MEDICAL UNIVERSITY – PLOVDIV
FACULTY OF PUBLIC HEALTH
Department of Social Medicine and Public Health

MEDICAL STATISTICS

A Practical Handbook for Exercises

$\bar{X}$ SD $95\%\,CI$ p χ^2 r $\hat{Y}$

seven exercises • solved examples • self-study

Chief Assist. Prof. Kostadin Kostadinov, MD, PhD, MPH, MEcon

A study aid for students of
Medicine and Dental Medicine, Year 2

Academic year 2026/2027
Plovdiv

Medical Statistics. A Practical Handbook for Exercises

Author: Chief Assist. Prof. Kostadin Kostadinov, MD, PhD, MPH, MEcon

A study aid for the practical classes in medical statistics. The handbook follows the course syllabus. It is intended for preparation before each practical class, for in-class work, and for self-study before the colloquium and the semester examination.

All computations, tables, and figures in this handbook were produced reproducibly with the statistical environment R and the kkstatfun package. The data in the examples are teaching data: they are constructed to illustrate the method and do not describe real patients.

Comments, corrections, and suggestions are welcome.

Contents

Preface 6
- Why Study Statistics at All? 6
- How the Handbook Is Organized 6
- How to Work with It 7
- Notation 7
- References and Further Reading 8

1 Absolute and Relative Values. Direct Standardization 9
- 1.1 Absolute Values 9
- 1.2 Relative Values 9
- 1.3 Why Crude Rates Are Not Compared Directly 11
- 1.4 Worked Example 12
- 1.5 Check: The Same Task with the Other Standard 17
- 1.6 Summary of the Method 18
- 1.7 Common Mistakes 18
- 1.8 Self-Study Tasks 19
- 1.9 Short Quiz 21

2 Descriptive Statistics 22
- 2.1 Sample and Population 22
- 2.2 Statistical Variables and Scales of Measurement 23
- 2.3 Measures of Central Tendency 24
- 2.4 Measures of Dispersion 27
- 2.5 Which Measure When 32
- 2.6 Common Mistakes 32
- 2.7 Self-Study Tasks 32
- 2.8 Short Quiz 35

3 Inferential Statistics. Confidence Interval 36
- 3.1 Statistic and Parameter 36
- 3.2 Sampling Distribution and Standard Error 36
- 3.3 Confidence Interval for the Mean 38
- 3.4 Confidence Interval for a Proportion 41
- 3.5 What Changes the Width of the Interval 42
- 3.6 Comparing Two Groups with Confidence Intervals 43
- 3.7 Common Mistakes 43
- 3.8 Self-Study and Homework Tasks 44
- 3.9 Short Quiz 45

4 Hypothesis Testing. Z-Test and t-Test 47
- 4.1 Hypotheses 47
- 4.2 The p-Value 48
- 4.3 Type I and Type II Errors 49

4.4 Power . . . 50
4.5 Two-Sided and One-Sided Alternatives . . . 50
4.6 Choosing a Statistical Test . . . 50
4.7 Z-Test for Comparing Two Proportions . . . 51
4.8 t-Test for Comparing Two Independent Samples . . . 53
4.9 Mandatory Exam Question . . . 55
4.10 Statistical and Clinical Significance . . . 56
4.11 Common Mistakes . . . 56
4.12 Self-Study and Homework Tasks . . . 57
4.13 Short Quiz . . . 58

5 Association between Categorical Variables. The χ^2-Test **60**
5.1 When the χ^2-Test Is Used . . . 60
5.2 Contingency Table . . . 61
5.3 The Logic of the Test . . . 62
5.4 Worked Example: a 3 × 3 Table . . . 62
5.5 Worked Example: a 2 × 2 Table and the φ Coefficient . . . 68
5.6 What the χ^2-Test Does Not Do . . . 70
5.7 Common Mistakes . . . 71
5.8 Self-Study and Homework Tasks . . . 71
5.9 Short Quiz . . . 74

6 Correlation Analysis **75**
6.1 When Which Analysis . . . 75
6.2 The First Step Is the Graph . . . 76
6.3 The Idea behind the Coefficient . . . 77
6.4 The Pearson Correlation Coefficient . . . 77
6.5 Worked Example . . . 79
6.6 What the Coefficient Does Not Show . . . 82
6.7 Correlation and Causality . . . 84
6.8 Common Mistakes . . . 84
6.9 Self-Study and Homework Tasks . . . 85
6.10 Short Quiz . . . 87

7 Regression Analysis **88**
7.1 Essence of the Regression Model . . . 88
7.2 The Line of Best Fit . . . 89
7.3 Worked Example . . . 90
7.4 Prediction . . . 95
7.5 General Rule of Interpretation . . . 96
7.6 More about Regression . . . 96
7.7 Common Mistakes . . . 97
7.8 Self-Study and Homework Tasks . . . 97
7.9 Short Quiz . . . 100

8 Colloquium **101**
8.1 Format and Assessment . . . 101
8.2 How the Type of Task Is Recognised . . . 101

8.3 General Course of the Solution . . . 102
8.4 Four Sample Variants . . . 104
8.5 Solutions of the Variants . . . 107
8.6 Preparation Test, Part 1 . . . 115
8.7 Preparation Test, Part 2 . . . 118
8.8 Preparation Test, Part 3 . . . 121
8.9 Open Questions . . . 124

A Additional Concepts for the Examination **126**
A.1 Choosing a Graph . . . 126
A.2 Sampling Designs . . . 127
A.3 Sample Size . . . 128
A.4 Robustness of Parametric Tests . . . 129
A.5 One-Sided Tests . . . 129
A.6 The Binomial Distribution . . . 130
A.7 Autocorrelation . . . 131
A.8 Interpolation and Extrapolation . . . 132

B Formulas **133**
Relative Measures and Standardization . . . 133
Central Tendency . . . 133
Dispersion . . . 134
Estimation of Population Parameters . . . 134
Hypothesis Testing . . . 135
Association between Categorical Characteristics . . . 136
Association between Quantitative Characteristics . . . 136

C Statistical Tables **138**
Critical Values of the t-Distribution . . . 138
Critical Values of the Chi-Square Distribution . . . 140
Critical Values of the Correlation Coefficient . . . 140

D Answers **142**
D.1 To Chapter 1 . . . 142
D.2 To Chapter 2 . . . 142
D.3 To Chapter 3 . . . 143
D.4 To Chapter 4 . . . 144
D.5 To Chapter 5 . . . 145
D.6 To Chapter 6 . . . 146
D.7 To Chapter 7 . . . 146
D.8 To the Colloquium . . . 147

E Glossary of Basic Terms **153**

Preface

This handbook is a concise study guide for Year 2 students of Medicine and Dental Medicine who study **medical statistics**. It follows the syllabus of the seven practical exercises and combines explanatory text, worked examples, and self-study tasks. Its aim is to connect the lecture concepts with the calculations and interpretations needed during the practical classes, the colloquium, and the examination.

Why Study Statistics at All?

Statistics provides the methods by which data are turned into information, and information supports decisions. For the physician this has three immediate applications.

First, it supports **informed choice** in everyday practice. When you decide whether to apply a new treatment, you need to understand how large the observed effect is, how uncertain the estimate is, and how applicable the result is to the particular patient.

Second, it shows **how scientific evidence is built**. Clinical studies, recommendations in guidelines, and the evaluation of medicines use statistical analysis to distinguish a robust result from random fluctuation.

Third, it allows the medical literature to be read **critically**. The ability to recognise an inappropriate method, a misleading figure, or an overconfident conclusion is more valuable than memorising a single numerical result.

> **What you will not learn in this course**
>
> Modern statistics is the foundation of "data science" and of artificial intelligence systems. In this course you will not program and you will not build complex models. The goal is more modest and more important: to understand the **logic** of statistical analysis and the **language** in which it is spoken — so that you can read a result, judge whether to believe it, and explain to a patient what it means.

How the Handbook Is Organized

The handbook contains seven chapters — one for each practical exercise — a chapter on the colloquium, and four appendices.

Table 1: Contents of the practical exercises

Ch.	Exercise	Main method
1	Absolute and relative values	direct standardization
2	Descriptive statistics	mean, median, SD, IQR

Ch.	Exercise	Main method
3	Inferential statistics	standard error, confidence interval
4	Hypothesis testing	Z-test and t-test
5	Association between categorical variables	χ^2-test, φ, Cramér's V
6	Correlation analysis	Pearson coefficient r
7	Regression analysis	simple linear regression

Each chapter is built in the same way:

- **Aim of the exercise** — what you should be able to do by the end of the class;
- **What you should know beforehand** — the link to the previous exercises;
- **Theory in brief** — definitions and formulas, every symbol explained;
- **Worked example** — step by step, with all intermediate numbers;
- **Common mistakes** — what is most often confused and why;
- **Self-study tasks** — with short answers in Section D;
- **Short quiz** — questions with one correct answer to check your understanding.

How to Work with It

Read the theoretical part **before** the practical class. During the class, follow the steps of the worked examples and write down the intermediate calculations. After the class, solve the tasks and the short quiz without looking at the answers; check your results only after you have also formulated a verbal conclusion.

> **!** Three rules that apply throughout the course
>
> 1. **Write the steps, not only the answer.** At the colloquium and at the examination the course of the solution is assessed. An answer without steps carries no points.
> 2. **Every number has a unit of measurement.** "The mean survival is 8" is not an answer; "8 months" is.
> 3. **Every conclusion is formulated in words.** The number $p < 0.01$ is not a conclusion. The conclusion is the sentence that follows from it.

Notation

Throughout the handbook the same symbols are used. Latin letters denote measures computed **in the sample** (statistics), and Greek letters — the corresponding unknown measures **in the population** (parameters).

Table 2: Main notation

Symbol	Meaning
n	sample size (number of observations)
$\bar{X}$	sample mean (statistic)
μ	population mean (parameter)
S (or SD)	sample standard deviation
$S_{\bar{X}}$ (or SE)	standard error of the mean
$\hat{p}$	sample proportion, %
P	population proportion, %
f	frequency (number of observations in a group or interval)
α	significance level (acceptable risk of a type I error)
p	p-value
df	degrees of freedom

About the decimal sign

In English-language scientific practice the decimal sign is a **point** (8.33), while in Bulgarian practice a **comma** is used ($8{,}33$). This handbook follows the English convention with a point. Both notations are correct; the important thing is not to mix them within a single solution.

References and Further Reading

In English

- Kirkwood, B. R., Sterne, J. A. C. (2003). *Essential Medical Statistics*. 2nd ed. Blackwell.
- Bland, M. (2015). *An Introduction to Medical Statistics*. 4th ed. Oxford University Press.
- Altman, D. G. (1991). *Practical Statistics for Medical Research*. Chapman & Hall/CRC.
- Motulsky, H. (2018). *Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking*. 4th ed. Oxford University Press.
- Diez, D., Çetinkaya-Rundel, M., Barr, C. (2019). *OpenIntro Statistics*. 4th ed. OpenIntro.

In Bulgarian (in Bulgarian)

- Сепетлиев, Д. (1980). *Медицинска статистика*. Трето преработено и допълнено издание. Медицина и физкултура.
- Димитров, И. (1996). *Медицинска статистика*. Второ преработено и допълнено издание. Пигмалион.
- Искров, Г. (2025). *Медицинска статистика*. УИ „Паисий Хилендарски“.

All computations, tables, and figures were produced with the statistical environment **R** and the **kkstatfun** package. The code is not shown — it is not part of the examination material — but it is available on request for students who would like to examine it.

1 Absolute and Relative Values. Direct Standardization

> **Aim of the exercise**
>
> After this exercise you should be able to:
>
> 1. distinguish an absolute value from a relative value;
> 2. recognise an **extensive** (structural) measure from an **intensive** (frequency) measure and compute each of them;
> 3. explain why two rates should not be compared directly when the groups differ in composition;
> 4. carry out **direct standardization** in four steps and formulate the conclusion.

1.1 Absolute Values

Absolute values are numbers that characterize quantitatively the size of a statistical population or of parts of it, as well as the value of a specific statistical characteristic.

They have two mandatory properties: they are always **named** — accompanied by a unit of measurement — and they are always obtained by direct measurement or by counting.

The systolic blood pressure, measured in mmHg, is an absolute value: it has a numerical value, it has a unit of measurement, and it characterizes a specific characteristic quantitatively. The number of patients who passed through a department in a year is also an absolute value.

> **Important**
>
> A statistical study usually **starts** with absolute values, but they are **not sufficient** for comparisons in time and space. We cannot say that a hospital with 400 deaths performs worse than a hospital with 40 deaths before we learn how many patients passed through each of them.

1.2 Relative Values

Relative values are obtained by dividing two absolute values. They are presented as a coefficient (ratio), as a percentage (multiplying by 100), as a permille (multiplying by 1,000), or per 100,000 population.

An example from clinical practice. The stroke volume is the amount of blood that enters the aorta after a single heart contraction — an absolute value in millilitres. The heart of a large athlete ejects more millilitres than the heart of a first-grader, but this does not mean that it works better. That is why the **ejection fraction** is used — the ratio of the stroke volume to the volume of the ventricle

before the contraction. It is a relative value and it is precisely what allows the comparison between the two.

In epidemiology three kinds of relative measure are distinguished: **ratios**, **proportions** and **rates**. The Bulgarian teaching tradition uses two terms for the last two: *extensive* measures for proportions and *intensive* measures for rates. Both sets of names are given below, so that the lecture and the international literature can be read with the same understanding.

Table 1.1: The three kinds of relative measure

Measure	Numerator	Denominator	Range	Traditional name
Ratio	one quantity	another, unrelated quantity	any positive number	—
Proportion	a part of the whole	the whole	0 to 1 (0—100 %)	extensive
Rate	events	population and time at risk	0 to infinity	intensive

A **ratio** relates two quantities that are not part of one another, for example the male-to-female ratio among the admitted patients. Ratios are not discussed further in this chapter.

1.2.1 Proportions (Structural or "Extensive" Measures)

A **proportion** shows how a whole is distributed among mutually exclusive component parts at a given time and place. The numerator is **contained in** the denominator: the denominator is the total number of units in the whole under consideration, and the numerator is the number in one of its categories.

$$\text{proportion} = \frac{\text{part of the whole}}{\text{the whole}} \times 100$$

A proportion is dimensionless and can never exceed 100 %. When several proportions describe the same whole, they form its **composition** or **structure**.

> **!** A property you will use constantly
>
> The sum of all proportions describing one whole is always equal to **1 (100 %)**. If you do not get 100 %, there is an error somewhere.

If we divide the patients with hepatitis A into age groups and divide the number in each group by the total number of patients, we obtain the age **structure** of the cases — a set of proportions whose sum is 100 %.

1.2.2 Rates (Frequency or "Intensive" Measures)

A **rate** shows how often an event occurs in the population in which it can occur. The numerator contains the number of events; the denominator contains the population at risk, and in a true rate also the time over which it was observed (**person-time**). Unlike a proportion, a rate does not describe the composition of a whole but the frequency of an event in its source population.

$$\text{rate} = \frac{\text{number of events}}{\text{population at risk} \times \text{time}} \times 100$$

> **i** Rate, risk and the everyday use of the word
>
> Strictly, a **rate** has time in its denominator (cases per 1,000 person-years), whereas a **risk** is a proportion of persons who develop the outcome over a stated period (cases per 1,000 persons followed for one year). In routine practice the word *rate* is used loosely for both, and this handbook follows that convention. What must never be omitted is the **period of observation** and the **denominator**: "12 per 1,000" is meaningless until both are stated.

Four measures you should be able to distinguish:

- **Case fatality** (case-fatality risk) — deaths from a given disease relative to the number of **diagnosed cases** of it. A case fatality of 5 % for measles in children means that of 100 children who fell ill, 5 died. It measures the severity of the disease.
- **Mortality rate** — deaths relative to the **whole population** at risk. The difference from case fatality lies in the denominator: for case fatality the denominator holds the cases, for mortality — everyone.
- **Incidence** — the number of **new cases** arising in the population at risk during a stated period. It measures the appearance of new disease.
- **Prevalence** — the number of **existing cases**, new and old, at a given moment. Incidence measures flow; prevalence measures stock.

> **💡** How to tell them apart in a second
>
> Look at what the **denominator** means. If it is a whole that we divide into categories, the measure is a **proportion** (extensive). If the denominator is the population at risk in which we count events, the measure is a **rate** (intensive). The mere fact that the persons in the numerator also appear in the denominator is not enough: in case fatality the deceased are among the cases, yet the measure describes the frequency of an outcome among the cases, not a structure by categories.

1.3 Why Crude Rates Are Not Compared Directly

A rate computed for a whole population, without regard to its composition, is called a **crude rate**. Its value depends on the **structure of the population** in which it was measured. The birth rate is higher where people aged 20–30 predominate — not because they are more fertile, but because

such is the composition of the population. The case-fatality rate of a hospital is higher if mainly oncological patients pass through it.

Therefore, when crude rates from populations of different composition are compared, the comparison can mislead. Age, or another structural variable, is mixed up with the difference under study; in epidemiological language it acts as a **confounder**. The method that removes this mixing by imposing a common composition is called **standardization**.

> ⚠ Definition
>
> **Standardization** is a procedure that removes the influence of age, or of another difference in composition, from the comparison of crude rates. An **age-standardized rate** shows what the overall rate of a group would be if that group had the age composition of a chosen **standard population**.

Two methods exist. **Direct standardization** applies the age-specific rates of each study group to one common standard population, and is the method used in this course. **Indirect standardization** applies the age-specific rates of a standard population to the composition of the study group and yields a standardized ratio, such as the standardized mortality ratio (SMR); it is not part of this course.

1.4 Worked Example

> 💡 Setting
>
> Two groups of miners were studied — those who **did not consume** alcohol and those who **consumed** alcohol. For each age group, the number of miners who underwent a preventive examination N and the number of cases C are known.
> **Task.** Compare the *age-standardized* incidence rates of the two groups, taking the age composition of the second group (the alcohol consumers) as the **standard population**.

The initial data are presented in Table 1.2.

Table 1.2: Initial data. N_1, C_1 — examined and cases among the non-consumers of alcohol; N_2, C_2 — among the consumers

Age	N_1	C_1	N_2	C_2
< 24	500	250	4,000	2,400
25–34	500	300	2,000	1,400
35–44	1,000	700	1,000	800
45+	2,000	1,600	1,000	900
Total	**4,000**	**2,850**	**8,000**	**5,500**

1.4.1 Step 1. Computing the crude rates

We compute the incidence separately for each age group and overall for each of the two groups:

$$IR = \frac{C}{N} \times 100$$

For the miners under 24 who did not consume alcohol:

$$IR_1 = \frac{250}{500} \times 100 = 50\,\%$$

Overall for the non-consumers:

$$IR_1^{\text{overall}} = \frac{2\,850}{4\,000} \times 100 = 71.25\,\%$$

Overall for the consumers:

$$IR_2^{\text{overall}} = \frac{5\,500}{8\,000} \times 100 = 68.75\,\%$$

The complete results are in Table 1.3.

Table 1.3: Step 1 — crude incidence rates

Age	N_1	C_1	IR_1, %	N_2	C_2	IR_2, %
< 24	500	250	50.00	4,000	2,400	60.00
25–34	500	300	60.00	2,000	1,400	70.00
35–44	1,000	700	70.00	1,000	800	80.00
45+	2,000	1,600	80.00	1,000	900	90.00
Total	**4,000**	**2,850**	**71.25**	**8,000**	**5,500**	**68.75**

> **!** Important
>
> The overall **crude** rate is **not** the sum of the rates by subgroup. It is computed anew, from the overall numbers: $2\,850/4\,000$. If you add 50 + 60 + 70 + 80, you get 260 % — nonsense.

The result of step 1 looks paradoxical and is shown in Figure 1.1. Viewed overall, the miners who did not consume alcohol fall ill **more often** (71.25 % vs. 68.75 %). Viewed by age group, however, the consumers fall ill more often in **every single** age group.

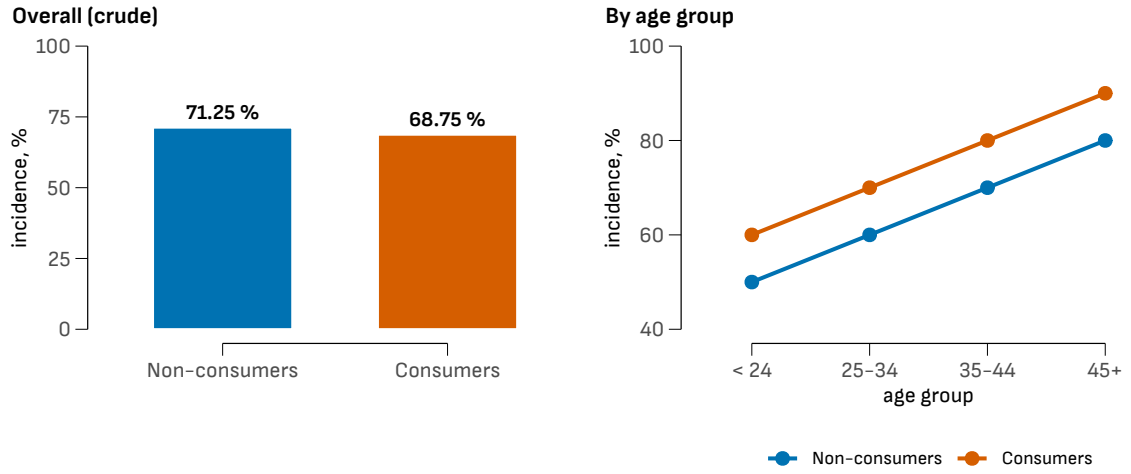

Figure 1.1: Left — the overall crude rates: the non-consumers appear sicker. Right — the same data by age group: the consumers are sicker in every single group. The contradiction is due to the different age structure.

The explanation lies in the age structure (Figure 1.2): among the non-consumers, **older** miners predominate, and among the consumers — **younger** ones. Age is associated with both the factor and the outcome — it is a **confounder**. The reversal of the conclusion is a classic example of Simpson's paradox.

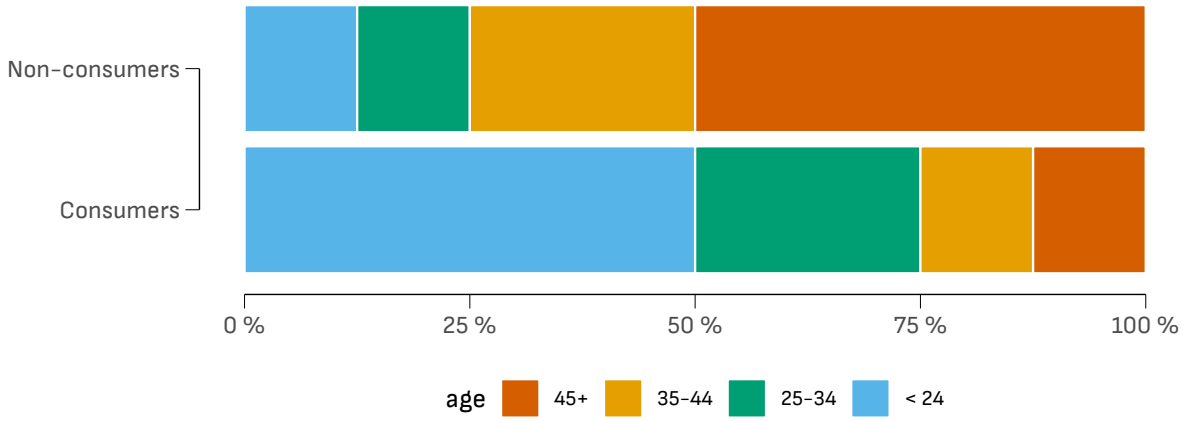

Figure 1.2: Age structure of the two groups. Among the non-consumers of alcohol half of the miners are over 45, while among the consumers half are under 24. It is precisely this difference that distorts the comparison of the overall measures.

1.4.2 Step 2. Computing the standard

The standard is the **age structure** that we take as common for the two groups. Here, by the problem statement, it is the structure of the second group (the consumers of alcohol):

$$S = \frac{N_2^{\text{group}}}{N_2^{\text{total}}} \times 100$$

For the age group under 24:

$$S_{<24} = \frac{4\,000}{8\,000} \times 100 = 50\,\%$$

Table 1.4: Step 2 — computing the standard

Age	N_2	S, %
< 24	4,000	50.0
25–34	2,000	25.0
35–44	1,000	12.5
45+	1,000	12.5
Total	**8,000**	**100.0**

> **Check**
>
> The standard is always an **proportion** — its sum must be exactly 100 %. If it is not, go back and check the division.

Which standard should be chosen? In textbook tasks it is usually given. In a real comparison, a common external population or the pooled structure of the compared groups is preferred, so that the results are comparable. The chosen standard affects the concrete values and can sometimes affect the conclusion as well; therefore it is always stated explicitly. In Section 1.5 we will check what happens in the present example with another standard.

1.4.3 Step 3. Computing the standardized rates

For each age group we multiply the **crude age-specific rate** by the **quota of the standard** for the same age group:

$$AIR = \frac{IR \times S}{100}$$

For the non-consumers of alcohol under 24:

$$AIR_1 = \frac{50 \times 50}{100} = 25\,\%$$

For the consumers of alcohol under 24:

$$AIR_2 = \frac{60 \times 50}{100} = 30\,\%$$

Table 1.5: Step 3 — age-standardized incidence rates

Age	IR_1, %	IR_2, %	S, %	AIR_1, %	AIR_2, %
< 24	50	60	50.0	25.00	30.00
25–34	60	70	25.0	15.00	17.50
35–44	70	80	12.5	8.75	10.00
45+	80	90	12.5	10.00	11.25
Total	**71.25**	**68.75**	**100.0**	**58.75**	**68.75**

! The two summation rules

- Crude rates are **NOT** summed.
- Standardized measures **ARE** summed — their sum is the overall standardized measure.

i Useful check

The group whose structure was taken as the standard **keeps** the value of its overall measure before and after the standardization. Here the standard is the structure of the consumers, and their measure remains 68.75 % in both columns. If this does not happen, look for an arithmetic error.

1.4.4 Step 4. Conclusion

The age-standardized incidence rates of the two groups are **58.75 %** (non-consumers) and **68.75 %** (consumers). The incidence among the miners who consume alcohol is **10 percentage points higher**.

The conclusion is **opposite** to the one from the crude data. After equalizing the age structure, the comparison shows a higher standardized measure among the consumers of alcohol. By itself, this comparison does not prove that alcohol is the cause of the difference.

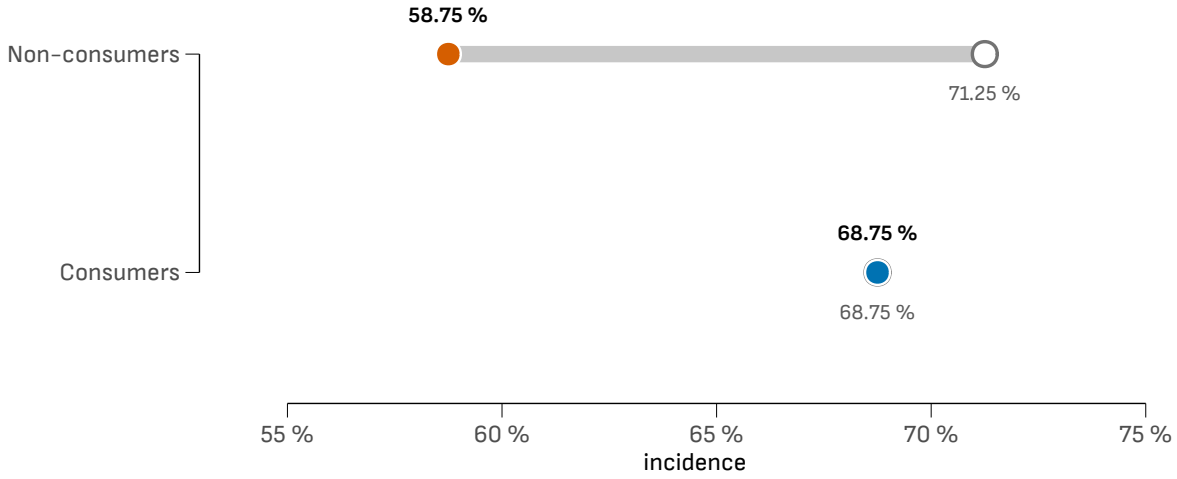

Figure 1.3: Before and after standardization. The crude rates suggest that the non-consumers fall ill more often; after removing the influence of age, the picture is reversed.

> ⚠ How to interpret standardized rates
>
> Standardized measures are **not real frequencies**. Nobody has observed 58.75 % incidence among the non-consumers of alcohol — the actually observed one is 71.25 %. The standardized rate answers the question: *"what would the incidence be if the two groups had the same age structure?"* Therefore it is used **only for comparison with each other**, not as an independent estimate of frequency.

1.5 Check: The Same Task with the Other Standard

To check the sensitivity of the result to the choice of standard, we solve the task again, this time with the structure of the **first** group (the non-consumers) as the standard.

Table 1.6: The same task, with the structure of the non-consumers as the standard

Age	N_1	S, %	IR_1, %	IR_2, %	AIR_1, %	AIR_2, %
< 24	500	12.5	50	60	6.25	7.50
25–34	500	12.5	60	70	7.50	8.75
35–44	1,000	25.0	70	80	17.50	20.00
45+	2,000	50.0	80	90	40.00	45.00
Total	**4,000**	**100.0**	**71.25**	**68.75**	**71.25**	**81.25**

The numbers are different — **71.25 %** vs. **81.25 %** — but the difference is again **10 percentage points** and the direction is the same: the consumers of alcohol fall ill more often. Note also that now the first group keeps its measure (71.25 %), because its structure is the standard.

What to remember from the check

In this example the choice of standard changes the **concrete values** but not the **conclusion**, because the rate among the consumers is 10 percentage points higher in every age group. This is not a universal rule. If the differences by subgroup go in different directions, the choice of standard can change the final comparison as well. For this reason a standardized rate is never reported without the standard used.

1.6 Summary of the Method

Direct standardization in four steps:

1. **Crude and age-specific rates** — for each subgroup separately and overall for each of the compared groups.
2. **Standard** — the structure of one of the groups (or an external population); the sum of the quotas is 100 %.
3. **Standardized rates by subgroup** — $AIR = IR \times S/100$.
4. **Summation and conclusion** — the sum of the standardized rates is the overall standardized rate; the two overall rates are compared.

1.7 Common Mistakes

Table 1.7: Common mistakes in standardization

#	Mistake	How to avoid it
1	Comparing crude rates of groups with a different structure	Always look first at the structure of the groups
2	Summing crude rates to obtain the overall one	The overall crude rate is computed anew, from the overall numbers
3	Adding instead of multiplying in step 3	$AIR = IR \times S/100$ — **multiplication**
4	A standard whose quotas do not sum to 100 %	Check the sum before you continue
5	Interpreting the standardized rate as a real frequency	It is “what would be if…”, not an observation
6	Confusing case fatality and mortality	Look at the denominator: cases or population

1.8 Self-Study Tasks

Task 1.1

Determine the age structure of the patients with hepatitis A treated at hospital X.

Age	0–1	1–4	5–9	10–19	Total
Cases of hepatitis A	70	41	102	87	300
Age structure, %					

What kind of measure do you compute and how will you check your result?

Task 1.2

Compute the standardized fertility measures of regions A and B. Take the age composition of the women in region B as the standard.

Age	Women (A)	Live births (A)	Women (B)	Live births (B)
15–20	1,000	18	1,200	22
21–30	9,000	225	7,000	175
31–49	8,000	128	10,000	160
Total	**18,000**	**371**	**18,200**	**357**

Task 1.3

Compute the standardized case-fatality measures of two city hospitals. Take the composition of the patients in hospital B as the standard.

Disease	Treated (A)	Deaths (A)	Treated (B)	Deaths (B)
Hypertensive disease	180	4	200	4
Stomach cancer	100	30	90	27
Myocardial infarction	120	8	160	10
Total	**400**	**42**	**450**	**41**

Task 1.4

The crude case-fatality rates of hospitals A and B are 10 ‰ and 12 ‰, respectively. After standardization by the structure of hospital A, the case fatality in hospital A is 10 ‰, and in hospital B — 8 ‰.
What conclusion follows? If the structure of hospital B is used as the standard, can the new standardized values be computed from the given information alone? Justify your answer.

Task 1.5 (for reasoning)

The overall mortality in Japan is higher than the overall mortality in Bangladesh. Does this mean that health care in Japan is worse? Explain your answer using the terms of this exercise.

Task 1.6

During one year, 240 patients were treated in a hospital: 96 in the cardiology department, 84 in the surgery department, and 60 in the neurology department. 18 of all treated patients died.

1. Compute the structure of the patients by department.
2. Compute the hospital case fatality.
3. State which of the two results is an extensive and which is an intensive measure. For each result, explain what the denominator means.

1.9 Short Quiz

For each question choose **one** correct answer. The key is in Section D.1.

1. Which of the following is a proportion (an extensive measure)?

    a) 28 new cases per 100,000 people per year;
    b) 35 % of hospitalised patients were admitted as emergencies;
    c) 4 deaths per 100 cases;
    d) 12 births per 1,000 women of fertile age.

