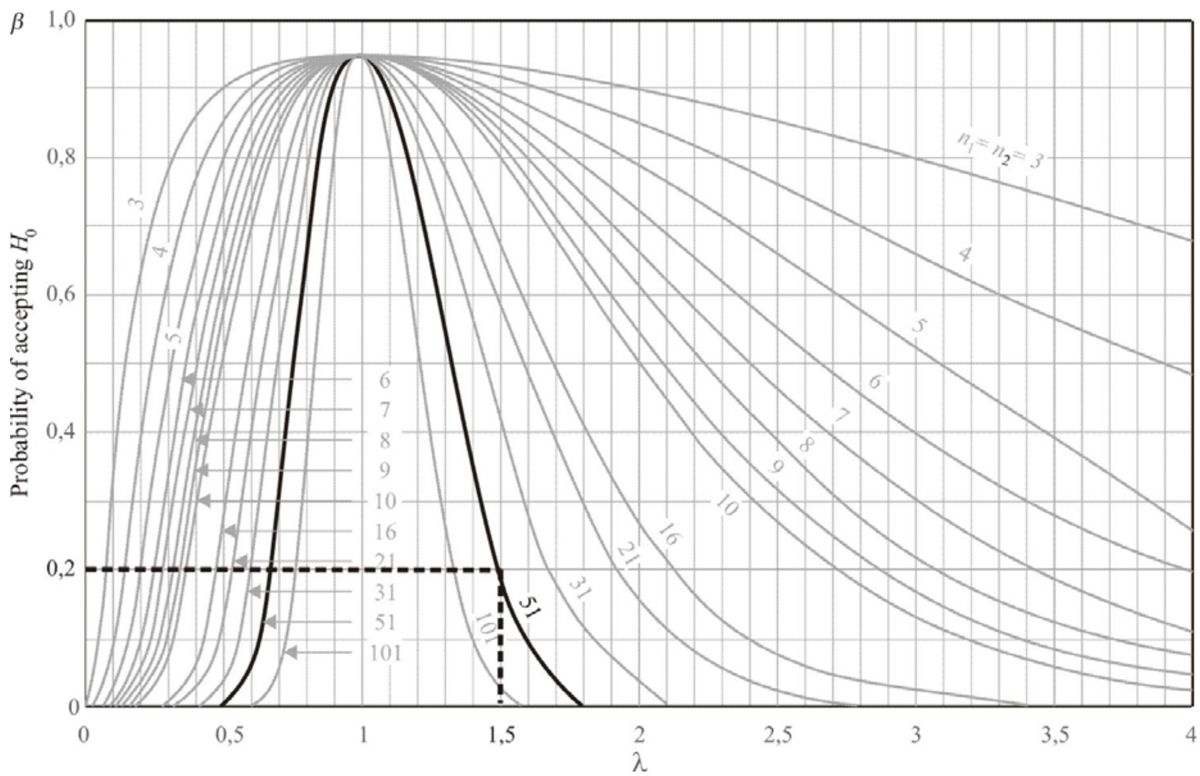
2. *Sample size determination*

Determine the sample size $n(= n_1 = n_2)$ required for this two-sided *F*-test to detect the ratio of σ_1 to σ_2 as high as 1,5 with 0,8 of power. Apply an appropriate OC curve.

To design a two-sided *F*-test at $\alpha = 0,05$, OC−o chart is applicable with the parameter

$$\lambda = \frac{\sigma_1}{\sigma_2} = 1,5$$

By using $\lambda = 1,5$ and $\beta = 0,2$ (because power $= 1 - \beta = 0,8$), the sample size required is determined $n(= n_1 = n_2) = 50$ as displayed below.



OC−o curves for the two–sided *F-test* with different values of *n* and $\alpha = 0,05$.

3. *Confidence interval on σ_1^2 / σ_2^2; two-sided CI*

Construct a 95% two-sided confidence interval on σ_1^2 / σ_2^2. Based on this 95% CI on σ_1^2 / σ_2^2 test $H_0: \sigma_1^2 / \sigma_2^2 = 1$ vs. $H_1: \sigma_1^2 / \sigma_2^2 \neq 1$ at $\alpha = 0,05$.

$$P(l \le \frac{\sigma_1^2}{\sigma_2^2} \le u) = 0,95 = 1-\alpha \quad \Rightarrow \quad \alpha = 0,05$$

95 % two-sided CI on σ_1^2 / σ_2^2:

$$\frac{S_1^2}{S_2^2} \frac{1}{f(n_1-1, n_2-1, \alpha/2)} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2} \frac{1}{f(n_1-1, n_2-1, 1-\alpha/2)}$$

$$\frac{40,0164^2}{30,0048^2} \frac{1}{f(29;24;0,025)} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{40,0164^2}{30,0048^2} \frac{1}{f(29;24;0,975)}$$

$$\frac{40,0164^2}{30,0048^2} \frac{1}{2,22} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{40,0164^2}{30,0048^2} \frac{1}{0,46}$$

$$0,801201 \le \frac{\sigma_1^2}{\sigma_2^2} \le 3,86663$$

$$0,801 \le \frac{\sigma_1^2}{\sigma_2^2} \le 3,867$$

Since this 95% two-sided CI on σ_1^2 / σ_2^2 include the hypothesized value unity ($\frac{\sigma_{1,0}^2}{\sigma_{2,0}^2} = 1$), fail to reject H_0 at $\alpha = 0,05$.

8.5 Inference on two population proportions

Learning goals

- Test a hypothesis on $p_1 - p_2$ (*z*-test).
- Determine the sample size of a *z*-test for statistical inference on $p_1 - p_2$ by using an appropriate sample size formula.
- Establish a $100(1-\alpha)\%$ confidence interval (CI) on $p_1 - p_2$.

Inference context

Parameter of interest: $p_1 - p_2$

Point estimator of $p_1 - p_2$: $\hat{P}_1 - \hat{P}_2 \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$, where

$X_1 \sim B(n_1, p_1)$, $X_2 \sim B(n_2, p_2)$;

$n_1 p_1(1-p_1) > 9$ and $n_2 p_2(1-p_2) > 9$;

X_1 and X_2 are independent;

$\hat{P}_1 = \dfrac{X_1}{n_1} \sim N\left(p_1, \dfrac{p_1(1-p_1)}{n_1}\right)$ and $\hat{P}_2 = \dfrac{X_2}{n_2} \sim N\left(p_2, \dfrac{p_2(1-p_2)}{n_2}\right)$.

Test statistic of $p_1 - p_2$: The test statistic of $p_1 - p_2$ depends on the equality of p_1 and p_2 as follows:

Case 1: Unequal proportions ($p_1 \neq p_2$)

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} \sim N(0,1)$$

Case 2: Equal proportions ($p_1 = p_2 = p$)

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{p(1-p)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim N(0,1), \text{ where } \hat{P} = \frac{X_1 + X_2}{n_1 + n_2} \quad \text{(estimator of } p\text{)}$$

Test procedure (*z*-test):

Step 1: State the **null hypothesis** H_0 **and alternative hypothesis** H_1.

H_0: $p_1 - p_2 = \delta_0$ $\qquad$ H_1: $p_1 - p_2 \neq \delta_0$ for two–sided test

$p_1 - p_2 < \delta_0$ for lower–sided test

$p_1 - p_2 > \delta_0$ for upper–sided test

Step 2: Determine a **test statistic and its value**.

Case 1: Unequal proportions ($p_1 \neq p_2$)

$$Z_0 = \frac{(\hat{P}_1 - \hat{P}_2) - \delta_0}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} = \frac{(\hat{P}_1 - \hat{P}_2) - \delta_0}{\sqrt{\dfrac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \dfrac{\hat{P}_2(1-\hat{P}_2)}{n_2}}}$$

Case 2: Equal proportions ($p_1 = p_2 = p$)

$$Z_0 = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad \hat{P} = \frac{X_1 + X_2}{n_1 + n_2}$$

Step 3: Determine a **critical value(s) for** α .

k_α for two–sided test

$k_{2\alpha}$ for one–sided test

Step 4: Make a **conclusion**. Reject H_0 if

$|z_0| > k_\alpha$ for two–sided test

$z_0 < -k_{2\alpha}$ for lower–sided test

$z_0 > k_{2\alpha}$ for upper–sided test

Sample size formula

For a hypothesis test on $p_1 - p_2$, the following formulas are applied to determine:

$$n = \left(\frac{k_\alpha\sqrt{(p_1 + p_2)(q_1 + q_2)/2} + k_{2\beta}\sqrt{p_1q_1 + p_2q_2}}{p_1 - p_2}\right)^2 \quad \text{for two–sided test}$$

$$n = \left(\frac{k_{2\alpha}\sqrt{(p_1 + p_2)(q_1 + q_2)/2} + k_{2\beta}\sqrt{p_1q_1 + p_2q_2}}{p_1 - p_2}\right)^2 \quad \text{for one–sided test}$$

where $n_1 = n_2 = n$, $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$.

Confidence interval formula

Like the test statistic on $p_1 - p_2$, a $100(1-\alpha)\%$ CI on $p_1 - p_2$ depends on the equality of p_1 and p_2 as follows:

Case 1: Unequal proportions ($p_1 \neq p_2$)

$$(\hat{P}_1 - \hat{P}_2) - k_\alpha\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}} \leq$$

$$\leq p_1 - p_2 \leq (\hat{P}_1 - \hat{P}_2) + k_\alpha\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}} \quad \text{for two–sided CI}$$

$$(\hat{P}_1 - \hat{P}_2) - k_{2\alpha}\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}} \le p_1 - p_2 \quad \text{for lower–sided CI}$$

$$p_1 - p_2 \le (\hat{P}_1 - \hat{P}_2) + k_{2\alpha}\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}} \quad \text{for upper–sided CI}$$

Case 2: Equal proportions ($p_1 = p_2 = p$)

$$(\hat{P}_1 - \hat{P}_2) - k_{\alpha}\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \le$$

$$\le p_1 - p_2 \le (\hat{P}_1 - \hat{P}_2) + k_{\alpha}\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad \text{for two–sided CI}$$

$$(\hat{P}_1 - \hat{P}_2) - k_{2\alpha}\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \le p_1 - p_2 \quad \text{for lower–sided CI}$$

$$p_1 - p_2 \le (\hat{P}_1 - \hat{P}_2) + k_{2\alpha}\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad \text{for upper–sided CI}$$

Example 8.5

Random samples of bridges are tested for metal corrosion in the A and B counties, resulting in the following:

County	Sample size (no. bridges)	X (no. corroded bridges)	Sample proportion ($\hat{p}_i = x_i / n_i$)
A	$n_1 = 40$	$x_1 = 28$	$\hat{p}_1 = 0,7$
B	$n_2 = 30$	$x_2 = 15$	$\hat{p}_2 = 0,5$

1. *Hypothesis test on* $p_1 - p_2$*;* $p_1 \ne p_2$*; unequal proportions; upper-sided test*

Assuming $p_1 \ne p_2$ test if the proportion of corroded bridges of the A county exceeds that of the B county by at least 0,1. Use $\alpha = 0,05$.

Since

$$n_1\hat{p}_1 = 40 \times 0,7 = 28, \; n_1(1-\hat{p}_1) = 40 \times 0,3 = 12$$

$$n_2\hat{p}_2 = 30 \times 0,5 = 15, \; n_2(1-\hat{p}_2) = 30 \times 0,5 = 15$$

are greater than nine, the sampling distributions of $\hat{P}_1$ and $\hat{P}_2$ are approximately normal.

Step 1: State the **null hypothesis** H_0 **and alternative hypothesis** H_1.

H_0: $p_1 - p_2 = 0{,}1$ $\quad\quad$ H_1: $p_1 - p_2 > 0{,}1$

Step 2: Determine a **test statistic and its value**.

$$z_0 = \frac{(\hat{p}_1 - \hat{p}_2) - \delta_0}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} = \frac{(0,7-0,5)-0,1}{\sqrt{\dfrac{0,7\times(1-0,7)}{40} + \dfrac{0,5\times(1-0,5)}{30}}} = 0,86$$

Step 3: Determine a **critical value(s) for** α

$$k_{2\alpha} = k_{0,1} = 1,645$$

Step 4: Make a **conclusion**.

Since $|z_0| = 0,86 < k_{0,1} = 1,645$, fail to reject H_0 at $\alpha = 0,05$.

2. *Sample size determination*

Suppose that $p_1 = 0,7$ and $p_2 = 0,5$. Determine the sample size $n\,(= n_1 = n_2)$ required for this two-sided *z*-test to detect the difference of the two proportions with power of 0,9.

$$\text{Power of test } = P(\text{reject}\, H_0 | H_0 \text{ is false}) = 1 - \beta = 0,9 \Rightarrow \beta = 0,1$$

$$q_1 = 1 - p_1 = 1 - 0,7 = 0,3 \text{ a } q_2 = 1 - p_2 = 1 - 0,5 = 0,5$$

$$n = \left(\frac{k_{2\alpha}\sqrt{(p_1+p_2)(q_1+q_2)/2} + k_{2\beta}\sqrt{p_1q_1 + p_2q_2}}{p_1 - p_2} \right)^2 =$$

$$= \left(\frac{k_{0,1}\sqrt{(0,7+0,5)(0,3+0,5)/2} + k_{0,2}\sqrt{0,7\times 0,3 + 0,5\times 0,5}}{0,7-0,5} \right)^2 =$$

$$= \left(\frac{1,645\times 0,69 + 1,28\times 0,68}{0,2} \right)^2 \approx 101$$

3. *Confidence interval on* $p_1 - p_2$*; unequal proportions; upper-confidence bound*

Assuming $p_1 \neq p_2$, construct a 95% upper-confidence bound on the difference of the two corroded bridge proportions ($p_1 - p_2$).

95 % one-sided CI on $p_1 - p_2$:

$$p_1 - p_2 \le (\hat{p}_1 - \hat{p}_2) + k_{2\alpha}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

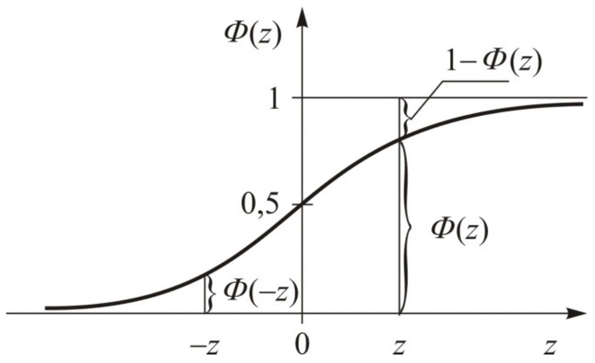
$$p_1 - p_2 \leq (0,7 - 0,5) + k_{0,1}\sqrt{\frac{0,7\times(1-0,7)}{40} + \frac{0,5\times(1-0,5)}{n_2}}$$

$$p_1 - p_2 \leq 0,2 + 1,645 \times 0,117$$

$$p_1 - p_2 \leq 0,39$$

APPENDIX

CUMULATIVE DISTRIBUTION FUNCTIONS
STANDARD NORMAL DISTRIBUTION



$$F(x)=\Phi(z),\ \text{where}\ z=\frac{x-\mu}{\sigma}$$

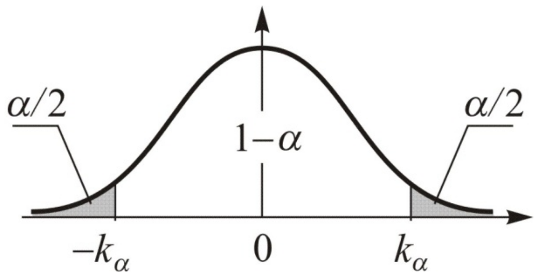
$$\Phi(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{t^2}{2}}dt$$

z	ϕ[z]	z	ϕ[z]	z	ϕ[z]	z	ϕ[z]	z	ϕ[z]
0.	0.50000	0.3	0.61791	0.6	0.72575	0.9	0.81594	1.2	0.88493
0.01	0.50399	0.31	0.62172	0.61	0.72907	0.91	0.81859	1.21	0.88686
0.02	0.50798	0.32	0.62552	0.62	0.73237	0.92	0.82121	1.22	0.88877
0.03	0.51197	0.33	0.62930	0.63	0.73565	0.93	0.82381	1.23	0.89065
0.04	0.51595	0.34	0.63307	0.64	0.73891	0.94	0.82639	1.24	0.89251
0.05	0.51994	0.35	0.63683	0.65	0.74215	0.95	0.82894	1.25	0.89435
0.06	0.52392	0.36	0.64058	0.66	0.74537	0.96	0.83147	1.26	0.89617
0.07	0.52790	0.37	0.64431	0.67	0.74857	0.97	0.83398	1.27	0.89796
0.08	0.53188	0.38	0.64803	0.68	0.75175	0.98	0.83646	1.28	0.89973
0.09	0.53586	0.39	0.65173	0.69	0.75490	0.99	0.83891	1.29	0.90147
0.1	0.53983	0.4	0.65542	0.7	0.75804	1.	0.84134	1.3	0.90320
0.11	0.54380	0.41	0.65910	0.71	0.76115	1.01	0.84375	1.31	0.90490
0.12	0.54776	0.42	0.66276	0.72	0.76424	1.02	0.84614	1.32	0.90658
0.13	0.55172	0.43	0.66640	0.73	0.76730	1.03	0.84849	1.33	0.90824
0.14	0.55567	0.44	0.67003	0.74	0.77035	1.04	0.85083	1.34	0.90988
0.15	0.55962	0.45	0.67364	0.75	0.77337	1.05	0.85314	1.35	0.91149
0.16	0.56356	0.46	0.67724	0.76	0.77637	1.06	0.85543	1.36	0.91309
0.17	0.56749	0.47	0.68082	0.77	0.77935	1.07	0.85769	1.37	0.91466
0.18	0.57142	0.48	0.68439	0.78	0.78230	1.08	0.85993	1.38	0.91621
0.19	0.57535	0.49	0.68793	0.79	0.78524	1.09	0.86214	1.39	0.91774
0.2	0.57926	0.5	0.69146	0.8	0.78814	1.1	0.86433	1.4	0.91924
0.21	0.58317	0.51	0.69497	0.81	0.79103	1.11	0.86650	1.41	0.92073
0.22	0.58706	0.52	0.69847	0.82	0.79389	1.12	0.86864	1.42	0.92220
0.23	0.59095	0.53	0.70194	0.83	0.79673	1.13	0.87076	1.43	0.92364
0.24	0.59483	0.54	0.70540	0.84	0.79955	1.14	0.87286	1.44	0.92507
0.25	0.59871	0.55	0.70884	0.85	0.80234	1.15	0.87493	1.45	0.92647
0.26	0.60257	0.56	0.71226	0.86	0.80511	1.16	0.87698	1.46	0.92785
0.27	0.60642	0.57	0.71566	0.87	0.80785	1.17	0.87900	1.47	0.92922
0.28	0.61026	0.58	0.71904	0.88	0.81057	1.18	0.88100	1.48	0.93056
0.29	0.61409	0.59	0.72240	0.89	0.81327	1.19	0.88298	1.49	0.93189