2. The main difference between mortality and case fatality is:

    a) the unit of measurement of the numerator;
    b) the observation period;
    c) the denominator of the measure;
    d) the age of the studied persons.

3. The sum of the quotas in the chosen standard structure must be:

    a) 1 %;
    b) 10 %;
    c) 100 %;
    d) equal to the number of subgroups.

4. The overall age-standardized rate is obtained as:

    a) the sum of the crude rates by subgroup;
    b) the sum of the standardized rates by subgroup;
    c) the mean of the overall observed measures;
    d) the product of the standardized rates.

5. Which statement about the standardized rate is true?

    a) It is a directly observed frequency;
    b) It can be compared independently of the standard used;
    c) It shows what the overall measure would be under a given common structure;
    d) It proves a causal relationship between the factor and the outcome.

2 Descriptive Statistics

Aim of the exercise

After this exercise you should be able to:

1. determine the type of the statistical variable and the scale on which it is measured;
2. group data into a **simple** and into a **grouped (interval)** frequency series;
3. compute the arithmetic mean, the median, and the mode for ungrouped and for grouped data;
4. compute the range, the standard deviation, and the interquartile range;
5. choose the correct pair of measures: $\bar{X} \pm SD$ or $Me\ (IQR)$.

2.1 Sample and Population

A statistical study requires time, money, and people. It is almost never possible to examine all the individuals that interest us. It is unthinkable to examine all patients with heart failure in order to establish whether a new drug lowers their pulse.

That is why a smaller part is examined — the **sample** — and statistical methods are used to draw conclusions about all the persons to whom the question is directed — the **population** (Figure 2.1).

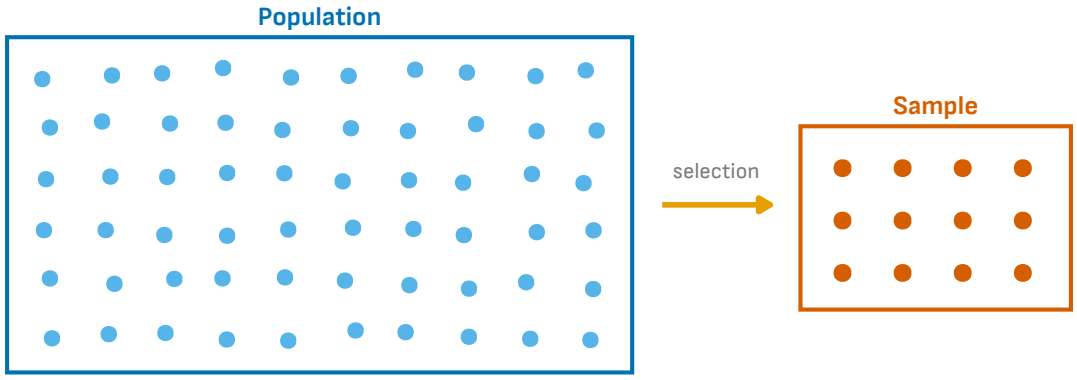

Figure 2.1: From the population a sample is selected. The arrow represents the selection procedure, not the movement of the same individuals.

For these conclusions to be reliable, the sample must be obtained through an appropriate selection procedure, must cover a correctly defined target population, and must not have substantial systematic loss of participants. Random selection reduces the risk of systematic error, but does not guarantee that every sample will reproduce exactly the structure of the population by sex, age, or another characteristic.

⚠ The two halves of statistics

Descriptive statistics describes what has been established **in the sample**.
Inferential statistics transfers this result to the **population**. We deal with it in Section 3.

2.2 Statistical Variables and Scales of Measurement

Statistical data are numerical or alphabetical values recorded after measuring or counting a certain characteristic. Because these characteristics take different values in different persons, they are called **variables**. If the characteristic is the same for everyone — it is a **constant**.

⚠ Definition

A **statistical variable** is a distinguishing feature, a characteristic of one or more units of observation.

Variables are divided into two main groups.

Quantitative variables are expressed with numbers for which arithmetic operations make sense. They are:

- *discrete* — they take separate, isolated values: number of live-born children, number of hospitalisations per year;
- *continuous* — they can be measured with arbitrary precision after the decimal point: weight, temperature, blood pressure.

Categorical (qualitative) variables distribute the observations into categories. Labels are used for them, for example blood group, type of disease, or presence of an allergy. If the categories are coded with numbers, the codes remain labels and do not automatically become quantitative values.

The scale on which a variable is measured determines which statistical methods are allowed (Table 2.1).

Table 2.1: Types of variables and measurement scales

Scale	Characteristic	Example
Nominal	categories, "names, labels"; no order	diagnosis, blood group
Dichotomous	*a kind of nominal scale with only two possible values*	*sex; alive/dead*
Ordinal	categories with a logical order (increasing or decreasing); the differences are not measurable	severity of heart failure (mild, moderate, severe); stage of an oncological disease
Rank	*a kind of ordinal scale — the units receive ranks from 1 to n*	*ranking by academic performance*
Interval	numbers; the value "0" does **not** mean absence of the characteristic; differences — yes, ratios — no	temperature in Celsius

Scale	Characteristic	Example
Ratio	numbers; the value "0" means absence of the characteristic; ratios are also allowed	weight, height, number of days

The nominal and the ordinal scale, including their varieties, are **non-metric**. The interval and the ratio scale are **metric**. The labels "weak" and "strong" describe what operations the scale allows, not whether the measurement was done well or badly.

> Why this matters
>
> If we say about a person that they are "tall", this is not the same as "height 185 cm". In the first case we measured on a weak scale, in the second — on a strong one. The stronger the scale, the more information the variable carries and the more methods are allowed.

2.3 Measures of Central Tendency

The descriptive analysis describes the sample by two groups of summarizing numerical characteristics: **central tendency** and **dispersion**.

The central tendency is a value that represents all observations "as a whole" — the typical expected representative of the sample. The dispersion shows how spread out the observations are around it.

For variables measured on an **interval** or **ratio** scale, the arithmetic mean and the median can be used, the choice also depending on the shape of the distribution. For **nominal** variables, frequencies, proportions, and the mode are reported. For **ordinal** variables, both the median and the quartiles can be used, when the ordering of the categories allows it.

2.3.1 Arithmetic Mean for Ungrouped Data

$$\bar{X} = \frac{\sum X}{n}$$

where $\bar{X}$ is the arithmetic mean, $\sum X$ — the sum of all observed values, and n — the number of observations.

> Example
>
> The heart rate of 10 second-year students is:
> 90, 80, 89, 72, 73, 74, 78, 86, 84, 82 bpm.
>
> $$\bar{X} = \frac{90 + 80 + 89 + 72 + 73 + 74 + 78 + 86 + 84 + 82}{10} = \frac{808}{10} = 80.8 \text{ bpm}$$

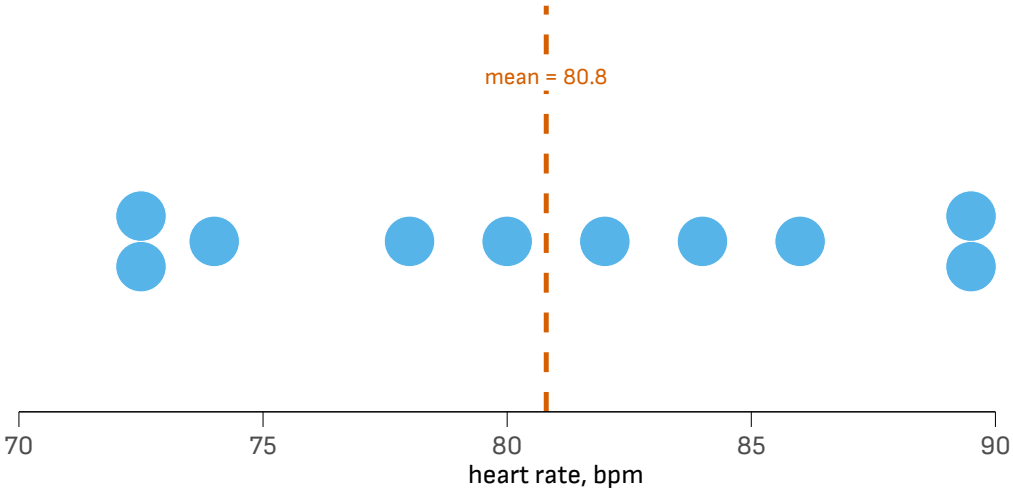

Figure 2.2: Distribution of the pulse of 10 students. Each dot is one observation; equal or close values are stacked on top of each other. The dashed line is the arithmetic mean.

2.3.2 Grouping the Data

With a larger volume of data it is convenient to present them grouped. The set of the different values (or intervals) of the variable together with their frequencies is called a **frequency distribution**.

Ungrouped frequency distribution. Each individual value of the variable is written, and opposite it — the number of observations with this value. With it **no** statistical precision is lost: we know the value of every observation.

Grouped (interval) frequency distribution. The values are grouped into intervals. With it precision is lost — we know only in which interval the observation falls, but not its exact value.

> **!** Two rules for constructing a grouped frequency distribution
>
> 1. The intervals are of **equal width**.
> 2. Every observation falls into **exactly one** interval. For integer data the notation 1—5, 6—10, 11—15 can be used. For continuous data the boundaries are set unambiguously, for example $[0; 5)$, $[5; 10)$ and $[10; 15]$.

2.3.3 Arithmetic Mean for Grouped Data

$$\bar{X} = \frac{\sum(X_u \cdot f)}{\sum f}$$

where:

- X_u is the **midpoint of the interval**: $X_u = \dfrac{\text{lower boundary} + \text{upper boundary}}{2}$;
- f is the frequency — the number of observations in the interval;
- $\sum f = n$ is the sample size.

Main example for the exercise

$n = 30$ patients treated with an experimental therapy were followed up. The survival in months is:
3, 11, 6, 9, 10, 10, 12, 15, 1, 4, 8, 12, 9, 9, 6, 5, 9, 10, 7, 11, 10, 12, 12, 6, 4, 15, 11, 9, 4, 5.
Grouped into a frequency distribution, the data look like this:

Survival X, months	Frequency f
$1-5$	7
$6-10$	14
$11-15$	9
Total	**30**

Step 1 — we find the midpoint of each interval and the product $X_u \cdot f$:

Table 2.2: Computing the arithmetic mean for a grouped frequency distribution

Survival X	f	X_u	$X_u \cdot f$
$1-5$	7	3	21
$6-10$	14	8	112
$11-15$	9	13	117
Total	**30**		**250**

Step 2 — we apply the formula:

$$\bar{X} = \frac{250}{30} = 8.33 \text{ months}$$

Grouping 'costs' a little precision

If we compute the mean from the original 30 values, we get 8.5 months. The grouped estimate gives 8.33 months. The difference of 0.17 months is the price of grouping — all observations in an interval are replaced by its midpoint.

2.3.4 Median and Mode

The **median** Me is the value that divides the data, arranged in ascending order, into two equal halves: 50 % of the observations are below it and 50 % — above it.

- With an **odd** number of observations the median is the value of the middle observation. For the data 6.0; 6.5; **7.0**; 7.5; 8.0 the median is 7.0.
- With an **even** number the median is the arithmetic mean of the two middle values. For the data 6.0; 6.5; **7.0; 7.5**; 8.0; 8.5 the median is $(7.0 + 7.5)/2 = 7.25$.

The **mode** M_0 is the most frequent value. It is found by counting. For the series 120; 120; 130; 135; 120; 140; 140 the mode is 120. It is possible that there is no mode or that there is more than one. In medical statistics the mode is rarely used.

2.3.5 Proportion

For qualitative variables the central tendency is described by the **proportion**:

$$\hat{p} = \frac{m}{n} \times 100$$

where m is the number of observations with the characteristic of interest, and n — the number of all observations.

For a group of 10 students, 4 of whom are girls:

$$\hat{p} = \frac{4}{10} \times 100 = 40\ \%$$

2.4 Measures of Dispersion

2.4.1 Range

The **range** R is the difference between the maximum and the minimum value:

$$R = X_{max} - X_{min}$$

For the pulse of the 10 students: $R = 90 - 72 = 18$ bpm.

The range is rarely used, because it depends only on two values, which are often extreme or even wrongly measured.

2.4.2 Standard Deviation for Ungrouped Data

The **standard deviation** S (or SD) is a measure of the mean level of dispersion relative to the arithmetic mean:

$$S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$$

where:

- X_i is the value of the i-th observation;
- $X_i - \bar{X}$ is the **deviation** of the observation from the mean;

- $(X_i - \bar{X})^2$ is the square of the deviation — squaring removes the signs, so that positive and negative deviations do not cancel each other out;
- $n - 1$ is the **Bessel's correction**: it is used when we work with a sample, not with the whole population;
- the square root returns the result to the original units of measurement.

Example — the pulse of the 10 students

The arithmetic mean is 80.8 bpm, but no student has exactly this value. Each one deviates by a few beats below or above it.

Pulse	Deviation $X_i - \bar{X}$	$(X_i - \bar{X})^2$
90	9.2	84.64
80	−0.8	0.64
89	8.2	67.24
72	−8.8	77.44
73	−7.8	60.84
74	−6.8	46.24
78	−2.8	7.84
86	5.2	27.04
84	3.2	10.24
82	1.2	1.44
Total	**0.0**	**383.60**

$$S = \sqrt{\frac{383.60}{10 - 1}} = \sqrt{42.62} = 6.53 \text{ bpm}$$

The result is written as $\bar{X} \pm S = 80.8 \pm 6.53$ bpm.

Why we square

The sum of the deviations is always zero — see the column in the example. If we did not square, the "mean deviation" would be zero for every sample and the measure would be useless.

2.4.3 Standard Deviation for a Grouped Series

$$S = \sqrt{\frac{\sum (X_u - \bar{X})^2 \cdot f}{n - 1}}$$

We continue the example with the 30 patients ($\bar{X} = 8.33$ months).

Table 2.3: Computing the standard deviation for a grouped frequency distribution

Survival	f	X_u	$X_u - \bar{X}$	$(X_u - \bar{X})^2$	$(X_u - \bar{X})^2 \cdot f$
$1 - 5$	7	3	−5.33	28.44	199.11
$6 - 10$	14	8	−0.33	0.11	1.56
$11 - 15$	9	13	4.67	21.78	196.00
Total	**30**				**396.67**

$$S = \sqrt{\frac{396.67}{30 - 1}} = \sqrt{13.68} = 3.70 \text{ months}$$

The survival in the sample is described as 8.33 ± 3.70 months.

2.4.4 The Three-Sigma Rule

If the variable is **normally distributed** — a symmetric, bell-shaped distribution in which the mean, the median, and the mode coincide (Figure 2.3) — the standard deviation describes what share of the observations lies at a given distance from the mean:

- $\bar{X} \pm 1 \cdot S$ covers **68 %** of the observations;
- $\bar{X} \pm 2 \cdot S$ covers **95 %** of the observations;
- $\bar{X} \pm 3 \cdot S$ covers **99.7 %** of the observations.

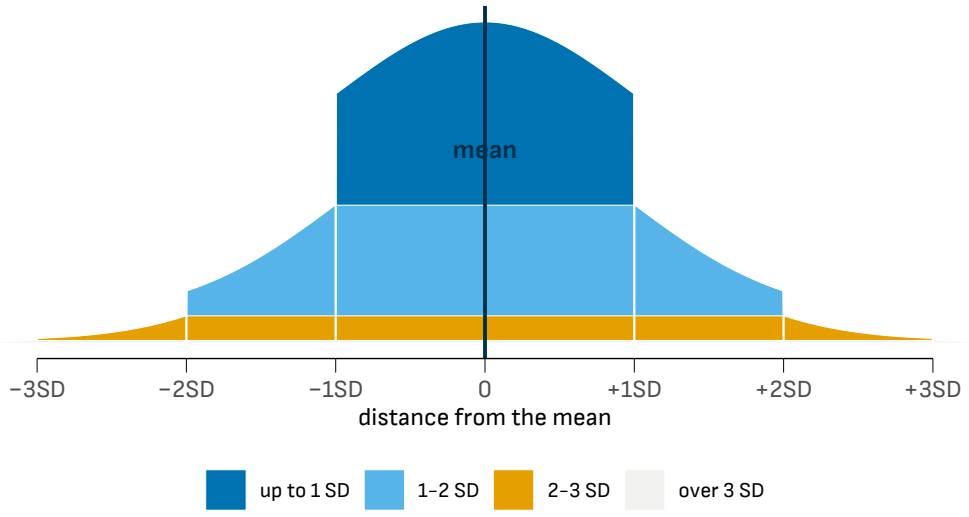

Figure 2.3: Normal distribution and the three-sigma rule. The coloured areas show the intervals from 0 to 1, from 1 to 2, and from 2 to 3 standard deviations from the mean; the distribution is symmetric.

Example

The weight of a sample of 1,000 people was measured: $\bar{X} \pm S = 80 \pm 10$ kg. We assume that the weight is normally distributed.
Then 95 % of the participants (950 people) have a weight in the interval $80 \pm 2 \cdot 10$, that is, **between 60 and 100 kg**.
The remaining 5 % (50 people) are distributed symmetrically at the two ends: about 25 people (2.5 %) weigh over 100 kg and about 25 people (2.5 %) — under 60 kg.

Caution — often confused

The three-sigma rule describes **how the individual observations are spread out**. It must not be confused with the confidence interval from Section 3, which describes **how precisely the mean measure is estimated**.

2.4.5 Interquartile Range

Sometimes the arithmetic mean does not describe the central tendency well. 22 patients were followed up and the duration of their hospitalisation in days was recorded:

24, 4, 2, 3, 4, 1, 6, 17, 4, 3, 2, 6, 25, 1, 6, 4, 2, 4, 3, 6, 2, 1.

The arithmetic mean is 5.91 days. However, this number is misleading: only 3 of 22 patients (13.6 %) stayed longer than 6 days. Their values are **extreme** and pull the mean upward (Figure 2.4).

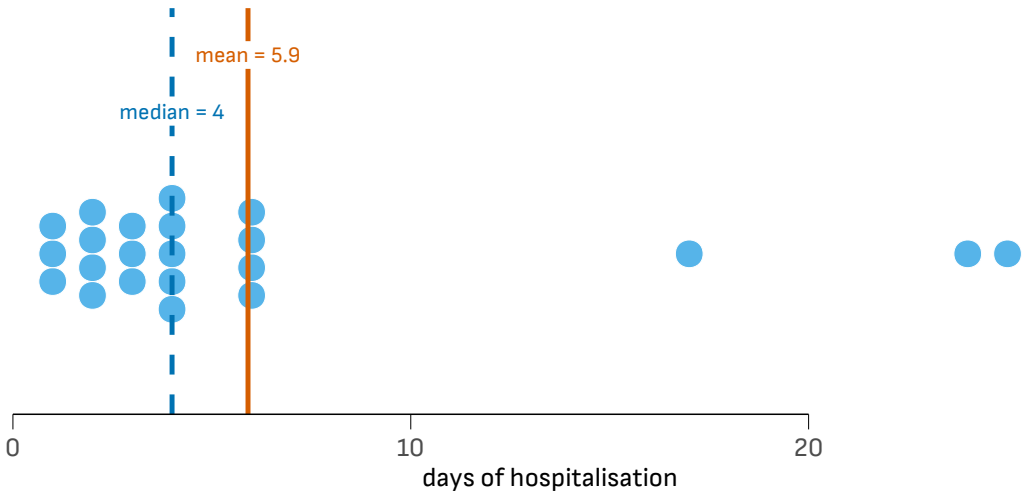

Figure 2.4: Right-skewed distribution. The arithmetic mean (solid line, 5.9 days) is shifted toward the extreme values. The median (dashed line, 4 days) describes the central tendency better.

Shape of the distribution and choice of measure

- With a **right-skewed** distribution (long tail to the right) the median is **smaller** than the arithmetic mean.

- With a **left-skewed** distribution the median is **larger** than the mean.
- With a **normal** distribution the two coincide or almost coincide.

When the distribution is skewed, **median and interquartile range** are reported, not mean and standard deviation.

The **interquartile range** IQR measures the dispersion relative to the median:

$$IQR = Q_3 - Q_1$$

where Q_1 and Q_3 are the first and the third quartile — the values that, together with the median Q_2, divide the ordered data into four equal parts.

We arrange the data in ascending order:

1, 1, 1, 2, 2, 2, **2**, 3, 3, 3, 4, **4**, 4, 4, 4, 6, **6**, 6, 6, 17, 24, 25.

With 22 observations:

- $Q_1 = 2$ days — below this value is the first quarter of the patients;
- $Q_2 = Me = 4$ days — the median;
- $Q_3 = 6$ days — below this value are three quarters of the patients.

$$IQR = Q_3 - Q_1 = 6 - 2 = 4 \text{ days}$$

The result is written as $Me\ (IQR) = 4\ (4)$ days or, in more detail, $Me = 4$ days; Q_1–Q_3: $2 - 6$ days.

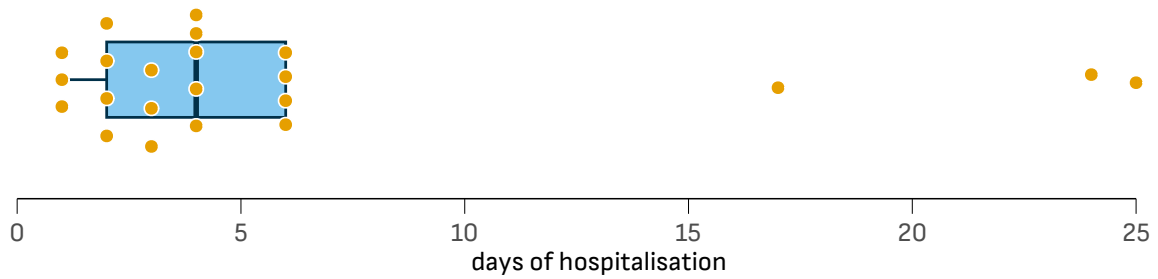

Figure 2.5: Boxplot of the duration of hospitalisation. The left side of the box is Q1, the right side is Q3, and the vertical line inside is the median. The points beyond the whiskers are extreme values.

How to read a boxplot

- The box covers the middle 50 % of the observations — its width is the IQR.
- The line inside the box is the **median**, not the arithmetic mean.
- The "whiskers" reach the most distant observation within $1.5 \cdot IQR$ from the box.
- The points beyond the whiskers are **extreme values (outliers)**.

- A longer whisker and more distant points on one side **suggest** skewness toward that side.

Extreme are all values higher than $Q_3 + 1.5 \cdot IQR$ or lower than $Q_1 - 1.5 \cdot IQR$. In the example the boundary is $6 + 1.5 \cdot 4 = 12$ days, that is, the values 17, 24, and 25 days are extreme.

2.5 Which Measure When

Table 2.4: Choice of measures according to the type of the variable

Type of variable / distribution	Central tendency	Dispersion
Quantitative, normally distributed	arithmetic mean $\bar{X}$	standard deviation S
Quantitative, skewed distribution	median Me	interquartile range IQR
Qualitative (nominal, ordinal)	proportion $\hat{p}$	–

2.6 Common Mistakes

Table 2.5: Common mistakes in descriptive analysis

#	Mistake	How to avoid it
1	Arithmetic mean with a strongly skewed distribution	Look at the distribution or the boxplot before choosing a measure
2	Dividing by n instead of $n - 1$ for the standard deviation	In a sample **always** $n - 1$
3	Forgetting the multiplication by f in a grouped frequency distribution	Each interval "weighs" according to the number of observations in it
4	Overlapping intervals (1–5, 5–10)	The boundaries are not repeated
5	Confusing $\bar{X} \pm 2S$ with a confidence interval	The first describes the observations, the second — the precision of the estimate
6	Reporting a number without a unit of measurement	"8.33 months", not "8.33"

2.7 Self-Study Tasks

Task 2.1

The pulse of 45 students was measured. The data are grouped into a frequency series:

Pulse	48–52	53–57	58–62	63–67	68–72	73–77	78–82
Number of students	4	5	8	11	6	6	5

Compute the arithmetic mean and the standard deviation.

Task 2.2

Can the standard deviation be equal to zero? And negative? Explain in both cases.

Task 2.3

Is it possible that in a given sample **all** individual deviations $X_i - \bar{X}$ are positive? And all zero? Explain.

Task 2.4

A sample of 2,000 newborns is included in a prospective study. The mean birth weight is 3,500 g with a standard deviation of 400 g. The weight is a normally distributed quantity.
How many of the newborns weigh more than 4,300 g or less than 3,100 g?

Task 2.5

Determine the type of the variable and the scale on which it is measured:

a) stage of kidney disease (I–V);
b) number of cigarettes per day;
c) blood group;
d) body temperature in °C;
e) score on an anxiety test (0–100 points);
f) presence of an allergy (yes/no).

Task 2.6

The boxplot below presents the overall survival in months.

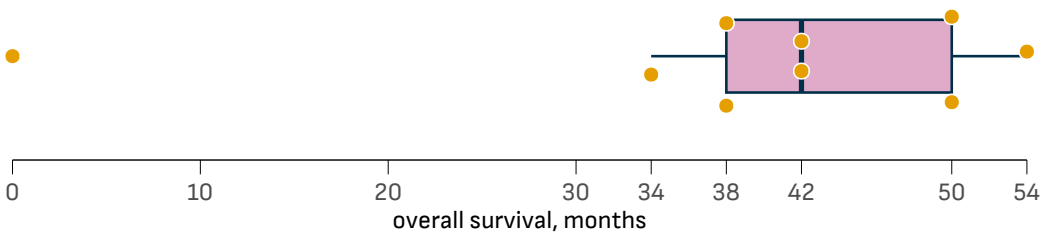

Figure 2.6: Boxplot for task 2.6.

What is the value of the interquartile range? What is the shape of the distribution? Are there extreme values and what are they?

Task 2.7

The concentration of C-reactive protein of nine patients is 2, 3, 3, 4, 4, 5, 6, 7, and 18 mg/L. Compute the arithmetic mean, the median, Q_1, Q_3, and IQR. Which pair of measures describes these data more appropriately and why?

Task 2.8

Two small samples have the following values:

- group A: 98, 99, 100, 101, 102;
- group B: 80, 90, 100, 110, 120.

Compare the arithmetic means and the standard deviations. Which group is more homogeneous and which measure supports the conclusion?

2.8 Short Quiz

For each question choose **one** correct answer. The key is in Section D.2.

1. The variable "blood group" is measured on:

    a) nominal;
    b) ordinal;
    c) interval;
    d) ratio scale.

2. Which measure is least affected by a single very high value?

    a) arithmetic mean;
    b) range;
    c) median;
    d) standard deviation.

3. With a strongly right-skewed distribution it is most appropriate to report:

    a) mean and standard deviation;
    b) median and interquartile range;
    c) mode and range;
    d) proportion and standard deviation.

4. The standard deviation is equal to zero when:

    a) the mean is zero;
    b) the sample contains negative values;
    c) all observations have the same value;
    d) the sample size is an even number.

5. With a normal distribution approximately 95 % of the observations are in:

    a) $\bar{X} \pm 1S$;
    b) $\bar{X} \pm 2S$;
    c) $\bar{X} \pm 3S$;
    d) $\bar{X} \pm 4S$.

6. The line inside the box of a standard boxplot shows:

    a) the mean;
    b) the mode;
    c) the median;
    d) the range.

3 Inferential Statistics. Confidence Interval

Aim of the exercise

After this exercise you should be able to:

1. distinguish a **statistic** from a **parameter**;
2. explain why different random samples give different estimates;
3. compute the standard error of the mean and of the proportion;
4. construct a 95 % confidence interval for a mean and for a proportion;
5. formulate correctly the interpretation of a confidence interval;
6. judge how the sample size, the dispersion, and the confidence level change the width of the interval.

What you should know beforehand

From Section 2: the arithmetic mean for grouped data, the standard deviation, and the proportion.

3.1 Statistic and Parameter

The measure computed **from the sample** is called a **statistic**. The corresponding measure for the **population** is called a **parameter** and is usually unknown. The sample statistic serves as an estimate of the population parameter.

Table 3.1: Statistics and parameters

In the sample (statistic)	In the population (parameter)
arithmetic mean $\bar{X}$	arithmetic mean μ
proportion $\hat{p}$	proportion P
standard deviation S	standard deviation σ

The mean pulse of 10 randomly selected students is a statistic. The mean pulse of all second-year students is a parameter. If we take a new random sample, we will probably get a slightly different mean. This natural difference between samples is called **random sampling error**. It does not necessarily mean that the measurement or the computation is wrong.

3.2 Sampling Distribution and Standard Error

Imagine that we repeatedly select random samples of the same size from one population. For each sample we compute an arithmetic mean. The obtained sample means form a **sampling distribution**.

Its centre is around the population mean μ, and its dispersion shows how much $\bar{X}$ varies from sample to sample (Figure 3.1).

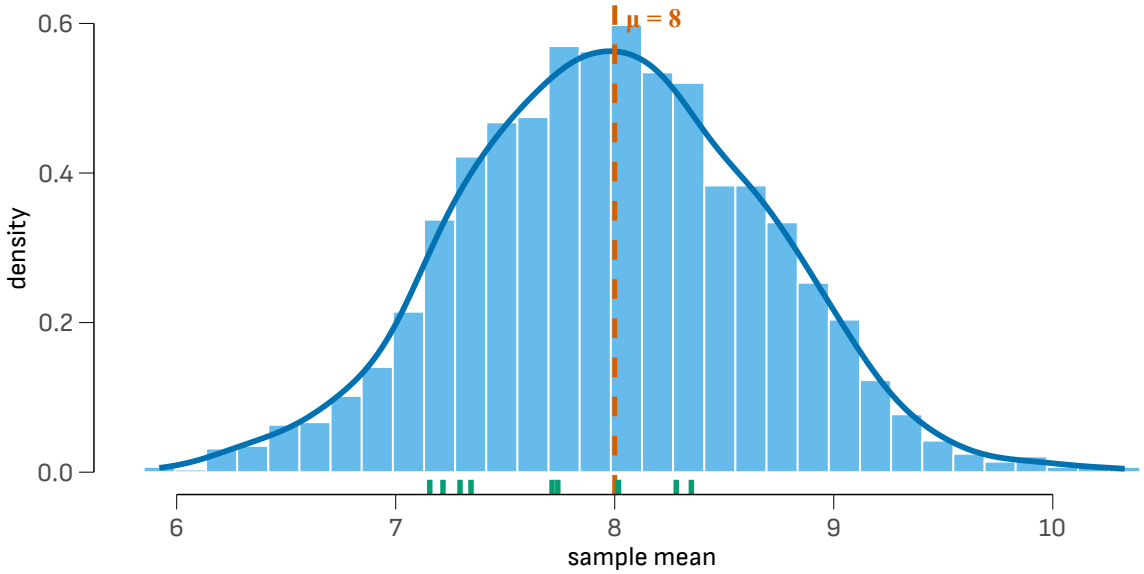

Figure 3.1: Sampling distribution of the arithmetic mean. The histogram shows the means of 2,000 random samples of 30 observations. The dashed line is the population mean, and the coloured ticks are the means of nine individual samples.

If the studied variable is normally distributed, the sampling distribution of the mean is also normal. With sufficiently large independent samples it is approximately normal even when the distribution of the individual observations is not normal. This is the practical consequence of the **central limit theorem** (conventionally, for sample sizes of roughly $n \geq 30$). As n increases, the sample means concentrate ever closer to μ.

The **standard error of the mean** measures the dispersion of the sample means and therefore the precision of $\bar{X}$ as an estimate of μ:

$$SE_{\bar{X}} = S_{\bar{X}} = \frac{S}{\sqrt{n}},$$

where S is the sample standard deviation and n is the sample size. A small standard error does not exclude systematic error: a biased sample can give a very precise but incorrect estimate.

! What the standard error depends on

- With a larger **standard deviation** the observations are more spread out and the standard error is larger.
- With a larger **sample size** the standard error is smaller. It decreases with $\sqrt{n}$: to halve it, we must increase the sample four times.

> ⚠ Do not confuse the standard deviation and the standard error
>
> The **standard deviation** S describes the dispersion of the **individual observations** around the mean. The **standard error** $SE_{\bar{X}}$ describes the dispersion of the **sample mean** from sample to sample. For $n > 1$ the standard error is smaller than the standard deviation.

3.3 Confidence Interval for the Mean

The point estimate $\bar{X}$ is a single number, but it does not show how precise it is. The **confidence interval** adds boundaries around it and thus presents the uncertainty of the estimate. Its general form is:

$$\text{estimate} \pm \text{critical value} \times \text{standard error}.$$

When the population standard deviation σ is known, the critical value Z is used:

$$CI = \bar{X} \pm Z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

In practice σ is almost always unknown and is replaced by S. Then for the mean the correct multiplier is a critical value from the t-distribution with $n - 1$ degrees of freedom:

$$CI = \bar{X} \pm t_{1-\alpha/2,\, n-1} \cdot \frac{S}{\sqrt{n}}.$$

This interval assumes independent, randomly selected observations. With a small sample it is also necessary that the distribution in the population is approximately normal, without strong asymmetry and extreme values. With larger samples the method is more robust to moderate deviations from normality.

With large samples t and Z are almost identical and the approximation $Z = 1.96$ is often used. With small and moderate samples t is slightly larger and gives a wider interval. Table 3.2 gives the values of Z that are also used as a large-sample approximation.

Table 3.2: Critical values of the standard normal distribution

Confidence level	α	$Z_{1-\alpha/2}$
90 %	0.10	1.645
95 %	**0.05**	**1.960**
99 %	0.01	2.576

3.3.1 What "95 % confidence" Means

Before the sample is selected, the boundaries of the interval are random, because they depend on the randomly selected observations. If the procedure is repeated many times, about 95 % of the intervals constructed in this way will contain the true parameter. In a particular series the share may not be exactly 95 % (Figure 3.2).

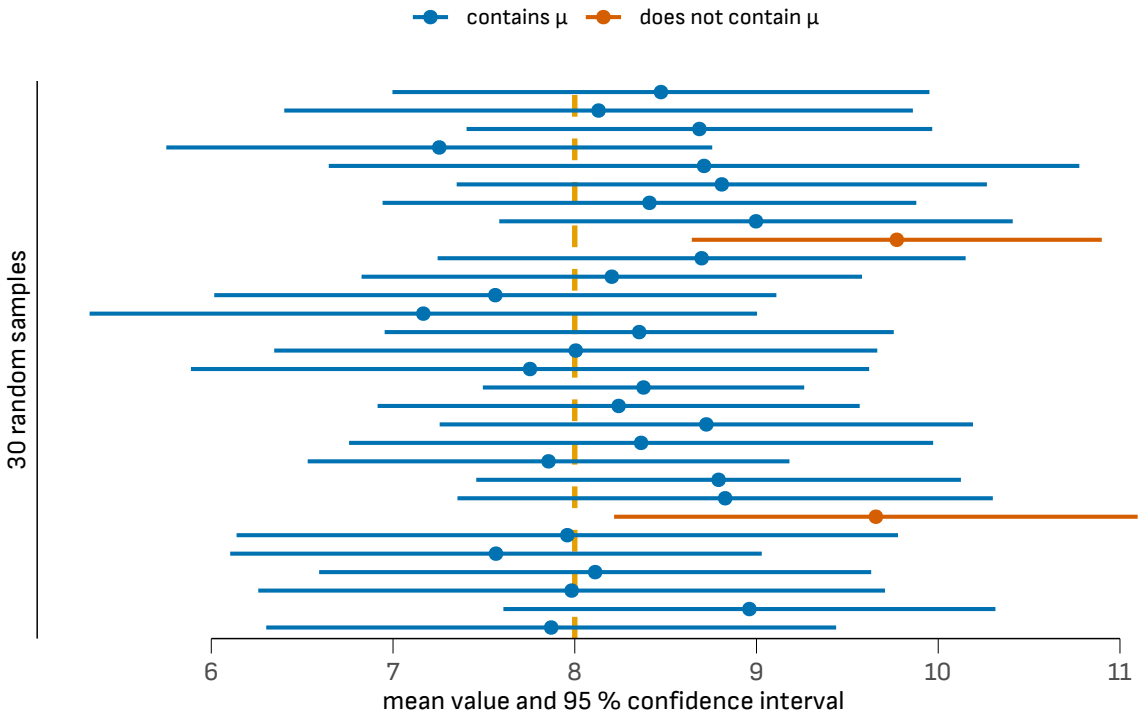

Figure 3.2: Thirty 95 % confidence intervals for the same population mean. The blue intervals contain the true value μ = 8, and the red ones do not. The 95 % level describes the reliability of the method in a long series, not a promise of an exact number of successful intervals in each series.

Once the concrete boundaries are computed, the parameter is one fixed, albeit unknown, value. Therefore, with this interpretation we do not say: "there is a 95 % probability that the parameter is in this particular interval". We say: "the computed 95 % confidence interval is from … to …" and state that the method used has 95 % coverage under repeated application and fulfilled assumptions.

! Why we do not use a 100 % confidence interval

A non-trivial finite interval with 100 % coverage cannot, as a rule, be constructed. The interval that includes all possible values would be certain, but practically useless. Statistics does not remove uncertainty — it measures it.

3.3.2 Worked Example: Confidence Interval for the Mean

Setting

$n = 30$ patients treated with an experimental therapy were followed up. Their survival is presented as a grouped frequency distribution:

Survival X, months	Frequency f
1–5	7
6–10	14
11–15	9
Total	**30**

Task. Compute a 95 % confidence interval for the population mean survival μ.

Step 1. Sample arithmetic mean. The midpoints of the intervals are 3, 8, and 13 months.

$$\bar{X} = \frac{3 \cdot 7 + 8 \cdot 14 + 13 \cdot 9}{30} = \frac{250}{30} = 8.33 \text{ months.}$$

Step 2. Sample standard deviation. For grouped data we use the midpoints of the intervals as representative values:

$$S = \sqrt{\frac{\sum (X_u - \bar{X})^2 f}{n - 1}} = \sqrt{\frac{396.67}{29}} = 3.70 \text{ months.}$$

Step 3. Standard error of the mean.

$$SE_{\bar{X}} = \frac{S}{\sqrt{n}} = \frac{3.70}{\sqrt{30}} = 0.68 \text{ months.}$$

Step 4. Critical value. Since σ is unknown and is replaced by S, we use $t_{0.975;29} = 2.045$.

Step 5. Confidence interval. We compute with the unrounded intermediate values and round only the final result:

$$95\,\%\, CI = 8.33 \pm 2.045 \cdot 0.675 = 8.33 \pm 1.38 = [6.95;\ 9.71] \text{ months.}$$

Conclusion. The estimated population mean survival is 8.33 months, and its 95 % confidence interval is from **6.95 to 9.71 months**. If we repeatedly take samples in the same way and construct intervals with the same procedure, about 95 % of them will contain the true mean μ.

⚠ Do not round the intermediate results too early

If we replace 8.33 with 8 and 0.675 with 0.7 already before the final computation, the centre and the boundaries of the interval shift. Keep at least three or four digits in the intermediate steps and round the final answer according to the precision of the measurement.

3.4 Confidence Interval for a Proportion

When $\hat{p}$ is written as a percentage, the standard error of the proportion is computed with:

$$SE_{\hat{p}} = S_p = \sqrt{\frac{\hat{p}(100 - \hat{p})}{n}}.$$

With a sufficiently large sample the 95 % interval by the normal approximation is:

$$CI = \hat{p} \pm 1.96 \cdot SE_{\hat{p}}.$$

The approximation is acceptable when the number of observations with the event and the number without the event are not small; a practical rule is that both should be at least 5. For rare events or small samples, a Wilson interval or an exact binomial interval, computed with statistical software, is preferred. Otherwise the simple formula above can give impossible boundaries below 0 % or above 100 %.

3.4.1 Worked Example: Confidence Interval for a Proportion

Setting

Type of diet	With improvement	Without improvement	Total
Gluten-free	56	44	100
Control	45	55	100
Total	**101**	**99**	**200**

Task. Compute a 95 % confidence interval for the population proportion with improvement on the gluten-free diet P_1.

Step 1. Sample proportion.

$$\hat{p}_1 = \frac{56}{100} \cdot 100 = 56\,\%.$$

Step 2. Standard error of the proportion.

$$SE_{\hat{p}_1} = \sqrt{\frac{56(100-56)}{100}} = \sqrt{24.64} = 4.96 \text{ percentage points.}$$

Step 3. Confidence interval.

$$95\,\%\; CI = 56 \pm 1.96 \cdot 4.96 = 56 \pm 9.73 = [46.27;\; 65.73]\,\%.$$

Conclusion. The sample proportion with improvement is 56 %, and its 95 % confidence interval is from **46.27 % to 65.73 %**. The conditions for the normal approximation are fulfilled, because there are 56 observations with improvement and 44 without improvement.

3.5 What Changes the Width of the Interval

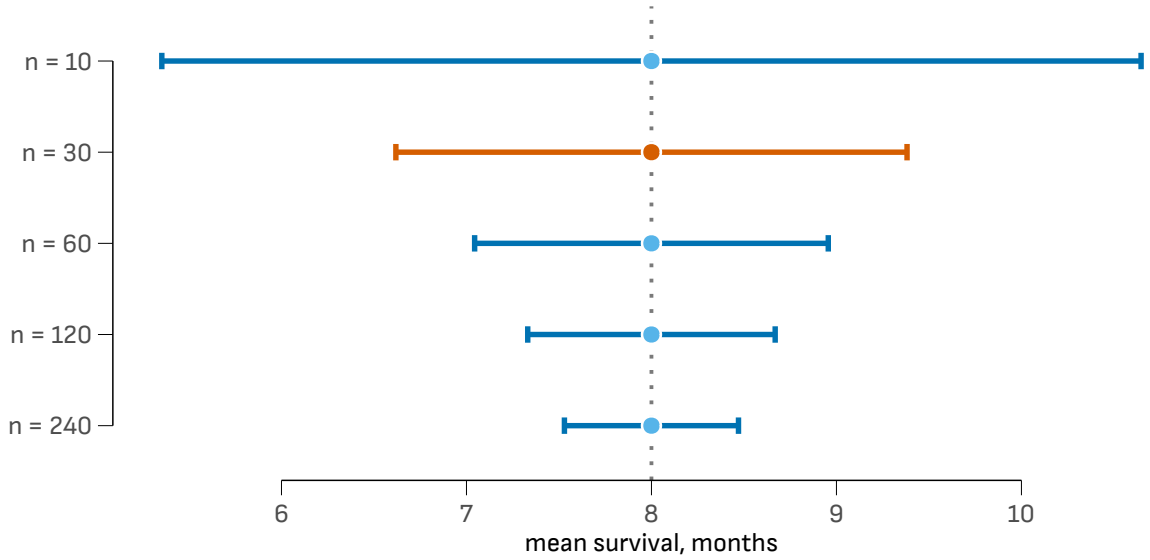

Figure 3.3: Influence of the sample size on the width of the 95 % confidence interval. The mean (8 months) and the standard deviation (3.7 months) are the same; only n changes. The red row is the example with n = 30.

Table 3.3: What the width of the confidence interval depends on

If we...	the width of the interval...	because...
increase the sample size n	**decreases**	the standard error decreases
increase the standard deviation S	**increases**	the standard error increases
choose 99 % instead of 95 %	**increases**	the critical value is larger
choose 90 % instead of 95 %	**decreases**	the critical value is smaller

⚠ The trade-off between coverage and precision

With the same data, the 99 % interval has higher coverage, but it is wider and gives a less precise estimate. Greater precision without reducing the confidence level is achieved through a larger sample or through more precise measurements that reduce the dispersion.

3.6 Comparing Two Groups with Confidence Intervals

Confidence intervals are useful for visual orientation, but the individual intervals of two groups are not a confidence interval of the **difference between the groups**.

❗ Visual orientation, not a formal test

For two independent groups, the absence of overlap between their 95 % confidence intervals is a strong indication of a difference. The overlap, however, does not prove that there is no difference. The categorical conclusion must be based on an appropriate hypothesis test or on a confidence interval for the difference itself (Section 4).

Example

The 95 % CI for the mean weight loss on diet A is [4.5; 7.2] kg, and on diet B — [2.2; 4.2] kg. The intervals do not overlap, and the data convincingly point to a larger mean loss on diet A. For a formal conclusion, an interval for the difference is computed or an appropriate test is applied. If the intervals are [4.5; 7.2] and [2.2; 5.0], they overlap. From this alone one cannot conclude either "there is a difference" or "there is no difference".

3.7 Common Mistakes

Table 3.4: Common mistakes with confidence intervals

#	Mistake	How to avoid it
1	"There is a 95 % probability that the true value is in this interval"	The level describes the long-term coverage of the method
2	"95 % of the patients have a value in the interval"	The confidence interval refers to a parameter, not to individual patients
3	Using S instead of $SE_{\bar{X}}$ in the formula	The interval is constructed with the standard error
4	Using Z with a small sample and unknown σ	For the mean use a critical value from the t-distribution
5	Confusing 0.56 and 56 in the formula for a proportion	The formula with $(100 - \hat{p})$ requires a percentage
6	Conclusion "no difference" when intervals overlap	A test or an interval of the difference is needed

#	Mistake	How to avoid it
7	Early rounding and a missing unit of measurement	Round at the end and write the unit at both boundaries

3.8 Self-Study and Homework Tasks

Task 3.1

Compute a 95 % confidence interval for the mean population incubation period of patients with hepatitis A. Use a t-critical value and interpret the result with a unit of measurement.

Incubation period, days	1–15	16–30	31–45	46–60	61–75	Total
Frequency f, number of patients	40	120	75	14	1	250

Task 3.2

Compute a 95 % confidence interval for the mean population survival of patients with lung cancer treated with an experimental therapy.

Survival, months	1–5	6–10	11–15	16–20	21–25	Total
Frequency f, number of patients	35	65	95	70	35	300

Task 3.3

Use the data from the gluten-free diet example (56 of 100 with improvement), but compute a **99 %** confidence interval. How do the point estimate and the width of the interval change relative to the 95 % CI?

Task 3.4

In a prospective study of patients with myasthenia gravis, severe myasthenic crises were observed in 23 of 75 patients: $\hat{p} = 30.7\ \%$ and the 95 % CI is [20.1 %; 40.3 %]. Which statement is the most correct?

a) In the population the frequency is necessarily greater than 30.7 %.

b) Every individual patient has exactly a 30.7 % probability of developing a severe crisis.

c) There is a 95 % probability that the sample proportion is between 20.1 % and 40.3 %.

d) The interval was obtained with a method that, under repeated application, would cover the true population proportion in about 95 % of the cases.

Task 3.5

With the same S and confidence level, a researcher increases the sample from 50 to 200 observations. How many times does the standard error change? How does this affect the width of the interval?

Task 3.6

In a pilot study, improvement was observed in 2 of 20 patients. Check the condition for the normal approximation. Is it appropriate to use the simple interval $\hat{p} \pm 1.96SE_{\hat{p}}$? Suggest a more appropriate choice.

Task 3.7

Two independent groups have 95 % confidence intervals [21 %; 38 %] and [34 %; 48 %]. What can and what cannot be concluded from their overlap alone?

3.9 Short Quiz

For each question choose **one** correct answer. The key is in Section D.3.

1. Which is a parameter?

   a) the mean pulse of 40 examined students;
   b) the proportion with improvement among 100 included patients;
   c) the true mean pulse of all second-year students;
   d) the standard deviation in a particular sample.

2. If the sample size is increased four times with unchanged S, the standard error:

   a) increases four times;
   b) increases two times;
   c) decreases two times;
   d) does not change.

3. Which interpretation of a 95 % confidence interval is correct?

   a) 95 % of the observations are between the two boundaries;
   b) the method will give intervals containing the true parameter in about 95 % of the repeated applications;
   c) there is a 95 % probability that the sample mean is in the interval;
   d) the parameter varies randomly between the two boundaries.

4. Which action widens the confidence interval with the same data?

   a) increasing n;
   b) moving from 95 % to 90 % confidence;
   c) moving from 95 % to 99 % confidence;

d) decreasing the standard deviation.

5. With a small sample and an unknown population standard deviation, the interval for the mean uses:

   a) a t-critical value with $n - 1$ degrees of freedom;
   b) always $Z = 1.96$;
   c) the standard deviation without dividing by $\sqrt{n}$;
   d) only the sample mean.

6. Two 95 % confidence intervals overlap. Which conclusion is admissible?

   a) the groups certainly do not differ;
   b) the groups certainly differ;
   c) the overlap alone is not enough; an interval of the difference or an appropriate test is needed;
   d) both means are equal to zero.

4 Hypothesis Testing. Z-Test and t-Test

Aim of the exercise

After this exercise you should be able to:

1. formulate the null and the alternative hypothesis for a specific task;
2. explain what a p-value is and what it is **not**;
3. distinguish a type I from a type II error and connect the power with them;
4. choose an appropriate statistical test according to the type of the variable, the number of groups, and the design;
5. apply the Z-test for two proportions and the t-test for two independent means and formulate the conclusion.

What you should know beforehand

From Section 2: arithmetic mean, standard deviation, proportion. From Section 3: standard error and confidence interval.

4.1 Hypotheses

A statistical hypothesis is a testable statement about a parameter or about an association in the population. The sample data are used to assess whether they are compatible with this statement. In the tests considered in this chapter, the procedure starts with the assumption that the **null hypothesis** is true.

Definitions

The **null hypothesis** H_0 in tests for a difference states that the population difference is zero. For example $H_0 : \mu_1 - \mu_2 = 0$ or $H_0 : P_1 - P_2 = 0$.
The **alternative hypothesis** H_1 states that the population difference is not zero. In a two-sided test: $H_1 : \mu_1 - \mu_2 \neq 0$.
The statistical test does not require the researcher to believe in H_0. It is accepted temporarily as a starting assumption, in order to compute how unusual the observed data are in the absence of a population difference.

Example

A new antihypertensive drug is being tested. One group receives the drug, the other — a placebo. At the end of the study, the blood pressure in the experimental group is on average 10 mmHg lower.
H_0: The population mean change in blood pressure is the same for the drug and the placebo: $\mu_1 - \mu_2 = 0$.

H_1: The population mean change differs between the two groups: $\mu_1 - \mu_2 \neq 0$.
The difference of 10 mmHg is a sample estimate. The test shows how compatible it is with H_0, but by itself it does not prove a causal effect. The causal conclusion also depends on the design, the randomisation, the adherence to the treatment, and the possible systematic errors.

4.2 The p-Value

Definition

The p-value is the probability, **if H_0 and the assumptions of the test are true**, of obtaining a test statistic at least as far from what is expected under H_0 as the observed one.

p is a number between 0 and 1. A small p-value means that the data are poorly compatible with H_0. A large p-value means that the data do not give sufficient reason to reject H_0; it does not show that H_0 is probably true.

The **significance level** α is a pre-selected threshold for the decision. In medical research $\alpha = 0.05$ is often used. The threshold must be set before the analysis, not changed after the results are obtained.

Decision rule

1. If $p < \alpha$ we **reject** H_0. The result is statistically significant at the chosen α.
2. If $p \geq \alpha$ we **do not reject** H_0. The data are not sufficient for a statistically significant conclusion.

In part of the teaching materials the second decision is written as "we accept H_0". The more accurate formulation is "we do not reject", because a non-significant result does not prove equality.

How to read $p = 0.09$

If H_0 and the assumptions of the test are true, then under repeated application of the same design in about 9 % of the runs a test statistic at least as extreme as the observed one will be obtained. Since $0.09 > 0.05$, we do not reject H_0. This does not mean that the probability that H_0 is true is 9 %.

What the p-value is NOT

- p is **not** the probability that the null hypothesis is true.
- p is **not** the probability that the result will be reproduced in a new study.
- p does **not measure** the size of the effect. A very small p with a huge sample can accompany a clinically unimportant effect.
- $p \geq 0.05$ does **not prove** that the groups are identical or that there is no effect.

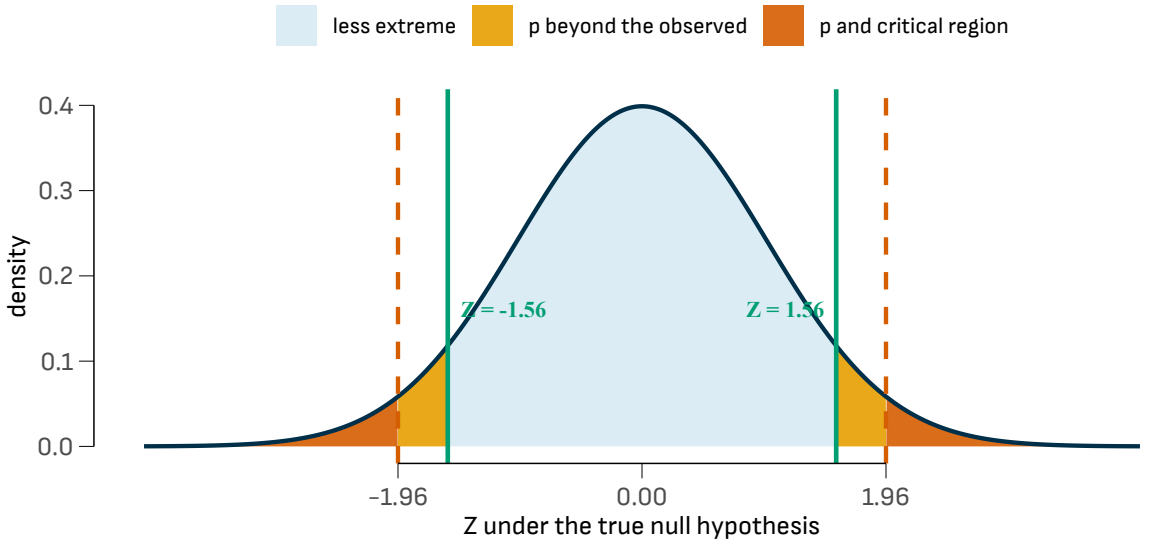

Figure 4.1: Two-sided p-value at Z = 1.56. The orange and red tails together contain the values at least as extreme as the observed one; their total area is p = 0.12. The dashed lines are the critical boundaries ±1.96 at α = 0.05.

4.3 Type I and Type II Errors

With every decision made on the basis of a sample, two kinds of errors are possible (Table 4.1).

Table 4.1: Types of errors in hypothesis testing

	H_0 is **true**	H_0 is **false**
We **reject** H_0	type I error α	correct decision
We **do not reject** H_0	correct decision	type II error β

The **type I error** α means declaring as a "scientific discovery" something that in reality does not exist. The pre-specified α limits the probability of such an error when H_0 is true.

The **type II error** β means missing a difference that really exists. It depends on the size of the effect, the variability, the sample size, the chosen α, and the statistical test.

> **!** The two errors are related
>
> With unchanged sample size, effect, and test, decreasing α from 0.05 to 0.01 reduces the risk of a type I error, but **increases** the risk of a type II error. Reducing both risks simultaneously usually requires a larger sample or more precise measurements.

4.4 Power

The **power** is the probability that the test rejects H_0 when a pre-specified real difference actually exists:

$$\text{power} = 1 - \beta$$

It depends on four factors:

1. **Significance level** α. Increasing α (from 0.01 to 0.05) decreases β and increases the power.
2. **Effect size** Δ. The larger the real difference, the easier it is to demonstrate.
3. **Variability**. The less the observations vary, the smaller the standard error and the greater the power.
4. **Sample size** n. More observations → smaller standard error → higher power.

A typical exam question

All other things being equal, the smallest number of units of observation will be needed when: the population is homogeneous and the clinical effect is large in size. Homogeneity reduces the dispersion, and a large effect is easier to detect — both increase the power, so fewer observations suffice.

4.5 Two-Sided and One-Sided Alternatives

In a **two-sided test** the alternative hypothesis allows a difference in both directions: $H_1 : \mu_1 - \mu_2 \neq 0$. In a **one-sided test** the direction is specified in advance, for example $H_1 : \mu_1 - \mu_2 > 0$. A one-sided test is not chosen after the direction of the result is already known. All worked examples in this chapter are two-sided.

4.6 Choosing a Statistical Test

The choice starts with the type of the outcome: quantitative, ordinal, or categorical. Then the number of groups is determined, and whether the observations are independent or paired. Finally, the assumptions of the particular test are checked.

Table 4.2: A guide to choosing a statistical test

Outcome and design	One group	Two groups	Three or more groups
quantitative, independent	one-sample t-test	Welch's t-test	ANOVA
quantitative, paired	—	paired t-test	repeated-measures ANOVA

Outcome and design	One group	Two groups	Three or more groups
quantitative/ordinal, non-parametric	Wilcoxon signed-rank test	Mann-Whitney for independent; Wilcoxon for paired	Kruskal-Wallis for independent; Friedman for paired
categorical, independent	binomial test or Z-test for one proportion	Z-test for two proportions or χ^2-test	χ^2-test
dichotomous, paired	—	McNemar's test	Cochran's Q test

"Non-parametric" does not automatically mean "the same test for the mean without normality". For example, the Mann-Whitney test compares distributions and, under additional assumptions, can be interpreted as a difference in location. For two independent means, Welch's t-test is often robust to moderate deviation from normality, especially with sufficiently large groups and in the absence of extreme values.

> **How to recognise a paired design**
>
> Ask yourself: *was the same unit of observation measured more than once?* "Change from baseline", "before and after treatment", "left and right eye of one patient" — all of this is a paired design.

4.7 Z-Test for Comparing Two Proportions

4.7.1 Worked Example

> **Setting**

Gluten-free diet regimen	With improvement	Without improvement	Total
Yes	56	44	100
No	45	55	100
Total	**101**	**99**	**200**

$\hat{p}_1 = 56\,\%$, $\hat{p}_2 = 45\,\%$, difference (clinical effect) $\Delta = 11$ percentage points.
Task. Is there a difference between the two groups with respect to the proportion of cases with improvement?

Step 1. Hypotheses.

$H_0 : P_1 - P_2 = 0$ – the population proportions with improvement are the same.

$H_1 : P_1 - P_2 \neq 0$ – the population proportions with improvement differ.

The test is two-sided, and the pre-selected level is $\alpha = 0.05$.

Step 2. Computing the test statistic.

First the **pooled** proportion is computed:

$$\hat{p} = \frac{m_1 + m_2}{n_1 + n_2} \times 100 = \frac{56 + 45}{100 + 100} \times 100 = \frac{101}{200} \times 100 = 50.5\,\%$$

Then:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p} \cdot (100 - \hat{p}) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$Z = \frac{56 - 45}{\sqrt{50.5 \cdot 49.5 \cdot \left(\frac{1}{100} + \frac{1}{100}\right)}} = \frac{11}{\sqrt{49.995}} = \frac{11}{7.07} = 1.56$$

Step 3. Determining the p-value. We compare the absolute value $|Z|$ with the two-sided critical values of the normal distribution (Table 4.3).

Table 4.3: Critical values of Z

$p = 0.10$	$p = 0.05$	$p = 0.01$
$Z = 1.645$	$Z = 1.960$	$Z = 2.576$

The computed $|Z| = 1.56$ is smaller than 1.645. From the table we write $p > 0.10$; the statistical software gives a two-sided $p = 0.120$.

Step 4. Conclusion. Since $p = 0.120 > \alpha$, we do not reject H_0. The data do not give sufficient evidence for a difference between the population proportions with improvement. This does not prove that the two diets are equally effective.

Step 5. Size and precision of the effect. The sample difference is $56\% - 45\% = 11$ percentage points. For the confidence interval of the difference we use the unpooled standard error:

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{56 \cdot 44}{100} + \frac{45 \cdot 55}{100}} = 7.03 \text{ percentage points.}$$

$$95\,\%\,CI = 11 \pm 1.96 \cdot 7.03 = [-2.8;\ 24.8] \text{ percentage points.}$$

The interval includes zero and leads to the same statistical conclusion. It also shows the wide range of population differences compatible with the data.

Why the standard errors differ

In the Z-test the pooled proportion is used, because the computation is performed under $H_0 : P_1 = P_2$. In the confidence interval we do not impose equality and use the separate sample proportions. Do not mix the two standard errors.

4.8 t-Test for Comparing Two Independent Samples

Welch's t-test checks whether two independent population means differ. It does not require equal variances and is therefore a suitable default choice for two independent groups.

Conditions for application

1. The outcome is quantitative, and the observations in the two groups are independent.
2. The samples are random, or the design allows generalisation to the target population.
3. With small samples the distributions are approximately normal and without extreme values. With larger samples the test is robust to moderate asymmetry.
4. The observations within each group are independent. Repeated measurements on one patient require a paired test or another approach.

4.8.1 Worked Example

Setting

Type of treatment	Sample size	Mean survival	Standard deviation
Experimental	$n_1 = 30$	$\bar{X}_1 = 15$ months	$S_1 = 4.5$ months
Chemotherapy	$n_2 = 30$	$\bar{X}_2 = 11$ months	$S_2 = 1.5$ months

Difference (clinical effect) $\Delta = 15 - 11 = 4$ months.
Task. Is there a difference between the two groups with respect to the mean survival?

Step 1. Hypotheses.

$H_0 : \mu_1 - \mu_2 = 0$ – the population mean survivals are the same.

$H_1 : \mu_1 - \mu_2 \neq 0$ – the population mean survivals differ.

The test is two-sided and $\alpha = 0.05$.

Step 2. Computing the test statistic.

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$t = \frac{|15 - 11|}{\sqrt{\frac{4.5^2}{30} + \frac{1.5^2}{30}}} = \frac{4}{\sqrt{0.75}} = \frac{4}{0.866} = 4.62.$$

> **The same formula written with standard errors**
>
> Since $S_{\bar{X}} = S/\sqrt{n}$, it follows that $S_{\bar{X}}^2 = S^2/n$. Therefore the formula can also be written as:
>
> $$t = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{S_{\bar{X}_1}^2 + S_{\bar{X}_2}^2}}$$
>
> The two notations are **one and the same** expression. Use the second when the setting gives the standard errors directly (the notation $\bar{X} \pm S_{\bar{X}}$), and the first when the standard deviations and the sizes are given.

Step 3. Determining the p-value.

In Welch's test the degrees of freedom are computed with the Welch–Satterthwaite approximation:

$$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(S_2^2/n_2\right)^2}{n_2 - 1}} = 35.37.$$

For a manual search in the table (Section C) we use the nearest smaller integer row, for example $df = 35$:

df	$p = 0.10$	$p = 0.05$	$p = 0.01$
35	1.690	2.030	2.724

The computed $t = 4.62$ is larger than 2.724, therefore $p < 0.01$. The statistical software gives a two-sided $p < 0.001$.

Step 4. Conclusion. Since $p < 0.001 < \alpha$, we reject H_0. The data show a statistically significant difference in the mean survival between the two groups.

Step 5. Size and precision of the effect. The estimated difference is $15 - 11 = 4$ months. Its 95 % confidence interval is:

$$4 \pm t_{0.975;35.37} \cdot 0.866 = 4 \pm 2.03 \cdot 0.866 = [2.24;\ 5.76] \text{ months.}$$

The interval does not include zero and is consistent with the result of the test. The conclusion that the treatment causes better survival is justified only if the design and the conduct of the study allow a causal interpretation.

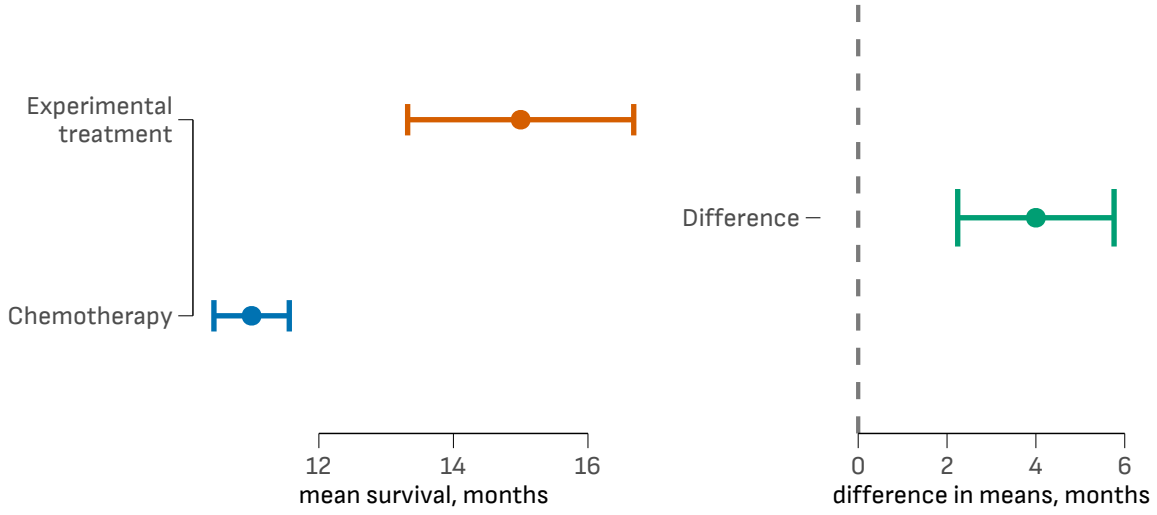

Figure 4.2: The mean survival and the 95 % confidence intervals for the two groups (left) and the estimated difference between the means (right). The interval of the difference does not include zero.

! Important relationships

All other things being equal:

- a larger **absolute difference** between the means → larger test statistic → smaller p-value;
- lower **variability** (small S) → smaller standard error → larger test statistic;
- larger **sample size** → smaller standard error → larger test statistic.

In summary: **the larger the computed test statistic, the smaller the p-value.**

4.9 Mandatory Exam Question

⚠ Format of the question

A study reaches a result $[Z;\ t,\ df;\ \chi^2,\ df]$. *What is the* p*-value in this case and what does it mean? What conclusion can be drawn?*

Example question and answer

Question. A study of the level of obesity among smokers and non-smokers reaches a result $t = 3.45$ $(df = 50)$. What is the p-value, what does it mean, and what conclusion follows?
Answer.

- *What is the* p*-value?* With $df = 50$ the critical value for a two-sided $p = 0.01$ is 2.678. Since $3.45 > 2.678$, from the table it follows that $p < 0.01$; the exact value is

approximately 0.001.