z	ϕ[z]	z	ϕ[z]	z	ϕ[z]	z	ϕ[z]	z	ϕ[z]
1.5	0.93319	1.8	0.96407	2.1	0.98214	2.4	0.99180	2.7	0.99653
1.51	0.93448	1.81	0.96485	2.11	0.98257	2.41	0.99202	2.71	0.99664
1.52	0.93574	1.82	0.96562	2.12	0.98300	2.42	0.99224	2.72	0.99674
1.53	0.93699	1.83	0.96638	2.13	0.98341	2.43	0.99245	2.73	0.99683
1.54	0.93822	1.84	0.96712	2.14	0.98382	2.44	0.99266	2.74	0.99693
1.55	0.93943	1.85	0.96784	2.15	0.98422	2.45	0.99286	2.75	0.99702
1.56	0.94062	1.86	0.96856	2.16	0.98461	2.46	0.99305	2.76	0.99711
1.57	0.94179	1.87	0.96926	2.17	0.98500	2.47	0.99324	2.77	0.99720
1.58	0.94295	1.88	0.96995	2.18	0.98537	2.48	0.99343	2.78	0.99728
1.59	0.94408	1.89	0.97062	2.19	0.98574	2.49	0.99361	2.79	0.99736
1.6	0.94520	1.9	0.97128	2.2	0.98610	2.5	0.99379	2.8	0.99744
1.61	0.94630	1.91	0.97193	2.21	0.98645	2.51	0.99396	2.81	0.99752
1.62	0.94738	1.92	0.97257	2.22	0.98679	2.52	0.99413	2.82	0.99760
1.63	0.94845	1.93	0.97320	2.23	0.98713	2.53	0.99430	2.83	0.99767
1.64	0.94950	1.94	0.97381	2.24	0.98745	2.54	0.99446	2.84	0.99774
1.65	0.95053	1.95	0.97441	2.25	0.98778	2.55	0.99461	2.85	0.99781
1.66	0.95154	1.96	0.97500	2.26	0.98809	2.56	0.99477	2.86	0.99788
1.67	0.95254	1.97	0.97558	2.27	0.98840	2.57	0.99492	2.87	0.99795
1.68	0.95352	1.98	0.97615	2.28	0.98870	2.58	0.99506	2.88	0.99801
1.69	0.95449	1.99	0.97670	2.29	0.98899	2.59	0.99520	2.89	0.99807
1.7	0.95543	2.	0.97725	2.3	0.98928	2.6	0.99534	2.9	0.99813
1.71	0.95637	2.01	0.97778	2.31	0.98956	2.61	0.99547	2.91	0.99819
1.72	0.95728	2.02	0.97831	2.32	0.98983	2.62	0.99560	2.92	0.99825
1.73	0.95818	2.03	0.97882	2.33	0.99010	2.63	0.99573	2.93	0.99831
1.74	0.95907	2.04	0.97932	2.34	0.99036	2.64	0.99585	2.94	0.99836
1.75	0.95994	2.05	0.97982	2.35	0.99061	2.65	0.99598	2.95	0.99841
1.76	0.96080	2.06	0.98030	2.36	0.99086	2.66	0.99609	2.96	0.99846
1.77	0.96164	2.07	0.98077	2.37	0.99111	2.67	0.99621	2.97	0.99851
1.78	0.96246	2.08	0.98124	2.38	0.99134	2.68	0.99632	2.98	0.99856
1.79	0.96327	2.09	0.98169	2.39	0.99158	2.69	0.99643	2.99	0.99861

z	ϕ[z]	z	ϕ[z]	z	ϕ[z]	z	ϕ[z]	z	ϕ[z]
3.	0.99865	3.5	0.99977	4.	0.99996833	4.5	0.99999660	5.	0.99999971
3.1	0.99903	3.6	0.99984	4.1	0.99997934	4.6	0.99999789	5.1	0.99999983
3.2	0.99931	3.7	0.99989	4.2	0.99998665	4.7	0.99999870	5.2	0.99999990
3.3	0.99952	3.8	0.99993	4.3	0.99999146	4.8	0.99999921	5.3	0.99999994
3.4	0.99966	3.9	0.99995	4.4	0.99999459	4.9	0.99999952	5.4	0.99999997
3.5	0.99977	4.	0.99997	4.5	0.99999660	5.	0.99999971	5.5	0.99999998

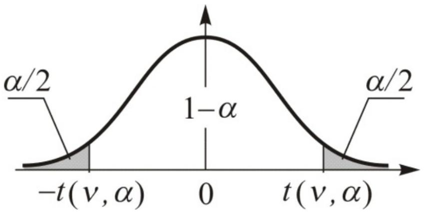
CRITICAL VALUES OF NORMAL DISTRIBUTION



$$P\left(|X| > k_\alpha\right) = \alpha$$

α	k_α	α	k_α	α	k_α
0,002	3,090	0,042	2,034	0,082	1,739
0,004	2,878	0,044	2,014	0,084	1,728
0,006	2,748	0,046	2,995	0,086	1,717
0,008	2,652	0,048	1,977	0,088	1,706
0,010	2,576	0,050	1,960	0,090	1,695
0,012	2,512	0,052	1,943	0,092	1,685
0,014	2,457	0,054	1,927	0,094	1,675
0,016	2,409	0,056	1,911	0,096	1,665
0,018	2,366	0,058	1,896	0,098	1655
0,020	2,326	0,060	1,881	0,100	1,645
0,022	2,290	0,062	1,866	0,110	1,598
0,024	2,257	0,064	1,852	0,120	1,555
0,026	2,226	0,066	1,838	0,130	1,514
0,028	2,197	0,068	1,825	0,140	1,476
0,030	2,170	0,070	1,812	0,150	1,440
0,032	2,144	0,072	1,799	0,160	1,405
0,034	2,120	0,074	1,787	0,170	1,372
0,036	2,097	0,076	1,774	0,180	1,341
0,038	2,075	0,078	1,762	0,190	1,311
0,040	2,054	0,080	1,751	0,200	1,282

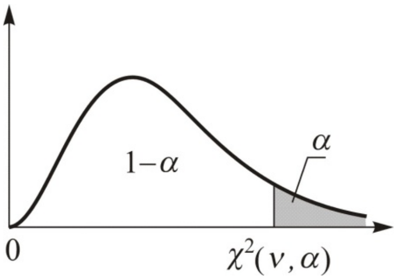
CRITICAL VALUES OF t – DISTRIBUTION



$$P\left(\left|T\right| > t(\nu,\alpha)\right) = \alpha$$

ν	**α = 0,20**	**α = 0,10**	**α = 0,05**	**α = 0,02**	**α = 0,01**
1	3,080	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	6,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
40	1,303	1,684	2,021	2,426	2,704
60	1,296	1,671	2,000	2,390	2,660
120	1,289	1,658	1,980	2,358	2,617
∞	1,282	1,645	1,960	2,326	2,576

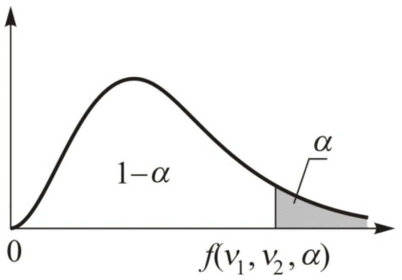
CRITICAL VALUES OF χ^2 – DISTRIBUTION



$$P\left(X^2 > \chi^2(\nu,\alpha)\right) = \alpha$$

ν \ α	0,995	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010	0,005
1		0,0002	0,0010	0,0039	0,0158	2,71	3,84	5,02	6,63	7,88
2	0,0100	0,0201	0,0506	0,1030	0,2110	4,61	5,99	7,38	9,21	10,60
3	0,0717	0,1150	0,2160	0,3250	0,5840	6,25	7,82	9,35	11,30	12,80
4	0,2070	0,2970	0,4840	0,7110	1,0600	7,78	9,49	11,10	13,30	14,90
5	0,4120	0,5540	0,8310	1,1500	1,6100	9,24	11,10	12,80	15,10	16,70
6	0,6760	0,8720	1,2400	1,6400	2,2000	10,60	12,60	14,40	16,80	18,50
7	0,9890	1,2400	1,6900	2,1700	2,8300	12,00	14,10	16,00	18,50	20,30
8	1,3400	1,6500	2,1800	2,7300	3,4900	13,40	15,50	17,50	20,10	22,00
9	1,7300	2,0900	2,7000	3,3300	4,1700	14,70	16,90	19,00	21,70	23,60
10	2,1600	2,5600	3,2500	3,9400	4,8700	16,00	18,30	20,50	23,20	25,20
11	2,6000	3,0500	3,8200	4,5700	5,5800	17,30	19,70	21,90	24,70	26,80
12	3,0700	3,5700	4,4000	5,2300	6,3000	18,50	21,00	23,30	26,20	28,30
13	3,5700	4,1100	5,0100	5,8900	7,0400	19,80	22,40	24,70	27,70	29,80
14	4,0700	4,6600	5,6300	6,5700	7,7900	21,10	23,70	26,10	29,10	31,30
15	4,6000	5,2300	6,2600	7,2600	8,5500	22,30	25,00	27,50	30,60	32,80
16	5,1400	5,8100	6,9100	7,9600	9,3100	23,50	26,30	28,80	32,00	34,30
17	5,7000	6,4100	7,6500	8,6700	10,1000	24,80	27,60	30,20	33,40	35,70
18	6,2600	7,0100	8,2300	9,3900	10,9000	26,00	28,90	31,50	34,80	37,20
19	6,8400	7,6300	8,9100	10,1000	11,7000	27,20	30,10	32,90	36,20	38,60
20	7,4300	8,2600	9,5900	10,9000	12,4000	28,40	31,40	34,20	37,60	40,00
21	8,0300	8,9000	10,3000	11,6000	13,2000	29,60	32,70	35,50	38,90	41,40
22	8,6400	9,5400	11,0000	12,3000	14,0000	30,80	33,90	36,80	40,30	42,80
23	9,2600	10,2000	11,7000	13,1000	14,8000	32,00	35,20	38,10	41,60	44,20
24	9,8900	10,9000	12,4000	13,8000	15,7000	33,20	36,40	39,40	43,00	45,60
25	10,5000	11,5000	13,1000	14,6000	16,5000	34,40	37,70	40,60	44,30	46,90
26	11,2000	12,2000	13,8000	15,4000	17,3000	35,50	38,90	41,90	45,60	48,30
27	11,8000	12,9000	14,6000	16,2000	18,1000	36,70	40,10	43,20	47,00	49,60
28	12,5000	13,6000	15,3000	16,9000	18,9000	37,90	41,30	44,50	48,30	51,00
29	13,1000	14,3000	16,0000	17,7000	19,8000	39,10	42,60	45,70	49,60	52,30
30	13,8000	15,0000	16,8000	18,5000	20,6000	40,30	43,80	47,00	50,90	53,70

CRITICAL VALUES OF F – DISTRIBUTION



$$P\big(F > f(\nu_1, \nu_2, \alpha)\big) = \alpha$$

$\alpha = 0{,}01$									
ν_2 \ ν_1	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
4	15,977020	15,52186	15,206860	14,975760	14,798890	14,65913	14,545900	14,452280	14,373590
5	11,391930	10,96702	10,672250	10,455510	10,289310	10,15776	10,051020	9,962648	9,888275
6	9,148301	8,745895	8,466125	8,259995	8,101651	7,976121	7,874119	7,789570	7,718333
7	7,846645	7,460435	7,191405	6,992833	6,840049	6,718752	6,620063	6,538166	6,469091
8	7,006077	6,631825	6,370681	6,177624	6,028870	5,910619	5,814294	5,734275	5,666719
9	6,422085	6,056941	5,801770	5,612865	5,467123	5,351129	5,256542	5,177890	5,111431
10	5,994339	5,636326	5,385811	5,200121	5,056693	4,942421	4,849147	4,771518	4,705870
11	5,668300	5,316009	5,069210	4,886072	4,744468	4,631540	4,539282	4,462436	4,397401
12	5,411951	5,064343	4,820574	4,639502	4,499365	4,387510	4,296054	4,219820	4,155258
13	5,205330	4,861621	4,620363	4,440997	4,302062	4,191078	4,100267	4,024518	3,960326
14	5,035378	4,694964	4,455820	4,277882	4,139946	4,029680	3,939396	3,864039	3,800141
15	4,893210	4,555614	4,318273	4,141546	4,004453	3,894788	3,804940	3,729902	3,666240
16	4,772578	4,437420	4,201634	4,025947	3,889572	3,780415	3,690931	3,616157	3,552687
17	4,668968	4,335939	4,101505	3,926719	3,790964	3,682242	3,593066	3,518512	3,455198
18	4,579036	4,247882	4,014637	3,840639	3,705422	3,597074	3,508162	3,433793	3,370608
19	4,500258	4,170767	3,938573	3,765269	3,630525	3,522503	3,433817	3,359605	3,296527
20	4,430690	4,102685	3,871427	3,698740	3,564412	3,456676	3,368186	3,294108	3,231120
21	4,368815	4,042144	3,811725	3,639590	3,505632	3,398147	3,309830	3,235867	3,172953
22	4,313429	3,987963	3,758301	3,586660	3,453034	3,345773	3,257606	3,183742	3,120891
23	4,263567	3,939195	3,710218	3,539024	3,405695	3,298634	3,210599	3,136822	3,074025
24	4,218445	3,895070	3,666717	3,495928	3,362867	3,255985	3,168069	3,094367	3,031615
25	4,177420	3,854957	3,627174	3,456754	3,323937	3,217217	3,129406	3,055771	2,993056
26	4,139960	3,818336	3,591075	3,420993	3,288399	3,181824	3,094108	3,020530	2,957848
27	4,105622	3,784770	3,557991	3,388219	3,255827	3,149385	3,061754	2,988228	2,925573
28	4,074032	3,753895	3,527559	3,358073	3,225868	3,119547	3,031992	2,958512	2,895881
29	4,044873	3,725399	3,499475	3,330252	3,198219	3,092009	3,004524	2,931084	2,868472
30	4,017877	3,699019	3,473477	3,304499	3,172624	3,066516	2,979094	2,905690	2,843095

$\alpha = 0{,}01$									
v_2 \ v_1	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
31	3,992811	3,674528	3,449341	3,280591	3,148863	3,042849	2,955484	2,882112	2,819532
32	3,969477	3,651731	3,426876	3,258338	3,126746	3,020818	2,933506	2,860163	2,797595
33	3,947701	3,630458	3,405914	3,237573	3,106108	3,000261	2,912997	2,839680	2,777122
34	3,927333	3,610562	3,386309	3,218154	3,086807	2,981033	2,893814	2,820521	2,757971
35	3,908241	3,591914	3,367935	3,199952	3,068716	2,963012	2,875833	2,802561	2,740018
36	3,890308	3,574399	3,350677	3,182858	3,051726	2,946086	2,858945	2,785692	2,723155
37	3,873433	3,557918	3,334440	3,166774	3,035738	2,930159	2,843053	2,769817	2,707284
38	3,857524	3,542383	3,319133	3,151612	3,020668	2,915145	2,828072	2,754851	2,692322
39	3,842502	3,527713	3,304681	3,137296	3,006438	2,900968	2,813925	2,740719	2,678192
40	3,828294	3,513840	3,291012	3,123757	2,992981	2,887560	2,800545	2,727352	2,664827
41	3,814835	3,500699	3,278067	3,110934	2,980234	2,874861	2,787871	2,714690	2,652167
42	3,802069	3,488235	3,265787	3,098771	2,968144	2,862814	2,775850	2,702679	2,640156
43	3,789942	3,476396	3,254125	3,087218	2,956661	2,851373	2,764431	2,691269	2,628747
44	3,778409	3,465137	3,243033	3,076232	2,945740	2,840491	2,753570	2,680418	2,617896
45	3,767427	3,454416	3,232472	3,065771	2,935341	2,830129	2,743229	2,670084	2,607562
46	3,756957	3,444196	3,222404	3,055798	2,925427	2,820251	2,733369	2,660232	2,597709
47	3,746964	3,434442	3,212796	3,046281	2,915966	2,810823	2,723960	2,650829	2,588305
48	3,737417	3,425123	3,203617	3,037188	2,906927	2,801816	2,714969	2,641845	2,579319
49	3,728286	3,416211	3,194838	3,028492	2,898283	2,793202	2,706371	2,633253	2,570725
50	3,719545	3,407680	3,186434	3,020168	2,890008	2,784956	2,698139	2,625026	2,562497
55	3,680897	3,369962	3,149283	2,983369	2,853424	2,748497	2,661744	2,588651	2,526110
60	3,649047	3,338884	3,118674	2,953049	2,823280	2,718454	2,631751	2,558670	2,496116
65	3,622349	3,312836	3,093020	2,927638	2,798015	2,693272	2,606607	2,533535	2,470966
70	3,599647	3,290689	3,071209	2,906032	2,776533	2,671859	2,585226	2,512158	2,449575
75	3,580106	3,271628	3,052437	2,887437	2,758044	2,653429	2,566821	2,493756	2,431158
80	3,563110	3,255049	3,036111	2,871265	2,741964	2,637398	2,550812	2,477747	2,415136
85	3,548191	3,240499	3,021782	2,857072	2,727851	2,623328	2,536759	2,463695	2,401070
90	3,534992	3,227626	3,009106	2,844515	2,715364	2,610879	2,524326	2,451260	2,388623
95	3,523230	3,216156	2,997811	2,833327	2,704238	2,599787	2,513246	2,440179	2,377530
100	3,512684	3,205872	2,987684	2,823295	2,694263	2,589841	2,503311	2,430242	2,367582
105	3,503174	3,196599	2,978553	2,814250	2,685268	2,580872	2,494352	2,421281	2,358610
110	3,494555	3,188194	2,970278	2,806052	2,677115	2,572743	2,486232	2,413158	2,350478
115	3,486707	3,180542	2,962743	2,798588	2,669692	2,565341	2,478838	2,405762	2,343072
120	3,479531	3,173545	2,955854	2,791764	2,662906	2,558574	2,472077	2,398999	2,336300
125	3,472945	3,167124	2,949531	2,785500	2,656676	2,552362	2,465871	2,392791	2,330083
129	3,468053	3,162354	2,944835	2,780848	2,652050	2,547748	2,461261	2,388179	2,325465

$\alpha = 0,01$									
v_2 \ v_1	**15**	**20**	**24**	**30**	**40**	**50**	**60**	**80**	**100**
4	14,19820	14,01961	13,92906	13,83766	13,74538	13,68958	13,65220	13,60526	13,57699
5	9,722219	9,552646	9,466471	9,379329	9,291189	9,237811	9,202015	9,157029	9,129907
6	7,558994	7,395832	7,312721	7,228533	7,143222	7,091475	7,056737	7,013037	6,986667
7	6,314331	6,155438	6,074319	5,992010	5,908449	5,857682	5,823566	5,780605	5,754657
8	5,515125	5,359095	5,279264	5,198130	5,115610	5,065398	5,031618	4,989038	4,963296
9	4,962078	4,807995	4,728998	4,648582	4,566649	4,516715	4,483087	4,440656	4,414980
10	4,558140	4,405395	4,326929	4,246933	4,165287	4,115452	4,081855	4,039422	4,013719
11	4,250867	4,099046	4,020910	3,941132	3,859573	3,809716	3,776071	3,733533	3,707744
12	4,009619	3,858433	3,780485	3,700789	3,619181	3,569222	3,535473	3,492763	3,466845
13	3,815365	3,664609	3,586753	3,507042	3,425293	3,375176	3,341287	3,298357	3,272282
14	3,655697	3,505222	3,427387	3,347596	3,265641	3,215328	3,181274	3,138094	3,111842
15	3,522194	3,371892	3,294029	3,214110	3,131906	3,081371	3,047135	3,003683	2,977242
16	3,408947	3,258737	3,180811	3,100733	3,018248	2,967476	2,933046	2,889308	2,862669
17	3,311694	3,161518	3,083502	3,003241	2,920458	2,869437	2,834806	2,790774	2,763932
18	3,227286	3,077097	2,998974	2,918516	2,835420	2,784144	2,749309	2,704978	2,677930
19	3,153343	3,003109	2,924866	2,844201	2,760786	2,709251	2,674211	2,629578	2,602323
20	3,088041	2,937735	2,859363	2,778485	2,694749	2,642954	2,607708	2,562774	2,535313
21	3,029951	2,879556	2,801050	2,719955	2,635896	2,583844	2,548393	2,503160	2,475492
22	2,977946	2,827447	2,748802	2,667490	2,583111	2,530803	2,495149	2,449619	2,421747
23	2,931118	2,780504	2,701720	2,620191	2,535496	2,482935	2,447081	2,401258	2,373184
24	2,888732	2,737997	2,659072	2,577329	2,492321	2,439512	2,403461	2,357349	2,329076
25	2,850186	2,699325	2,620260	2,538305	2,452990	2,399937	2,363691	2,317296	2,288826
26	2,814982	2,663991	2,584787	2,502624	2,417007	2,363715	2,327279	2,280604	2,251941
27	2,782703	2,631580	2,552239	2,469872	2,383960	2,330434	2,293812	2,246863	2,218009
28	2,753000	2,601744	2,522268	2,439701	2,353501	2,299745	2,262941	2,215723	2,186682
29	2,725577	2,574188	2,494579	2,411817	2,325335	2,271355	2,234372	2,186890	2,157666
30	2,700180	2,548659	2,468921	2,385967	2,299211	2,245012	2,207854	2,160114	2,130710
31	2,676594	2,524942	2,445077	2,361937	2,274913	2,220500	2,183171	2,135178	2,105597
32	2,654632	2,502850	2,422861	2,339539	2,252253	2,197632	2,160136	2,111895	2,082141
33	2,634132	2,482222	2,402111	2,318613	2,231072	2,176247	2,138588	2,090105	2,060180
34	2,614952	2,462916	2,382687	2,299016	2,211227	2,156203	2,118384	2,069664	2,039573
35	2,596969	2,444810	2,364466	2,280626	2,192595	2,137377	2,099403	2,050450	2,020195
36	2,580074	2,427794	2,347337	2,263334	2,175068	2,119661	2,081534	2,032354	2,001938
37	2,564172	2,411773	2,331207	2,247044	2,158548	2,102957	2,064681	2,015278	1,984705
38	2,549177	2,396662	2,315989	2,231671	2,142952	2,087180	2,048759	1,999138	1,968411
39	2,535014	2,382385	2,301608	2,217140	2,128202	2,072255	2,033692	1,983858	1,952979
40	2,521616	2,368876	2,287998	2,203382	2,114232	2,058113	2,019411	1,969368	1,938341

$\alpha = 0,01$									
v_2 \ v_1	**15**	**20**	**24**	**30**	**40**	**50**	**60**	**80**	**100**
41	2,508922	2,356074	2,275097	2,190338	2,100981	2,044695	2,005857	1,955609	1,924436
42	2,496878	2,343924	2,262851	2,177953	2,088394	2,031944	1,992974	1,942526	1,911210
43	2,485436	2,332378	2,251211	2,166177	2,076423	2,019813	1,980713	1,930069	1,898612
44	2,474552	2,321392	2,240134	2,154968	2,065022	2,008257	1,969029	1,918193	1,886599
45	2,464185	2,310926	2,229580	2,144285	2,054151	1,997234	1,957883	1,906859	1,875129
46	2,454300	2,300945	2,219512	2,134091	2,043775	1,986709	1,947237	1,896028	1,864166
47	2,444863	2,291414	2,209897	2,124354	2,033860	1,976649	1,937058	1,885669	1,853677
48	2,435846	2,282305	2,200705	2,115043	2,024376	1,967023	1,927316	1,875749	1,843630
49	2,427220	2,273589	2,191910	2,106132	2,015295	1,957803	1,917982	1,866242	1,833997
50	2,418961	2,265243	2,183485	2,097593	2,006592	1,948964	1,909032	1,857122	1,824753
55	2,382427	2,2283000	2,146180	2,059761	1,967989	1,909727	1,869272	1,816559	1,783606
60	2,352297	2,197806	2,115364	2,028479	1,936018	1,877187	1,836259	1,782816	1,749328
65	2,327023	2,172206	2,089479	2,002175	1,909099	1,849753	1,808397	1,754286	1,720305
70	2,305517	2,150410	2,067425	1,979748	1,886115	1,826304	1,784557	1,729835	1,695398
75	2,286997	2,131626	2,048411	1,960396	1,866260	1,806024	1,763920	1,708635	1,673777
80	2,270879	2,115271	2,031847	1,943526	1,848932	1,788309	1,745877	1,690072	1,654822
85	2,256726	2,100901	2,017288	1,928688	1,833677	1,772697	1,729964	1,673677	1,638062
90	2,244198	2,088176	2,004390	1,915536	1,820141	1,758834	1,715821	1,659088	1,623133
95	2,233031	2,076829	1,992884	1,903797	1,808050	1,746440	1,703168	1,646019	1,609745
100	2,223015	2,066646	1,982556	1,893254	1,797181	1,735292	1,691780	1,634242	1,597669
105	2,213979	2,057458	1,973234	1,883733	1,787360	1,725210	1,681474	1,623572	1,586719
110	2,205788	2,049125	1,964777	1,875093	1,778440	1,716047	1,672102	1,613860	1,576742
115	2,198327	2,041533	1,957070	1,867216	1,770302	1,707684	1,663542	1,604980	1,567613
120	2,191504	2,034588	1,950018	1,860005	1,762849	1,700018	1,655693	1,596830	1,559227
125	2,185240	2,028210	1,943540	1,853380	1,755996	1,692967	1,648469	1,589322	1,551495
129	2,180586	2,023471	1,938726	1,848454	1,750899	1,687720	1,643091	1,583728	1,545731

$\alpha = 0{,}05$									
v_2 \ v_1	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
4	6,388233	6,256057	6,163132	6,094211	6,041044	5,998779	5,964371	5,935813	5,911729
5	5,192168	5,050329	4,950288	4,875872	4,818320	4,772466	4,735063	4,703967	4,677704
6	4,533677	4,387374	4,283866	4,206658	4,146804	4,099016	4,059963	4,027442	3,999935
7	4,120312	3,971523	3,865969	3,787044	3,725725	3,676675	3,636523	3,603037	3,574676
8	3,837853	3,687499	3,580580	3,500464	3,438101	3,388130	3,347163	3,312951	3,283939
9	3,633089	3,481659	3,373754	3,292746	3,229583	3,178893	3,137280	3,102485	3,072947
10	3,478050	3,325835	3,217175	3,135465	3,071658	3,020383	2,978237	2,942957	2,912977
11	3,356690	3,203874	3,094613	3,012330	2,947990	2,896223	2,853625	2,817930	2,787569
12	3,259167	3,105875	2,996120	2,913358	2,848565	2,796375	2,753387	2,717331	2,686637
13	3,179117	3,025438	2,915269	2,832098	2,766913	2,714356	2,671024	2,634650	2,603661
14	3,112250	2,958249	2,847726	2,764199	2,698672	2,645791	2,602155	2,565497	2,534243
15	3,055568	2,901295	2,790465	2,706627	2,640797	2,587626	2,543719	2,506806	2,475313
16	3,006917	2,852409	2,741311	2,657197	2,591096	2,537667	2,493513	2,456369	2,424660
17	2,964708	2,809996	2,698660	2,614299	2,547955	2,494291	2,449916	2,412561	2,380654
18	2,927744	2,772853	2,661305	2,576722	2,510158	2,456281	2,411702	2,374156	2,342067
19	2,895107	2,740058	2,628318	2,543534	2,476770	2,422699	2,377934	2,340210	2,307954
20	2,866081	2,710890	2,598978	2,514011	2,447064	2,392814	2,347878	2,309991	2,277581
21	2,840100	2,684781	2,572712	2,487578	2,420462	2,366048	2,320953	2,282916	2,250362
22	2,816708	2,661274	2,549061	2,463774	2,396503	2,341937	2,296696	2,258518	2,225831
23	2,795539	2,639999	2,527655	2,442226	2,374812	2,320105	2,274728	2,236419	2,203607
24	2,776289	2,620654	2,508189	2,422629	2,355081	2,300244	2,254739	2,216309	2,183380
25	2,758710	2,602987	2,490410	2,404728	2,337057	2,282097	2,236474	2,197929	2,164891
26	2,742594	2,586790	2,474109	2,388314	2,320527	2,265453	2,219718	2,181067	2,147926
27	2,727765	2,571886	2,459108	2,373208	2,305313	2,250131	2,204292	2,165540	2,132303
28	2,714076	2,558127	2,445259	2,359260	2,291264	2,235982	2,190044	2,151197	2,117869
29	2,701399	2,545386	2,432434	2,346342	2,278251	2,222874	2,176844	2,137908	2,104493
30	2,689628	2,533555	2,420523	2,334344	2,266163	2,210697	2,164580	2,125559	2,092063
31	2,678667	2,522538	2,409432	2,323171	2,254906	2,199355	2,153156	2,114054	2,080482
32	2,668437	2,512255	2,399080	2,312741	2,244396	2,188766	2,142488	2,103311	2,069665
33	2,658867	2,502635	2,389394	2,302982	2,234562	2,178856	2,132504	2,093254	2,059539
34	2,649894	2,493616	2,380313	2,293832	2,225340	2,169562	2,123140	2,083822	2,050040
35	2,641465	2,485143	2,371781	2,285235	2,216675	2,160829	2,114300	2,074956	2,041111
36	2,633532	2,477169	2,363751	2,277143	2,208518	2,152607	2,106054	2,066608	2,032703
37	2,626052	2,469650	2,356179	2,269512	2,200826	2,144853	2,098239	2,058734	2,024771
38	2,618988	2,462548	2,349027	2,262304	2,193559	2,137528	2,090856	2,051294	2,017276
39	2,612306	2,455831	2,342262	2,255485	2,186685	2,130597	2,083869	2,044253	2,010183
40	2,605975	2,449466	2,335852	2,249024	2,180107	2,124029	2,077248	2,037580	2,003459

$\alpha = 0{,}05$									
v_2 \ v_1	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
41	2,599969	2,443429	2,329771	2,242894	2,173989	2,117797	2,070965	2,031247	1,997078
42	2,594263	2,437693	2,323994	2,237070	2,168117	2,111875	2,064994	2,025229	1,991013
43	2,588836	2,432236	2,318498	2,231530	2,162530	2,106241	2,059313	2,019502	1,985242
44	2,583667	2,427040	2,313264	2,226253	2,157208	2,100873	2,053901	2,014046	1,979743
45	2,578739	2,422085	2,308273	2,221221	2,152133	2,095755	2,048739	2,008842	1,974498
46	2,574035	2,417356	2,303509	2,216417	2,147288	2,090868	2,043811	2,003873	1,969490
47	2,569540	2,412837	2,298956	2,211827	2,142658	2,086198	2,039101	1,999124	1,964702
48	2,565241	2,408514	2,294601	2,207436	2,138229	2,081730	2,034595	1,994580	1,960121
49	2,561124	2,404375	2,290432	2,203232	2,133988	2,077452	2,030279	1,990228	1,955734
50	2,557179	2,400409	2,286436	2,199202	2,129923	2,073351	2,026143	1,986056	1,951528
55	2,539689	2,382823	2,268717	2,181333	2,111894	2,055161	2,007792	1,967547	1,932863
60	2,525215	2,368270	2,254053	2,166541	2,096968	2,040098	1,992592	1,952212	1,917396
65	2,513040	2,356028	2,241716	2,154095	2,084407	2,027419	1,979796	1,939300	1,904370
70	2,502656	2,345586	2,231192	2,143478	2,073690	2,016601	1,968875	1,928278	1,893248
75	2,493696	2,336576	2,222110	2,134314	2,064439	2,00726	1,959445	1,918759	1,883642
80	2,485885	2,328721	2,214193	2,126324	2,056373	1,999115	1,95122	1,910456	1,875262
85	2,479015	2,321812	2,207229	2,119296	2,049276	1,991949	1,943984	1,903149	1,867886
90	2,472927	2,315689	2,201056	2,113067	2,042986	1,985595	1,937567	1,896669	1,861344
95	2,467494	2,310225	2,195548	2,107506	2,037370	1,979923	1,931838	1,890884	1,855503
100	2,462615	2,305318	2,190601	2,102513	2,032328	1,974829	1,926692	1,885687	1,850255
105	2,458210	2,300888	2,186134	2,098005	2,027774	1,970229	1,922045	1,880993	1,845515
110	2,454213	2,296868	2,182082	2,093913	2,023641	1,966054	1,917827	1,876732	1,841212
115	2,450571	2,293205	2,178387	2,090184	2,019874	1,962247	1,913982	1,872847	1,837288
120	2,447237	2,289851	2,175006	2,086770	2,016426	1,958763	1,910461	1,869290	1,833695
125	2,444174	2,286771	2,171900	2,083634	2,013257	1,955562	1,907226	1,866022	1,830394
129	2,441897	2,284481	2,169591	2,081303	2,010902	1,953182	1,904821	1,863592	1,827939

$\alpha = 0,05$									
v_2 \ v_1	**15**	**20**	**24**	**30**	**40**	**50**	**60**	**80**	**100**
4	5,857805	5,802542	5,774389	5,745877	5,716998	5,699492	5,687744	5,672973	5,664064
5	4,618759	4,558131	4,527153	4,495712	4,463793	4,444406	4,431380	4,414982	4,405081
6	3,938058	3,874189	3,841457	3,808164	3,774286	3,753668	3,739797	3,722314	3,711745
7	3,510740	3,444525	3,410494	3,375808	3,340430	3,318856	3,304323	3,285983	3,274885
8	3,218406	3,150324	3,115240	3,079406	3,042778	3,020398	3,005303	2,986230	2,974674
9	3,006102	2,936455	2,900474	2,863652	2,825933	2,802843	2,787249	2,767522	2,755557
10	2,845017	2,774016	2,737248	2,699551	2,660855	2,637124	2,621077	2,600753	2,588412
11	2,718640	2,646445	2,608974	2,570489	2,530905	2,506587	2,490123	2,469246	2,456555
12	2,616851	2,543588	2,505482	2,466279	2,425880	2,401018	2,384166	2,362772	2,349753
13	2,533110	2,458882	2,420196	2,380334	2,339180	2,313811	2,296596	2,274716	2,261387
14	2,463003	2,387896	2,348678	2,308207	2,266350	2,240507	2,222950	2,200611	2,186988
15	2,403447	2,327535	2,287826	2,246789	2,204276	2,177985	2,160105	2,137331	2,123428
16	2,352223	2,275570	2,235405	2,193841	2,150711	2,123999	2,105813	2,082625	2,068455
17	2,307693	2,230354	2,189766	2,147708	2,103998	2,076888	2,058411	2,034828	2,020401
18	2,268622	2,190648	2,149665	2,107143	2,062885	2,035397	2,016643	1,992682	1,978010
19	2,234063	2,155497	2,114143	2,071186	2,026410	1,998561	1,979544	1,955221	1,940314
20	2,203274	2,124155	2,082454	2,039086	1,993819	1,965628	1,946358	1,921689	1,906554
21	2,175670	2,096033	2,054004	2,010248	1,964515	1,935997	1,916486	1,891483	1,876131
22	2,150778	2,070656	2,028319	1,984195	1,938018	1,909188	1,889445	1,864123	1,848559
23	2,128217	2,047638	2,005009	1,960537	1,913938	1,884809	1,864844	1,839213	1,823446
24	2,107673	2,026664	1,983760	1,938957	1,891955	1,862539	1,842360	1,816432	1,800468
25	2,088887	2,007471	1,964306	1,919188	1,871801	1,842111	1,821727	1,795512	1,779357
26	2,071642	1,989842	1,946428	1,901010	1,853255	1,823301	1,802719	1,776228	1,759888
27	2,055755	1,973590	1,929940	1,884236	1,836129	1,805922	1,785149	1,75839	1,741871
28	2,041071	1,958561	1,914686	1,868709	1,820263	1,789813	1,768857	1,741838	1,725146
29	2,027458	1,944620	1,900531	1,854293	1,805523	1,774838	1,753704	1,726435	1,709574
30	2,014804	1,931653	1,887360	1,840872	1,791790	1,760879	1,739574	1,712062	1,695037
31	2,003009	1,919561	1,875073	1,828345	1,778964	1,747835	1,726363	1,698616	1,681432
32	1,991990	1,908258	1,863582	1,816625	1,766956	1,735616	1,713984	1,686009	1,668670
33	1,981671	1,897669	1,852814	1,805636	1,755689	1,724147	1,702359	1,674162	1,656673
34	1,971988	1,887727	1,842701	1,795311	1,745097	1,713358	1,691420	1,663007	1,645371
35	1,962884	1,878375	1,833184	1,785591	1,735119	1,703190	1,681106	1,652484	1,634706
36	1,954308	1,869562	1,824213	1,776424	1,725703	1,693590	1,671365	1,642539	1,624621
37	1,946216	1,861242	1,815742	1,767764	1,716803	1,684511	1,662149	1,633125	1,615072
38	1,938568	1,853375	1,807729	1,759569	1,708376	1,675911	1,653416	1,624200	1,606014
39	1,931327	1,845925	1,800138	1,751803	1,700385	1,667753	1,645128	1,615724	1,597409
40	1,924463	1,838859	1,792937	1,744432	1,692797	1,660003	1,637252	1,607666	1,589224