- *What does it mean?* If H_0 and the assumptions of the test are true, the probability of obtaining $|t| \geq 3.45$ is about 0.1 %.
- *What conclusion follows?* There is a statistically significant difference in the level of obesity between smokers and non-smokers, and we reject H_0 at $\alpha = 0.05$. Without an estimate and a confidence interval we cannot determine the direction, the magnitude, or the practical importance of the difference.

4.10 Statistical and Clinical Significance

A statistically significant difference is not automatically clinically significant, and a statistically non-significant result does not prove the absence of a clinically important effect.

⚠ Definitions

Statistical significance means that p is below the pre-selected α under the model used and its assumptions.
Clinical significance is determined by the size of the effect, its confidence interval, the benefits, the harms, the costs, and a pre-specified minimal clinically important difference.

A difference of 2 mmHg in systolic blood pressure can have a very small p-value with a huge sample, but its practical importance must be assessed in the concrete clinical and population context. Conversely, a clinically important effect can remain statistically unproven if the sample is too small.

4.11 Common Mistakes

Table 4.4: Common mistakes in hypothesis testing

#	Mistake	How to avoid it
1	Formulating H_0 as "there is a difference"	In the tests considered, H_0 sets a zero population difference
2	"We proved that there is no difference" when $p > 0.05$	We do not reject H_0; absence of evidence is not evidence of absence
3	Interpreting p as the probability that H_0 is true	p is the probability of such data, **if** H_0 is true
4	t-test for paired groups ("before and after")	Check the design before choosing a test
5	t-test with a small sample with strong asymmetry and extreme values	Examine the data and use an appropriate robust or non-parametric method
6	Computing Z without the pooled proportion	The pooled proportion is a mandatory step
7	A conclusion that ends with the number p	The conclusion is a sentence with the names of the variables

4.12 Self-Study and Homework Tasks

Task 4.1

400 patients with multiple sclerosis are included in a clinical study. 25 of 200 treated with the experimental therapy relapse within 6 months after the end of the therapeutic course, while with the conventional therapy this number is 45 of 200.
Is there a statistically significant difference between the two types of therapy with respect to the proportion of relapsing patients?

Task 4.2

600 patients with lung cancer are included in a clinical study. 250 of 300 treated with the experimental therapy survive at least 1 year, while with the conventional therapy this number is 190 of 300.
Is there a statistically significant difference with respect to the proportion of patients surviving at least 1 year?

Task 4.3

Obese smokers are included in a clinical study of a new cholesterol-lowering product. Three months after the start, the mean value (± standard error) is 213 ± 6 mg/dL for the experimental group and 228 ± 7 mg/dL for the control group.
Compute the test statistic. Can the exact p-value be determined if the sample sizes or the degrees of freedom are not given?

Task 4.4

Patients with gastric adenocarcinoma are divided into 3 groups according to the therapy. The 4-week changes in tumour size relative to baseline are compared; the quantity is not normally distributed. Which test should be applied?

a) Kruskal-Wallis H-test
b) ANOVA
c) Friedman test for paired samples
d) repeated-measures ANOVA

Task 4.5

All other things being equal, if we decrease the significance level from $\alpha = 0.05$ to $\alpha = 0.01$:

a) the probability of a type I error will increase, and of a type II error will decrease;
b) the probability of a type II error will increase, and of a type I error will decrease;
c) both probabilities will increase;
d) both probabilities will decrease.

Task 4.6

Patients with lung cancer are divided into 2 groups according to the therapy. Those treated with the experimental therapy survive on average 10.1 months, and those treated with chemotherapy — 8.7 months ($p = 0.08$). Which of the following statements is true?

a) We accept the null hypothesis that there is a statistically significant difference.
b) The t-test is appropriate, because the means are approximately equal.
c) Since the result is not significant, one may proceed to a paired t-test.
d) An appropriate alternative hypothesis is that the two groups have different survival depending on the type of treatment.

Task 4.7

A study of the level of physical activity (hours of training per week) among smokers and non-smokers reaches $t = 2.96$ with $df = 40$. Answer according to the scheme of the mandatory exam question.

Task 4.8

The power of a research design depends on four factors. Give an example of a combination of them that achieves the greatest possible power.

Task 4.9

In 40 patients the concentration of C-reactive protein is measured before and after treatment. The "after minus before" differences are approximately normally distributed. Which test is appropriate and how is H_0 formulated?

Task 4.10

In a large study the difference in the mean systolic blood pressure is 1.8 mmHg, 95 % CI [1.2; 2.4] mmHg, and $p < 0.001$. Distinguish the statistical conclusion from the assessment of the clinical significance.

4.13 Short Quiz

For each question choose **one** correct answer. The key is in Section D.4.

1. What is the p-value?

   a) the probability that H_0 is true;
   b) the probability, if H_0 is true, of obtaining an at least as extreme test statistic;
   c) the probability that the result will be reproduced;
   d) the size of the clinical effect.

2. With $p = 0.18$ and $\alpha = 0.05$:

a) we prove that H_0 is true;
b) we reject H_0;
c) we do not reject H_0;
d) we prove that the groups are equivalent.

3. Which increases the power, all other things being equal?

a) A smaller sample;
b) A smaller real effect;
c) greater variability;
d) A larger sample.

4. Which test is appropriate for a normally distributed quantitative variable measured before and after treatment in the same patients?

a) paired t-test;
b) welch's t-test for independent samples;
c) mann-Whitney test;
d) Z-test for two proportions.

5. Decreasing α from 0.05 to 0.01 with unchanged sample size and effect:

a) increases the type I error and decreases the type II error;
b) decreases the type I error and increases the type II error;
c) decreases both errors;
d) does not change the power.

6. What should be presented together with the p-value?

a) only the number of significant tests;
b) an estimate of the effect and a confidence interval;
c) only the chosen α;
d) only the direction of the alternative hypothesis.

5 Association between Categorical Variables. The χ^2-Test

Aim of the exercise

After this exercise you should be able to:

1. recognise when the χ^2-test is applied and distinguish it from the t-test and from correlation analysis;
2. build a contingency table from a verbal problem statement and find the marginal totals;
3. compute the expected frequencies under the true null hypothesis and check the conditions for applying the test;
4. compute the χ^2 statistic and the degrees of freedom and determine the p-value from the table;
5. assess the strength of the association with the φ coefficient or Cramér's V and formulate a complete conclusion.

What you should know beforehand

From Section 2: categorical variables, nominal and ordinal scale, proportion. From Section 4: null and alternative hypothesis, p-value, significance level, degrees of freedom.

5.1 When the χ^2-Test Is Used

The choice of a statistical method is determined by the type of the **two** variables whose association we study (Table 5.1).

Table 5.1: Choice of method according to the type of the two variables

First variable	Second variable	Appropriate method
quantitative	categorical with 2 categories	t-test (Section 4)
categorical with 2 categories	categorical with 2 categories	Z-test for two proportions or χ^2-test
categorical	categorical (any number of categories)	χ^2**-test** (this chapter)
quantitative	quantitative	correlation and regression (Section 6, Section 7)

When **both** variables are categorical, an arithmetic mean cannot be computed and the Pearson correlation coefficient is not applicable. We have only the **number of cases** in each combination

of categories. It is precisely on such data that **Pearson's test of independence**, also known as the **χ^2-test of association**, works.

Typical questions it answers:

- Is there an association between the frequency of physiotherapy and the degree of recovery?
- Is there an association between the type of antibiotic regimen and the success of eradication?
- Is there an association between the duration of work experience and the presence of neurological complaints?

> **i** The relationship between the Z-test and the χ^2-test
>
> For a 2 × 2 table, the two-sided Z-test for two proportions and the χ^2-test lead to the same conclusion, because $\chi^2 = Z^2$. For example, $Z = 1.56$ corresponds to $\chi^2 = 1.56^2 = 2.43$ at $df = 1$, and both values give $p > 0.10$.
> The χ^2-test is more general: it works also with more than two categories, where the Z-test is not applicable.

5.2 Contingency Table

> **⚠** Definition
>
> The **contingency table** distributes the units of observation simultaneously by two categorical variables. In its cells stand the **observed frequencies** O, and at the edges — the row totals R_i, the column totals C_j, and the overall total N, called **marginal totals**.

The general form of a table with R rows and C columns is presented in Table 5.2.

Table 5.2: General form of a contingency table

	Category B_1	Category B_2	Row total
Category A_1	O_{11}	O_{12}	R_1
Category A_2	O_{21}	O_{22}	R_2
Column total	C_1	C_2	N

> **!** Three rules for constructing the table
>
> 1. Every unit of observation falls into **exactly one** cell. The sum of all cells equals the total number of studied persons.
> 2. The categories of each variable are **exhaustive** and **mutually exclusive**.
> 3. In the cells stand **counts**, not percentages. The percentages are computed additionally, only for the verbal description of the table.

5.3 The Logic of the Test

The χ^2-test compares two pictures of the same data:

- **Observed frequencies** O – what was actually counted. These are the **facts**.
- **Expected frequencies** E – what we would count if there were **no association at all** between the two variables. These are the **expectations under the true null hypothesis**.

If the two pictures coincide, there is no reason to claim an association between the variables. The larger the discrepancy between them, the stronger the evidence against the null hypothesis.

> The intuition behind the expected frequencies
>
> If the frequency of physiotherapy has no importance for the outcome, the proportion of fully recovered should be the **same** under every regimen and coincide with the overall proportion in the whole sample. The expected frequencies are exactly this overall proportion, applied separately to each row.

5.4 Worked Example: a 3 × 3 Table

> Setting
>
> 500 patients receiving different physiotherapy regimens after joint replacement were followed up. The achieved degree of recovery was recorded.
>
> | Frequency of physiotherapy | Unsatisfactory | Moderate | Complete | Total |
> |---|---|---|---|---|
> | 1 week once | 65 | 70 | 15 | 150 |
> | 1 week monthly | 100 | 110 | 40 | 250 |
> | 2 weeks monthly | 15 | 40 | 45 | 100 |
> | **Total** | **180** | **220** | **100** | **500** |
>
> **Task.** Is there an association (dependence) between the frequency of the physiotherapy treatment and the degree of recovery?

Before the computations the data are described verbally. The share of fully recovered is $15/150 = 10.0\ \%$ under the weakest regimen, $40/250 = 16.0\ \%$ under the medium one, and $45/100 = 45.0\ \%$ under the most intensive one (Figure 5.1). The differences look large, but a test is needed for a conclusion.

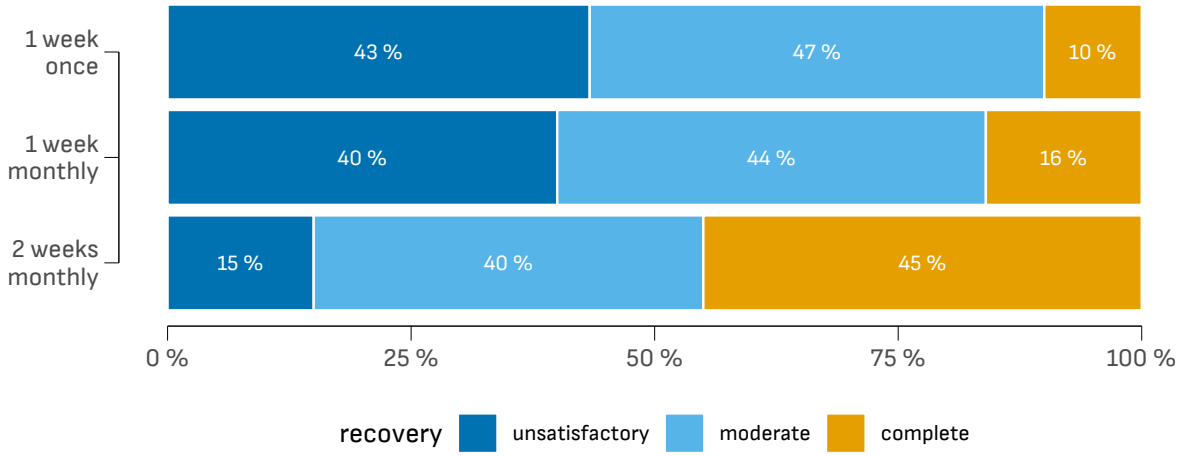

Figure 5.1: Distribution of the outcome under the three physiotherapy regimens. The share of fully recovered increases from 10 % under the weakest to 45 % under the most intensive regimen.

5.4.1 Step 1. Hypotheses and Significance Level

H_0: There is no association (dependence) between the frequency of the physiotherapy treatment and the degree of recovery. The two variables are independent, and the frequency of physiotherapy is not a factor for the outcome.

H_1: There is an association between the frequency of the physiotherapy treatment and the degree of recovery. The two variables are not independent.

The hypothesis under test is the null one; the pre-selected significance level is $\alpha = 0.05$.

> **i** Why the alternative has no direction
>
> The χ^2-test does not check whether one particular regimen is better than another. It checks only whether the distribution of the outcome is the **same** under all regimens. Therefore the alternative hypothesis contains no direction and the test has no one-sided version in the sense of Section 4.

5.4.2 Step 2. Expected Frequencies

$$E_{ij} = \frac{R_i \times C_j}{N}$$

where R_i is the total of the i-th row, C_j — the total of the j-th column, and N — the overall number of observations.

For the cell “1 week once × unsatisfactory”:

$$E_{11} = \frac{150 \times 180}{500} = \frac{27{,}000}{500} = 54.0$$

For the cell “2 weeks monthly × complete”:

$$E_{33} = \frac{100 \times 100}{500} = 20.0$$

The complete set of expected frequencies is given in Table 5.3.

Table 5.3: Expected (theoretical) frequencies under the true H_0

Frequency of physiotherapy	Unsatisfactory	Moderate	Complete	Total
1 week once	54.0	66.0	30.0	150
1 week monthly	90.0	110.0	50.0	250
2 weeks monthly	36.0	44.0	20.0	100
Total	**180**	**220**	**100**	**500**

Two checks and one saving

- The marginal totals of the expected frequencies **coincide** with those of the observed ones. If they do not coincide for you, there is an arithmetic error.
- Therefore the last cell in each row and each column can be obtained by subtraction: $150 - 54.0 - 66.0 = 30.0$.
- Note that under the true H_0 the share of fully recovered is the same in all rows: $30/150 = 50/250 = 20/100 = 20\,\%$. This is exactly what “no association” means.

The comparison of the two pictures is presented in Figure 5.2.

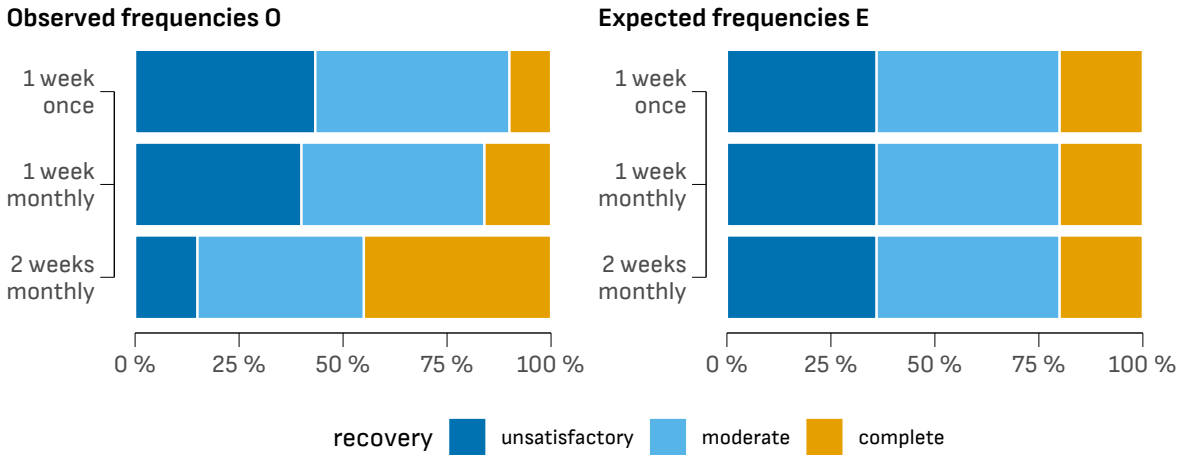

Figure 5.2: Observed frequencies (left) and expected frequencies under the true null hypothesis (right). Under independence the distribution of the outcome would be the same in all three regimens; the observed data deviate from this picture.

5.4.3 Step 3. Checking the Conditions

> **!** Conditions for applying the χ^2-test
>
> 1. The observations are **independent** — each person falls into only one cell.
> 2. In the cells stand **counts**, not percentages or mean values.
> 3. At least **80 %** of the expected frequencies are ≥ 5.
> 4. **There is no** expected frequency smaller than 1.
>
> If condition 3 or 4 is not fulfilled, for a 2 × 2 table Fisher's exact test is used, and for larger tables neighbouring categories are combined. The combining is decided **before** the result is seen.

In the example the smallest expected frequency is 20.0. All four conditions are fulfilled.

5.4.4 Step 4. Computing the χ^2 Statistic

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where the summation is over all cells of the table, and for each cell:

- $O - E$ is the difference between the fact and the expectation;
- $(O - E)^2$ is the square of the difference — squaring removes the signs, so that positive and negative deviations do not cancel each other out;
- the division by E weighs the difference relative to the scale of the cell: a difference of 10 cases is a lot with 20 expected and little with 500 expected.

Table 5.4: Computing χ^2 by cells

Cell	O	E	$O - E$	$(O - E)^2$	$(O - E)^2/E$
once × unsatisfactory	65	54.0	11.0	121.0	2.24
once × moderate	70	66.0	4.0	16.0	0.24
once × complete	15	30.0	−15.0	225.0	7.50
monthly × unsatisfactory	100	90.0	10.0	100.0	1.11
monthly × moderate	110	110.0	0.0	0.0	0.00
monthly × complete	40	50.0	−10.0	100.0	2.00
2 weeks × unsatisfactory	15	36.0	−21.0	441.0	12.25
2 weeks × moderate	40	44.0	−4.0	16.0	0.36
2 weeks × complete	45	20.0	25.0	625.0	31.25
Total	**500**	**500**	**0.0**		**56.96**

$$\chi^2 = 2.24 + 0.24 + 7.50 + 1.11 + 0.00 + 2.00 + 12.25 + 0.36 + 31.25 = 56.96$$

> **Check**
>
> The sum of the $O - E$ column is always zero. If it is not, some expected frequency has been computed wrongly.

The contribution of the individual cells shows **where** the association lies (Figure 5.3).

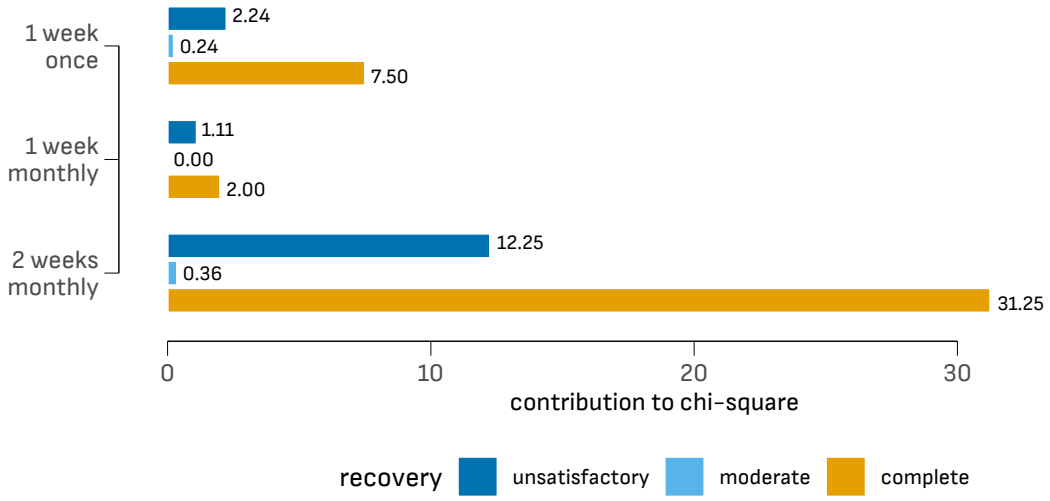

Figure 5.3: Contribution of the individual cells to the value of chi-square. About three quarters of the overall value come from the two cells of the most intensive regimen: many more complete recoveries and many fewer unsatisfactory results are observed than expected under independence.

5.4.5 Step 5. Degrees of Freedom and p-Value

$$df = (R - 1) \times (C - 1)$$

where R is the number of categories by rows, and C — the number of categories by columns. The degrees of freedom do **not depend** on the sample size, but only on the size of the table.

In the example the table is 3 × 3:

$$df = (3 - 1) \times (3 - 1) = 2 \times 2 = 4$$

The degrees of freedom determine which row of the table for the χ^2-distribution is used (Section C).

Table 5.5: Critical values of χ^2 at $df = 4$

df	$p = 0.10$	$p = 0.05$	$p = 0.01$
4	7.779	9.488	13.277

The computed $\chi^2 = 56.96$ is larger than 13.277 — the rightmost critical value. From the table we write:

$$\chi^2 = 56.96 \text{ at } df = 4 \Rightarrow p < 0.01$$

The statistical software gives $p < 0.001$.

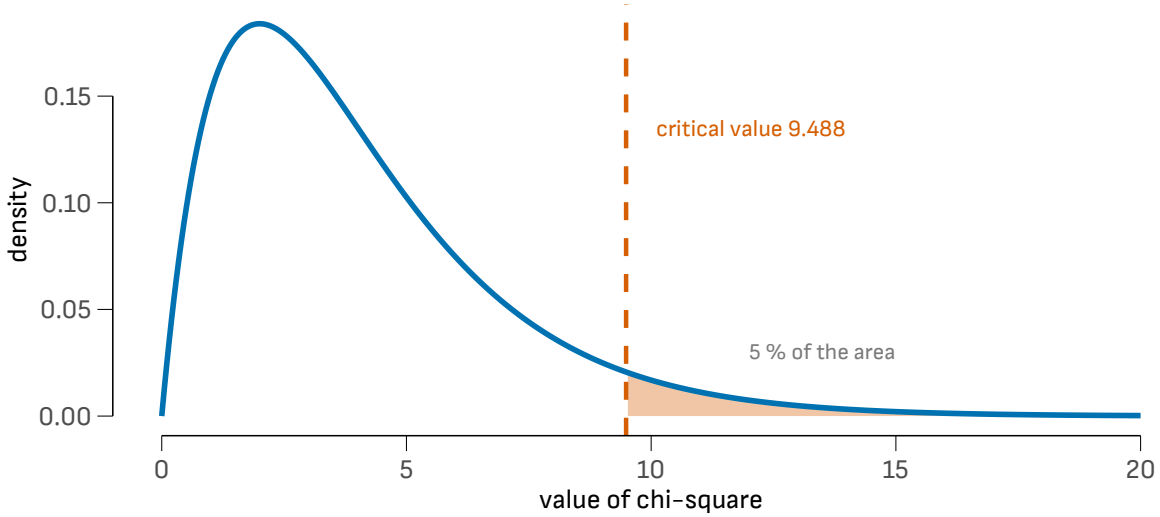

Figure 5.4: Chi-square distribution at df = 4. The shaded area to the right of the critical value 9.488 corresponds to a 5 % probability. The computed value 56.96 remains far outside the range of the graph, that is, the p-value is very small.

> **!** The direction of the test
>
> The larger the computed χ^2 statistic, the smaller the p-value and the stronger the evidence against the null hypothesis. Rejection is obtained only for **large** values, therefore the critical region is always at the right end of the distribution.

5.4.6 Step 6. Strength of the Association

The χ^2 statistic answers the question *is there an association*, but not *how strong is it*. Its value grows proportionally to the sample size: with a sufficiently large N even a negligible association becomes statistically significant. Therefore the result is always accompanied by a **coefficient of association**.

$$V = \sqrt{\frac{\chi^2}{N \cdot (k-1)}}$$

where N is the overall number of observations and $k = \min(R,\ C)$ — the smaller of the number of rows and the number of columns.

In the example $k = \min(3,\ 3) = 3$, therefore $k - 1 = 2$:

$$V = \sqrt{\frac{56.96}{500 \cdot 2}} = \sqrt{0.0570} = 0.24$$

V takes values from 0 (complete independence) to 1 (complete dependence). The indicative boundaries depend on $k - 1$ (Table 5.6).

Table 5.6: Indicative interpretation of Cramér's V

Strength of the association	$k - 1 = 1$	$k - 1 = 2$	$k - 1 = 3$
weak	0.10	0.07	0.06
moderate	0.30	0.21	0.17
strong	0.50	0.35	0.29

At $k - 1 = 2$ the value 0.24 corresponds to a **moderate** association.

> **Full reporting of the result**
>
> There is a statistically significant, moderate association between the frequency of the physiotherapy treatment and the degree of recovery ($\chi^2 = 56.96$; $df = 4$; $p < 0.001$; $V = 0.24$; $N = 500$).
> The analysis of the contributions shows that the association is due above all to the regimen "2 weeks monthly", under which the complete recoveries are more and the unsatisfactory results — fewer than expected.

5.5 Worked Example: a 2 × 2 Table and the φ Coefficient

For a 2 × 2 table the φ coefficient is used. It has the same meaning as V, but also carries a **sign**, which shows the direction of the association, and is interpreted as a correlation coefficient (Table 6.2).

$$\varphi = \frac{(a \cdot d) - (b \cdot c)}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}$$

where a, b, c, and d are the four observed frequencies arranged as in Table 5.7.

Table 5.7: Cell notation for a 2 × 2 table

	Outcome "yes"	Outcome "no"
Factor "yes"	a	c
Factor "no"	b	d

Setting

A maternity ward records data on 200 singleton pregnancies. For each mother, the smoking during the pregnancy and the weight of the newborn are known.

	Low weight (< 2,500 g)	Normal weight	Total
Smoker	a = 30	c = 50	80
Non-smoker	b = 15	d = 105	120
Total	**45**	**155**	**200**

Task. Is there an association between smoking during pregnancy and low birth weight, and how strong is it?

Step 1. Hypotheses. H_0: smoking and birth weight are independent. H_1: they are not independent. Significance level $\alpha = 0.05$.

Step 2. Description. Among the 80 smokers, 30 gave birth to a child with low weight ($37.5\ \%$), and among the 120 non-smokers — 15 ($12.5\ \%$).

Step 3. Expected frequencies.

$$E_{11} = \frac{80 \times 45}{200} = 18.0 \qquad E_{12} = \frac{80 \times 155}{200} = 62.0$$

$$E_{21} = \frac{120 \times 45}{200} = 27.0 \qquad E_{22} = \frac{120 \times 155}{200} = 93.0$$

All expected frequencies are ≥ 5 and the conditions are fulfilled.

Step 4. Test statistic.

$$\chi^2 = \frac{(30-18)^2}{18} + \frac{(50-62)^2}{62} + \frac{(15-27)^2}{27} + \frac{(105-93)^2}{93}$$

$$\chi^2 = 8.00 + 2.32 + 5.33 + 1.55 = 17.20$$

Step 5. Degrees of freedom and p-value.

$$df = (2-1) \times (2-1) = 1$$

df	$p = 0.10$	$p = 0.05$	$p = 0.01$
1	2.706	3.841	6.635

Since $17.20 > 6.635$, it follows that $p < 0.01$; the software gives $p < 0.001$.

Step 6. Strength and direction of the association.

$$\varphi = \frac{(30 \cdot 105) - (15 \cdot 50)}{\sqrt{(30+15)(50+105)(30+50)(15+105)}}$$

$$\varphi = \frac{3{,}150 - 750}{\sqrt{45 \cdot 155 \cdot 80 \cdot 120}} = \frac{2{,}400}{\sqrt{66{,}960{,}000}} = \frac{2{,}400}{8{,}182.9} = 0.29$$

Conclusion. There is a statistically significant, moderate association between smoking during pregnancy and low birth weight ($\chi^2 = 17.20$; $df = 1$; $p < 0.001$; $\varphi = 0.29$). The positive sign shows that smoking is associated with a higher frequency of low birth weight.

i The relationship between φ, V, and Z

For a 2 × 2 table $k - 1 = 1$ holds and then

$$V = \sqrt{\frac{\chi^2}{N}} = |\varphi|$$

A check for the example: $\sqrt{17.20/200} = \sqrt{0.086} = 0.29$.
Moreover $Z = \sqrt{\chi^2} = \sqrt{17.20} = 4.15$ — the same value that the Z-test for the two proportions 37.5 % and 12.5 % gives.
That is, Cramér's V is the generalisation of φ for tables of arbitrary size.

5.6 What the χ^2-Test Does Not Do

⚠ Four limitations

1. **Does not prove causality.** A significant result means only that the two variables are not independent. The causal conclusion requires an appropriate design, not a smaller p-value.
2. **Does not indicate a direction for tables larger than 2 × 2.** To understand where the association is, the per-cell contributions and the row percentages are examined.
3. **Does not use the ordering of the categories.** If one variable is ordinal (stage I–IV), the test treats it as nominal and loses information. For such cases, tests for trend exist.
4. **Is not applicable to paired observations.** When the same persons are measured twice, McNemar's test is used.

! The grouping affects the conclusion

The conclusion sometimes depends on the way the categories are grouped. Combining two neighbouring categories can change both the value of χ^2 and the conclusion. Therefore the grouping is determined by substantive considerations **before** the analysis, not after the result is seen.

5.7 Common Mistakes

Table 5.8: Common mistakes with the χ^2-test

#	Mistake	How to avoid it
1	Computing χ^2 on percentages	The test works only with counts
2	Forgetting the division by E	Each cell is divided by **its own** expected frequency
3	df computed from the sample size	df depends only on the size of the table
4	Applying the test with small expected frequencies	Check the 80 % rule and $E \geq 1$
5	Applying the test to paired observations	For repeated measurements use McNemar's test
6	Conclusion about causality	χ^2 shows an association, not a cause
7	Reporting χ^2 without the strength of the association	Always add V or φ
8	Conclusion about a single category in a 3 × 3 table	First look at the per-cell contributions

5.8 Self-Study and Homework Tasks

Task 5.1

A retrospective study analyses the complications after an abortion performed by two methods. Is there an association between the method of abortion and the type of complication?

Method of abortion	Bleeding	Infection	Infertility	Total
Classical	10	26	24	60
Aspiration	12	13	15	40
Total	**22**	**39**	**39**	**100**

Compute the expected frequencies, χ^2, df, and V, and formulate a conclusion.

Task 5.2 (homework)

Is there an association between the duration of work experience of IT specialists and the presence of neurological complaints?

Duration of experience	With complaints	Without complaints	Total
Up to 5 y.	4	46	50
6–10 y.	15	85	100
11–15 y.	27	73	100
Over 15 y.	28	22	50

Total	**74**	**226**	**300**

Besides the test, also indicate which category contributes the most to the result.

Task 5.3 (homework)

Is there an association between the duration of work experience of IT specialists and the presence of **visual** complaints?

Duration of experience	With visual complaints	Without complaints	Total
Up to 5 y.	4	27	31
6–10 y.	6	46	52
11–15 y.	13	85	98
Over 15 y.	21	84	105
Total	**44**	**242**	**286**

Compare the conclusion with that of task 5.2 and explain the difference.

Task 5.4

In a study of 250 patients with an acute coronary syndrome, it was recorded whether the reperfusion treatment started in the first 2 hours and whether the patient developed heart failure.

	Heart failure	Without failure	Total
Reperfusion ≤ 2 h.	18	132	150
Reperfusion > 2 h.	32	68	100
Total	**50**	**200**	**250**

Compute χ^2, df, and φ and formulate a conclusion, including the direction of the association.

Task 5.5

For a 4 × 3 table $\chi^2 = 18.0$ was obtained with $N = 400$.

a) How many degrees of freedom are there?
b) Is the result significant at $\alpha = 0.05$?
c) Compute Cramér's V and describe the strength of the association.

Task 5.6

A study with 60 participants gives a 2 × 2 table in which the smallest **expected** frequency is 3.2. Can the χ^2-test be applied? What is the appropriate choice?

Task 5.7 (for reasoning)

A study with 100,000 participants establishes an association between two variables with $\chi^2 = 12.0$, $df = 1$, and $V = 0.011$. The result is statistically significant at $\alpha = 0.05$. Is it clinically significant? Explain the difference between the two concepts in this particular case.

Task 5.8 (for reasoning)

A researcher checks sequentially 20 independent associations between categorical variables at the same level $\alpha = 0.05$ and reports only the two for which he obtained $p < 0.05$. Why is this way of presentation misleading?

5.9 Short Quiz

For each question choose **one** correct answer. The key is in Section D.5.

1. The χ^2-test of independence is applied when:

    a) both variables are quantitative;
    b) one is quantitative and the other is categorical;
    c) both variables are categorical;
    d) the variable is one, but it is measured twice.

2. The expected frequency of a cell is computed as:

    a) row total plus column total, divided by N;
    b) product of the row total and the column total, divided by N;
    c) the overall number of observations, divided by the number of cells;
    d) the observed frequency, multiplied by N.

3. For a 4 × 3 table the degrees of freedom are:

    a) 12;
    b) 7;
    c) 6;
    d) equal to $N - 1$.