$\alpha = 0{,}05$									
v_2 \ v_1	**15**	**20**	**24**	**30**	**40**	**50**	**60**	**80**	**100**
41	1,917946	1,832149	1,786096	1,737427	1,685582	1,652631	1,629757	1,599993	1,581428
42	1,911751	1,825767	1,779588	1,730762	1,678713	1,645608	1,622615	1,592678	1,573993
43	1,905855	1,819691	1,773391	1,724411	1,672165	1,638912	1,615803	1,585696	1,566893
44	1,900236	1,813898	1,767481	1,718354	1,665916	1,632518	1,609296	1,579024	1,560106
45	1,894875	1,808370	1,761839	1,712569	1,659945	1,626407	1,603075	1,572642	1,553612
46	1,889755	1,803089	1,756448	1,707039	1,654235	1,620560	1,597122	1,566531	1,547390
47	1,884859	1,798038	1,751291	1,701748	1,648769	1,614961	1,591417	1,560673	1,541425
48	1,880175	1,793202	1,746353	1,696679	1,643530	1,609593	1,585947	1,555053	1,535699
49	1,875687	1,788569	1,741620	1,691820	1,638505	1,604442	1,580697	1,549656	1,530199
50	1,871384	1,784125	1,737080	1,687157	1,633682	1,599495	1,575654	1,544469	1,524911
55	1,852280	1,764379	1,716893	1,666408	1,612191	1,577435	1,553142	1,521285	1,501251
60	1,836437	1,747984	1,700117	1,649141	1,594273	1,559011	1,534314	1,501853	1,481386
65	1,823086	1,734152	1,685951	1,634544	1,579098	1,543385	1,518326	1,485316	1,464455
70	1,811681	1,722325	1,673829	1,622040	1,566078	1,52996	1,504572	1,471064	1,449840
75	1,801825	1,712096	1,663338	1,611207	1,554782	1,518297	1,492612	1,458647	1,437090
80	1,793222	1,703160	1,654168	1,601730	1,544887	1,508069	1,482111	1,447728	1,425862
85	1,785647	1,695287	1,646084	1,593369	1,536147	1,499025	1,472817	1,438048	1,415896
90	1,778927	1,688298	1,638904	1,585937	1,528369	1,490968	1,464531	1,429404	1,406986
95	1,772924	1,682051	1,632483	1,579288	1,521402	1,483745	1,457096	1,421637	1,398970
100	1,767530	1,676434	1,626708	1,573302	1,515125	1,477231	1,450386	1,414618	1,391720
105	1,762656	1,671357	1,621485	1,567886	1,509441	1,471327	1,444299	1,408244	1,385127
110	1,758230	1,666744	1,616739	1,562962	1,504268	1,465951	1,438753	1,402428	1,379106
115	1,754193	1,662536	1,612407	1,558465	1,499540	1,461034	1,433676	1,397099	1,373585
120	1,750497	1,658680	1,608437	1,554343	1,495202	1,456519	1,429013	1,392198	1,368503
125	1,747099	1,655135	1,604786	1,550549	1,491208	1,452360	1,424714	1,387676	1,363808
129	1,744573	1,652498	1,602069	1,547725	1,488234	1,449260	1,421509	1,384301	1,360303

Values of ***k*** for two-sided statistical tolerance interval									
1 − α	**0,90**			**0,95**			**0,99**		
n* \ *p	**0,90**	**0,95**	**0,99**	**0,90**	**0,95**	**0,99**	**0,90**	**0,95**	**0,99**
2	15,5124	18,2208	23,4235	31,0923	36,5192	46,7452	155,5690	182,7200	234,8769
3	5,7881	6,8233	8,8186	8,3060	9,7888	12,6471	18,7825	22,1308	28,5857
4	4,1571	4,9127	6,3722	5,3681	6,3411	8,2207	9,4162	11,1178	14,4054
5	3,4993	4,1425	5,3868	4,2907	5,0769	6,5980	6,6550	7,8698	10,2201
6	3,1406	3,7226	4,8498	3,7326	4,4222	5,7578	5,3832	6,3735	8,2916
7	2,9128	3,4558	4,5085	3,3896	4,0196	5,2411	4,6576	5,5196	7,1907
8	2,7542	3,2699	4,2707	3,1561	3,7456	4,8893	4,1887	4,9677	6,4790
9	2,6368	3,1323	4,0945	2,9861	3,5459	4,6328	3,8602	4,5810	5,9802
10	2,5460	3,0258	3,9580	2,8564	3,3935	4,4370	3,6167	4,2942	5,6102
11	2,4734	2,9406	3,8488	2,7537	3,2728	4,2818	3,4286	4,0726	5,3242
12	2,4140	2,8707	3,7591	2,6703	3,1747	4,1556	3,2786	3,8959	5,0960
13	2,3643	2,8123	3,6840	2,6011	3,0932	4,0506	3,1561	3,7514	4,9093
14	2,3220	2,7625	3,6201	2,5425	3,0242	3,9617	3,0538	3,6309	4,7535
15	2,2855	2,7196	3,5649	2,4922	2,9650	3,8853	2,9672	3,5286	4,6212
16	2,2537	2,6821	3,5166	2,4486	2,9135	3,8189	2,8926	3,4406	4,5074
17	2,2257	2,6491	3,4741	2,4103	2,8684	3,7606	2,8278	3,3641	4,4084
18	2,2008	2,6197	3,4362	2,3764	2,8283	3,7089	2,7708	3,2968	4,3212
19	2,1785	2,5934	3,4122	2,3461	2,7926	3,6626	2,7203	3,2371	4,2439
20	2,1584	2,5697	3,3716	2,3188	2,7604	3,6210	2,6752	3,1838	4,1748
21	2,1401	2,5482	3,3437	2,2942	2,7313	3,5834	2,6347	3,1359	4,1126
22	2,1235	2,5285	3,3183	2,2718	2,7048	3,5490	2,5979	3,0924	4,0563
23	2,1083	2,5105	3,2951	2,2516	2,6806	3,5177	2,5645	3,0529	4,0050
24	2,0943	2,4940	3,2736	2,2325	2,6583	3,4888	2,5340	3,0168	3,9580
25	2,0813	2,4787	3,2538	2,2151	2,6378	3,4622	2,5060	2,9836	3,9149
30	2,0289	2,4166	3,1734	2,1452	2,5549	3,3546	2,3940	2,8510	3,7425
40	1,9611	2,4479	3,0688	2,0624	2,4484	3,2160	2,2529	2,6836	3,5144
50	1,9184	2,3948	3,0027	1,9991	2,3816	3,1288	2,1660	2,5805	3,3898
60	1,8885	2,2500	2,9564	1,9599	2,3351	3,0681	2,1063	2,5095	3,2970
70	1,8662	2,2236	2,9218	1,9308	2,3005	3,0228	2,0623	2,4571	3,2284
80	1,8489	2,2029	2,8947	1,9082	2,2736	2,9875	2,0282	2,4165	3,1753
90	1,8348	2,1862	2,8729	1,8899	2,2519	2,9591	2,0009	2,3840	3,1327
100	1,8232	2,1724	2,8548	1,8749	2,2339	2,9356	1,9784	2,3573	3,0976

Values of ***k*** for one-sided statistical tolerance interval									
1 − α	**0,90**			**0,95**			**0,99**		
n \ ***p***	**0,90**	**0,95**	**0,99**	**0,90**	**0,95**	**0,99**	**0,90**	**0,95**	**0,99**
2	10,2528	13,0898	18,5001	20,5815	25,2597	37,0936	103,0287	131,4263	185,6170
3	4,2582	5,3115	7,3405	6,1553	7,6560	10,5528	13,9955	17,3702	23,8956
4	3,1879	3,9566	5,4383	4,1620	5,1439	7,0424	7,3799	9,0835	12,3873
5	2,7424	3,3999	4,6660	3,4067	4,2027	5,7411	5,3618	6,5784	8,9391
6	2,4937	3,0919	4,2426	3,0063	3,7077	5,0620	4,4111	5,4056	7,3346
7	2,3327	2,8938	3,9721	2,7555	3,3995	4,6418	3,8592	4,7279	6,4120
8	2,2186	2,7543	3,7826	2,5820	3,1873	4,3539	3,4973	4,2853	5,8118
9	2,1329	2,6500	3,6415	2,4538	3,0313	4,1431	3,2405	3,9723	5,3889
10	2,0657	2,5684	3,5317	2,3547	2,9110	3,9812	3,0480	3,7384	5,0738
11	2,0113	2,5027	3,4435	2,2754	2,8150	3,8524	2,8977	3,5562	4,8291
12	1,9662	2,4483	3,3707	2,2102	2,7364	3,7471	2,7768	3,4100	4,6331
13	1,9281	2,4025	3,3095	2,1555	2,6706	3,6592	2,6770	3,2896	4,4721
14	1,8954	2,3632	3,2572	2,1088	2,6145	3,5846	2,5932	3,1886	4,3372
15	1,8669	2,3290	3,2119	2,0684	2,5661	3,5202	2,5215	3,1024	4,2224
16	1,8418	2,2990	3,1721	2,0330	2,5237	3,4640	2,4595	3,0279	4,1233
17	1,8195	2,2725	3,1369	2,0018	2,4863	3,4145	2,4051	2,9628	4,0367
18	1,7996	2,2487	3,1055	1,9738	2,4530	3,3704	2,3571	2,9052	3,9604
19	1,7816	2,2273	3,0772	1,9487	2,4231	3,3309	2,3142	2,8539	3,8925
20	1,7653	2,2078	3,0516	1,9260	2,3961	3,2952	2,2757	2,8079	3,8316
21	1,7503	2,1901	3,0283	1,9054	2,3715	3,2628	2,2409	2,7663	3,7767
22	1,7367	2,1739	3,0069	1,8865	2,3490	3,2332	2,2092	2,7286	3,7268
23	1,7241	2,1590	2,9873	1,8691	2,3284	3,2061	2,1802	2,6941	3,6813
24	1,7124	2,1452	2,9692	1,85300	2,3093	3,1811	2,1536	2,6624	3,6396
25	1,70161	2,1323	2,9524	1,8382	2,2917	3,1580	2,1291	2,6332	3,6011
30	6571	2,0799	2,8838	1,7774	2,2199	3,0640	2,0299	2,5155	3,4466
40	1,5979	2,0103	2,7932	1,6972	2,1255	2,9410	1,9018	2,3642	3,2486
50	1,5595	1,9653	2,7349	1,6456	2,0650	2,8625	1,8208	2,2689	3,1247
60	1,5321	1,9333	2,6936	1,6090	2,0222	2,8071	1,7641	2,2024	3,0383
70	1,5113	1,9091	2,6623	1,5813	1,9899	2,7654	1,7216	2,1527	2,9740
80	1,4948	1,8899	2,6377	1,5594	1,9645	2,7327	1,6883	2,1138	2,9238
90	1,4813	1,8743	2,6177	1,5416	1,9438	2,7061	1,6614	2,0824	2,8832
100	1,4701	1,8613	2,6010	1,5268	1,9266	2,6840	1,6390	2,0563	2,8497

SHAPIRO – WILK TEST – coefficients $a_i(n)$

i \ n	7	8	9	10	11	12	13	14
1	0,6233	0,6052	0,5888	0,5739	0,5601	0,5475	0,5359	0,5251
2	0,3031	0,3164	0,3244	0,3291	0,3315	0,3325	0,3325	0,3318
3	0,1401	0,1743	0,1976	0,2141	0,2260	0,2347	0,2412	0,2460
4	0	0,0561	0,0947	0,1224	0,1429	0,1586	0,1707	0,1802
5	0	0	0	0,0399	0,0695	0,0922	0,1099	0,1240
6	0	0	0	0	0	0,0303	0,0539	0,0727
7	0	0	0	0	0	0	0	0,0240

i \ n	15	16	17	18	19	20	21	22
1	0,5150	0,5056	0,4968	0,4886	0,4808	0,4734	0,4643	0,4590
2	0,3306	0,3290	0,3273	0,3253	0,3232	0,3211	0,3185	0,3156
3	0,2495	0,2521	0,2540	0,2553	0,2565	0,2565	0,2578	0,2571
4	0,1878	0,1939	0,1988	0,2027	0,2085	0,2085	0,2119	0,2131
5	0,1353	0,1447	0,1524	0,1587	0,1686	0,1686	0,1736	0,1764
6	0,0880	0,1005	0,1109	0,1197	0,1334	0,1334	0,1399	0, 1443
7	0,0433	0,0593	0,0725	0,0837	0,1013	0,1013	0,1092	0,1150
8	0	0,0196	0,0359	0,0496	0,0711	0,0711	0,0804	0,0878
9	0	0	0	0,0163	0,1422	0,0422	0,0530	0,0618
10	0	0	0	0	0	0,0140	0,0263	0,0368
11	0	0	0	0	0	0	0	0,0122

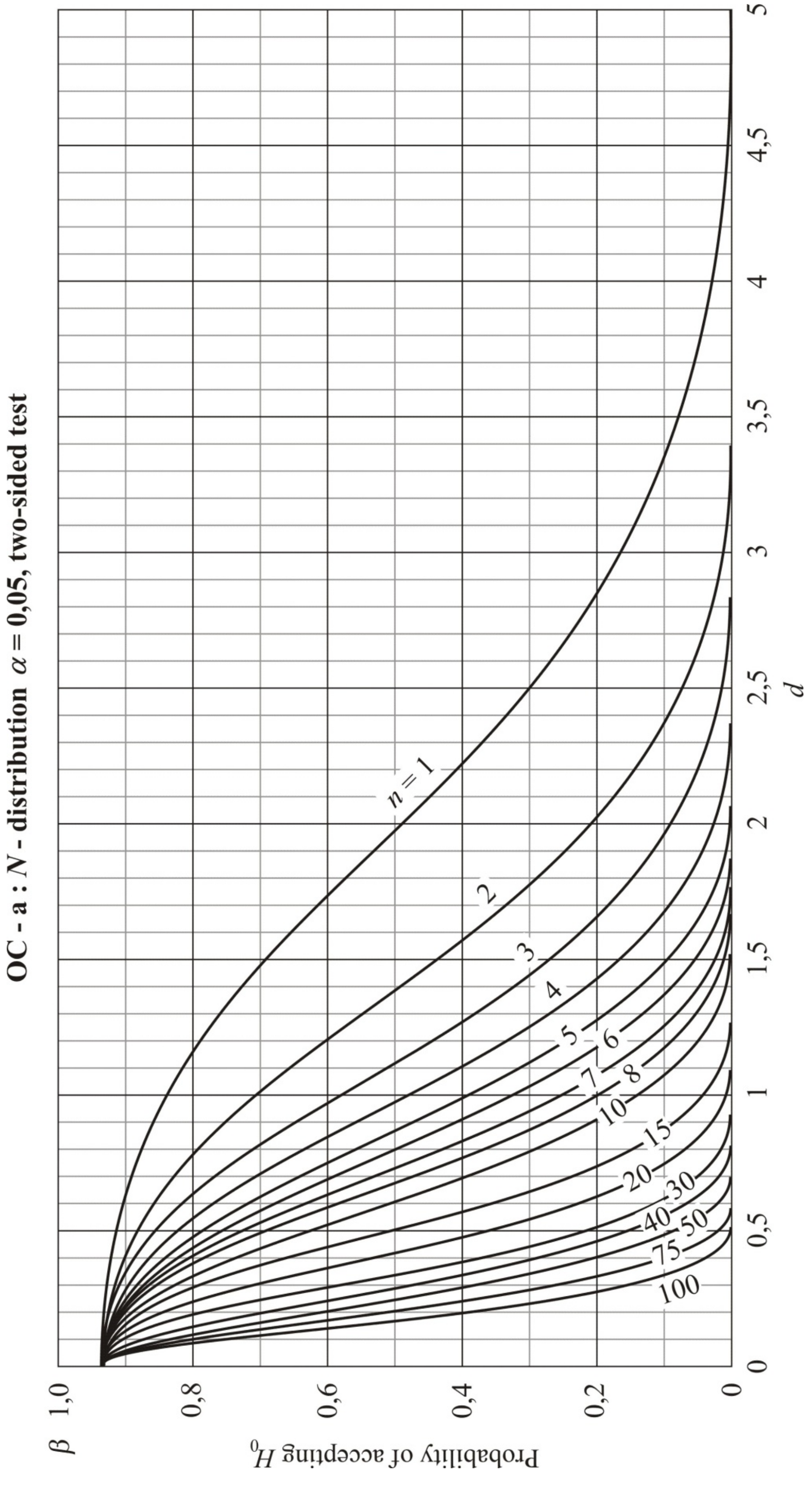
i \ n	23	24	25	26	27	28	29	30
1	0,4542	0,4493	0,4450	0,4407	0,4366	0,4328	0,4291	0,4254
2	0,3126	0,3098	0,3069	0,3043	0,3018	0,2992	0,2968	0,2944
3	0,2563	0,2554	0,2543	0,2533	0,2522	0,2510	0,2499	0,2487
4	0,2139	0,2145	0,2148	0,2151	0,2152	0,2151	0,2150	0,2148
5	0,1787	0,1807	0,1822	0,1836	0,1848	0,1857	0,1864	0,1870
6	0,1480	0,1512	0,1539	0,1563	0,1584	0,1601	0,1616	0,1630
7	0,1201	0,1245	0,1283	0,1316	0,1346	0,1372	0,1395	0,1415
8	0,0941	0,0997	0,1046	0,1089	0,1128	0,1162	0,1192	0,1219
9	0,0696	0,0764	0,0823	0,0876	0,0923	0,0965	0,1002	0,1036
10	0,0459	0,0539	0,0610	0,0672	0,0728	0,0778	0,0822	0,0862
11	0,0228	0,0320	0,0403	0,0476	0,0540	0,0598	0,0650	0,0697
12	0	0,0107	0,0200	0,0284	0,0358	0,0424	0,0483	0,0537
13	0	0	0	0,0094	0,0178	0,0253	0,0320	0,0381
14	0	0	0	0	0	0,0084	0,0159	0,0227
15	0	0	0	0	0	0	0	0,0076