4. Which statement about the χ^2 statistic is true?

    a) It can take negative values;
    b) It does not depend on the sample size;
    c) All other things being equal, it grows with the increase of the sample;
    d) It directly measures the strength of the association.

5. Cramér's V is reported because:

    a) it replaces the p-value;
    b) it measures the strength of the association independently of the sample size;
    c) it shows the direction of the association in a 3 × 3 table;
    d) it proves a causal dependence.

6. For a 2 × 2 table with $N = 200$, $\chi^2 = 8.0$ was obtained. The value of V is approximately:

    a) 0.04;
    b) 0.20;
    c) 0.40;
    d) it cannot be computed.

6 Correlation Analysis

Aim of the exercise

After this exercise you should be able to:

1. recognise when correlation analysis is applied and when it is inappropriate;
2. construct and read a scatter diagram;
3. compute Pearson's correlation coefficient r step by step;
4. describe verbally the strength and the direction of the association;
5. check the statistical significance of the correlation and formulate a complete conclusion;
6. explain why correlation is not evidence of causality.

What you should know beforehand

From Section 2: arithmetic mean, deviation from the mean, sum of the squares of the deviations. From Section 4: null and alternative hypothesis, degrees of freedom, p-value.

6.1 When Which Analysis

The choice of a method for studying an association is determined by the type of the two variables (Table 6.1).

Table 6.1: Choice of method according to the type of the two variables

First variable	Second variable	Method
categorical	categorical	χ^2-test, φ, Cramér's V (Section 5)
quantitative	categorical with 2 categories	t-test (Section 4)
quantitative	quantitative	**correlation analysis** — strength and direction (this chapter)
quantitative	quantitative	**regression analysis** — a model for prediction (Section 7)

Definition

Correlation analysis is a statistical method for studying the relationship between two **quantitative** variables. It assesses the extent to which the two variables vary together (change in a coordinated manner).

Correlation or regression

The two methods work with the same data, but answer different questions.

- **Correlation** answers the question *how strong and in what direction is the association*. The two variables are on an equal footing: the coefficient between X and Y is the same as the one between Y and X.
- **Regression** answers the question *how to predict one variable from the other*. Here there is already a dependent and an independent variable, and swapping them changes the result.

6.2 The First Step Is the Graph

Correlation analysis **begins** with a scatter diagram, not with the formula. Each observation is plotted as a point with coordinates (X_i, Y_i). The graph shows three things that the number by itself does not show: the **shape** of the association, the presence of **extreme values**, and the **homogeneity** of the group.

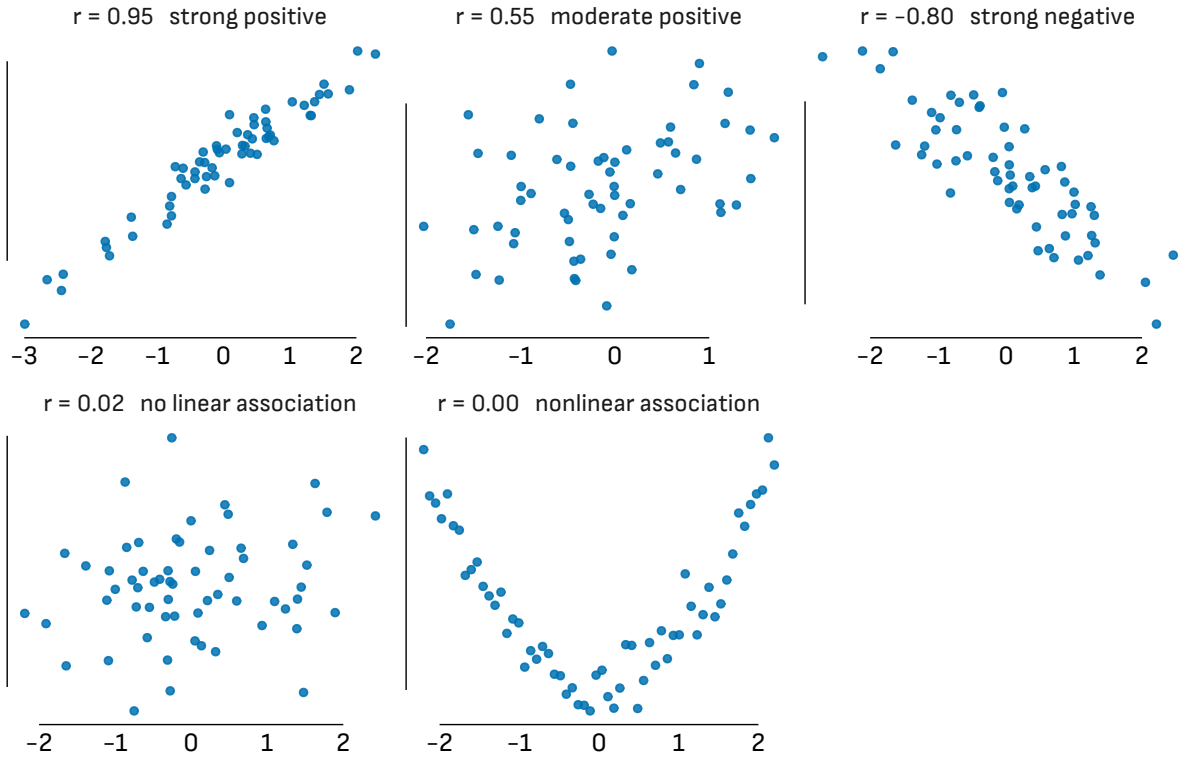

Figure 6.1: Scatter diagrams for different values of r. In the lower right panel the two variables have a clear, but nonlinear, association; Pearson's coefficient reports it as almost zero. This is why the graph is examined before the computation.

6.3 The Idea behind the Coefficient

Plot the arithmetic means $\bar{X}$ and $\bar{Y}$ as two lines on the diagram. They divide the field into four quadrants (Figure 6.2).

For each observation the two deviations are computed:

$$d_x = X_i - \bar{X}, \qquad d_y = Y_i - \bar{Y}$$

and their **cross product** $d_x \cdot d_y$.

- In the upper right quadrant both deviations are positive, so the product is **positive**.
- In the lower left quadrant both deviations are negative, so the product is again **positive**.
- In the other two quadrants the signs differ and the product is **negative**.

The sum $\sum(d_x \cdot d_y)$ is positive when the points cluster along the diagonal from lower left to upper right (positive association), and negative under the opposite arrangement.

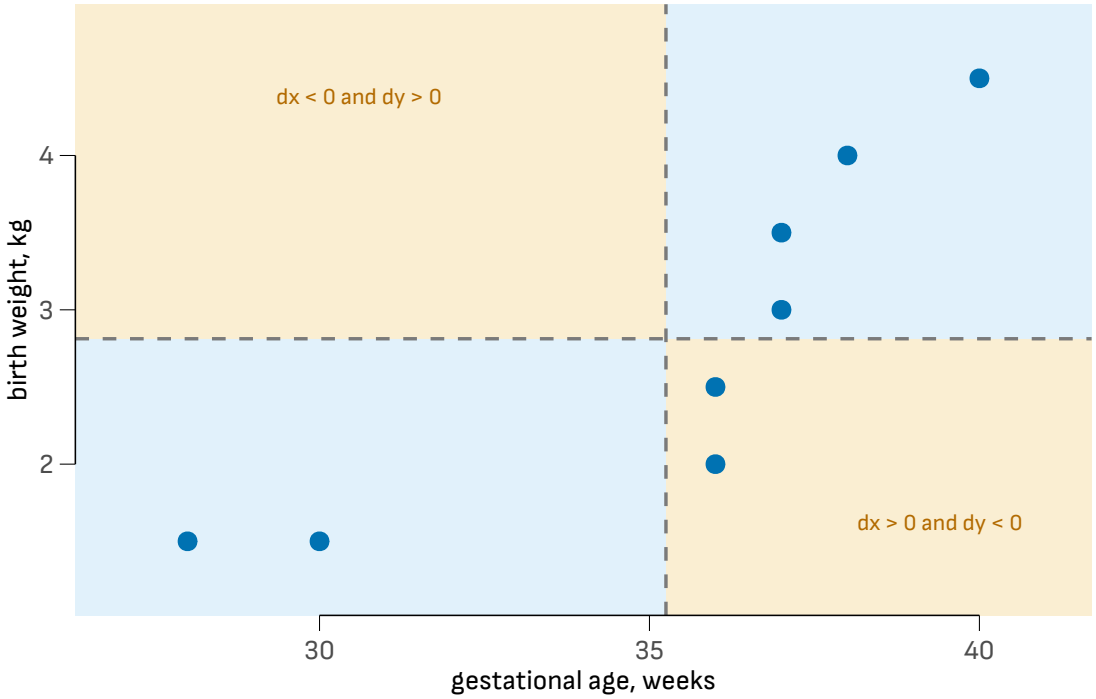

Figure 6.2: The four quadrants around the arithmetic means. The points in the blue quadrants give a positive contribution to the sum of the cross products, and the points in the orange ones — a negative one. In a positive association the blue quadrants prevail.

6.4 The Pearson Correlation Coefficient

If we used only the sum $\sum(d_x \cdot d_y)$, its value would depend on the units of measurement: the same data, expressed in grams instead of kilograms, would give a number a thousand times larger. That is why the sum is divided by an expression that removes the scale:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \cdot \sum(Y_i - \bar{Y})^2}} = \frac{\sum(d_x \cdot d_y)}{\sqrt{\sum d_x^2 \cdot \sum d_y^2}}$$

where:

- X_i and Y_i are the values of the two variables for the i-th observation;
- $\bar{X}$ and $\bar{Y}$ — the arithmetic means of the two variables;
- $d_x \cdot d_y$ — the cross product for each observation;
- $\sum(d_x \cdot d_y)$ — the sum of the cross products, that is, the numerator;
- $\sum d_x^2$ and $\sum d_y^2$ — the sums of the squares of the deviations.

In short: the numerator is the **covariance** of the two variables, and the denominator — the product of their **standard deviations**.

! Properties of the coefficient

1. r is a **dimensionless** number and does not depend on the units of measurement.
2. r takes values only in the interval from -1 to $+1$.
3. The sign shows the **direction**, and the absolute value — the **strength** of the association.
4. The r between X and Y equals the r between Y and X.
5. r measures only a **linear** association.

Direction of the association

Positive (direct) correlation: the two variables change in the same direction. Example: gestational age and birth weight.
Negative (inverse) correlation: an increase in one is associated with a decrease in the other. Example: the distance walked in a 6-minute walk test and the level of NT-proBNP.

Table 6.2: Interpretation of Pearson's coefficient

r	Strength and direction of the association
$+1$	perfect (functional) positive
from $+0.67$ to $+1$	strong positive
from $+0.34$ to $+0.66$	moderate positive
from 0 to $+0.33$	weak positive
0	**no linear association**
from 0 to -0.33	weak negative
from -0.34 to -0.66	moderate negative
from -0.67 to -1	strong negative
-1	perfect (functional) negative

! Two notes on $r = 0$ and $r = \pm 1$

Perfect (functional) dependence means that every definite change in one variable is associated with a strictly definite change in the other, that is, an exact formula $Y = f(X)$ can be derived. In medicine such associations practically do not occur.
$r = 0$ does **not** mean that there is no association at all between the two variables. It means that if an association exists, it is **not linear**. Exactly for this reason the scatter diagram is examined before computing r.

⚠ Conditions for applying Pearson's coefficient

1. Both variables are quantitative, measured on an interval or ratio scale.
2. The observations are **independent** — each pair of values belongs to a different person.
3. The association is approximately **linear**, which is checked graphically.
4. There are no single extreme values that by themselves determine the result.
5. For the significance test it is desirable that the two variables are approximately normally distributed.

If condition 3 or 5 is not fulfilled, a rank coefficient is used, for example Spearman's.

6.5 Worked Example

Setting

The gestational age at birth X (weeks) and the birth weight Y (kg) of 8 newborns were recorded.

Newborn №	Gestational age X, weeks	Weight Y, kg
1	28	1.5
2	30	1.5
3	36	2.0
4	36	2.5
5	37	3.0
6	37	3.5
7	38	4.0
8	40	4.5

Task. To assess the strength and the direction of the association between gestational age and birth weight.

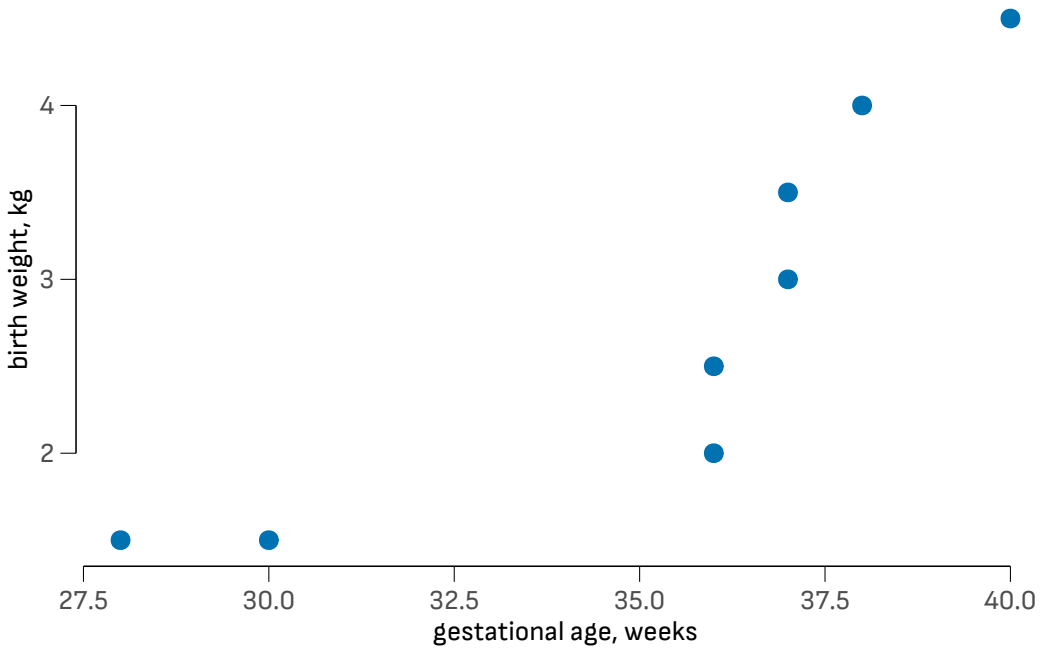

Figure 6.3: Scatter diagram for the eight newborns. The points cluster around an ascending line and no single observation stands out that would by itself determine the picture. The conditions for Pearson's coefficient are fulfilled.

6.5.1 Step 1. Arithmetic Means

$$\bar{X} = \frac{\sum X}{n} = \frac{282}{8} = 35.25 \text{ weeks}$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{22.5}{8} = 2.8125 \text{ kg}$$

> **Do not round the means**
>
> The two values are exact, not approximate. If $\bar{X}$ is rounded to 35 and $\bar{Y}$ to 3, the coefficient becomes 0.85 instead of 0.87 and no longer matches the value we will obtain from the regression model in Section 7. Work with the exact values and round only the final result.

6.5.2 Step 2. Deviations, Cross Products, and Squares

$$d_x = X - \bar{X}, \qquad d_y = Y - \bar{Y}$$

Table 6.3: Working table for the correlation coefficient

№	X	Y	d_x	d_y	$d_x \cdot d_y$	d_x^2	d_y^2
1	28	1.5	−7.25	−1.3125	9.5156	52.5625	1.7227

№	X	Y	d_x	d_y	$d_x \cdot d_y$	d_x^2	d_y^2
2	30	1.5	−5.25	−1.3125	6.8906	27.5625	1.7227
3	36	2.0	0.75	−0.8125	−0.6094	0.5625	0.6602
4	36	2.5	0.75	−0.3125	−0.2344	0.5625	0.0977
5	37	3.0	1.75	0.1875	0.3281	3.0625	0.0352
6	37	3.5	1.75	0.6875	1.2031	3.0625	0.4727
7	38	4.0	2.75	1.1875	3.2656	7.5625	1.4102
8	40	4.5	4.75	1.6875	8.0156	22.5625	2.8477
Total	**282**	**22.5**	**0.00**	**0.00**	**28.3750**	**117.5000**	**8.9688**

Two checks

- The sum of the d_x and the sum of the d_y are always zero. If you obtain some other number, one of the means has been computed wrongly.
- The sums $\sum d_x^2$ and $\sum d_y^2$ are always positive. A negative value is a sure sign of an error.

6.5.3 Step 3. Computing r

$$r = \frac{\sum(d_x \cdot d_y)}{\sqrt{\sum d_x^2 \cdot \sum d_y^2}} = \frac{28.3750}{\sqrt{117.5000 \cdot 8.9688}}$$

$$r = \frac{28.3750}{\sqrt{1{,}053.83}} = \frac{28.3750}{32.4627} = 0.87$$

6.5.4 Step 4. Checking the Statistical Significance

The computed coefficient describes **only these 8 newborns**. To check whether the association also exists in the general population, a hypothesis test is performed.

H_0: The population correlation coefficient is zero ($\rho = 0$) — there is no linear association between gestational age and birth weight.

H_1: The population correlation coefficient is different from zero ($\rho \neq 0$).

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}, \qquad df = n - 2$$

$$t = \frac{0.87 \cdot \sqrt{8-2}}{\sqrt{1-0.87^2}} = \frac{0.87 \cdot 2.4495}{\sqrt{1-0.7640}} = \frac{2.1411}{0.4858} = 4.41$$

$$df = 8 - 2 = 6$$

Table 6.4: Critical values of t at $df = 6$

df	$p = 0.10$	$p = 0.05$	$p = 0.01$
6	1.943	2.447	3.707

The computed $|t| = 4.41$ is larger than 3.707, therefore $p < 0.01$; the software gives $p = 0.005$.

> **i** Why a small sample requires a high coefficient
>
> At $n = 8$ and therefore $df = 6$, significance is reached only at $|r| \approx 0.71$. At $n = 30$ the same threshold is about 0.36, and at $n = 100$ — about 0.20. That is, **one and the same** value of r can be significant in a large sample and insignificant in a small one. For this reason both r, n, and p are reported.

6.5.5 Step 5. Conclusion

$r = 0.87$. There is a **strong positive** association between gestational age and birth weight in these 8 newborns: with the increase of gestational age, birth weight also increases. The association is statistically significant ($t = 4.41$; $df = 6$; $p < 0.01$).

6.6 What the Coefficient Does Not Show

A single number cannot replace the graph. The four panels of Figure 6.4 have almost identical correlation coefficients, but the data behind them are completely different.

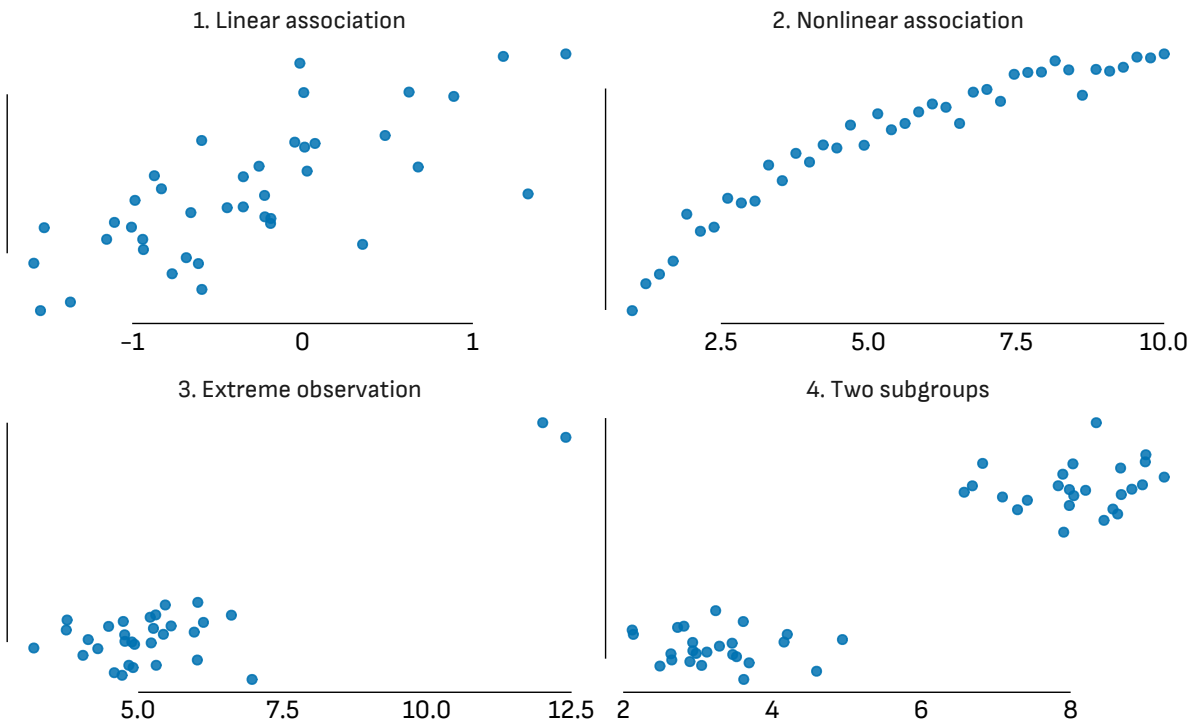

Figure 6.4: Four data sets with almost identical correlation coefficients around 0.8. Only in the first panel does the number describe the data correctly: in the remaining ones it is the result, respectively, of nonlinearity, of a single extreme observation, and of two separated subgroups.

⚠ Five traps

1. **Nonlinearity.** The coefficient measures only the linear part of the association.
2. **Extreme observations.** A single distant observation can create or destroy a high coefficient.
3. **Mixed subgroups.** If the sample contains two different populations, the coefficient describes the distance between them, not the association within them.
4. **Restricted range.** If we measure gestational age only between 39 and 41 weeks, the coefficient will be close to zero, although in the whole range the association is strong.
5. **Ecological fallacy.** A correlation established between **mean values for countries or regions** does not automatically apply to the individual persons.

6.6.1 Extrapolation

⚠ Definition

Extrapolation is the extension of an established association to values **outside** the range of the data on which it was built.

In a sample consisting mostly of children, a strong association between weight and height is observed. This association, however, should not be carried over to adults: in them the growth of height

is finished, while the weight may continue to increase. The association holds only within the range of the available data.

6.7 Correlation and Causality

> **!** The most important sentence in this chapter
>
> Proving a statistical dependence does **not** mean proving a causal relationship.

The presence of a correlation between X and Y admits at least five explanations:

1. X causes Y;
2. Y causes X (reverse causation);
3. a third variable Z causes both (confounding factor, see Section 1.3);
4. the association is due to the way the sample was selected;
5. the association is due to chance.

Statistics by itself cannot choose among these explanations. The probability that the association is causal is greater when:

1. the association is **strong**;
2. the association is **biologically plausible** — a known mechanism exists;
3. the association is confirmed in **other, independent studies**;
4. the **temporal sequence** is correct: the cause precedes the effect;
5. a **dose–response relationship** is observed — greater exposure leads to a greater effect.

> **i** The absence of a mechanism does not refute causality
>
> If the mechanism is not known, this only means that knowledge is incomplete. The association between smoking and lung carcinoma was established statistically decades before the mechanism was clarified. Conversely, the existence of a plausible mechanism by itself does not prove causality.

6.8 Common Mistakes

Table 6.5: Common mistakes in correlation analysis

#	Mistake	How to avoid it
1	Correlation between categorical characteristics	Pearson requires two quantitative variables
2	Conclusion "no association" at $r = 0$	It means absence only of a **linear** association
3	Computing r without a look at the graph	The scatter diagram is built first
4	Claim of causality	Use "is associated with", not "causes"

#	Mistake	How to avoid it
5	Skipping the significance test	r describes the sample; a conclusion about the population requires a test
6	Crude rounding of the means	Leads to a different r and to a contradiction with R^2
7	Reporting r without n and p	The same r means different things at different sample sizes
8	Extrapolation beyond the range of the data	The association holds only in the observed range

6.9 Self-Study and Homework Tasks

Task 6.1

Measure the strength and direction of the association between age at diagnosis and survival in patients with lung carcinoma.

Patient №	1	2	3	4	5	6	7	8
Age X, years	71	71	73	74	75	76	76	78
Survival Y, months	135	96	66	44	16	22	11	5

Check also the statistical significance of the obtained coefficient.

Task 6.2 (homework)

The table records the distance walked in a 6-minute walk test (6MWT) and the level of NT-proBNP in 8 patients with incipient heart failure. Measure the strength and direction of the association.

Patient №	1	2	3	4	5	6	7	8
Distance X, km	0.7	0.7	0.6	0.5	0.4	0.4	0.3	0.3
NT-proBNP Y, pg/µl	110	120	140	160	180	200	240	300

Task 6.3 (homework)

Construct a scatter diagram for the data from task 6.2 and indicate:

a) what is the shape of the association;
b) is there an observation that deviates from the general trend;
c) is Pearson's coefficient appropriate for these data.

Task 6.4

Is it possible that in a given sample all individual deviations d_x are positive? And all equal to zero? What would the correlation coefficient be in the second case and why?

Task 6.5

A study reports $r = -0.45$ between the duration of sleep and the level of cortisol in 60 participants.

a) Describe with words the strength and direction of the association.
b) Check the significance at $\alpha = 0.05$.
c) How would the conclusion change if the same coefficient was obtained with 10 participants?

Task 6.6 (for reasoning)

A study establishes $r = 0.62$ between the number of ice creams eaten per day and the number of drownings during the summer months. Does this mean that ice cream causes drownings? Which of the five explanations of correlation is the most likely here?

Task 6.7 (for reasoning)

A researcher measures the association between height and weight only among professional basketball players and obtains $r = 0.12$. He concludes that in adults height and weight are not associated. Which of the five traps has been committed?

Task 6.8

Two studies report the same value $r = 0.30$. The first one has $n = 20$, the second one — $n = 200$. Compute the test statistic for both and explain why the conclusions differ.

6.10 Short Quiz

For each question choose **one** correct answer. The key is in Section D.6.

1. Pearson's coefficient is applied when:

    a) both variables are categorical;
    b) both variables are quantitative;
    c) one is categorical and the other is quantitative;
    d) the variable is one, but it is measured twice.

2. The value $r = 0$ means that:

    a) there is no association at all between the two variables;
    b) there is no linear association between the two variables;
    c) the two variables are independent;
    d) the sample is too small.

3. The coefficient $r = -0.82$ describes an association that is:

    a) weak positive;
    b) moderate negative;
    c) strong negative;
    d) perfect negative.

4. Which **does not** change the value of the correlation coefficient?

    a) Switching from kilograms to grams;
    b) Adding one extreme observation;
    c) Restricting the range of one variable;
    d) Pooling two different subgroups.

5. At $n = 8$ and $r = 0.87$ the degrees of freedom for the significance test are:

    a) 8;
    b) 7;
    c) 6;
    d) 2.

6. A statistically significant correlation between two characteristics means that:

    a) one characteristic causes the other;
    b) the association most likely is not due to chance alone;
    c) the association is clinically important;
    d) the association is necessarily linear and strong.

7 Regression Analysis

Aim of the exercise

After this exercise you should be able to:

1. distinguish the dependent from the independent variable and justify the choice;
2. explain what the line of best fit minimises;
3. compute the regression coefficient b and the intercept a;
4. write down the regression equation and use it for prediction;
5. interpret the coefficient of determination R^2 and derive from it the correlation coefficient together with its sign;
6. indicate the limits within which the model is valid.

What you should know beforehand

From Section 6: scatter diagram, deviations from the mean, strength and direction of the association, Pearson's coefficient.

7.1 Essence of the Regression Model

Correlation analysis answers the question *how strong is the association*. Regression analysis goes one step further and answers the question *how to predict one variable from the other*. For this purpose the association is reduced to an equation.

Definitions

Dependent (outcome) variable Y — the one whose values are predicted.
Independent (explanatory) variable X — the one used for the prediction; it is also called a predictor or a factor.

The equation of the **simple (univariate) linear regression** is:

$$Y = b \cdot X + a + \varepsilon$$

where:

- b is the **regression coefficient** — the mean change in Y when X increases by **one unit**;
- a is the **intercept** (constant, the Y-intercept) — the value of Y at $X = 0$;
- ε is the **residual component** — the part of Y that the model does not explain.

The predicted value is denoted by $\hat{Y}$ (or Y_t) and is obtained without the residual component:

$$\hat{Y} = b \cdot X + a$$

How to recognise which variable is dependent

Ask yourself what is known in advance and what we want to predict. The gestational age is known before the birth, and the weight is measured after it, therefore the weight is the dependent variable. The opposite arrangement is mathematically possible, but has no practical meaning.
Unlike correlation, in regression swapping X and Y **changes** the result: a different line is obtained.

Types of regression models

According to the form of the dependence: linear, in which the regression is a straight line, and nonlinear, in which it is a curve. Only the linear ones are considered in this course.
According to the number of factors: univariate (simple) with one predictor, and multivariate (multiple) with two or more predictors, each with its own regression coefficient.

7.2 The Line of Best Fit

Through a cloud of points infinitely many lines can be drawn. Regression analysis chooses **one** of them by a clear criterion.

For each observation the vertical distance between the actual value Y_i and the value on the line $\hat{Y}_i$ is called a **residual**:

$$e_i = Y_i - \hat{Y}_i$$

The criterion of least squares

The line of best fit is the one for which the **sum of the squares of the residuals** $\sum e_i^2$ is as small as possible.
Squaring serves the same purpose as in the standard deviation: without it the positive and the negative residuals would cancel each other out and every line passing through the centre of the data would look equally good.

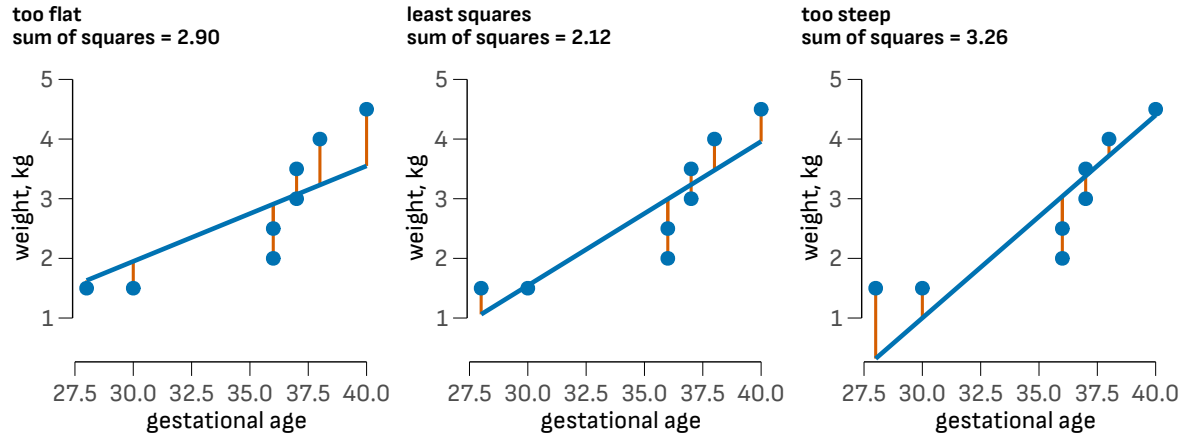

Figure 7.1: Three possible lines through the same data. For each one the sum of the squares of the residuals is shown. The method of least squares chooses the middle line, for which this sum is the smallest.

7.3 Worked Example

Setting

The gestational age at birth X (weeks) and the birth weight Y (kg) of 8 newborns were recorded — the same data with which we worked in Section 6.
Task. To model the dependence of the birth weight on the gestational age and to assess the quality of the model.

7.3.1 Step 1. Arithmetic Means

$$\bar{X} = \frac{282}{8} = 35.25 \text{ weeks}, \qquad \bar{Y} = \frac{22.5}{8} = 2.8125 \text{ kg}$$

Accuracy of the intermediate computations

In regression, **rounding of the intermediate results is not recommended**. Rounding $\bar{X}$ to 35 or b to one decimal changes noticeably the intercept and the predictions. Round only the final answer.

7.3.2 Step 2. Auxiliary Table

Table 7.1: Auxiliary table for the regression analysis

№	X	Y	$X \cdot Y$	X^2
1	28	1.5	42.0	784

№	X	Y	$X \cdot Y$	X^2
2	30	1.5	45.0	900
3	36	2.0	72.0	1,296
4	36	2.5	90.0	1,296
5	37	3.0	111.0	1,369
6	37	3.5	129.5	1,369
7	38	4.0	152.0	1,444
8	40	4.5	180.0	1,600
$n = 8$	**282**	**22.5**	**821.5**	**10,058**

7.3.3 Step 3. Regression Coefficient b

$$b = \frac{n \cdot \sum(X \cdot Y) - (\sum X)\,(\sum Y)}{n \cdot \sum X^2 - (\sum X)^2}$$

Numerator:

$$8 \cdot 821.5 - 282 \cdot 22.5 = 6{,}572 - 6{,}345 = 227$$

Denominator:

$$8 \cdot 10{,}058 - 282^2 = 80{,}464 - 79{,}524 = 940$$

$$b = \frac{227}{940} = 0.2415 \approx 0.24 \text{ kg per week}$$

The positive value of b means an **ascending slope** of the line, that is, a positive (direct) association.