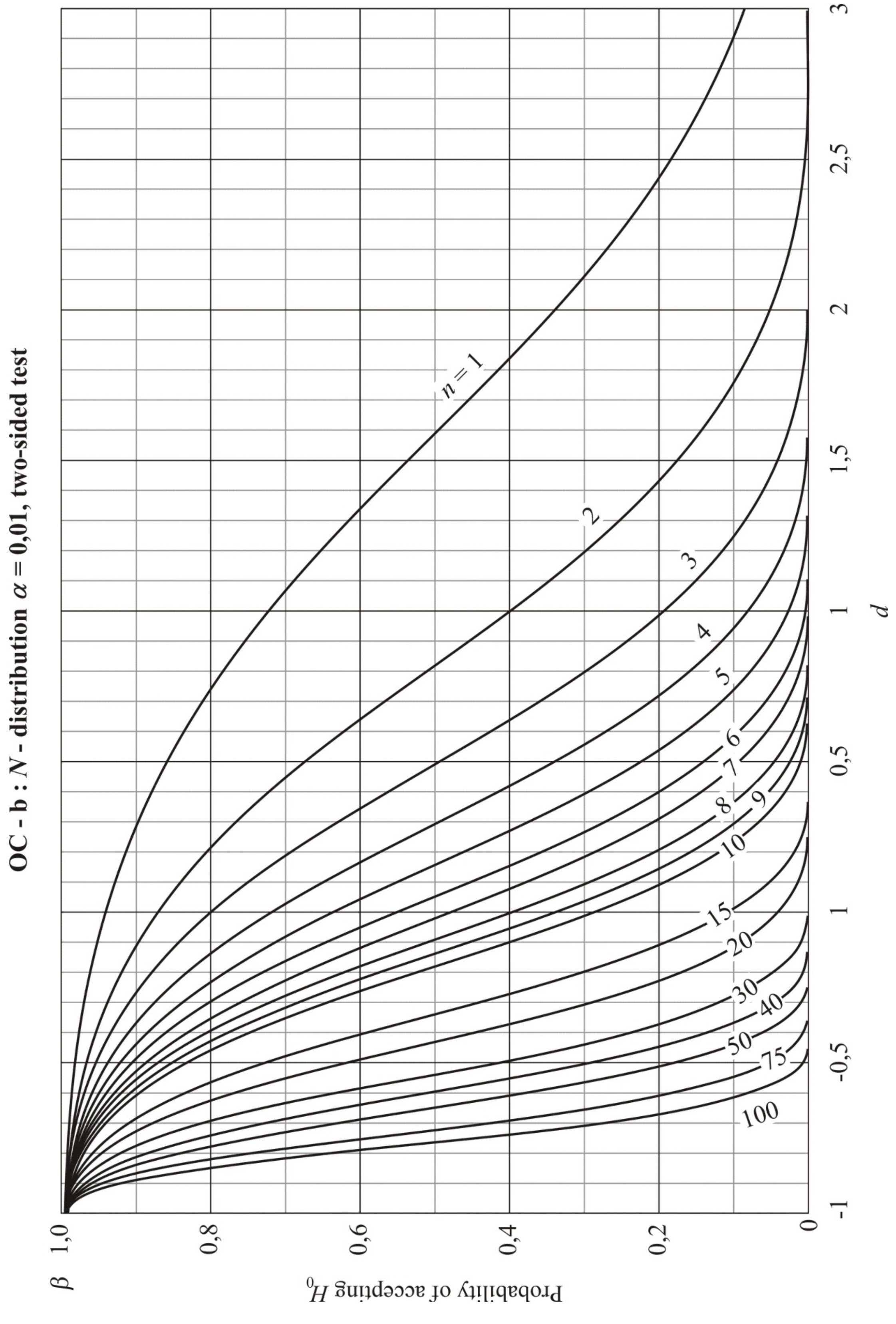
SHAPIRO – WILK TEST

Percentiles $w_\alpha(n)$ of Shapiro – Wilk statistic *W*: $P(W(n) \le W_\alpha(n)) = \alpha$

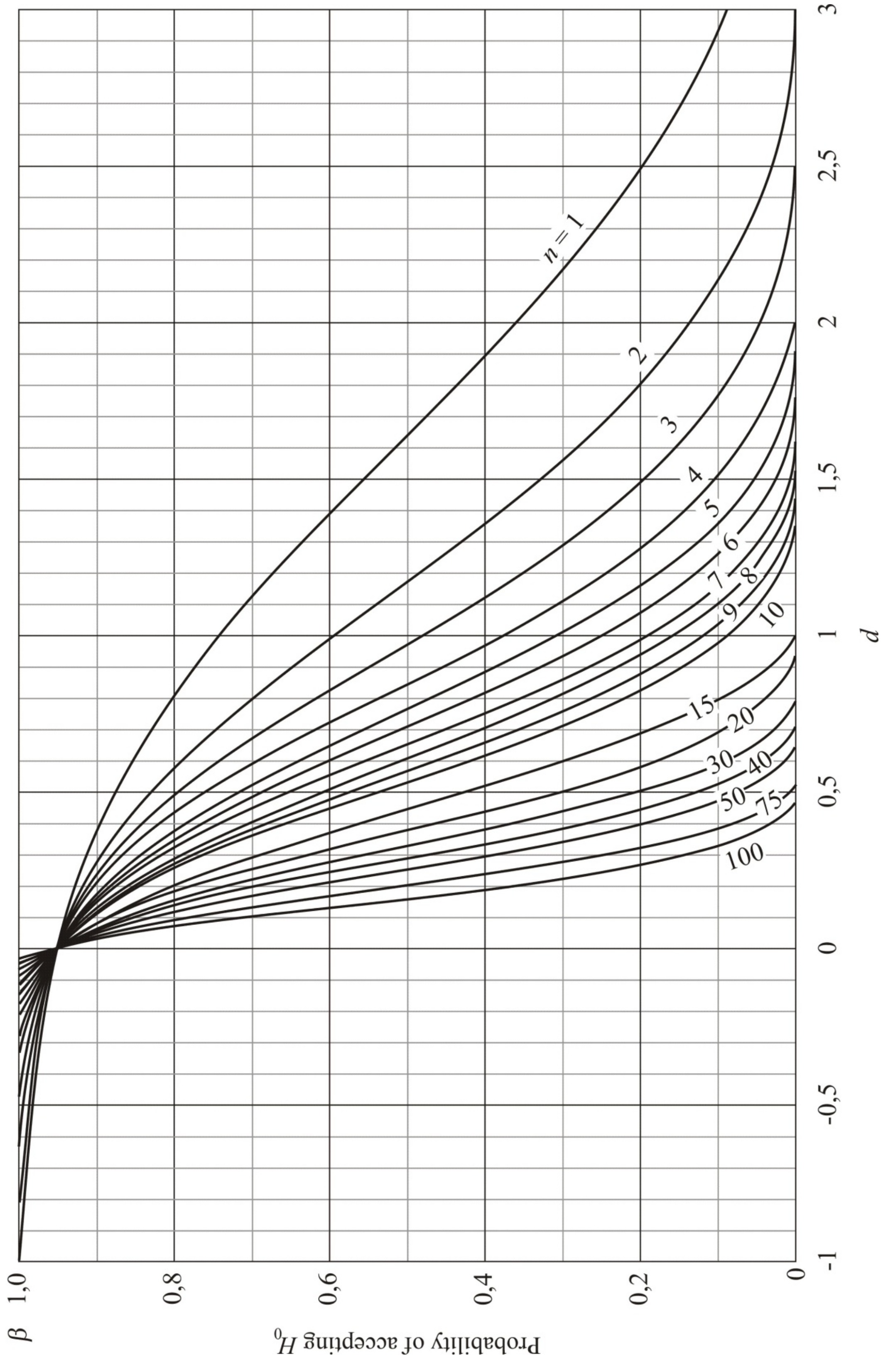
n	$\alpha = 0,01$	$\alpha = 0,05$	***n***	$\alpha = 0,01$	$\alpha = 0,05$	***n***	$\alpha = 0,01$	$\alpha = 0,05$
7	0,730	0,803	**15**	0,835	0,881	**23**	0,881	0,914
8	0,749	0,818	**16**	0,844	0,887	**24**	0,884	0,916
9	0,764	0,826	**17**	0,851	0,892	**25**	0,888	0,918
10	0,781	0,842	**18**	0,858	0,897	**26**	0,891	0,920
11	0,792	0,850	**19**	0,863	0,901	**27**	0,894	0,923
12	0,805	0,859	**20**	0,868	0,905	**28**	0,896	0,924
13	0,814	0,866	**21**	0,873	0,908	**29**	0,898	0,926
14	0,825	0,874	**22**	0,878	0,911	**30**	0,900	0,927

OPERATING CHARACTERISTIC CURVES

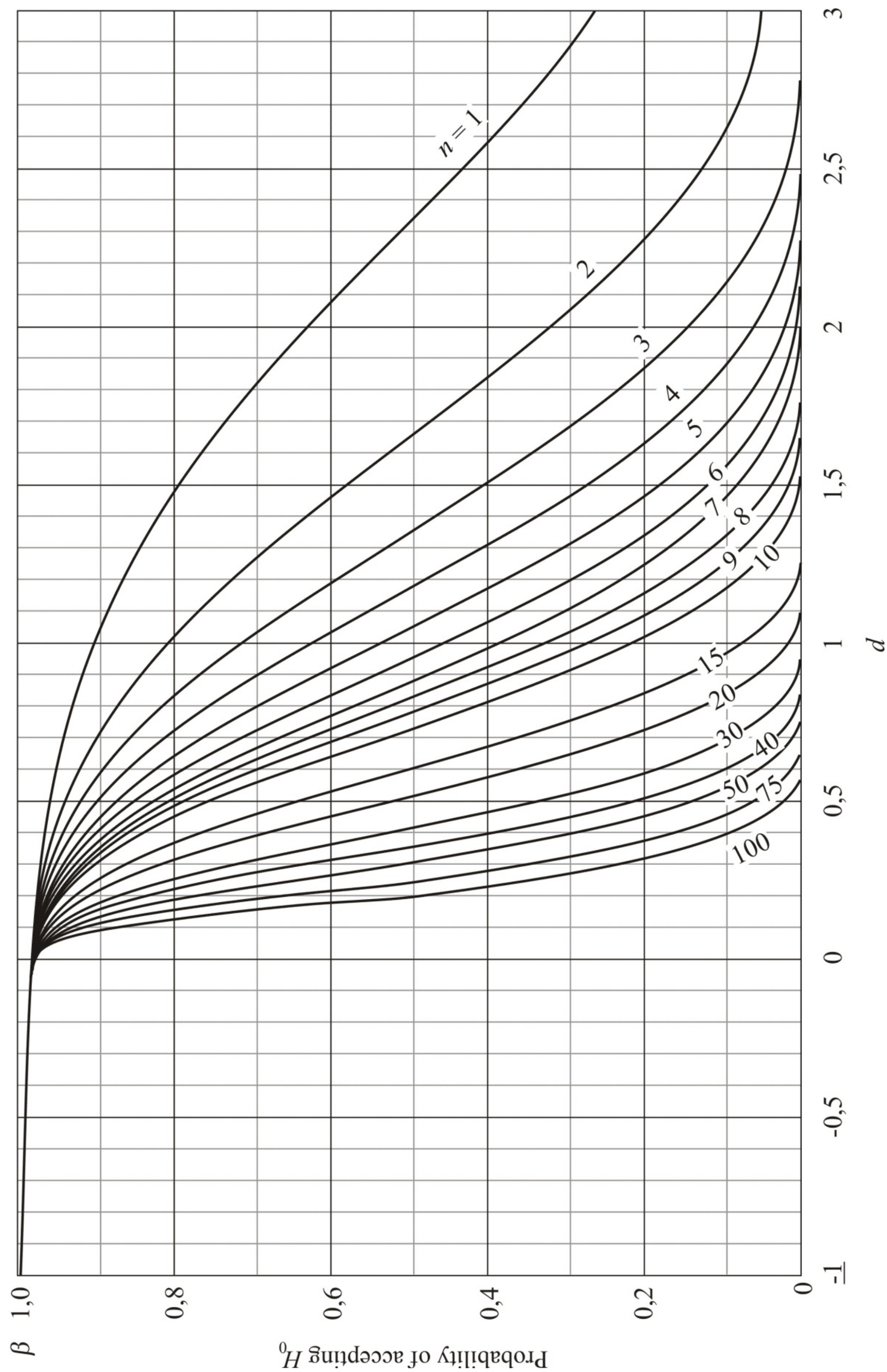


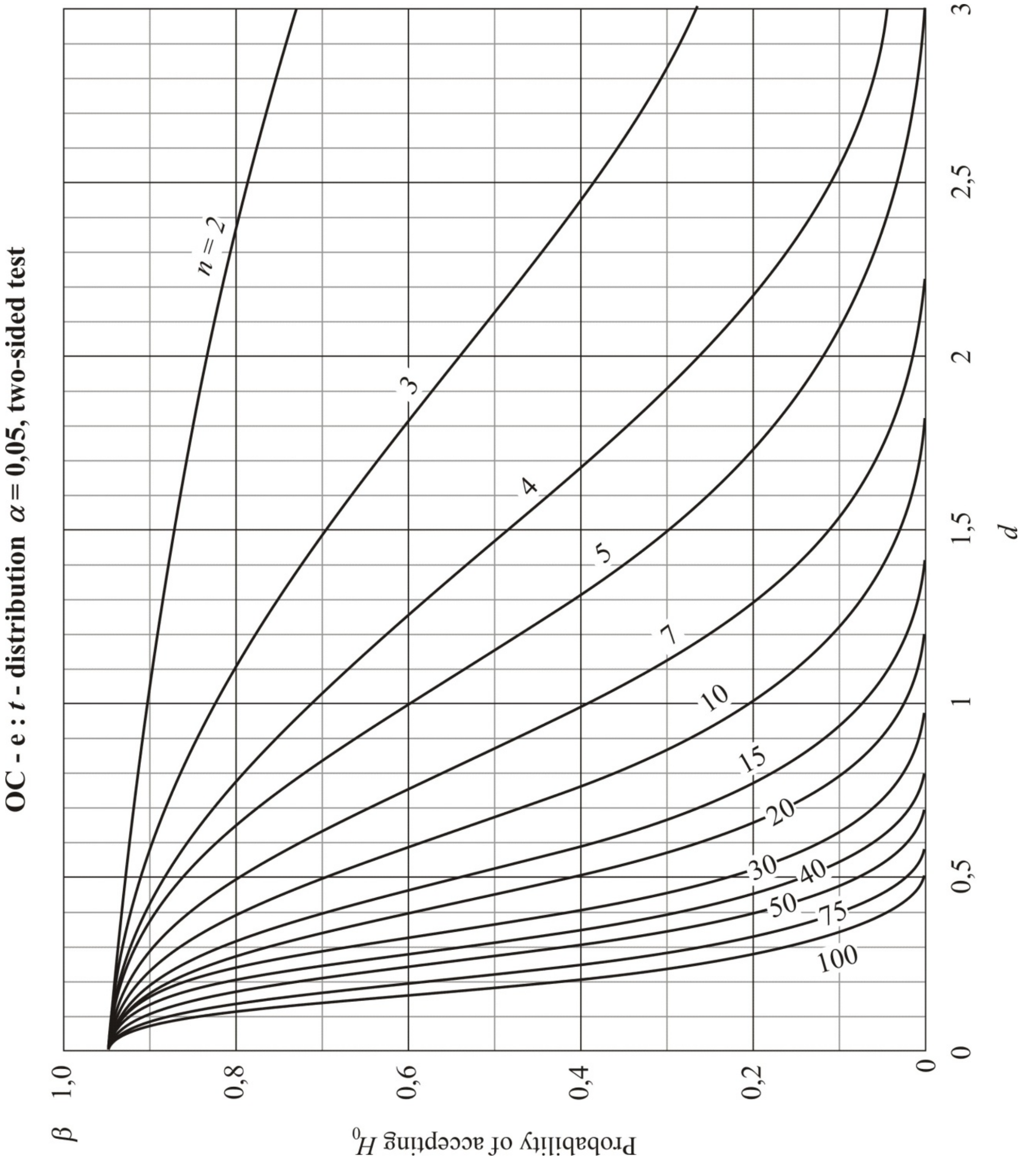

OC - b : N - distribution α = 0,01, two-sided test
β
Probability of accepting H_0
1,0
0,8
0,6
0,4
0,2
0
-1
-0,5
1
0,5
1
1,5
2
2,5
3
d
n = 1
2
3
4
5
6
7
8
9
10
15
20
30
40
50
75
100