> ⚠ Careful with the denominator
>
> $\sum X^2$ and $(\sum X)^2$ are **different** quantities. The first one is the sum of the squares (10,058), the second one — the square of the sum ($282^2 = 79{,}524$). Swapping them is the most common mistake in this step.

7.3.4 Step 4. Intercept a

$$a = \bar{Y} - (b \cdot \bar{X})$$

$$a = 2.8125 - (0.2415 \cdot 35.25) = 2.8125 - 8.5125 = -5.70$$

A negative intercept — is it a problem?

No. Formally $a = -5.70$ means "weight at a gestational age of zero weeks", which is meaningless. The intercept has no independent clinical interpretation when the value $X = 0$ is outside the range of the data. It determines where the line starts and is necessary for the predictions in the observed range.

A property worth knowing

The line of best fit always passes through the point $(\bar{X},\ \bar{Y})$. Check for the example: $0.2415 \cdot 35.25 - 5.70 = 2.8125$ — exactly the mean value of Y.

7.3.5 Step 5. Regression Equation

$$\hat{Y} = b \cdot X + a$$

$$\boxed{\hat{Y} = 0.2415 \cdot X - 5.70}$$

Two points are sufficient for drawing the line:

Table 7.2: Two points for drawing the line

Gestational age X	Predicted weight $\hat{Y} = 0.2415X - 5.70$
28 weeks	$0.2415 \cdot 28 - 5.70 = 1.06$ kg
40 weeks	$0.2415 \cdot 40 - 5.70 = 3.96$ kg

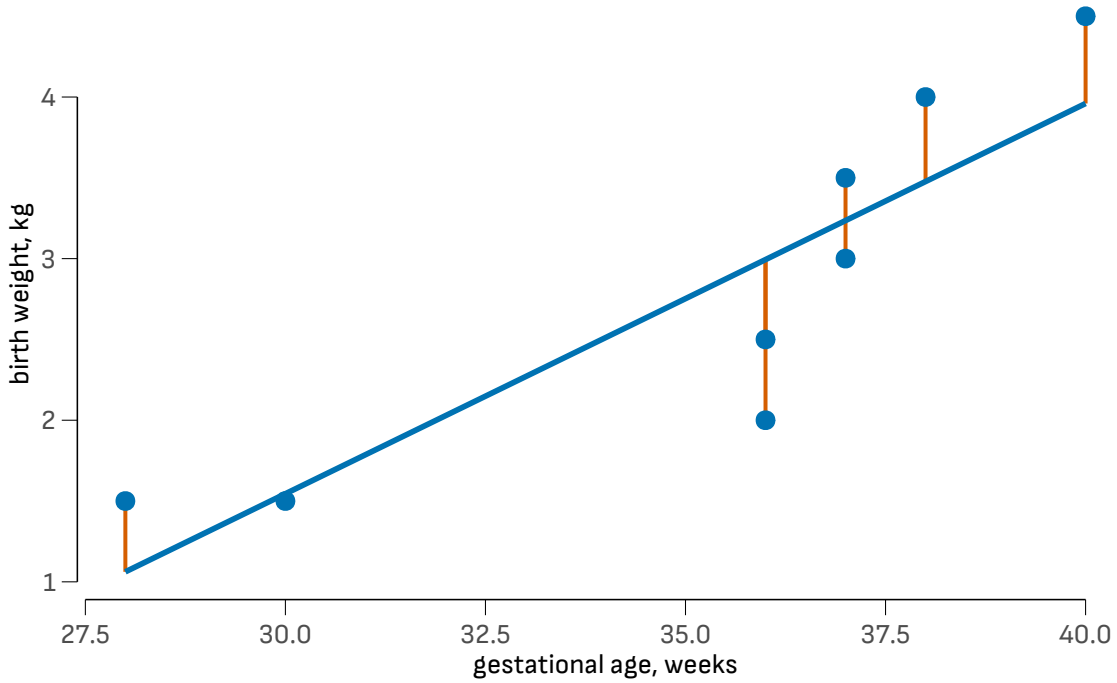

Figure 7.2: The line of best fit. The red segments are the residuals — the differences between the actual and the predicted weight. The sum of their squares is smaller than for any other line.

7.3.6 Step 6. Residuals

The residuals show how well the model copes with each individual observation.

Table 7.3: Predicted values and residuals

№	X	Y	$\hat{Y}$	Residual $e = Y - \hat{Y}$
1	28	1.5	1.06	0.44
2	30	1.5	1.54	−0.04
3	36	2.0	2.99	−0.99
4	36	2.5	2.99	−0.49
5	37	3.0	3.24	−0.24
6	37	3.5	3.24	0.26
7	38	4.0	3.48	0.52
8	40	4.5	3.96	0.54
Total		**22.5**	**22.5**	**0.00**

Check

In a correctly computed model the sum of the residuals is zero, and the sum of the predicted values equals the sum of the observed ones. This is a convenient control point.

7.3.7 Step 7. Coefficient of Determination R^2

Every regression model is assessed by its ability to predict correctly. This assessment is performed through the **coefficient of determination** R^2.

The idea is a decomposition of the total variation of the dependent variable:

$$\underbrace{\sum (Y_i - \bar{Y})^2}_{\text{total variation}} = \underbrace{\sum (\hat{Y}_i - \bar{Y})^2}_{\text{explained by the model}} + \underbrace{\sum (Y_i - \hat{Y}_i)^2}_{\text{unexplained (residual)}}$$

$$R^2 = \frac{\text{explained variation}}{\text{total variation}}$$

For the example the total variation is $\sum d_y^2 = 8.9688$ (computed in Section 6), and the unexplained one is $\sum e_i^2 = 2.1165$. Therefore:

$$R^2 = \frac{8.9688 - 2.1165}{8.9688} = \frac{6.8523}{8.9688} = 0.764$$

! Properties of R^2

1. It takes values between 0 and 1 (or between 0 % and 100 %).
2. It shows **what part** of the variation of the dependent variable is explained by the action of the factor.
3. The closer it is to 1, the more accurately the model predicts.
4. In simple linear regression $R^2 = r^2$ holds, and therefore $|r| = \sqrt{R^2}$.

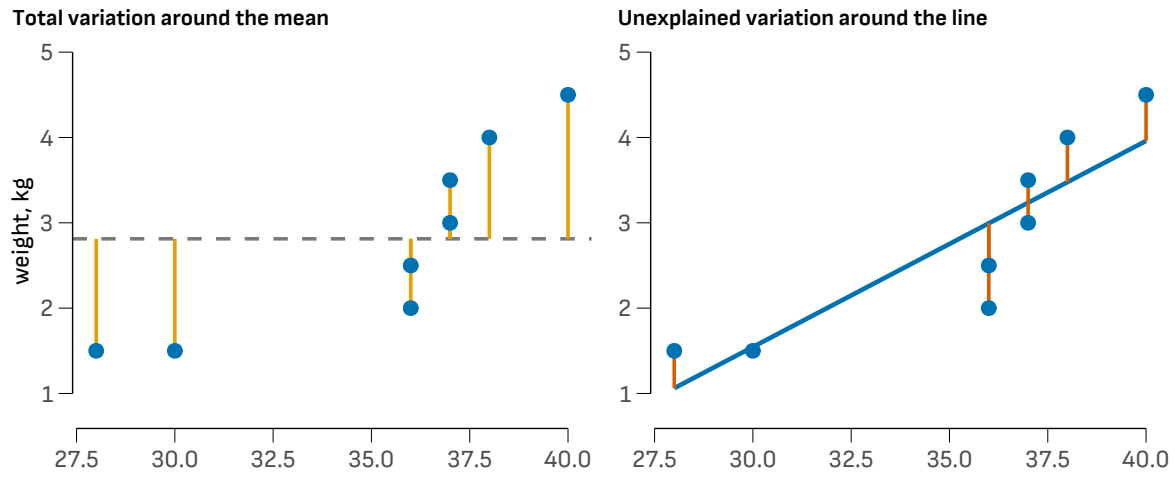

Figure 7.3: Decomposition of the variation. On the left the deviations of the observations from the overall mean (total variation), on the right the deviations from the regression line (unexplained variation). The shorter the segments on the right compared with those on the left, the larger R2 is.

7.3.8 Step 8. Conclusion

- Every single week of increase of the gestational age is associated with a **0.24 kg** increase of the expected birth weight.
- **76.4 %** of the variability of the birth weight can be explained by the gestational age. The remaining 23.6 % are due to other factors and to random variation.
- The correlation coefficient is $r = +\sqrt{0.764} = +0.87$ — a strong positive association, which coincides with the result from Section 6.

! The sign of r comes from b

Taking the square root of R^2 gives only the **absolute value** $|r|$. The sign is determined by the sign of the regression coefficient b, that is, by the slope of the line:

- $b > 0$ — the line goes up — r is **positive**;
- $b < 0$ — the line goes down — r is **negative**.

Therefore, when asked "what is the value of r", first look at the slope.

7.4 Prediction

The model is used by substituting the known value of X into the equation.

Example

A woman in labour arrives in the 34th gestational week. What weight is expected?

$$\hat{Y} = 0.2415 \cdot 34 - 5.70 = 8.211 - 5.70 = 2.51 \text{ kg}$$

Three limitations of the prediction

1. **Range of the data.** The model is built on observations from the 28th to the 40th gestational week. A prediction for the 20th week is extrapolation: the formula will give a number ($0.2415 \cdot 20 - 5.70 = -0.87$ kg), but it is meaningless.
2. **Individual versus group.** $\hat{Y}$ is the **mean expected** value for persons with a given X, not the exact weight of the particular newborn. The individual values scatter around the line.
3. **Causality.** Regression describes an association. The claim that a change in X will lead to a change in Y requires an appropriate design, not a higher R^2.

Regression models underlie many clinical instruments. The scales for assessing the cardiovascular risk and the risk of stroke in atrial fibrillation are derived precisely from regression models, in which several factors are combined into a single prediction.

7.5 General Rule of Interpretation

1. The **regression coefficient** b shows how the outcome changes when the factor increases by **one unit** — with its name: week, milligram, minute.
2. A **positive** b means a positive association, a **negative** one — an inverse association.
3. R^2 shows what share of the variation of the outcome variable is explained by the factor.
4. $\sqrt{R^2} = |r|$.
5. The sign before r is determined by the sign of b.

⚠ Mandatory exam question

What conclusion can be drawn on the basis of the regression model presented on the graph? How is the coefficient of determination interpreted in this case? What is the value of the correlation coefficient?
Expected answer for the example of this chapter:

- Every 1 week of increase of the gestational age is associated with a 0.24 kg increase of the expected birth weight.
- 76.4 % of the variability of the birth weight can be explained by the gestational age.
- $r = \sqrt{0.764} = +0.87$; the sign is positive because the slope of the line is positive.

7.6 More about Regression

ℹ Multiple regression

The presented model contains only one explanatory variable. In reality, almost always several predictors are used to explain an outcome — this is **multiple regression**:

$$\hat{Y} = a + b_1X_1 + b_2X_2 + \cdots + b_kX_k$$

Each coefficient b_j is interpreted "all other things being equal", that is, with unchanged values of the remaining predictors. Exactly this allows multiple regression to account for confounding factors — a task which in Section 1 we solved by standardization.
In these models the phenomenon of **collinearity** occurs: two or more predictors are in strong correlation with each other, whereby the estimates of their coefficients become unstable and difficult to interpret.

ℹ When the dependent variable is categorical

If the outcome is dichotomous (recovered / not recovered, alive / deceased), the straight line is not appropriate: its predictions can go outside the interval from 0 to 1. In such cases **logistic regression** is used, whose coefficients are presented as odds ratios. It is not part of the present course, but it is the most common regression model in the medical literature.

7.7 Common Mistakes

Table 7.4: Common mistakes in regression analysis

#	Mistake	How to avoid it
1	Swapped dependent and independent variable	Y is what we predict
2	$\sum X^2$ instead of $(\sum X)^2$	One is a sum of squares, the other — a square of a sum
3	Rounding the means or b before the computations	Leads to different a and predictions
4	Clinical interpretation of the intercept a	a is a mathematical starting point
5	Prediction outside the range of the data	Extrapolation gives numbers without meaning
6	Reporting r without a sign	The sign comes from the slope b
7	Interpreting $\hat{Y}$ as an exact individual value	$\hat{Y}$ is the mean expected value
8	Conclusion about causality from a high R^2	R^2 measures fit, not cause

7.8 Self-Study and Homework Tasks

Task 7.1

Compose a model of the dependence between the age at diagnosis and survival in patients with lung carcinoma (the data from task 6.1).

Patient №	1	2	3	4	5	6	7	8
Age X, years	71	71	73	74	75	76	76	78
Survival Y, months	135	96	66	44	16	22	11	5

What survival is expected for a patient diagnosed at the age of 72?

Task 7.2 (homework)

Compose a model of the dependence between the distance walked in a 6-minute walk test and the level of NT-proBNP (the data from task 6.2).

Patient №	1	2	3	4	5	6	7	8
Distance X, km	0.7	0.7	0.6	0.5	0.4	0.4	0.3	0.3
NT-proBNP Y, pg/µl	110	120	140	160	180	200	240	300

Interpret the regression coefficient with its exact unit of measurement.

Task 7.3 (homework)

A study measures the degree of anxiety among 10 medical students by means of a questionnaire with a scale from 0 to 100 and records the time in minutes spent reading statistics the previous day.

Reading time X, min	45	90	20	120	100	62	70	69	85	15
Anxiety Y, points	21	84	14	95	84	32	59	43	71	4

a) Construct the regression model.
b) What anxiety is expected for a student who reads 30 minutes a day?
c) At $R^2 = 0.93$ what is the correlation coefficient and what is its sign?
d) Can it be claimed that reading statistics causes anxiety?

Task 7.4

A regression model for the association between the daily salt intake (g) and the systolic pressure (mmHg) gives $b = 1.8$ and $R^2 = 0.25$.

a) How is b interpreted?
b) What is the correlation coefficient and what sign does it have?
c) Is this model appropriate for prediction at individual level? Explain.

Task 7.5

Two models predict the same dependent variable. The first one has $R^2 = 0.81$, the second one — $R^2 = 0.36$.

a) Which model predicts better?
b) What are the correlation coefficients in the two cases?
c) What else is needed to determine their sign?

Task 7.6

A regression model for the association between age and the maximal heart rate gives $\hat{Y} = 208 - 0.7 \cdot X$.

a) Interpret the slope.
b) What maximal heart rate is expected for a 60-year-old patient?
c) Does the intercept make sense in this case?

Task 7.7 (for reasoning)

A researcher builds a regression model on data for patients aged 40–60 years and uses it to predict a value for a 15-year-old child. What mistake does he commit and what is it called?

Task 7.8 (for reasoning)

A model with $R^2 = 0.98$ is built on 5 observations. Another model with the same dependent variable has $R^2 = 0.45$, but is built on 5,000 observations. Which of the two models would you prefer and why?

7.9 Short Quiz

For each question choose **one** correct answer. The key is in Section D.7.

1. The line of best fit minimises:

    a) the sum of the residuals;
    b) the sum of the squares of the residuals;
    c) the largest residual;
    d) the number of the residuals.

2. The regression coefficient b shows:

    a) the value of Y at $X = 0$;
    b) the share of the explained variation;
    c) the mean change in Y when X changes by one unit;
    d) the strength of the association in the interval from -1 to $+1$.

3. At $R^2 = 0.49$ and a negative slope the correlation coefficient is:

    a) $+0.70$;
    b) -0.70;
    c) -0.49;
    d) -0.24.

4. The intercept a:

    a) always has a clinical interpretation;
    b) equals the mean value of Y;
    c) shows the value of $\hat{Y}$ at $X = 0$;
    d) is always positive.

5. The use of the model outside the range of the observed values is called:

    a) interpolation;
    b) extrapolation;
    c) standardization;
    d) determination.

6. Which statement about R^2 is true?

    a) R^2 can be negative;
    b) R^2 proves a causal association;
    c) R^2 shows the share of the explained variation of the dependent variable;
    d) R^2 equals the regression coefficient.

8 Colloquium

Aim of this chapter

This chapter introduces no new material. It summarizes the seven exercises into one working algorithm and offers four sample variants with complete solutions, a preparation test of 45 multiple-choice questions and 9 open questions. Some of the questions require the concepts from Section A: types of graph, sampling designs, robustness, the binomial distribution and extrapolation.

8.1 Format and Assessment

The colloquium contains **two computational tasks**, each covering one of the studied methods. Each task is solved completely: from the source data to the verbal conclusion.

What is assessed

1. **Recognition of the method** — correctly determined type of the task according to the type of the variables and the question in the setting.
2. **Formulation of the hypotheses**, when the task is a hypothesis test.
3. **Written formula** before substituting into it.
4. **Intermediate steps** — the tables and the computations, not only the final number.
5. **Unit of measurement** of every obtained quantity.
6. **Verbal conclusion** that answers the question from the setting with the names of the studied variables.

An answer consisting only of a number does not carry the full number of points even when the number is correct.

Practical advice

- Bring a calculator. The formulas are provided, but the arithmetic is yours.
- Write the steps as they are shown in the solved examples in the handbook.
- If you make a mistake in an intermediate step but continue correctly, the following steps are assessed.
- Leave time for the last sentence — the conclusion carries points.

8.2 How the Type of Task Is Recognised

The first and most important step is to determine what the question is and what the variables are. Table 8.1 summarizes the features by which each of the studied methods is recognised.

Table 8.1: Recognition of the type of task by the setting

Key words in the setting	Type of data	Method	Chapter
"standardized rates", "take as a standard"	two groups with a subgroup structure	direct standardization	Section 1
"structure", "what percentage of"	one whole by categories	proportions	Section 1
"confidence interval", "population mean"	one sample, quantitative characteristic	$\bar{X} \pm t \cdot SE$	Section 3
"confidence interval for the proportion"	one sample, categorical characteristic	$\hat{p} \pm Z \cdot SE_{\hat{p}}$	Section 3
"is there a difference", two groups, mean values	quantitative characteristic, two independent groups	t-test	Section 4
"is there a difference", two groups, percentages or counts	categorical characteristic, two independent groups	Z-test for two proportions	Section 4
"is there an association / dependence", table with frequencies	two categorical characteristics	χ^2-test, φ or V	Section 5
"the strength and direction of the dependence"	two quantitative characteristics	Pearson's coefficient	Section 6
"compose a model", "to model", "to predict"	two quantitative characteristics	linear regression	Section 7

⚠ The two most common mistakes in recognition

1. **"Is there a difference" versus "is there an association".** The first compares two groups by one characteristic (t- or Z-test). The second checks an association between two characteristics (χ^2-test or correlation).
2. **Correlation versus regression.** "To measure the strength and direction" means correlation. "To compose a model" or "to predict" means regression. The two tasks can be built on the same data.

8.3 General Course of the Solution

Regardless of the method, the solution follows one and the same skeleton:

1. **Determining the type of the task** and the type of the variables.
2. **Hypotheses** (H_0 and H_1) and **significance level** α — in all hypothesis tests.
3. **Descriptive computations** — means, proportions, sums, auxiliary tables.
4. **Formula** — written before substituting into it.
5. **Test statistic** or **interval**.
6. **Degrees of freedom** and **p-value** from the table.

7. **Conclusion in words**, answering the question from the setting.
8. **Strength of the association**, when the method requires it (V, φ, r, R^2).

8.4 Four Sample Variants

8.4.1 Variant 1

Task 1

Compute the standardized case-fatality measures in two regional hospitals. Take the structure by departments in **hospital B** as the standard.

Department	Passed through (A)	Deceased (A)	Passed through (B)	Deceased (B)
Infectious diseases	500	10	1,500	35
Surgery	1,500	50	2,000	70
Oncology	1,000	85	1,500	120
Total	**3,000**	**145**	**5,000**	**225**

Task 2

A retrospective study analyses the observed complications after an abortion performed by two different methods. Is there an association between the method of abortion and the type of complication?

Method of abortion	Bleeding	Infection	Infertility	Total
Classical	10	26	24	60
Aspiration	12	13	15	40
Total	**22**	**39**	**39**	**100**

8.4.2 Variant 2

Task 1

Compute a 95 % confidence interval for the mean population incubation period in patients with hepatitis A.

Incubation period, days	Frequency f, number of patients
1 – 15	40
16 – 30	120
31 – 45	75
46 – 60	14
61 – 75	1
Total	**250**

Task 2

Using the data from the table, measure the strength and direction of the association between the age at diagnosis and survival in patients with lung carcinoma.

Patient №	1	2	3	4	5	6	7	8
Age X, years	71	71	73	74	75	76	76	78
Survival Y, months	135	96	66	44	16	22	11	5

8.4.3 Variant 3

Task 1

400 patients with multiple sclerosis are included in a clinical study. 25 of 200 treated with an experimental therapy relapse within 6 months after the end of the therapeutic course, while with the conventional therapy this number is 45 of 200.
Is there a statistically significant difference between the two types of therapy with respect to the proportion of relapsing patients?

Task 2

The table records the distance walked in a 6-minute walk test (6MWT) and the level of NT-proBNP in 8 patients with incipient heart failure. Compose a model of the dependence between the walked distance and the level of NT-proBNP.

Patient №	1	2	3	4	5	6	7	8
Distance X, km	0.7	0.7	0.6	0.5	0.4	0.4	0.3	0.3
NT-proBNP Y, pg/µl	110	120	140	160	180	200	240	300

8.4.4 Variant 4

Task 1

Patients with non-small-cell lung carcinoma are distributed in two independent groups according to the applied therapy. Is there a difference between the two groups with respect to the mean survival?

Type of treatment	Sample size	Mean survival	Standard deviation
Experimental	$n_1 = 40$	$\bar{X}_1 = 18.5$ months	$S_1 = 5.2$ months
Conventional	$n_2 = 40$	$\bar{X}_2 = 15.0$ months	$S_2 = 4.0$ months

Task 2

A study of 300 patients with type 2 diabetes records the adherence to therapy and the achievement of a target value of HbA1c. Is there an association between the adherence to therapy and the achievement of the target value?

Adherence to therapy	Target HbA1c	Without target HbA1c	Total
Good	96	84	180
Poor	36	84	120
Total	**132**	**168**	**300**

8.5 Solutions of the Variants

8.5.1 Solution of variant 1, task 1 (standardization)

Step 1. Crude and age-specific case-fatality rates.

$$IR = \frac{\text{deceased}}{\text{passed through}} \times 100$$

Table 8.2: Crude and department-specific case-fatality rates

Department	IR_A, %	IR_B, %
Infectious diseases	$10/500 \times 100 = 2.00$	$35/1{,}500 \times 100 = 2.33$
Surgery	$50/1{,}500 \times 100 = 3.33$	$70/2{,}000 \times 100 = 3.50$
Oncology	$85/1{,}000 \times 100 = 8.50$	$120/1{,}500 \times 100 = 8.00$
Total	$145/3{,}000 \times 100 =$ **4.83**	$225/5{,}000 \times 100 =$ **4.50**

Viewed overall, hospital A has a **higher** case fatality (4.83 % versus 4.50 %), although by departments the differences are small and go in different directions.

Step 2. Standard — the structure by departments in hospital B.

$$S = \frac{\text{passed through the department (B)}}{\text{all passed through (B)}} \times 100$$

Table 8.3: Computing the standard

Department	Passed through (B)	S, %
Infectious diseases	1,500	30.0
Surgery	2,000	40.0
Oncology	1,500	30.0
Total	**5,000**	**100.0**

Step 3. Standardized measures.

$$AIR = \frac{IR \times S}{100}$$

Table 8.4: Standardized measures

Department	S, %	AIR_A, %	AIR_B, %
Infectious diseases	30.0	$2.00 \cdot 30/100 = 0.60$	$2.33 \cdot 30/100 = 0.70$
Surgery	40.0	$3.33 \cdot 40/100 = 1.33$	$3.50 \cdot 40/100 = 1.40$

Department	S, %	AIR_A, %	AIR_B, %
Oncology	30.0	$8.50 \cdot 30/100 = 2.55$	$8.00 \cdot 30/100 = 2.40$
Total	**100.0**	**4.48**	**4.50**

Step 4. Conclusion. The standardized case-fatality measures are **4.48 %** for hospital A and **4.50 %** for hospital B. After equalizing the structure by departments, the difference between the two hospitals practically disappears. The higher crude rate in hospital A is due to the different composition of the patients who passed through. In hospital A the share of oncological patients, among whom the case fatality is the highest, is 33.3 % versus 30.0 % in hospital B, and the share of infectious-disease patients, among whom the case fatality is the lowest, is only 16.7 % versus 30.0 % in hospital B.

Check: hospital B, whose structure was taken as the standard, keeps its overall measure of 4.50 % before and after the standardization.

8.5.2 Solution of variant 1, task 2 (χ^2-test)

Step 1. Hypotheses. H_0: there is no association between the method of abortion and the type of complication. H_1: there is an association. Significance level $\alpha = 0.05$.

Step 2. Expected frequencies.

$$E_{ij} = \frac{R_i \times C_j}{N}$$

$$E_{11} = \frac{60 \times 22}{100} = 13.2 \qquad E_{12} = \frac{60 \times 39}{100} = 23.4 \qquad E_{13} = \frac{60 \times 39}{100} = 23.4$$

$$E_{21} = \frac{40 \times 22}{100} = 8.8 \qquad E_{22} = \frac{40 \times 39}{100} = 15.6 \qquad E_{23} = \frac{40 \times 39}{100} = 15.6$$

All expected frequencies are ≥ 5 — the conditions for applying the test are fulfilled.

Step 3. Test statistic.

Table 8.5: Computing χ^2

Cell	O	E	$(O-E)^2/E$
Classical × bleeding	10	13.2	0.776
Classical × infection	26	23.4	0.289
Classical × infertility	24	23.4	0.015
Aspiration × bleeding	12	8.8	1.164
Aspiration × infection	13	15.6	0.433
Aspiration × infertility	15	15.6	0.023
Total	**100**	**100**	**2.70**

$$\chi^2 = 0.776 + 0.289 + 0.015 + 1.164 + 0.433 + 0.023 = 2.70$$

Step 4. Degrees of freedom and p-value.

$$df = (R-1)(C-1) = (2-1)(3-1) = 2$$

The critical value at $df = 2$ and $p = 0.10$ is 4.605. Since $2.70 < 4.605$, it follows that $p > 0.10$.

Step 5. Strength of the association. $k = \min(2,\ 3) = 2$, that is, $k - 1 = 1$:

$$V = \sqrt{\frac{2.70}{100 \cdot 1}} = \sqrt{0.027} = 0.16$$

Step 6. Conclusion. $p > 0.10 > \alpha$. We do not reject H_0. The data do not give sufficient evidence of an association between the method of abortion and the type of complication ($\chi^2 = 2.70$; $df = 2$; $p > 0.10$; $V = 0.16$). The absence of statistical significance at $N = 100$ does not prove that there is no association — the sample is small and the power is low.

8.5.3 Solution of variant 2, task 1 (confidence interval)

Step 1. Midpoints of the intervals and arithmetic mean.

Table 8.6: Working table

Period, days	f	X_u	$X_u \cdot f$	$X_u - \bar{X}$	$(X_u - \bar{X})^2 \cdot f$
1 – 15	40	8	320	−18.96	14,379.26
16 – 30	120	23	2,760	−3.96	1,881.79
31 – 45	75	38	2,850	11.04	9,141.12
46 – 60	14	53	742	26.04	9,493.14
61 – 75	1	68	68	41.04	1,684.28
Total	**250**		**6,740**		**36,579.60**

$$\bar{X} = \frac{\sum(X_u \cdot f)}{\sum f} = \frac{6\,740}{250} = 26.96 \text{ days}$$

Step 2. Standard deviation.

$$S = \sqrt{\frac{\sum(X_u - \bar{X})^2 \cdot f}{n-1}} = \sqrt{\frac{36\,579.60}{249}} = \sqrt{146.91} = 12.12 \text{ days}$$

Step 3. Standard error of the mean.

$$SE_{\bar{X}} = \frac{S}{\sqrt{n}} = \frac{12.12}{\sqrt{250}} = \frac{12.12}{15.81} = 0.77 \text{ days}$$

Step 4. Confidence interval. At $n = 250$ the critical value of the t-distribution practically coincides with $Z = 1.96$:

$$95\ \%\ CI = 26.96 \pm 1.96 \cdot 0.77 = 26.96 \pm 1.51$$

- lower bound: 25.45 days;
- upper bound: 28.47 days.

Step 5. Conclusion. With 95 % confidence, the mean incubation period in patients with hepatitis A in the population lies between **25.45 and 28.47 days**.

8.5.4 Solution of variant 2, task 2 (correlation)

Step 1. Arithmetic means.

$$\bar{X} = \frac{594}{8} = 74.25 \text{ years}, \qquad \bar{Y} = \frac{395}{8} = 49.375 \text{ months}$$

Step 2. Deviations, cross products, and squares.

Table 8.7: Working table for the correlation

№	X	Y	d_x	d_y	$d_x \cdot d_y$	d_x^2	d_y^2
1	71	135	−3.25	85.625	−278.28	10.5625	7,331.64
2	71	96	−3.25	46.625	−151.53	10.5625	2,173.89
3	73	66	−1.25	16.625	−20.78	1.5625	276.39
4	74	44	−0.25	−5.375	1.34	0.0625	28.89
5	75	16	0.75	−33.375	−25.03	0.5625	1,113.89
6	76	22	1.75	−27.375	−47.91	3.0625	749.39
7	76	11	1.75	−38.375	−67.16	3.0625	1,472.64
8	78	5	3.75	−44.375	−166.41	14.0625	1,969.14
Total	**594**	**395**	**0.00**	**0.00**	**−755.75**	**43.50**	**15,115.88**

Step 3. Pearson's coefficient.

$$r = \frac{\sum (d_x \cdot d_y)}{\sqrt{\sum d_x^2 \cdot \sum d_y^2}} = \frac{-755.75}{\sqrt{43.50 \cdot 15{,}115.88}} = \frac{-755.75}{\sqrt{657{,}540.8}} = \frac{-755.75}{810.89} = -0.93$$

Step 4. Statistical significance.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.93 \cdot \sqrt{6}}{\sqrt{1-0.8649}} = \frac{-2.278}{0.3676} = -6.20$$

$$df = n - 2 = 6$$

The critical value at $df = 6$ and $p = 0.01$ is 3.707. Since $|-6.20| > 3.707$, it follows that $p < 0.01$.

Step 5. Conclusion. $r = -0.93$. There is a **strong negative** association between the age at diagnosis and survival in patients with lung carcinoma: the higher age at diagnosis is associated with a shorter survival. The association is statistically significant ($t = -6.20$; $df = 6$; $p < 0.01$). The conclusion holds for the observed age range 71–78 years and does not prove a causal dependence.

8.5.5 Solution of variant 3, task 1 (Z-test for two proportions)

Step 1. Hypotheses. H_0: there is no difference between the population proportions of relapsing patients under the two therapies. H_1: there is a difference. $\alpha = 0.05$.

Step 2. Sample proportions.

$$\hat{p}_1 = \frac{25}{200} \times 100 = 12.5\,\% \qquad \hat{p}_2 = \frac{45}{200} \times 100 = 22.5\,\%$$

The clinical effect is $\Delta = 22.5 - 12.5 = 10$ percentage points.

Step 3. Pooled proportion.

$$\hat{p} = \frac{25 + 45}{200 + 200} \times 100 = \frac{70}{400} \times 100 = 17.5\,\%$$

Step 4. Test statistic.

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(100 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{12.5 - 22.5}{\sqrt{17.5 \cdot 82.5 \cdot \left(\frac{1}{200} + \frac{1}{200}\right)}}$$

$$Z = \frac{-10}{\sqrt{1{,}443.75 \cdot 0.01}} = \frac{-10}{\sqrt{14.44}} = \frac{-10}{3.80} = -2.63$$

Step 5. p-value. $|Z| = 2.63 > 2.576$, therefore $p < 0.01$ (the software gives $p = 0.008$).

Step 6. Conclusion. $p < 0.01 < \alpha$. We reject H_0 and accept H_1. There is a statistically significant difference between the two types of therapy with respect to the proportion of relapsing patients. Under the experimental therapy the share of relapsed is 12.5 % versus 22.5 % under the conventional one, that is, 10 percentage points lower.

8.5.6 Solution of variant 3, task 2 (regression)

Step 1. Determining the variables. The walked distance X is the independent (explanatory) variable, and the level of NT-proBNP Y — the dependent (outcome) one.

Step 2. Auxiliary table.