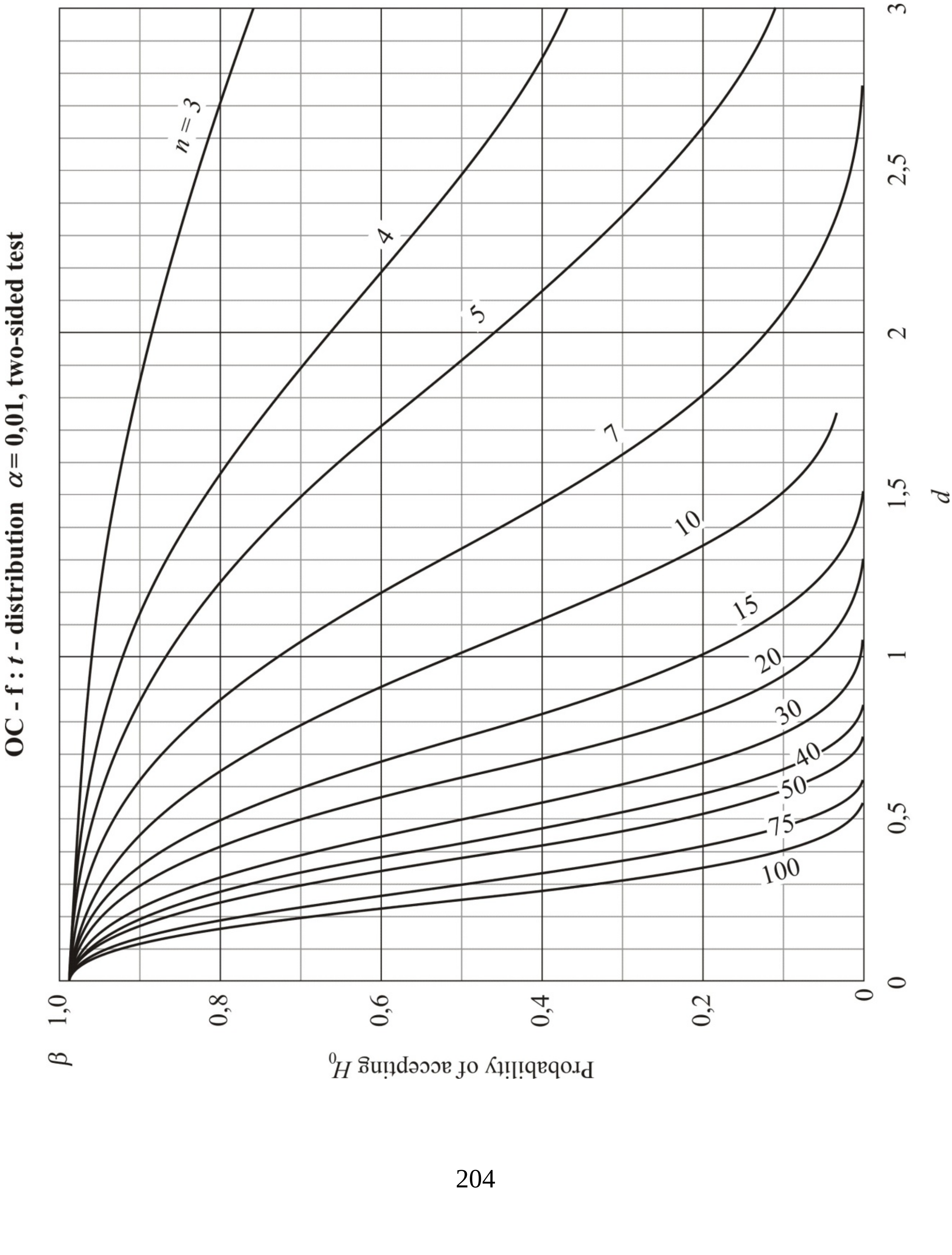
OC - c : N - distribution $\alpha = 0{,}05$, one-sided test



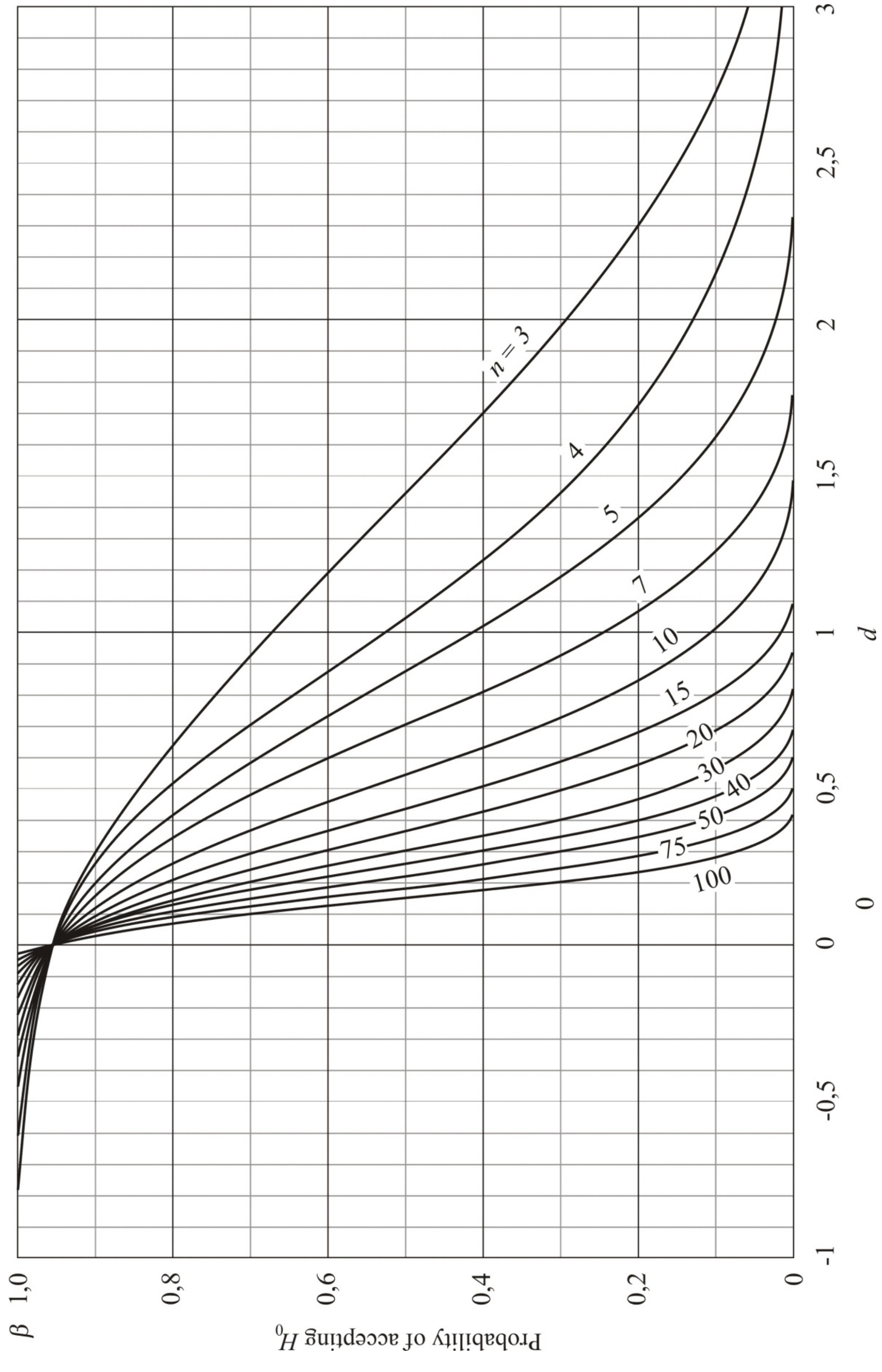
OC - d : N - distribution α = 0,01, one-sided test



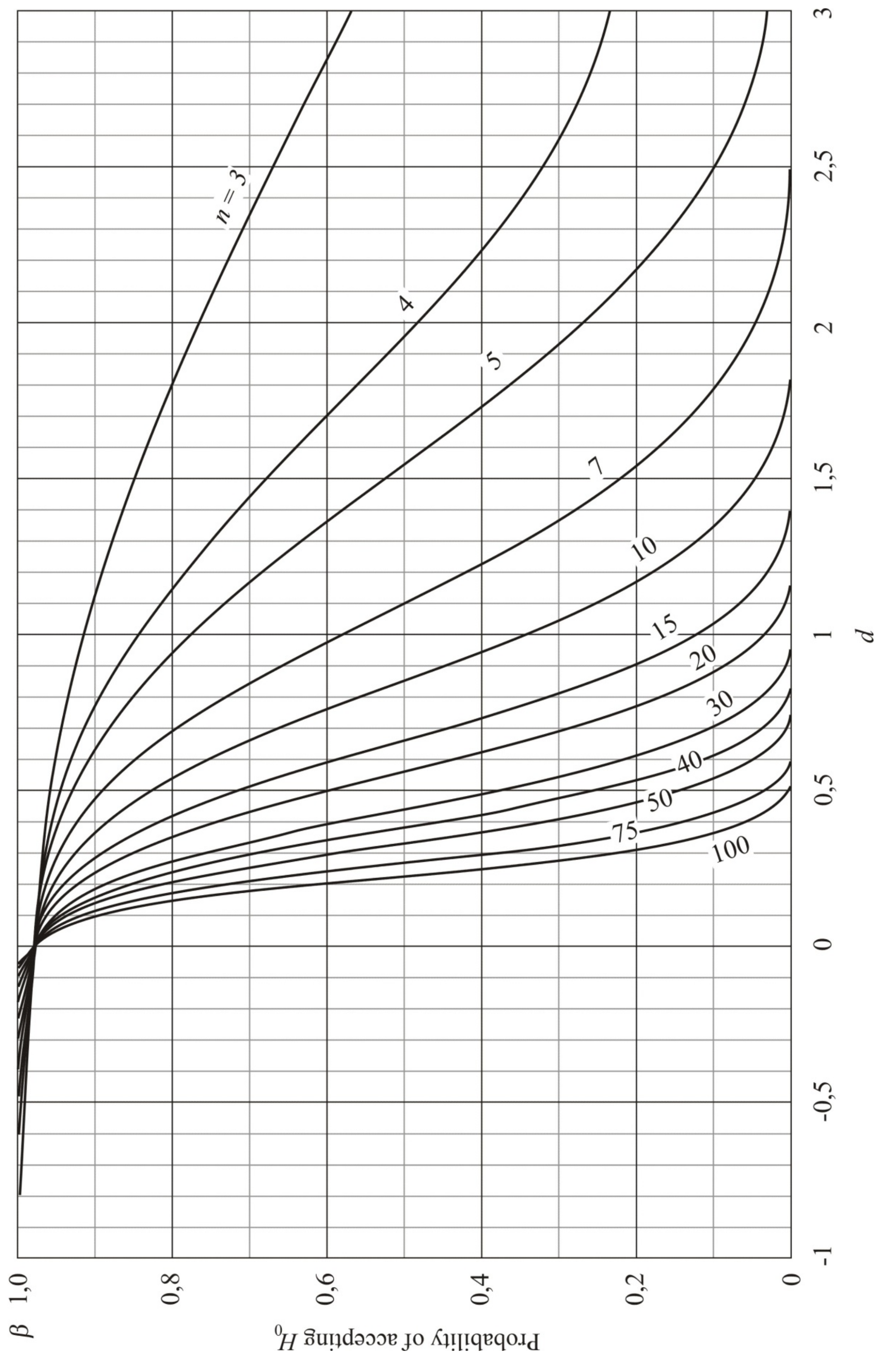

OC - e : t - distribution α = 0,05, two-sided test
β
Probability of accepting H0
1,0
0,8
0,6
0,4
0,2
0
0
0,5
1
1,5
2
2,5
3
d
n = 2
3
4
5
7
10
15
20
30
40
50
75
100


OC - f : t - distribution α = 0,01, two-sided test
β
Probability of accepting H_0
1,0
0,8
0,6
0,4
0,2
0
0
0,5
1
1,5
2
2,5
3
d
n = 3
4
5
7
10
15
20
30
40
50
75
100

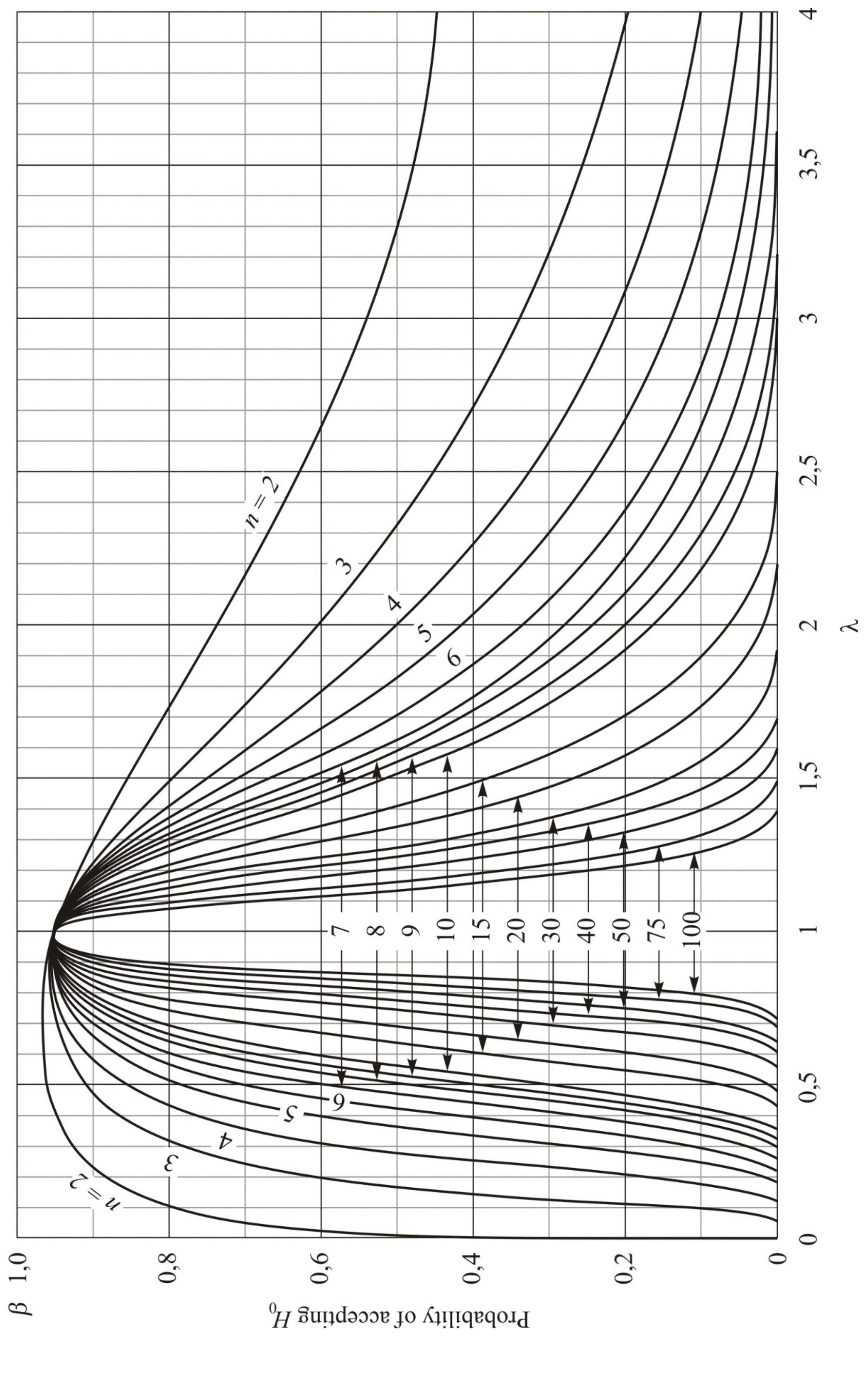
OC - g : *t* - distribution $\alpha = 0{,}05$, one-sided test



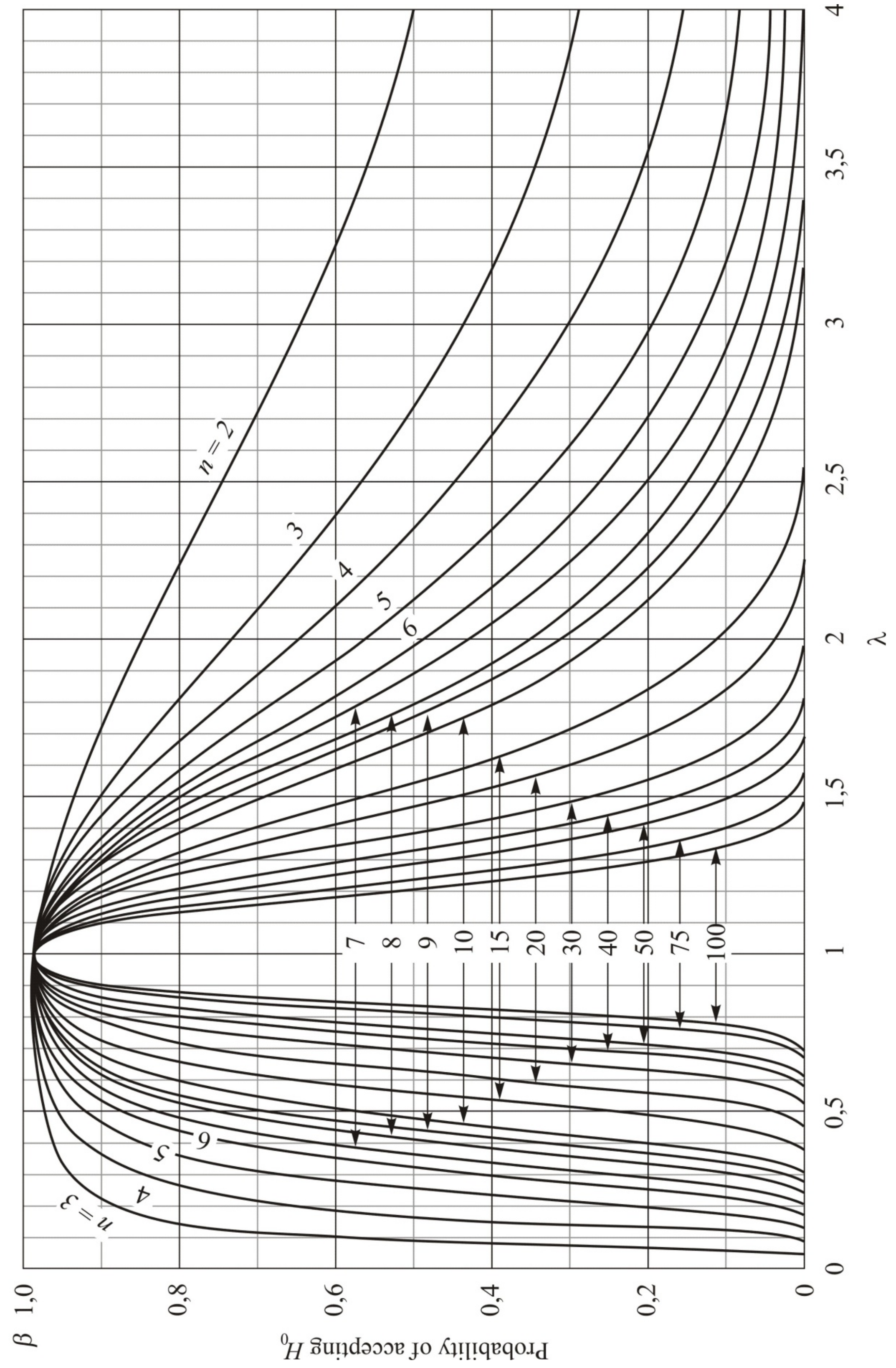
OC - h : t - distribution α = 0,01, one-sided test

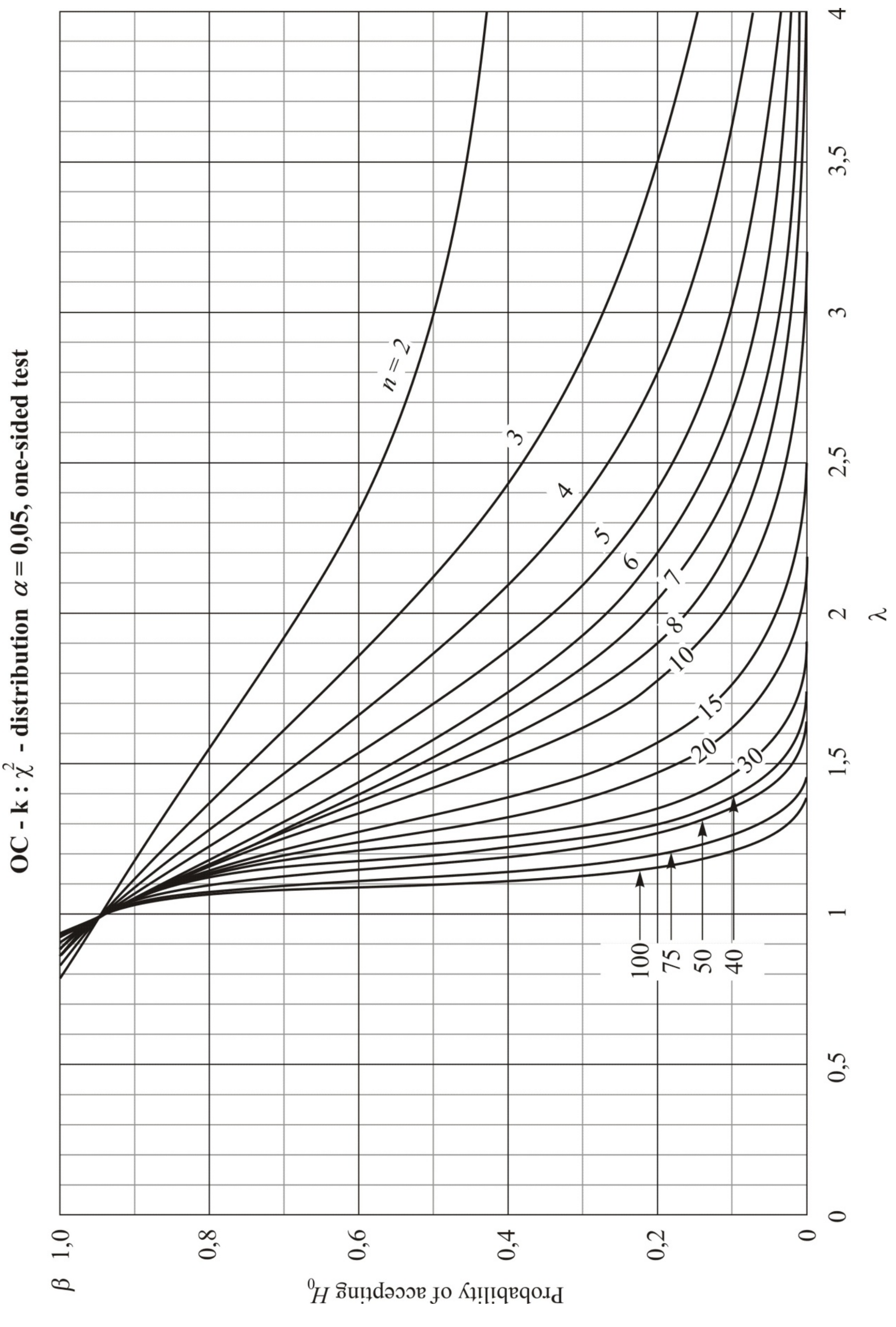


OC - i : χ^2 - distribution $\alpha = 0,05$, two-sided test

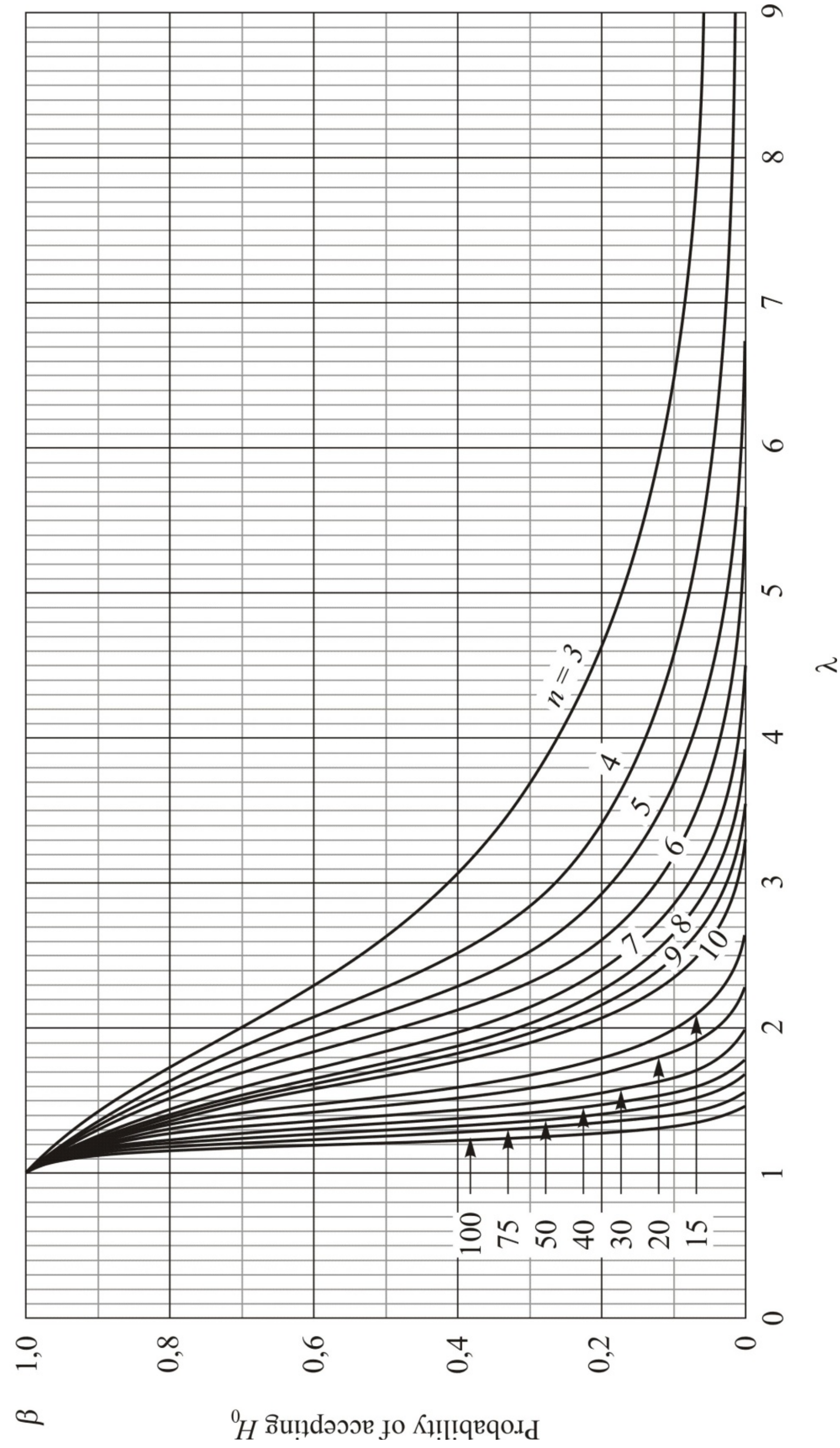


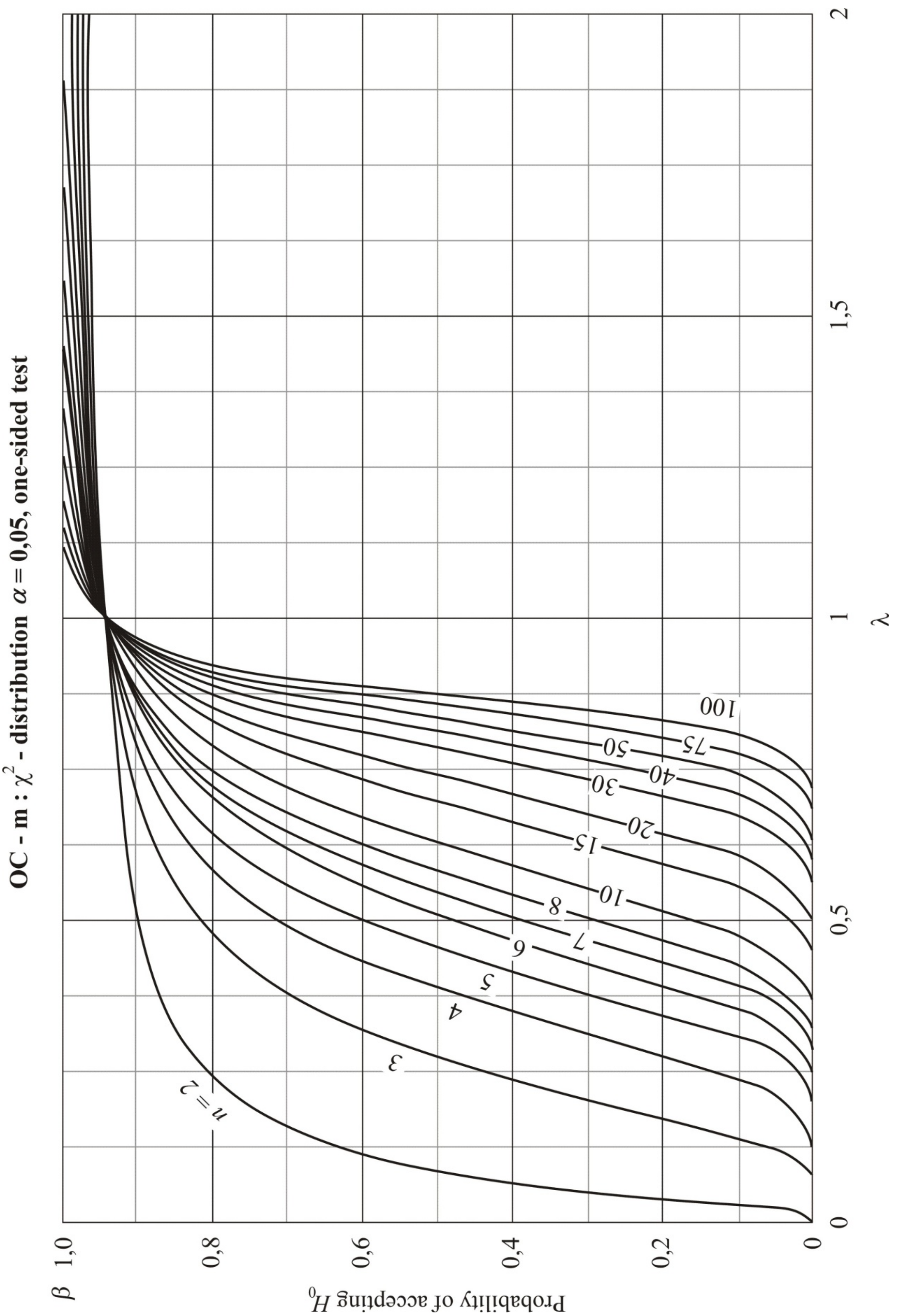
OC - j : χ^2 - distribution $\alpha = 0{,}01$, two-sided test

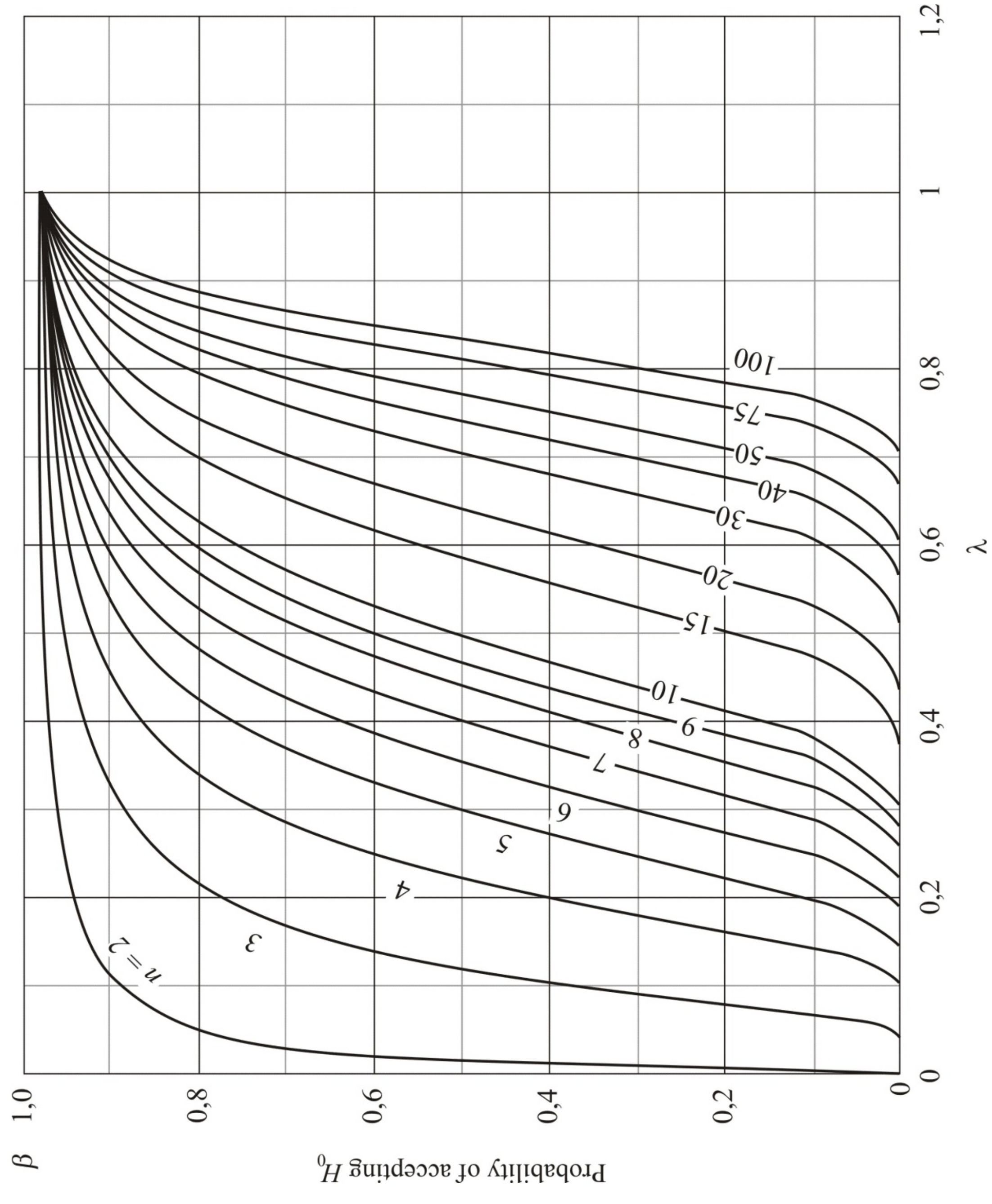



OC - k : χ² - distribution α = 0,05, one-sided test
n = 2
3
4
5
6
7
8
10
15
20
30
100
75
50
40
β 1,0
0,8
0,6
0,4
0,2
0
Probability of accepting H0
0
0,5
1
1,5
2
2,5
3
3,5
4
λ

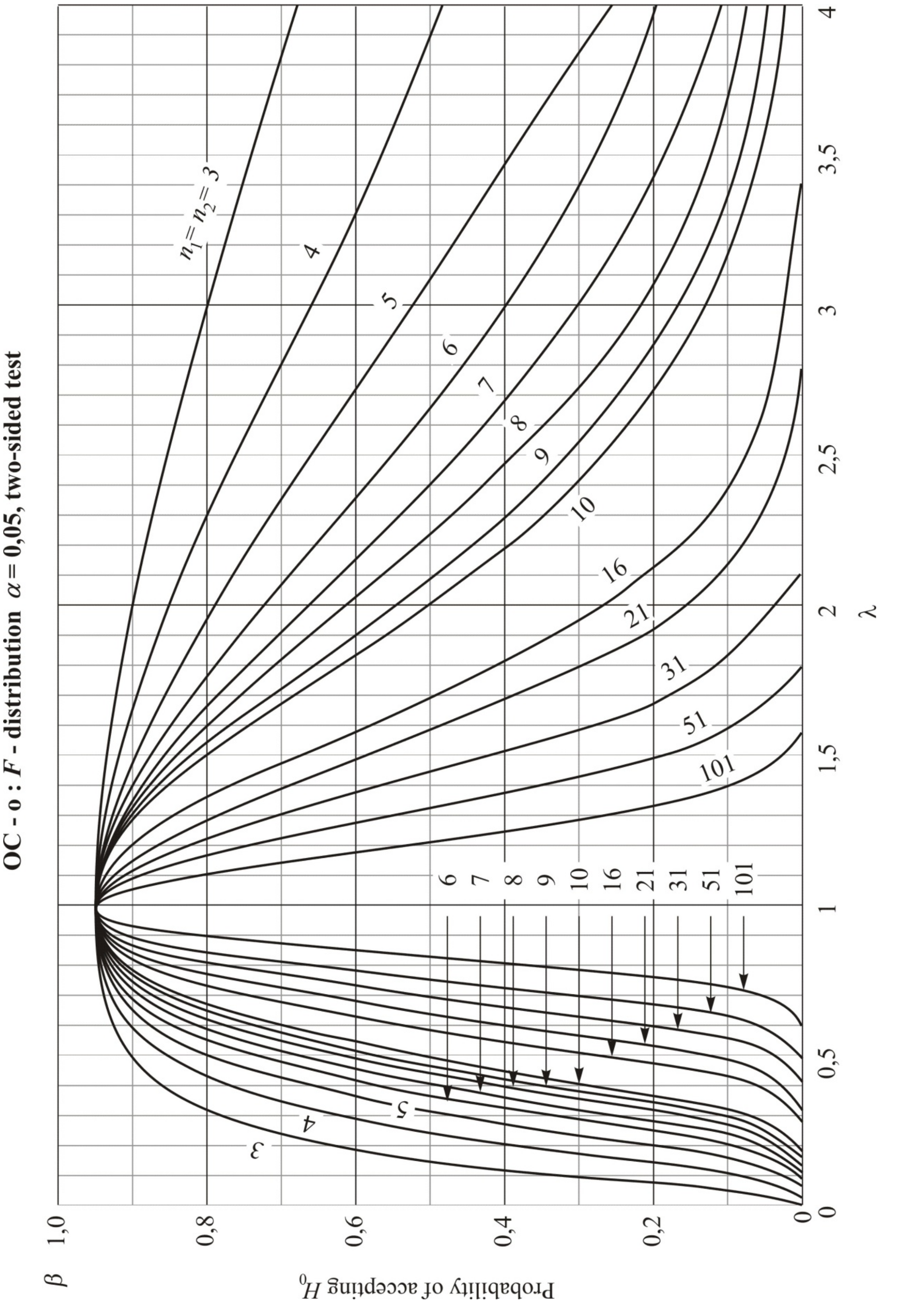
OC - 1 : χ^2 - distribution $\alpha = 0{,}01$, one-sided test