Table 8.8: Auxiliary table

№	X	Y	$X \cdot Y$	X^2
1	0.7	110	77.0	0.49
2	0.7	120	84.0	0.49
3	0.6	140	84.0	0.36
4	0.5	160	80.0	0.25
5	0.4	180	72.0	0.16
6	0.4	200	80.0	0.16
7	0.3	240	72.0	0.09
8	0.3	300	90.0	0.09
$n = 8$	**3.9**	**1,450**	**639.0**	**2.09**

Step 3. Regression coefficient.

$$b = \frac{n\sum(XY) - (\sum X)\,(\sum Y)}{n\sum X^2 - (\sum X)^2} = \frac{8 \cdot 639 - 3.9 \cdot 1{,}450}{8 \cdot 2.09 - 3.9^2}$$

$$b = \frac{5{,}112 - 5{,}655}{16.72 - 15.21} = \frac{-543}{1.51} = -359.60$$

Step 4. Intercept.

$$\bar{X} = \frac{3.9}{8} = 0.4875 \text{ km}, \qquad \bar{Y} = \frac{1{,}450}{8} = 181.25 \text{ pg/µl}$$

$$a = \bar{Y} - b \cdot \bar{X} = 181.25 - (-359.60 \cdot 0.4875) = 181.25 + 175.31 = 356.56$$

Step 5. Regression equation.

$$\hat{Y} = -359.60 \cdot X + 356.56$$

Step 6. Prediction. For a patient who walked 0.55 km:

$$\hat{Y} = -359.60 \cdot 0.55 + 356.56 = -197.78 + 356.56 = 158.78 \text{ pg/µl}$$

Step 7. Conclusion. Each additional walked kilometre in the 6-minute walk test is associated with a mean **360 pg/µl lower** level of NT-proBNP. Expressed in a more practical unit: each additional

100 metres corresponds to about 36 pg/µl lower value. At $R^2 = 0.845$, 84.5 % of the variability of NT-proBNP is explained by the walked distance, and the correlation coefficient is $r = -\sqrt{0.845} = -0.92$; the sign is negative because the slope of the line is negative.

8.5.7 Solution of variant 4, task 1 (t-test)

Step 1. Hypotheses. H_0: there is no difference between the two population mean survivals. H_1: there is a difference. $\alpha = 0.05$.

Step 2. Clinical effect.

$$\Delta = \bar{X}_1 - \bar{X}_2 = 18.5 - 15.0 = 3.5 \text{ months}$$

Step 3. Test statistic.

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{|18.5 - 15.0|}{\sqrt{\frac{5.2^2}{40} + \frac{4.0^2}{40}}}$$

$$t = \frac{3.5}{\sqrt{\frac{27.04}{40} + \frac{16.00}{40}}} = \frac{3.5}{\sqrt{0.676 + 0.400}} = \frac{3.5}{\sqrt{1.076}} = \frac{3.5}{1.04} = 3.37$$

Step 4. Degrees of freedom and p-value.

$$df \approx \min(n_1 - 1,\, n_2 - 1) = \min(39,\, 39) = 39$$

The critical value at $df = 39$ and $p = 0.01$ is 2.708. Since $3.37 > 2.708$, it follows that $p < 0.01$ (the software gives $p = 0.001$).

Step 5. Conclusion. $p < 0.01 < \alpha$. We reject H_0 and accept H_1. There is a statistically significant difference between the two groups with respect to the mean survival: under the experimental treatment it is 3.5 months higher (18.5 versus 15.0 months). Whether the difference is clinically significant is judged separately, against a pre-specified clinically important difference and against the profile of adverse reactions.

8.5.8 Solution of variant 4, task 2 (χ^2-test, 2 × 2 table)

Step 1. Hypotheses. H_0: there is no association between the adherence to therapy and the achievement of a target HbA1c. H_1: there is an association. $\alpha = 0.05$.

Step 2. Description. A target value is achieved by 96 of 180 patients with good adherence (53.3 %) and by 36 of 120 with poor adherence (30.0 %).

Step 3. Expected frequencies.

$$E_{11} = \frac{180 \times 132}{300} = 79.2 \qquad E_{12} = \frac{180 \times 168}{300} = 100.8$$

$$E_{21} = \frac{120 \times 132}{300} = 52.8 \qquad E_{22} = \frac{120 \times 168}{300} = 67.2$$

All expected frequencies are ≥ 5.

Step 4. Test statistic.

$$\chi^2 = \frac{(96 - 79.2)^2}{79.2} + \frac{(84 - 100.8)^2}{100.8} + \frac{(36 - 52.8)^2}{52.8} + \frac{(84 - 67.2)^2}{67.2}$$

$$\chi^2 = 3.56 + 2.80 + 5.35 + 4.20 = 15.91$$

Step 5. Degrees of freedom and p-value.

$$df = (2 - 1)(2 - 1) = 1$$

The critical value at $df = 1$ and $p = 0.01$ is 6.635. Since $15.91 > 6.635$, it follows that $p < 0.01$ (the software gives $p < 0.001$).

Step 6. Strength and direction of the association.

$$\varphi = \frac{(96 \cdot 84) - (36 \cdot 84)}{\sqrt{180 \cdot 120 \cdot 132 \cdot 168}} = \frac{8{,}064 - 3{,}024}{\sqrt{479{,}001{,}600}} = \frac{5{,}040}{21{,}886.1} = 0.23$$

Step 7. Conclusion. There is a statistically significant, but weak to moderate in strength association between the adherence to therapy and the achievement of a target HbA1c ($\chi^2 = 15.91$; $df = 1$; $p < 0.001$; $\varphi = 0.23$). The positive sign shows that good adherence is associated with more frequent achievement of the target value.

8.6 Preparation Test, Part 1

The first part covers the material of the seven exercises. For each question choose **one** correct answer. The key and the short explanations are in Section D.8.

1. The measure "35 % of the patients who passed through the department were admitted as emergencies" is:

    a) an absolute value;
    b) a proportion;
    c) a rate;
    d) a standardized rate.

2. After the standardization, the group whose structure was taken as the standard:

    a) always gets a lower rate;
    b) always gets a higher rate;
    c) keeps its overall rate unchanged;
    d) gets a rate equal to 100 %.

3. In a strongly right-skewed distribution of the hospital stay, the most appropriate is to report:

    a) arithmetic mean and standard deviation;
    b) median and interquartile range;
    c) mode and range;
    d) arithmetic mean and range.

4. A sample of 400 newborns has a mean weight of 3,400 g and a standard deviation of 500 g. Approximately how many newborns weigh between 2,400 g and 4,400 g?

    a) 272;
    b) 320;
    c) 380;
    d)399.

5. The standard error of the arithmetic mean:

    a) describes the scatter of the individual observations;
    b) grows with the increase of the sample size;
    c) decreases proportionally to $\sqrt{n}$;
    d) is always larger than the standard deviation.

6. A 95 % confidence interval for the mean survival is [6.6; 9.4] months. Which interpretation is correct?

    a) 95 % of the patients survive between 6.6 and 9.4 months;
    b) There is exactly 95 % probability that the true mean lies between these bounds;
    c) Upon repeated repetition of the procedure, about 95 % of the constructed intervals will contain the true mean;
    d) The mean in the sample lies between 6.6 and 9.4 months with probability 95 %.

7. A researcher lowers the significance level from $\alpha = 0.05$ to $\alpha = 0.01$ with unchanged sample size and effect size. The power of the test:

   a) increases;
   b) decreases;
   c) remains unchanged;
   d) becomes equal to β.

8. Patients are examined before and after a 12-week rehabilitation with respect to a normally distributed quantitative indicator. The appropriate test is:

   a) t-test for two independent samples;
   b) t-test for paired samples;
   c) Z-test for two proportions;
   d) χ^2-test.

9. The result of a study is $t = 1.42$ at $df = 60$. It follows that:

   a) $p < 0.01$ and we reject H_0;
   b) $p < 0.05$ and we reject H_0;
   c) $p > 0.10$ and we do not reject H_0;
   d) the p-value cannot be determined without α.

10. For a 3 × 4 contingency table the degrees of freedom are:

    a) 12;
    b) 7;
    c) 6;
    d) 5.

11. Which of the following is **not** a condition for applying the χ^2-test?

    a) The observations are independent;
    b) In the cells stand counts, not percentages;
    c) The studied variable is normally distributed;
    d) At least 80 % of the expected frequencies are ≥ 5.

12. Cramér's V is reported together with χ^2, because:

    a) it replaces the p-value;
    b) it measures the strength of the association independently of the sample size;
    c) it proves a causal dependence;
    d) it shows the direction of the association in a 3 × 3 table.

13. The scatter diagram shows a clear U-shaped association between two quantitative characteristics, and the computed Pearson coefficient is $r = 0.03$. The correct conclusion is:

    a) there is no association at all between the two variables;
    b) there is no linear association between the two variables, but there is a nonlinear one;
    c) the data contain an entry error;
    d) a larger sample is needed.

14. A regression model gives $\hat{Y} = 208 - 0.7 \cdot X$, where X is the age in years and Y — the maximal heart rate. The coefficient -0.7 means that:

    a) 70 % of the variability is explained by the age;
    b) the correlation coefficient is -0.7;
    c) with an increase of the age by 1 year the expected maximal heart rate decreases on average by 0.7 beats per minute;
    d) the maximal heart rate at birth is 0.7 beats per minute.

15. A regression model has $R^2 = 0.64$, and the slope of the line is negative. The correlation coefficient is:

    a) $+0.80$;
    b) -0.80;
    c) -0.64;
    d) -0.41.

8.7 Preparation Test, Part 2

The second part also covers concepts from Section A: types of graph, sampling designs, robustness and extrapolation. For each question choose **one** correct answer.

16. Which type of graph is recommended for displaying the distribution of a categorical variable measured on a **nominal** scale?

    a) pie chart;
    b) bar chart;
    c) histogram;
    d) boxplot.

17. The overall **crude** rate is:

    a) the sum of the crude rates of the subgroups;
    b) the sum of the standardized rates of the subgroups;
    c) the product of the crude rates of the subgroups;
    d) none of the above.

18. The variable "diagnosis", defined as "lung cancer, breast cancer, colorectal cancer or other", is measured on which scale?

    a) nominal;
    b) ordinal;
    c) interval;
    d) ratio.

19. An intercept $a = 1.85$ in a regression model means that:

    a) the slope of the line of best fit may be either positive or negative;
    b) the slope of the line of best fit is positive;
    c) the slope of the line of best fit is negative;
    d) none of the above.

20. Patients with multiple sclerosis were divided into 2 groups according to the therapy. The 95 % CI for the relapse rate under therapy A is between 21 % and 28 %, and under therapy B between 32 % and 42 %. What conclusion can be drawn?

    a) most probably there is no difference in the effectiveness of the two therapies;
    b) most probably therapy A is statistically significantly more effective than therapy B;
    c) most probably therapy B is statistically significantly more effective than therapy A;
    d) all conclusions are wrong.

21. Which answer correctly orders the sampling designs by the expected size of the **random** error, from smallest to largest?

    a) stratified sample, quota sample, cluster sample;
    b) cluster sample, quota sample, stratified sample;
    c) convenience sample, stratified sample, quota sample;

d) stratified sample, convenience sample, systematic sample.

22. Patients with lung cancer were divided into 2 groups according to the therapy. The relapse rate under therapy A is 18 % and under therapy B is 24 % ($p = 0.050$). What conclusion can be drawn?

    a) there is no ground to claim that the two therapies differ;
    b) therapy A is statistically significantly more effective than therapy B;
    c) therapy B is statistically significantly more effective than therapy A;
    d) all conclusions are wrong.

23. The presence of outliers in the data has the **weakest** effect on which measure?

    a) standard error of the mean;
    b) arithmetic mean;
    c) median;
    d) standard deviation.

24. In a study of mothers with multiple sclerosis the mean birth weight is 2 700 g for mothers in remission for more than 1 year and 2 100 g for mothers in remission for less than 1 year ($p = 0.006$); a t-test for two independent samples was used. Which statement is **true**?

    a) the result is statistically significant at $\alpha = 0.05$ and the null hypothesis may be rejected;
    b) McNemar's test could also be used in this case;
    c) the birth weight is most probably not normally distributed;
    d) the alternative hypothesis states that the mean birth weight is the same in the two groups.

25. Which of the following statements is **wrong**?

    a) interpolation is prediction inside the range of known values used to build the model;
    b) extrapolation is prediction outside that range;
    c) extrapolation achieves greater validity of the estimate obtained;
    d) extrapolation requires the observed relationship to hold outside the range of known values.

26. Parametric tests of hypotheses **lose** robustness with:

    a) small samples;
    b) the absence of outliers;
    c) samples of approximately equal size and dispersion;
    d) all answers are correct.

27. Under which condition does the t-distribution approach the Z-distribution in value?

    a) with normally distributed variables;
    b) with large samples;
    c) with quantitative variables measured on a ratio scale;
    d) with a paired-samples design.

28. In a study, postoperative complications were observed in 23 (10 %) of 230 patients; the 95 % CI for P is between 6 % and 14 %. Which statement is **wrong**?

    a) in the population the rate of postoperative complications is most probably greater than 6 %;
    b) in the sample studied the rate of postoperative complications may take a value between 6 % and 14 %;
    c) the authors would have obtained an estimate at a higher confidence level had they computed a 99 % CI;
    d) the authors would have obtained a more precise estimate had they included more patients.

29. Other things being equal, the **smallest** number of units of observation will be required with:

    a) a heterogeneous population and a high confidence level;
    b) a homogeneous population and a high confidence level;
    c) a heterogeneous population and a low confidence level;
    d) a homogeneous population and a low confidence level.

30. In a study with 121 followed-up patients the mean survival is 44 months with a standard deviation of 22 months. What is the standard error of the mean?

    a) 6 months;
    b) 4 months;
    c) 2 months;
    d) 1 month.

8.8 Preparation Test, Part 3

31. In a study of mothers with multiple sclerosis the mean birth weight is 2 700 g for mothers in remission for more than 1 year and 2 100 g for mothers in remission for less than 1 year ($p = 0.006$); a t-test for two independent samples was used. Which statement is **wrong**?

    a) the result is statistically significant at $\alpha = 0.05$ and the null hypothesis may be rejected;
    b) McNemar's test could also be used in this case;
    c) the birth weight is most probably normally distributed;
    d) the hypothesis under test states that the mean birth weight is the same in the two groups.

32. To use critical values from the Z-distribution, a large enough sample and a reliable estimate of which population parameter are required?

    a) the arithmetic mean;
    b) the standard error;
    c) the standard deviation;
    d) the kurtosis.

33. Other things being equal, the **largest** number of units of observation will be required with:

    a) a heterogeneous population and a high confidence level;
    b) a homogeneous population and a high confidence level;
    c) a heterogeneous population and a low confidence level;
    d) a homogeneous population and a low confidence level.

34. The variable "height in cm" is measured on which scale?

    a) nominal;
    b) ordinal;
    c) interval;
    d) ratio.

35. Patients with multiple sclerosis were divided into 2 groups according to the therapy. The 95 % CI for the relapse rate under therapy A is between 21 % and 28 %, and under therapy B between 24 % and 38 %. What conclusion can be drawn?

    a) most probably therapy A is statistically significantly more effective than therapy B;
    b) most probably therapy B is statistically significantly more effective than therapy A;
    c) the data do not allow a conclusion about a difference between the two therapies;
    d) the two therapies are proven to be equally effective.

36. Which answer correctly orders the sampling designs by the expected size of the **random** error, from largest to smallest?

    a) stratified sample, quota sample, cluster sample;
    b) cluster sample, quota sample, stratified sample;
    c) convenience sample, stratified sample, quota sample;
    d) stratified sample, convenience sample, systematic sample.

37. Patients with lung cancer were divided into 2 groups according to the therapy. The relapse rate under therapy A is 28 % and under therapy B is 24 % ($p = 0.0051$). What conclusion can be drawn?

    a) therapy A is statistically significantly more effective than therapy B;
    b) therapy B is statistically significantly more effective than therapy A;
    c) there is no difference in the effectiveness of the two therapies;
    d) all conclusions are wrong.

38. A regression coefficient $b = -1.85$ means that:

    a) the slope of the line may be either positive or negative;
    b) the slope of the line is positive;
    c) the slope of the line is negative;
    d) none of the above.

39. Which of the following statements is **true**?

    a) the ordinal scale has the highest precision of measurement;
    b) on the ordinal scale the distance between the categories is the same in every part of the scale;
    c) on the ordinal scale one unit may belong to more than one category;
    d) on the ordinal scale the categories have a logical ordering.

40. The presence of outliers in the data has the **strongest** effect on which measure?

    a) standard deviation;
    b) interquartile range;
    c) mode;
    d) significance level.

41. The overall **standardized** measure is:

    a) the product of the standardized rates of the subgroups;
    b) the sum of the crude rates of the subgroups;
    c) the sum of the standardized rates of the subgroups;
    d) none of the above.

42. In a study with 81 patients the mean survival is 36 months with a standard deviation of 9 months. What is the standard error of the mean?

    a) 9 months;
    b) 4 months;
    c) 2 months;
    d) 1 month.

43. Parametric tests of hypotheses **retain** robustness with:

    a) small samples;
    b) the presence of outliers;
    c) samples of approximately equal size and dispersion;

d) none of the above.

44. In a study, a prediabetic state was found in 27 of 91 children (29.7 %); the 95 % CI for P is between 20.3 % and 39.1 %. Which statement is **wrong**?

    a) in the population of children the rate of prediabetes is most probably between 20.3 % and 39.1 %;
    b) the probability that a randomly chosen child is in a prediabetic state is between 20.3 % and 39.1 %;
    c) in the sample studied the rate of prediabetes is 29.7 %;
    d) the authors would have obtained an estimate at a higher confidence level had they computed a 99 % CI.

45. Which type of graph is recommended for displaying the distribution of a categorical variable measured on an **ordinal** scale?

    a) pie chart;
    b) bar chart;
    c) histogram;
    d) boxplot.

8.9 Open Questions

The open questions require a short written answer rather than a choice among ready-made statements. An explanation in the terms of the course is expected.

Question 46

The crude incidence measures in two regions are 15 % for region A and 19 % for region B. Taking the age structure of region B as the standard, the standardized rates are 16 % for region A and 19 % for region B.
Fill in the missing values if the age structure of **region A** is taken as the standard.

	Crude measure	Standardized (standard — region A)	Standardized (standard — region B)
Region A	15 %		16 %
Region B	19 %		19 %

Question 47

What conclusion can be drawn from the regression model shown in the figure? How is the coefficient of determination interpreted in this case? What is the value of the correlation coefficient?

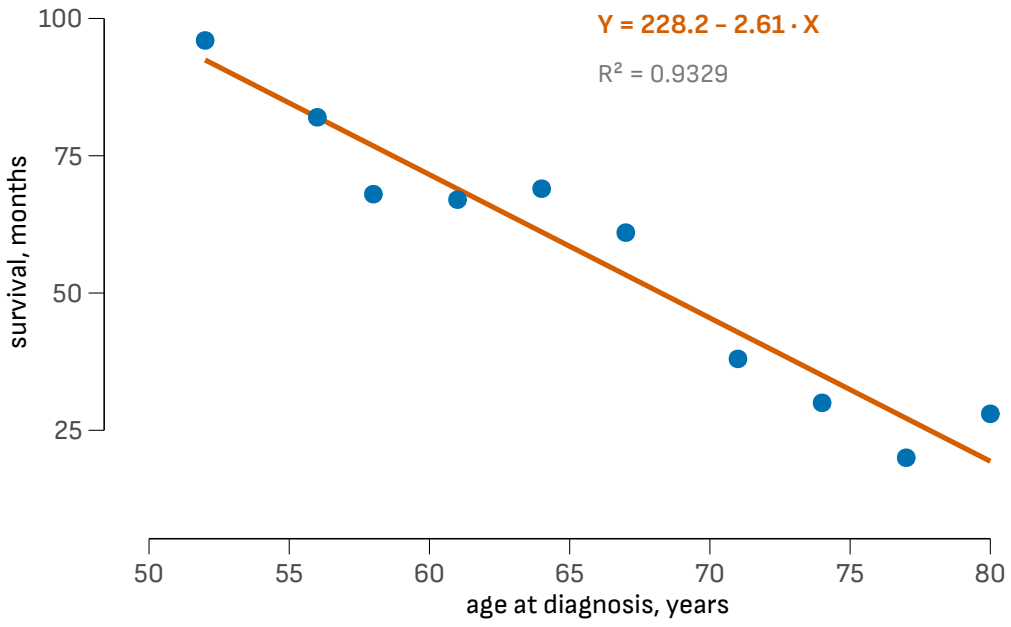

Figure 8.1: Figure for question 47: relationship between the age at diagnosis and survival in 10 patients.

Question 48

What are the main risks when a sample is selected by the snowball method?

Question 49

A study compares the relapse rate under two treatments. The result obtained is $t = 2.080$ with $df = 21$. What is the p-value in this case and what does it mean? What conclusion can be drawn?

Question 50

What is autocorrelation? How can its presence be checked?

Question 51

What is the binomial distribution? In which cases can it be approximated by the normal distribution?

Question 52

A researcher wants to compare the changes in the mean body mass index before and after a three-month ketogenic diet in patients with obesity. The body mass index is normally distributed. Which test of hypothesis should be applied?

Question 53

A sample of 100 newborns is included in a prospective study. The mean birth length is 49 cm with a standard deviation of 1 cm. Length is normally distributed.
How many of the newborns included have a length smaller than 49 cm or greater than 51 cm?

Question 54

What are one-sided tests of hypotheses? What is their main advantage and why?

A Additional Concepts for the Examination

This appendix collects concepts that appear in the examination tests but are not central to the computational exercises: choice of graph, sampling designs, robustness, the binomial distribution, autocorrelation and extrapolation.

A.1 Choosing a Graph

The type of graph is determined by the type of the variable, not by the taste of the author.

Table A.1: Choosing a graph according to the type of variable

Type of variable	Recommended graph	Unsuitable graph
categorical, **nominal**	pie chart; bar chart	histogram, boxplot
categorical, **ordinal**	**bar chart** (preserves the ordering)	pie chart
quantitative, one distribution	histogram; boxplot	pie chart
quantitative, by groups	boxplot split by group	pie chart
two quantitative	scatter plot	bar chart

! Pie chart or bar chart

A pie chart shows **shares of one whole** and therefore works only with nominal categories, in which the ordering carries no information.
With an **ordinal** variable the ordering does carry information. A pie chart loses it, because a circle has neither a beginning nor an end. A bar chart with the categories placed in their natural order is therefore preferred.

⚠ A histogram and a bar chart are not the same thing

A **bar chart** presents **categories**: the bars are separated by gaps, because there is no continuity between the categories.
A **histogram** presents a **quantitative** variable grouped into intervals: the bars touch each other, because the intervals are adjacent parts of one continuous scale.

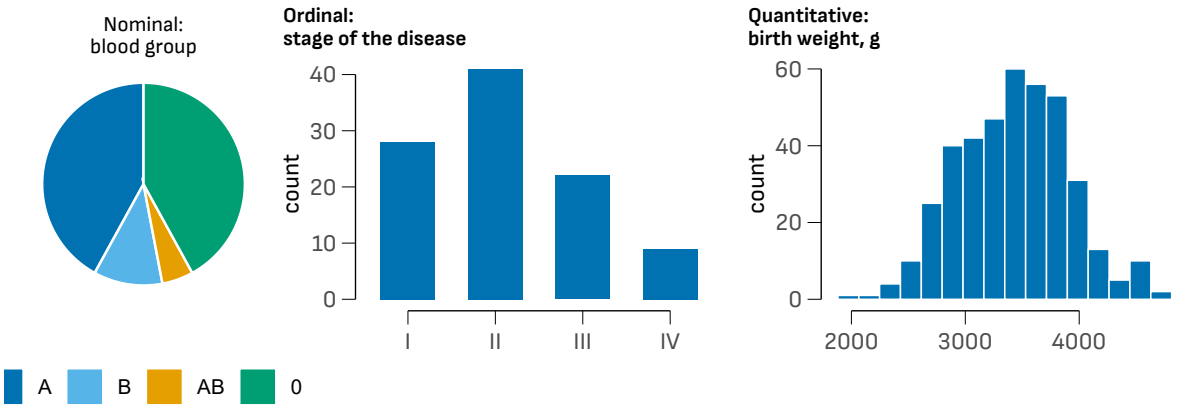

Figure A.1: Three graphs for three different types of variable. Left, a pie chart for a nominal variable; centre, a bar chart for an ordinal variable; right, a histogram for a quantitative variable.

A.2 Sampling Designs

The way the sample was selected determines both the possibility of generalisation and the expected size of the error.

> ⚠ The two kinds of error
>
> **Random (stochastic) error** arises because only a part of the target population is observed. It decreases as the sample size increases and is expressed numerically by the standard error. **Systematic (non-stochastic) error** arises from the way of selection, from inaccurate measurement or from the drop-out of participants. It does **not** decrease with sample size: a large but selectively recruited sample yields a precise but wrong estimate.

Probability designs — every unit has a known, non-zero probability of being included. Only for these can the random error be computed.

Table A.2: Probability sampling designs

Design	Essence	Expected random error
Simple random	each unit is drawn independently	baseline
Stratified (typological)	the population is split into homogeneous strata and units are drawn within each	**smallest**
Systematic (mechanical)	every k-th element of a list is drawn	close to simple random
Cluster	whole groups are drawn (schools, wards)	**largest**

Non-probability designs — the probability of inclusion is unknown. The random error cannot be assessed correctly and the risk of systematic error is high.

Table A.3: Non-probability sampling designs

Design	Essence	Main risk
Quota	selection until pre-set quotas are filled	selection within the quota is not random
Convenience	the most easily accessible persons are included	highest risk of systematic error
Purposive (expert)	the researcher picks "typical" cases	depends entirely on the researcher's judgement
Snowball	participants refer the next participants	chain recruitment of similar persons

! Ordering by expected random error

From **smallest** to **largest**:
stratified (typological) → simple random → quota → cluster → convenience.
A cluster sample has a larger error than a simple random sample of the same size, because the units within a cluster resemble each other and therefore carry less new information.

i The snowball method

It is used with hard-to-reach populations: people who inject drugs, patients with rare diseases, undocumented migrants. A few initial participants refer further participants, who refer still others, and the sample "grows".
There are three risks:

1. **Systematic error** if the initial participants are too similar to each other in place of residence, social status or attitudes — the whole chain then inherits their profile.
2. **High random error**, especially with a small number of initial participants.
3. **Interruption of the chain**: if a participant refuses to take part or does not refer anyone, recruitment in that branch stops.

A.3 Sample Size

The required number of units of observation depends on four things.

Table A.4: What the required sample size depends on

If...	the required size is...
the population is **homogeneous** (small S)	smaller
the effect sought is **large**	smaller
the desired confidence level is **high** (99 %)	larger
the desired precision is **high** (narrow interval)	larger

The two typical examination questions

The **smallest** number of observations is required with a **homogeneous population and a low confidence level** (equivalently, with a large expected effect).
The **largest** number of observations is required with a **heterogeneous population and a high confidence level** (equivalently, with a small expected effect).

A.4 Robustness of Parametric Tests

Robustness is the property of a test to give reliable results even when its assumptions are partly violated.

When parametric tests are robust

The t-test **retains** robustness when:

- the samples are large enough — the central limit theorem then makes the distribution of the sample means approximately normal;
- the two samples are approximately **equal in size** and in **dispersion**;
- there are no pronounced extreme values.

The t-test **loses** robustness when:

- the samples are **small**;
- the distribution is strongly skewed or contains outliers;
- the sizes and the variances of the two groups differ markedly.

When robustness is lost, a non-parametric test is used: Mann-Whitney for independent groups and Wilcoxon for paired ones.

A.5 One-Sided Tests

The difference between a two-sided and a one-sided alternative was introduced in Section 4. Here the reason why a one-sided test has greater power is summarised.

In a two-sided test the significance level $\alpha = 0.05$ is split into 2.5 % in each tail of the distribution. In a one-sided test the **whole** 5 % critical region is concentrated in one tail. The critical value is therefore lower: 1.645 instead of 1.960 for the normal distribution.

The lower threshold means that a real effect **in the expected direction** is detected more easily, that is, the test has higher power.

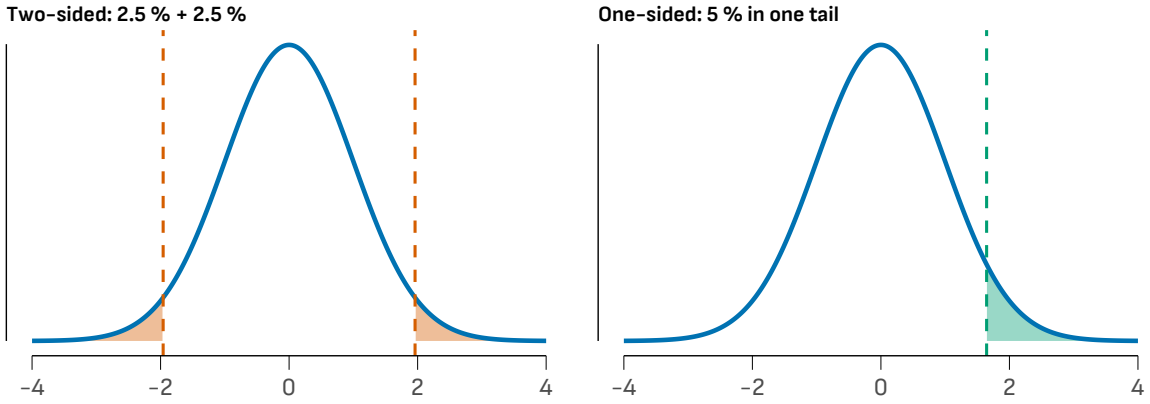

Figure A.2: Critical regions of a two-sided and a one-sided test at the same significance level. In the two-sided test 5 % are split into 2.5 % in each tail; in the one-sided test the whole critical region lies in one tail, which lowers the critical value from 1.960 to 1.645.

> ⚠ The price of a one-sided test
>
> A one-sided test **cannot** detect an effect in the opposite direction: if the new treatment turns out to be worse, the test simply fails to reject the null hypothesis. The direction is therefore declared **before** the data are seen. Switching from a two-sided to a one-sided test after the result has been seen inflates the actual risk of a type I error and is not acceptable.

A.6 The Binomial Distribution

When the outcome is **dichotomous** — "yes / no", "present / absent", "recovered / not recovered" — the number of cases with the outcome of interest is described by the **binomial distribution**.

> ⚠ Definition
>
> The binomial distribution describes the probability of observing exactly k successes in n independent trials in which the probability of success at each trial is the same and equal to p.

There are three requirements: a fixed number of trials n, independent trials and a constant probability p.

> ❗ When the binomial is approximated by the normal
>
> With a **large enough sample** and a probability p that is **not too close to 0 or 1**, the binomial distribution becomes almost symmetric and is well approximated by the normal distribution. The practical rule is that both counts — of cases with the event and of cases without it — should be at least 5, that is, $np \geq 5$ and $n(1-p) \geq 5$. The approximation is most accurate for p near 0.5.

It is precisely this approximation that allows the confidence interval for a proportion (Section 3) and the Z-test for two proportions (Section 4) to be built with critical values from the normal distribution.

With rare events or small samples the approximation breaks down and exact methods are used.

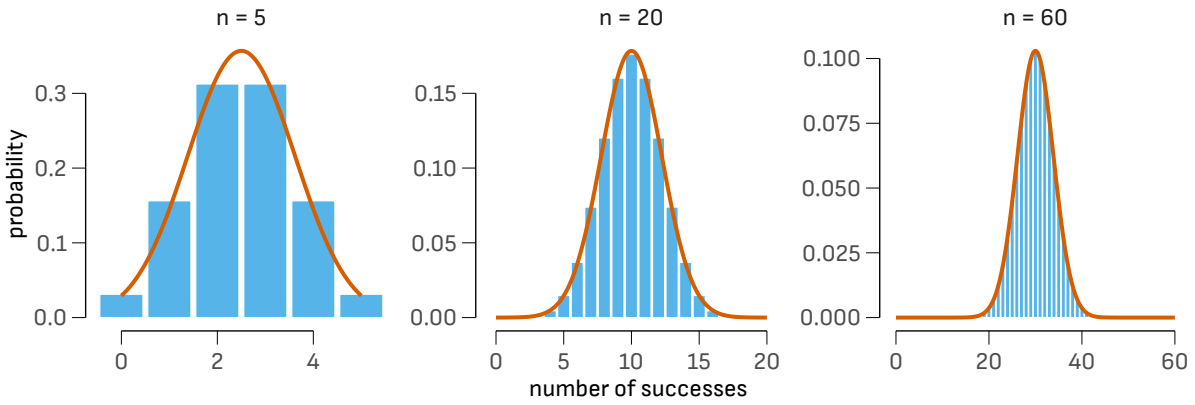

Figure A.3: The binomial distribution at p = 0.5 for three sample sizes. As n grows, the shape becomes more symmetric and closer to the normal curve shown as a solid line.

A.7 Autocorrelation

The regression model (Section 7) assumes that the residuals are **independent** of one another.

> ⚠ Definition
>
> **Autocorrelation** is the phenomenon in which the residuals are not independent but are related to their values at preceding observations: a positive residual is followed by a positive one, a negative one by a negative one.

It occurs most often with **time series** — monthly incidence, daily number of admissions — but also with spatial data or with systematically ordered observations.

The problem lies not in the estimates of the coefficients but in their **precision**: with autocorrelation the standard errors are underestimated, the confidence intervals look narrower than they are and the p-values look smaller. The result is an apparently significant effect.

> 💡 How it is checked
>
> The simplest check is **graphical**: the residuals are plotted in time order or in the order of the observations. If long waves of consecutive positive and negative residuals appear around the zero line instead of random scatter, there is autocorrelation. Formal tests also exist, for example the Durbin–Watson test.