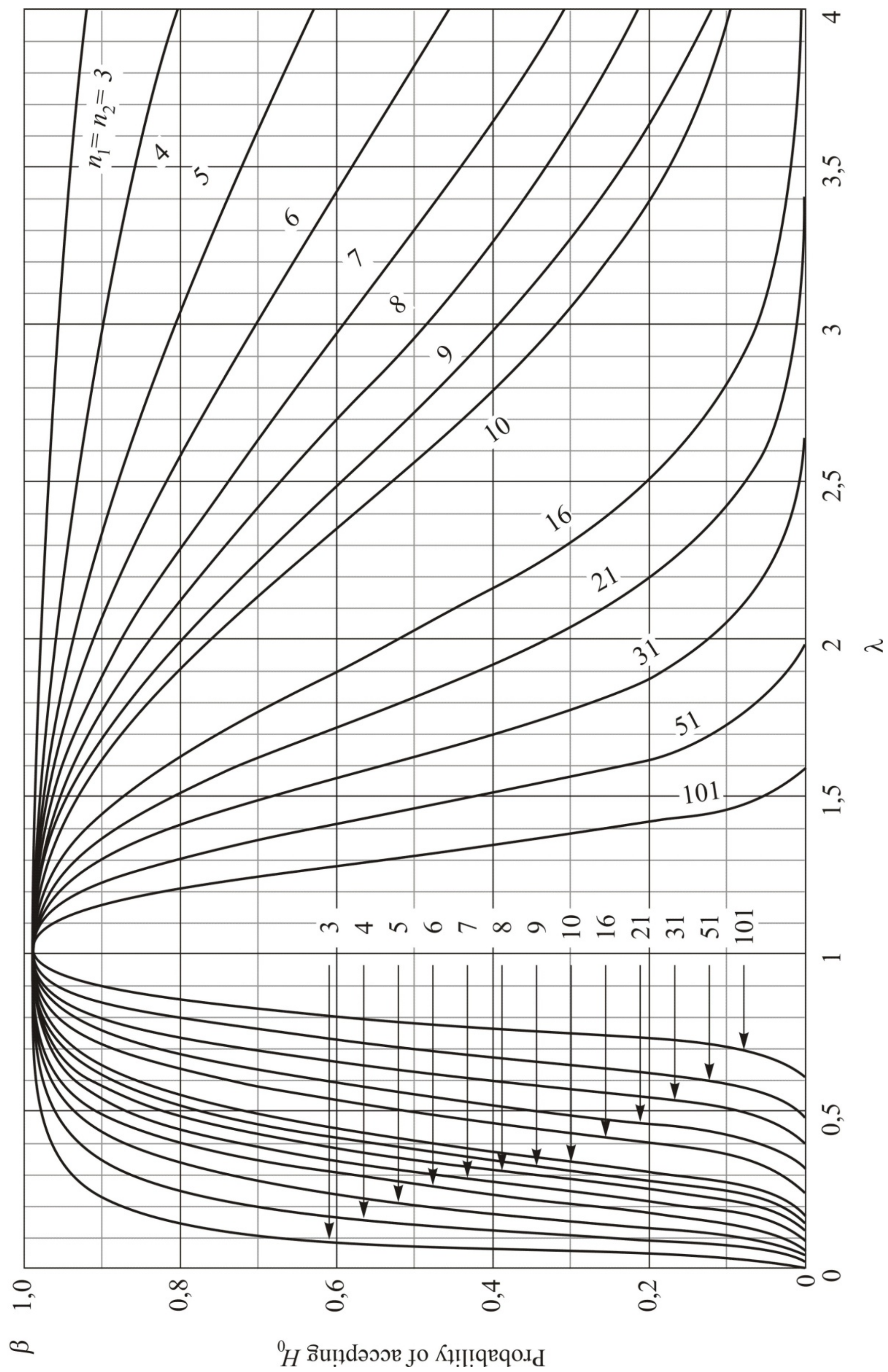

OC - m : χ^2 - distribution α = 0,05, one-sided test
β
Probability of accepting H_0
1,0
0,8
0,6
0,4
0,2
0
0
0,5
1
1,5
2
λ
n = 2
3
4
5
6
7
8
10
15
20
30
40
50
75
100



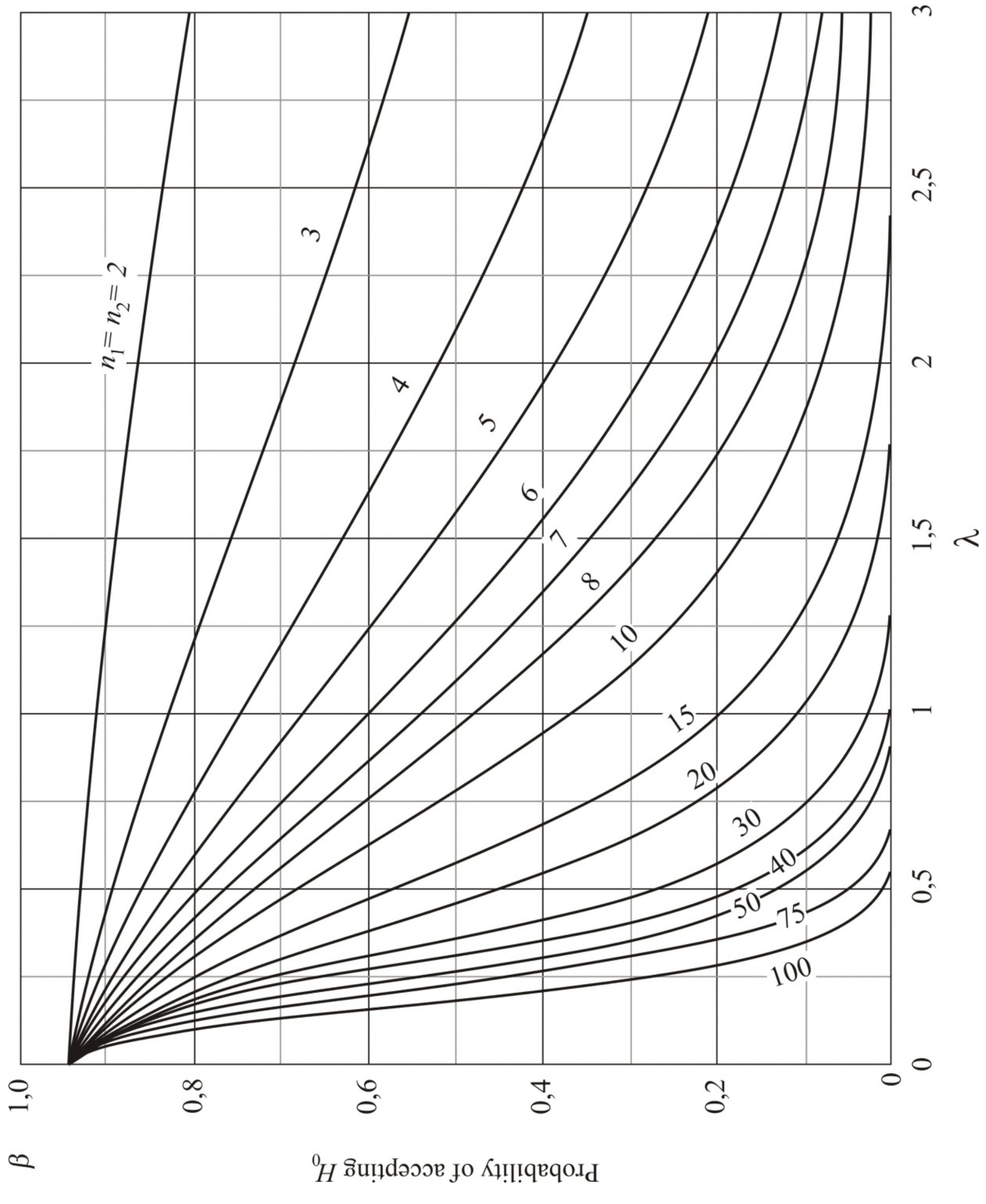
OC - n : χ^2 - distribution $\alpha = 0{,}01$, one-sided test

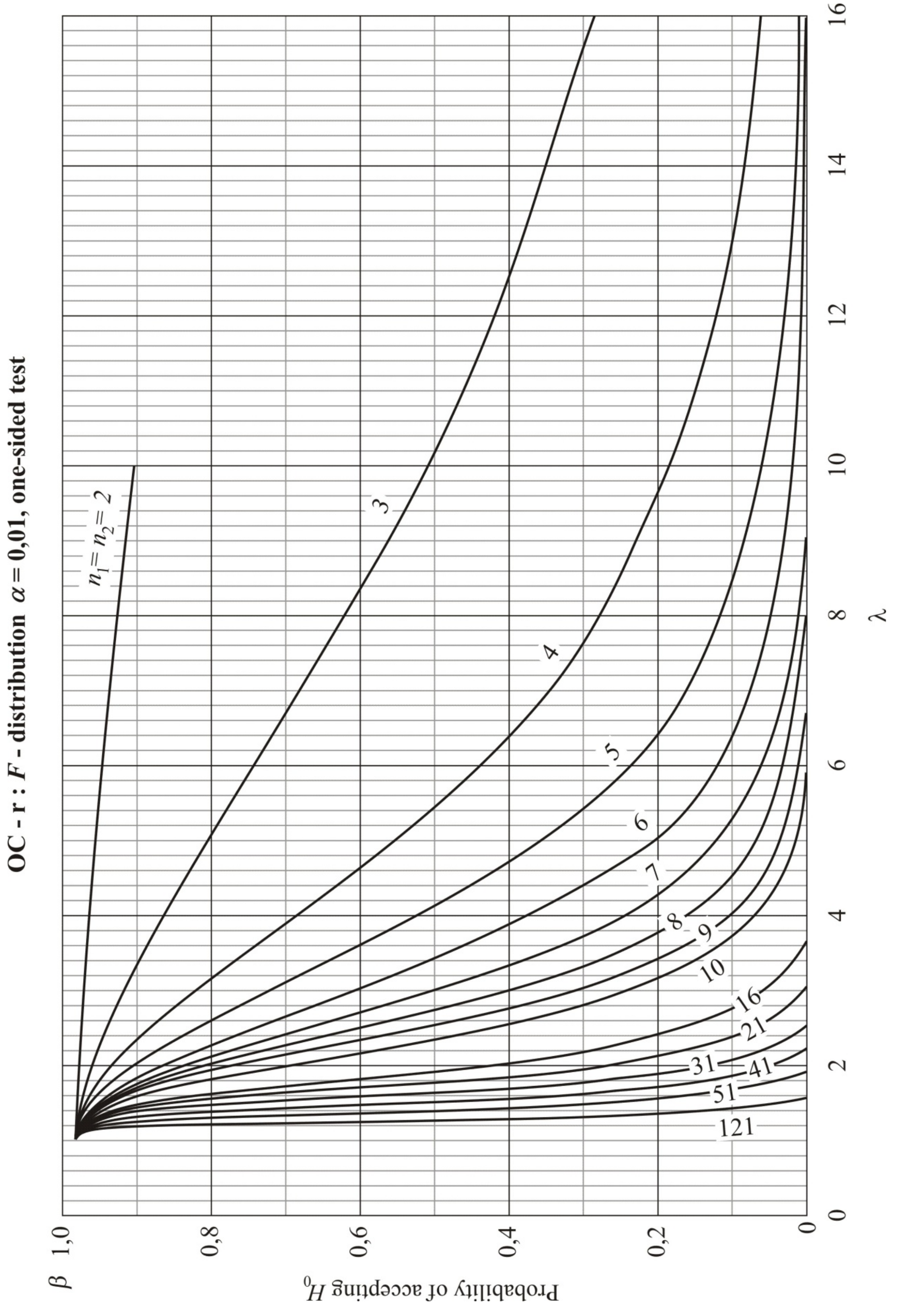

OC - o : F - distribution α = 0,05, two-sided test
β
Probability of accepting H0
λ
1,0
0,8
0,6
0,4
0,2
0
0
0,5
1
1,5
2
2,5
3
3,5
4
n1 = n2 = 3
4
5
6
7
8
9
10
16
21
31
51
101
3
4
5
6
7
8
9
10
16
21
31
51
101

OC - p : *F* - distribution $\alpha = 0{,}01$, two-sided test



OC - q : *F* - distribution $\alpha = 0{,}05$, one-sided test




OC - r : F - distribution α = 0,01, one-sided test
β
Probability of accepting H_0
1,0
0,8
0,6
0,4
0,2
0
λ
0
2
4
6
8
10
12
14
16
$n_1 = n_2 = 2$
3
4
5
6
7
8
9
10
16
21
31
41
51
121

BIBLIOGRAPHY

[1] GARAJ, I., JANIGA, I. 2002. *Dvojstranné tolerančné medze pre neznámu strednú hodnotu a rozptyl normálneho rozdelenia*. Bratislava, Vydavateľstvo STU, 2002, 147 s. ISBN 80-227-1779-7.

[2] GARAJ, I., JANIGA, I. 2004. *Dvojstranné tolerančné medze normálnych rozdelení s neznámymi strednými hodnotami a s neznámym spoločným rozptylom. Two Sided Tolerance Limits of Normal Distributions with Uknown Means and Uknown Common Variability.* Bratislava,Vydavateľstvo STU, 2004, 218 s. ISBN 80-227-2019-4.

[3] GARAJ, I., JANIGA, I. 2005. *Jednostranné tolerančné medze normálneho rozdelenia s neznámou strednou hodnotou a rozptylom. One Sided Tolerance Limits of Normal Distributions with Uknown Mean and Variability.* Bratislava, Vydavateľstvo STU, 2005, 214 s. ISBN 80-227-2218-9.

[4] HÁTLE, J., LIKEŠ, J. 1972. *Základy počtu pravděpodobnosti a matematické statistiky*. Praha, SNTL/ALFA, 1972, 463 s.

[5] JANIGA, I., 2013. *Aplikovaná pravdepodobnosť a štatistika pre inžinierov. 1.diel: Štatistická analýza jedného a dvoch súborov dát.* Vydavateľstvo STU, Bratislava, 2013

[6] JANIGA, I., MIKLÓŠ, R. Statistical tolerance intervals for a normal distribution. In *Measurement Science Review*. ISSN 13, 2001, vol. 1, no. 1, p. 29-32.

[7] JANIGA, I., STANISLAV, M., GABKOVÁ, J. 2012. Aplikácia DMAIC v procese kompletizácie. In *Forum Statisticum Slovacum*. ISSN 1336-7420, 2012, roč. VIII, č. 5. s. 52-59.

[8] JANIGA, I., STAREKOVÁ, A. 2001. *Základy pravdepodobnosti a štatistiky*. STU v Bratislave, 2001. 201 s. ISBN 80-227-1603-0.

[9] JÍLEK, M. 1988. *Statistické toleranční meze*. Praha, SNTL, 1988, 275 s.

[10] LAMOŠ, F., POTOCKÝ, R. 1989. *Pravdepodobnosť a matematická štatistika*. Vyd. ALFA, 1989. 342 s. ISBN 80-05-00115-0.

[11] LIKEŠ, J., LAGA, J. 1978. *Základní statistické tabulky*. Praha, SNTL, 1978, 488 s.

[12] MONTGOMERY, D.C., RUNGER, G.C. 2002. *Applied statistics and probability for engineers*. John Wiley & Sons, Inc., 2002. 706 p. ISBN 0-471-20454-4.

[13] MONTGOMERY, D.C., RUNGER, G.C. 2003. *Applied statistics and probability for engineers. Student workbook with solutions.* John Wiley & Sons, Inc., 2003. 706 p. ISBN 0-471-42682-2.

[14] PALENČÁR, R., RUIZ, J.M., JANIGA, I., HORNÍKOVÁ, A. 2001. *Štatistické metódy v metrologických a skúšobných laboratóriach.* Vyd. Grafické štúdio Ing. Peter Juriga, 2001. 366 s. ISBN 80-968449-3-8.

[15] VARGA, Š., KVASNIČKA, V. 1988. *Matematika III. Diferenciálne rovnice a matematická štatistika*. Edičné stredisko SVŠT v Bratislave, 1988. 167 s.

[16] VARGA, Š., KVASNIČKA, V. 1988. *Matematika III. Príklady.* Edičné stredisko SVŠT v Bratislave, 1988. 200 s.

[17] WIMMER, G. 1993. *Štatistické metódy v pedagogike*. Nakladateľstvo GAUDEAMUS, 1993. 154 s. ISBN 80-7041-864-8.

Doc. RNDr. Ivan Janiga, PhD.

BASICS OF STATISTICAL ANALYSIS

Vydala Slovenská technická univerzita v Bratislave v Nakladateľstve STU,
Bratislava, Vazovova 5, v roku 2014.

Edícia skrípt

Rozsah 218 strán, 45 obrázkov, 32 tabuliek, 8,768 AH, 9,092 VH,
1. vydanie, edičné číslo 5771, tlač Nakladateľstvo STU v Bratislave.

85 – 219 – 2014

ISBN 978-80-227-4161-3