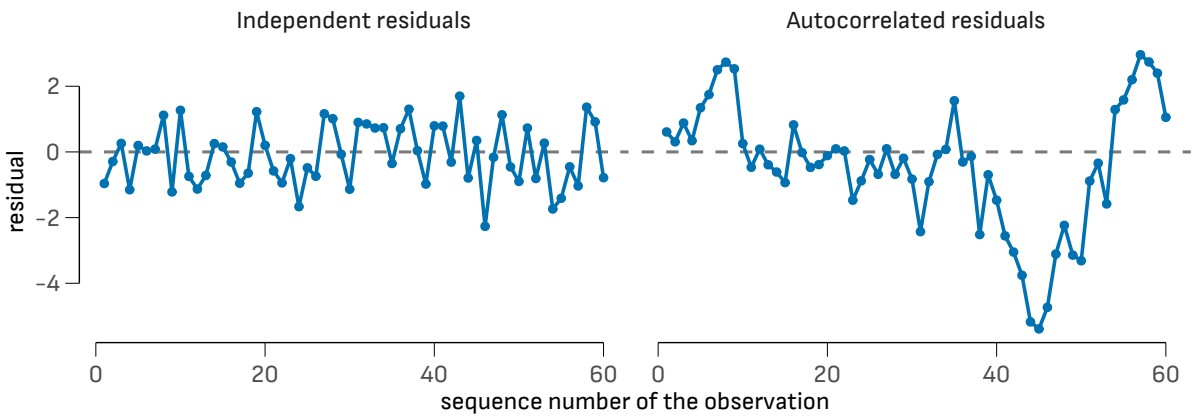

Figure A.4: Residuals plotted in time order. Left, the scatter is random and no pattern stands out; right, long waves of consecutive positive and negative residuals are visible – a sign of autocorrelation.

A.8 Interpolation and Extrapolation

Definitions

Interpolation is prediction **inside** the range of already known values that were used to build the model.
Extrapolation is prediction **outside** that range.

Interpolation is the usual and justified application of the model. Extrapolation requires an additional and unverifiable assumption: that the established relationship also holds outside the observed range.

A common misconception

Extrapolation does **not** increase the validity of an estimate; it decreases it. The further beyond the observed range one predicts, the more uncertain the result becomes, even though the formula keeps producing numbers.

B Formulas

The appendix collects all formulas of the course. Next to each one the chapter in which it is introduced and explained is indicated.

Relative Measures and Standardization

Proportion (structural or "extensive" measure) — Section 1

$$\text{proportion} = \frac{\text{part of the whole}}{\text{the whole}} \times 100$$

Rate (frequency or "intensive" measure) — Section 1

$$\text{rate} = \frac{\text{number of events}}{\text{population at risk} \times \text{time}} \times 100$$

Standard weight (quota of the standard population) — Section 1

$$S = \frac{\text{number in the subgroup of the standard population}}{\text{total number in the standard population}} \times 100$$

Age-standardized rate for a subgroup — Section 1

$$AIR = \frac{IR \times S}{100}$$

The overall standardized rate is the **sum** of the standardized rates by subgroups. The overall crude rate is **not** a sum of the crude rates.

Central Tendency

Arithmetic mean, ungrouped data — Section 2

$$\bar{X} = \frac{\sum X}{n}$$

Arithmetic mean, grouped data — Section 2

$$\bar{X} = \frac{\sum (X_u \cdot f)}{\sum f}, \qquad X_u = \frac{\text{lower bound} + \text{upper bound}}{2}$$

Proportion — Section 2

$$\hat{p} = \frac{m}{n} \times 100$$

Median Me — the value that divides the data, arranged in ascending order, into two equal parts.
Mode M_0 — the most frequently occurring value.

Dispersion

Range — Section 2

$$R = X_{max} - X_{min}$$

Standard deviation, ungrouped data — Section 2

$$S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$$

Standard deviation, grouped data — Section 2

$$S = \sqrt{\frac{\sum(X_u - \bar{X})^2 \cdot f}{n-1}}$$

Interquartile range — Section 2

$$IQR = Q_3 - Q_1$$

Extreme are the values outside the bounds $Q_1 - 1.5 \cdot IQR$ and $Q_3 + 1.5 \cdot IQR$.

Rule of the three sigmas (under a normal distribution) — Section 2

$$\bar{X} \pm 1S \approx 68\,\%, \qquad \bar{X} \pm 2S \approx 95\,\%, \qquad \bar{X} \pm 3S \approx 99.7\,\%$$

Estimation of Population Parameters

Standard error of the arithmetic mean — Section 3

$$SE_{\bar{X}} = S_{\bar{X}} = \frac{S}{\sqrt{n}}$$

Standard error of the proportion — Section 3

$$SE_{\hat{p}} = S_p = \sqrt{\frac{\hat{p}\,(100 - \hat{p})}{n}}$$

Confidence interval for a mean — Section 3

$$CI = \bar{X} \pm t_{1-\alpha/2;\,n-1} \cdot SE_{\bar{X}}$$

For a large sample the approximation $t \approx Z$ is used.

Confidence interval for a proportion — Section 3

$$CI = \hat{p} \pm Z_{1-\alpha/2} \cdot SE_{\hat{p}}$$

Table B.1: Confidence multipliers

Confidence level	90 %	95 %	99 %
$Z_{1-\alpha/2}$	1.645	1.960	2.576

Hypothesis Testing

Z-test for two proportions — Section 4

$$\hat{p} = \frac{m_1 + m_2}{n_1 + n_2} \times 100$$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\,(100 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

t-test for two independent samples — Section 4

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{SE_{\bar{X}_1}^2 + SE_{\bar{X}_2}^2}}$$

$$df \approx \min(n_1 - 1,\, n_2 - 1)$$

Confidence interval for the difference between two means — Section 4

$$CI = (\bar{X}_1 - \bar{X}_2) \pm t_{1-\alpha/2;\,df} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Power — Section 4

$$\text{power} = 1 - \beta$$

Association between Categorical Characteristics

Expected frequencies — Section 5

$$E_{ij} = \frac{R_i \times C_j}{N}$$

χ^2 **statistic** — Section 5

$$\chi^2 = \sum \frac{(O - E)^2}{E}, \qquad df = (R - 1)(C - 1)$$

Cramér's coefficient V — Section 5

$$V = \sqrt{\frac{\chi^2}{N \cdot (k - 1)}}, \qquad k = \min(R,\, C)$$

Coefficient φ **for a 2 × 2 table** — Section 5

$$\varphi = \frac{(a \cdot d) - (b \cdot c)}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}$$

For a 2 × 2 table $V = |\varphi|$ and $\chi^2 = Z^2$ hold.

Association between Quantitative Characteristics

Pearson's correlation coefficient — Section 6

$$r = \frac{\sum(d_x \cdot d_y)}{\sqrt{\sum d_x^2 \cdot \sum d_y^2}}, \qquad d_x = X - \bar{X}, \quad d_y = Y - \bar{Y}$$

Significance test for r — Section 6

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \qquad df = n-2$$

Regression coefficient — Section 7

$$b = \frac{n\sum(X \cdot Y) - (\sum X)\,(\sum Y)}{n\sum X^2 - (\sum X)^2}$$

Intercept — Section 7

$$a = \bar{Y} - (b \cdot \bar{X})$$

Regression equation and prediction — Section 7

$$\hat{Y} = b \cdot X + a$$

Coefficient of determination — Section 7

$$R^2 = r^2, \qquad |r| = \sqrt{R^2}$$

The sign before r coincides with the sign of the regression coefficient b.

C Statistical Tables

The tables give the **critical values** of the respective distributions under a two-sided alternative. The way of use is the same for all:

> **!** How to read a statistical table
>
> 1. Compute the test statistic and the degrees of freedom.
> 2. Find the row corresponding to your degrees of freedom.
> 3. Compare the **absolute value** of the statistic with the numbers in the row, starting from left to right.
> 4. If the statistic is **smaller** than the leftmost value, then $p > 0.10$.
> 5. If it lies between two neighbouring columns, p lies between their corresponding levels.
> 6. If it is **larger** than the rightmost value, p is smaller than the level of that column.
>
> If the exact value of df is missing in the table, the nearest **smaller** row is used — thus the conclusion remains on the safe side.

Critical Values of the t-Distribution

Table C.1: Critical values of the t-distribution under a two-sided alternative

df	$p = 0.10$	$p = 0.05$	$p = 0.01$
1	6.314	12.706	63.657
2	2.920	4.303	9.925
3	2.353	3.182	5.841
4	2.132	2.776	4.604
5	2.015	2.571	4.032
6	1.943	2.447	3.707
7	1.895	2.365	3.499
8	1.860	2.306	3.355
9	1.833	2.262	3.250
10	1.812	2.228	3.169
11	1.796	2.201	3.106
12	1.782	2.179	3.055
13	1.771	2.160	3.012
14	1.761	2.145	2.977
15	1.753	2.131	2.947
16	1.746	2.120	2.921
17	1.740	2.110	2.898
18	1.734	2.101	2.878
19	1.729	2.093	2.861
20	1.725	2.086	2.845

df	$p = 0.10$	$p = 0.05$	$p = 0.01$
21	1.721	2.080	2.831
22	1.717	2.074	2.819
23	1.714	2.069	2.807
24	1.711	2.064	2.797
25	1.708	2.060	2.787
26	1.706	2.056	2.779
27	1.703	2.052	2.771
28	1.701	2.048	2.763
29	1.699	2.045	2.756
30	1.697	2.042	2.750
35	1.690	2.030	2.724
40	1.684	2.021	2.704
45	1.679	2.014	2.690
50	1.676	2.009	2.678
60	1.671	2.000	2.660
80	1.664	1.990	2.639
100	1.660	1.984	2.626
120	1.658	1.980	2.617
over 120	1.645	1.960	2.576

The last row is the table of Z

For very large samples the t-distribution coincides with the normal one. Therefore the row "over 120" gives exactly the critical values of Z: **1.645** at $p = 0.10$, **1.960** at $p = 0.05$, and **2.576** at $p = 0.01$. These three numbers are used also in the Z-test for two proportions and in the confidence intervals.

Critical Values of the Chi-Square Distribution

Table C.2: Critical values of the χ^2-distribution

df	$p = 0.10$	$p = 0.05$	$p = 0.01$	$p = 0.001$
1	2.706	3.841	6.635	10.828
2	4.605	5.991	9.210	13.816
3	6.251	7.815	11.345	16.266
4	7.779	9.488	13.277	18.467
5	9.236	11.070	15.086	20.515
6	10.645	12.592	16.812	22.458
7	12.017	14.067	18.475	24.322
8	13.362	15.507	20.090	26.124
9	14.684	16.919	21.666	27.877
10	15.987	18.307	23.209	29.588
11	17.275	19.675	24.725	31.264
12	18.549	21.026	26.217	32.909
13	19.812	22.362	27.688	34.528
14	21.064	23.685	29.141	36.123
15	22.307	24.996	30.578	37.697
20	28.412	31.410	37.566	45.315
25	34.382	37.652	44.314	52.620
30	40.256	43.773	50.892	59.703

The most commonly used rows

For a 2 × 2 table $df = 1$ is used, for a 2 × 3 or 3 × 2 table — $df = 2$, for a 3 × 3 table — $df = 4$, for a 4 × 2 table — $df = 3$.

Critical Values of the Correlation Coefficient

The table allows the significance of r to be checked directly, without computing t. If the absolute value of the computed coefficient is larger than the tabulated one, the association is significant at the respective level.

Table C.3: Critical values of Pearson's correlation coefficient

n	$df = n - 2$	$p = 0.10$	$p = 0.05$	$p = 0.01$
5	3	0.805	0.878	0.959
6	4	0.729	0.811	0.917
7	5	0.669	0.754	0.875
8	6	0.621	0.707	0.834
9	7	0.582	0.666	0.798
10	8	0.549	0.632	0.765

n	$df = n - 2$	$p = 0.10$	$p = 0.05$	$p = 0.01$
12	10	0.497	0.576	0.708
15	13	0.441	0.514	0.641
20	18	0.378	0.444	0.561
25	23	0.337	0.396	0.505
30	28	0.306	0.361	0.463
40	38	0.264	0.312	0.403
50	48	0.235	0.279	0.361
60	58	0.214	0.254	0.330
80	78	0.185	0.220	0.286
100	98	0.165	0.197	0.256

⚠ Why this table is instructive

Read the column for $p = 0.05$ from top to bottom. At $n = 8$, $|r| \geq 0.71$ is necessary for the association to be significant, while at $n = 100$, $|r| \geq 0.20$ is enough.
Therefore **one and the same** number r can be significant in one study and insignificant in another. Significance depends on the sample size, and the strength of the association — not. Therefore all three are always reported: r, n, and p.

D Answers

The answers are given briefly, to serve for checking after independent solving. In the computational tasks of the colloquium the formula, the intermediate steps, the unit of measurement, and the verbal conclusion are assessed, not only the final number.

D.1 To Chapter 1

Task 1.1. The age structure is 23.33 %; 13.67 %; 34.00 % and 29.00 %. The measures are extensive, and their sum is 100 %.

Task 1.2. The crude fertility rates are 2.06 % in region A and 1.96 % in region B. With the structure of region B as the standard, the standardized rate for region A is 1.96 %, and that for region B remains 1.96 %. After equalizing the age structure practically no difference between the regions is observed; the unrounded values are 1.959 % and 1.962 %.

Task 1.3. The crude case-fatality rates are 10.50 % in hospital A and 9.11 % in hospital B. With the structure of hospital B as the standard, the standardized values are respectively 9.36 % and 9.11 %. The difference decreases after the standardization, but the rate in hospital A remains slightly higher.

Task 1.4. With the structure of hospital A as the standard the comparison is 10 ‰ versus 8 ‰, that is, the standardized rate is higher in hospital A. The new values with the structure of hospital B cannot be computed only from the given overall rates. The rates by subgroup and the composition of hospital B are necessary. Without them it cannot be guaranteed either that the direction of the comparison will remain the same.

Task 1.5. No. Japan has an older population structure, and age is strongly associated with mortality. The comparison of the overall crude measures mixes the influence of the age structure with the remaining differences between the countries. An age-standardized comparison is necessary.

Task 1.6. The structure by departments is 40.0 % cardiology, 35.0 % surgery, and 25.0 % neurology; these are proportions, because the denominator is the whole of 240 patients, divided into categories. The hospital case fatality is $18/240 \times 100 = 7.5$ %; this is an intensive measure, because the denominator represents the patients among whom the outcome “death” can occur.

Short quiz: 1 — b; 2 — c; 3 — c; 4 — b; 5 — c.

D.2 To Chapter 2

Task 2.1. The midpoints of the intervals are 50, 55, 60, 65, 70, 75, and 80 bpm. The weighted sum is 2,940, therefore $\bar{X} = 2\,940/45 = 65.33$ bpm. The sample standard deviation is $S = 8.88$ bpm.

Task 2.2. The standard deviation is zero only when all observations have the same value. It cannot be negative, because it is obtained as a square root of a non-negative quantity.

Task 2.3. All deviations cannot be positive, because the sum of $X_i - \bar{X}$ is always zero. All deviations are zero only when all observations equal the mean.

Task 2.4. The value 4,300 g lies two standard deviations above the mean, therefore above it are approximately 2.5 %, or 50 newborns. The value 3,100 g lies one standard deviation below the mean, therefore below it are approximately 16 %, or 320 newborns. The two areas do not overlap: about 370 newborns in total.

Task 2.5. a) categorical, ordinal scale; b) quantitative discrete, ratio scale; c) categorical, nominal scale; d) quantitative continuous, interval scale; e) quantitative discrete score, usually treated as an interval scale; f) categorical dichotomous, nominal scale.

Task 2.6. $IQR = Q_3 - Q_1 = 50 - 38 = 12$ months. The distribution is left-skewed. The value 0 months is extreme; the lower whisker reaches 34 months.

Task 2.7. $\bar{X} = 5.78$ mg/L, $Me = 4$ mg/L, $Q_1 = 3$ mg/L, $Q_3 = 6$ mg/L, and $IQR = 3$ mg/L. The value 18 mg/L pulls the mean to the right, therefore the median and the interquartile range describe the data more appropriately.

Task 2.8. Both groups have a mean of 100. The standard deviations are 1.58 in group A and 15.81 in group B. Group A is more homogeneous, because its values are considerably less scattered around the same mean.

Short quiz: 1 – a; 2 – c; 3 – b; 4 – c; 5 – b; 6 – c.

D.3 To Chapter 3

Task 3.1. The midpoints of the intervals are 8, 23, 38, 53, and 68 days. $\bar{X} = 26.96$ days, $S = 12.12$ days, and $SE_{\bar{X}} = 12.12/\sqrt{250} = 0.77$ days. At $t_{0.975;249} = 1.970$ we obtain the 95 % CI $[25.45;\ 28.47]$ days.

Task 3.2. The midpoints of the intervals are 3, 8, 13, 18, and 23 months. $\bar{X} = 13.08$ months, $S = 5.89$ months, and $SE_{\bar{X}} = 0.34$ months. At $t_{0.975;299} = 1.968$ we obtain the 95 % CI $[12.41;\ 13.75]$ months.

Task 3.3. The point estimate remains 56 %. At $Z = 2.576$ and $SE_{\hat{p}} = 4.96$ percentage points the 99 % CI is $56 \pm 2.576 \cdot 4.96 = [43.21;\ 68.79]$ %. The interval is wider than the 95 % CI, because the higher coverage requires a larger critical value.

Task 3.4. Correct answer: d. The level 95 % refers to the long-term coverage of the method, not to a probability for the particular already computed interval or for the sample proportion.

Task 3.5. The sample size grows four times, therefore the standard error decreases $\sqrt{4} = 2$ times. With approximately the same critical value the width of the interval also decreases approximately two times.

Task 3.6. There are 2 observations with improvement and 18 without improvement. The condition that both counts are at least 5 is not fulfilled, therefore the simple normal interval is not appropriate. A Wilson interval or an exact binomial interval is preferred.

Task 3.7. The intervals overlap between 34 % and 38 %. This is not enough either for proving or for rejecting a difference. A confidence interval for the difference between the proportions or an appropriate test is necessary.

Short quiz: 1 — c; 2 — c; 3 — b; 4 — c; 5 — a; 6 — c.

D.4 To Chapter 4

Task 4.1. The proportions with relapse are 12.5 % and 22.5 %, and the pooled proportion is 17.5 %. We obtain $Z = -2.63$ and a two-sided $p = 0.0085$. We reject H_0. Under the experimental therapy the sample proportion with relapse is 10 percentage points lower; the 95 % CI for the difference is $[-17.38;\ -2.62]$ percentage points.

Task 4.2. The proportions with one-year survival are 83.33 % and 63.33 %. $Z = 5.54$ and $p < 0.001$, therefore we reject H_0. The estimated difference is 20 percentage points, 95 % CI $[13.11;\ 26.89]$ percentage points.

Task 4.3. The standard error of the difference is $\sqrt{6^2 + 7^2} = 9.22$ mg/dL and $|t| = 15/9.22 = 1.63$. Without the sample sizes or the degrees of freedom the exact p-value from the t-distribution cannot be determined. The normal approximation gives $p \approx 0.104$ and does not lead to rejection of H_0 at $\alpha = 0.05$.

Task 4.4. Correct answer: a. The groups are independent, the outcome is not normally distributed, and three groups are compared, therefore the Kruskal–Wallis test is appropriate.

Task 4.5. Correct answer: b. With unchanged sample size, effect, and test, the lower α decreases the risk of a type I error, but increases the risk of a type II error and decreases the power.

Task 4.6. Correct answer: d. The appropriate two-sided alternative is that the population mean survivals are different. Since $p = 0.08 > 0.05$, we do not reject H_0, but we do not prove equality.

Task 4.7. At $t = 2.96$ and $df = 40$ the two-sided $p \approx 0.0052 < 0.01$. If H_0 and the assumptions are true, the probability of $|t| \geq 2.96$ is about 0.52 %. We reject H_0 at $\alpha = 0.05$ and conclude that there is a statistically significant difference in the physical activity between smokers and non-smokers. The direction and the magnitude require the descriptive estimates.

Task 4.8. The greatest power is obtained with a large true effect, low variability, a large sample, and a higher α, all other things being equal. In practice α is set in advance according to the acceptable risk of a type I error and is not increased only to obtain a significant result.

Task 4.9. The paired t-test is appropriate. $H_0 : \mu_{\text{after}-\text{before}} = 0$, that is, the mean within-individual change in the population is zero.

Task 4.10. The result is statistically significant: the interval does not include zero and $p < 0.001$. The estimated effect is 1.8 mmHg, and the interval [1.2; 2.4] mmHg shows its precision. The clinical significance is not determined by p; it requires a comparison with a pre-specified clinically important difference and an assessment of the benefits and harms.

Short quiz: 1 — b; 2 — c; 3 — d; 4 — a; 5 — b; 6 — b.

D.5 To Chapter 5

Task 5.1. The expected frequencies are 13.2; 23.4; 23.4 for the classical method and 8.8; 15.6; 15.6 for the aspiration one. The per-cell contributions are 0.776; 0.289; 0.015; 1.164; 0.433, and 0.023, whence $\chi^2 = 2.70$ at $df = (2-1)(3-1) = 2$. The critical value at $p = 0.10$ is 4.605, therefore $p > 0.10$ and we do not reject H_0. $V = \sqrt{2.70/(100 \cdot 1)} = 0.16$. The data do not give sufficient evidence of an association between the method of abortion and the type of complication; the sample is small and the absence of significance does not prove absence of an association.

Task 5.2. The expected frequencies are 12.33 and 37.67 for the first and the fourth group and 24.67 and 75.33 for the second and the third. $\chi^2 = 39.21$ at $df = (4-1)(2-1) = 3$; the critical value at $p = 0.01$ is 11.345, therefore $p < 0.01$, and the software gives $p < 0.001$. We reject H_0. $V = \sqrt{39.21/(300 \cdot 1)} = 0.36$ — a strong association at $k - 1 = 1$. The largest contribution comes from the cell "over 15 y. with complaints" — 19.90 of the total 39.21: 28 cases are observed against 12.33 expected. There is a statistically significant, strong association between the duration of the work experience and the presence of neurological complaints.

Task 5.3. $\chi^2 = 2.79$ at $df = 3$; the critical value at $p = 0.10$ is 6.251, therefore $p > 0.10$ and we do not reject H_0; $V = 0.10$. With the visual complaints the proportions increase weakly with the work experience (12.9 %; 11.5 %; 13.3 %; 20.0 %), while with the neurological ones the increase is sharp (8.0 %; 15.0 %; 27.0 %; 56.0 %). One and the same factor can be associated with one outcome and not associated with another.

Task 5.4. The expected frequencies are 30 and 120 for the early reperfusion and 20 and 80 for the late one. $\chi^2 = 15.0$ at $df = 1$; the critical value at $p = 0.01$ is 6.635, therefore $p < 0.01$, and the software gives $p < 0.001$. $\varphi = -0.24$. The negative sign shows that the early reperfusion is associated with a **lower** frequency of heart failure: 12.0 % versus 32.0 %. There is a statistically significant, moderate in strength association.

Task 5.5. a) $df = (4-1)(3-1) = 6$. b) The critical value at $df = 6$ and $p = 0.05$ is 12.592; since $18.0 > 12.592$, the result is significant. c) $k = \min(4,\ 3) = 3$, that is, $k - 1 = 2$, whence $V = \sqrt{18/(400 \cdot 2)} = 0.15$. At $k - 1 = 2$ this is a weak to moderate association: significant, but not strong.

Task 5.6. No. The condition requires the expected frequencies in a 2 × 2 table to be sufficiently large, and 3.2 is below 5. The appropriate choice is Fisher's exact test.

Task 5.7. The result is statistically significant, but not clinically significant. At $V = 0.011$ the association is practically zero: the significance is due entirely to the huge sample size, because χ^2 grows proportionally to N. Exactly for this reason the strength of the association is always reported together with the p-value.

Task 5.8. At $\alpha = 0.05$ and 20 independent checks, on average one significant finding is expected even when no association exists. By reporting only the significant results, the researcher presents an expected number of chance findings as discoveries. The correct behaviour is to declare all performed checks and, if necessary, to apply a correction for multiplicity.

Short quiz: 1 — c; 2 — b; 3 — c; 4 — c; 5 — b; 6 — b.

D.6 To Chapter 6

Task 6.1. $\bar{X} = 74.25$ years and $\bar{Y} = 49.375$ months; $\sum(d_x d_y) = -755.75$, $\sum d_x^2 = 43.50$, and $\sum d_y^2 = 15{,}115.88$, whence $r = -755.75/\sqrt{43.50 \cdot 15{,}115.88} = -0.93$. Significance test: $t = -6.20$ at $df = 6$; the critical value at $p = 0.01$ is 3.707, therefore $p < 0.01$. There is a strong negative, statistically significant association: the higher age at diagnosis is associated with a shorter survival.

Task 6.2. $\bar{X} = 0.4875$ km and $\bar{Y} = 181.25$ pg/µl; $r = -0.92$ and $t = -5.71$ at $df = 6$, therefore $p < 0.01$. A strong negative, statistically significant association: the greater walked distance is associated with a lower level of NT-proBNP.

Task 6.3. a) The association is descending and approximately linear in the observed range. b) Observation 8 (0.3 km; 300 pg/µl) is the farthest from the general trend, but it is not isolated and does not by itself determine the result. c) Pearson's coefficient is acceptable. If observation 8 is removed, the association remains strong and negative, which speaks for the robustness of the conclusion.

Task 6.4. All d_x cannot be positive, because their sum is always zero. All d_x are zero only when all values of X equal the mean, that is, the variable is a constant. Then $\sum d_x^2 = 0$ and the denominator of the formula becomes zero: the coefficient is not defined. In the absence of variability in X one cannot speak of a coordinated change with Y.

Task 6.5. a) A moderate negative association: the longer sleep is associated with a lower level of cortisol. b) At $n = 60$ and $df = 58$ we obtain $t = -3.84$; the critical value at $p = 0.01$ is about 2.66, therefore $p < 0.01$ and the association is significant. c) At $n = 10$ and $df = 8$ the same coefficient gives $t = -1.43$, which is below the critical 1.860 at $p = 0.10$; the association would not be statistically significant. The strength of the association does not change, but the evidential weight depends on the sample size.

Task 6.6. No. The most likely explanation is a third variable — the high outdoor temperature during the summer months, which simultaneously increases the consumption of ice cream and the number of bathers, and hence the drownings. This is an example of a confounding factor.

Task 6.7. The trap "restricted range" has been committed. Among professional basketball players the height varies in a very narrow range, and the artificial restriction of the range of one variable decreases the correlation coefficient. The conclusion about adults in general is unjustified.

Task 6.8. At $n = 20$: $t = 0.30 \cdot \sqrt{18}/\sqrt{1 - 0.09} = 1.33$ at $df = 18$; the critical value at $p = 0.10$ is 1.734, therefore $p > 0.10$ and the association is not significant. At $n = 200$: $t = 4.43$ at $df = 198$, therefore $p < 0.001$. The strength of the association is the same in the two studies; only the precision with which it is estimated differs.

Short quiz: 1 — b; 2 — b; 3 — c; 4 — a; 5 — c; 6 — b.

D.7 To Chapter 7

Task 7.1. $\sum X = 594$, $\sum Y = 395$, $\sum XY = 28{,}573$, and $\sum X^2 = 44{,}148$. The numerator is $8 \cdot 28{,}573 - 594 \cdot 395 = 228{,}584 - 234{,}630 = -6{,}046$, and the denominator is $8 \cdot 44{,}148 -$

$594^2 = 353{,}184 - 352{,}836 = 348$. Therefore $b = -6{,}046/348 = -17.37$ months per year and $a = 49.375 - (-17.37 \cdot 74.25) = 1{,}339.36$. The model is $\hat{Y} = -17.37 \cdot X + 1{,}339.36$. For a 72-year-old patient: $\hat{Y} = -17.37 \cdot 72 + 1{,}339.36 = 88.5$ months. At $R^2 = 0.869$ the model explains 86.9 % of the variability of the survival, and $r = -\sqrt{0.869} = -0.93$.

Task 7.2. $\sum X = 3.9$, $\sum Y = 1{,}450$, $\sum XY = 639$, and $\sum X^2 = 2.09$. Hence $b = -543/1.51 = -359.60$ and $a = 181.25 + 175.31 = 356.56$; the model is $\hat{Y} = -359.60 \cdot X + 356.56$. Each additional walked kilometre is associated with a mean 359.6 pg/µl lower level of NT-proBNP, that is, each additional 100 metres corresponds to about 36 pg/µl. At $R^2 = 0.845$ the correlation coefficient is $r = -0.92$.

Task 7.3. a) $\sum X = 676$, $\sum Y = 507$, $\sum XY = 43{,}761$, and $\sum X^2 = 55{,}880$; the numerator is $437{,}610 - 342{,}732 = 94{,}878$, and the denominator is $558{,}800 - 456{,}976 = 101{,}824$, whence $b = 0.93$ points per minute and $a = 50.7 - 0.93 \cdot 67.6 = -12.29$. The model is $\hat{Y} = 0.93 \cdot X - 12.29$. b) At 30 minutes of reading: $\hat{Y} = 0.93 \cdot 30 - 12.29 = 15.7$ points. c) $r = +\sqrt{0.93} = +0.96$; the sign is positive because the slope is positive. d) No. The observational design admits also reverse causation: the more anxious students may read more precisely because they are anxious.

Task 7.4. a) With an increase of the daily salt intake by 1 g the expected systolic pressure increases on average by 1.8 mmHg. b) $|r| = \sqrt{0.25} = 0.50$ and the sign is positive because $b > 0$, that is, $r = +0.50$. c) No. The explained part of the variability is only 25 %, that is, three quarters of the changes are due to other factors; for an individual prediction the accuracy is insufficient.

Task 7.5. a) The first model predicts better, because it explains 81 % versus 36 % of the variability. b) $|r| = \sqrt{0.81} = 0.90$ and $|r| = \sqrt{0.36} = 0.60$. c) The sign of the regression coefficient, that is, the slope of the line, is necessary.

Task 7.6. a) With an increase of the age by 1 year the expected maximal heart rate decreases on average by 0.7 beats per minute. b) $\hat{Y} = 208 - 0.7 \cdot 60 = 166$ beats per minute. c) Formally $a = 208$ means a maximal heart rate at an age of zero years. The value lies outside the range of the data, therefore it is interpreted only as a mathematical starting point.

Task 7.7. The prediction for a 15-year-old child is **extrapolation**: the model is built on ages 40–60 years and there is no reason to assume that the same line holds outside this range. The formula will give a number, but it is not justified.

Task 7.8. The second model. The high R^2 with 5 observations is weak evidence: through a small number of points a line can be drawn almost perfectly, without this meaning that the association will be confirmed with new data. The model on 5,000 observations gives a lower, but much more reliably estimated R^2.

Short quiz: 1 – b; 2 – c; 3 – b; 4 – c; 5 – b; 6 – c.

D.8 To the Colloquium

The complete solutions of the four variants are given in Section 8. Below are the answers to the three parts of the preparation test and to the open questions.

Preparation Test, Part 1

1 — b. The denominator is the whole (all patients who passed through), divided into categories; the sum of the proportions is 100 %.

2 — c. The group whose structure was taken as the standard keeps its overall measure. This is a convenient check of the computations.

3 — b. In a skewed distribution the mean is shifted towards the extreme values; the median and the interquartile range describe the data better.

4 — c. The bounds 2,400 g and 4,400 g lie two standard deviations from the mean, that is, they cover approximately 95 % of the observations: $0.95 \cdot 400 = 380$.

5 — c. The standard error describes the scatter of the sample mean and decreases with $\sqrt{n}$; it is smaller than the standard deviation.

6 — c. The level 95 % refers to the long-term coverage of the **method**, not to a probability for the particular already computed interval.

7 — b. The lower α decreases the risk of a type I error, but increases the risk of a type II error, that is, decreases the power.

8 — b. The same persons are measured twice, therefore the design is dependent.

9 — c. At $df = 60$ the critical value for $p = 0.10$ is 1.671. Since $1.42 < 1.671$, it follows that $p > 0.10$ and we do not reject H_0.

10 — c. $df = (3-1)(4-1) = 6$.

11 — c. The normal distribution is a condition for the parametric tests for quantitative characteristics, not for the χ^2-test, which works with frequencies.

12 — b. χ^2 grows with the sample size, while V is a standardized measure of the strength of the association.

13 — b. Pearson's coefficient measures only the linear part of the association. In a U-shaped association it can be close to zero despite the clear nonlinear association.

14 — c. The regression coefficient shows the mean change in Y when X increases by one unit.

15 — b. $|r| = \sqrt{0.64} = 0.80$, and the sign is determined by the slope, which is negative.

Preparation Test, Part 2

16 — a. A pie chart presents shares of one whole and is appropriate for a nominal scale, in which the ordering of the categories carries no information. A bar chart is also admissible, but the pie chart is the recommended one for a purely nominal variable.

17 — d. The overall crude rate is neither the sum nor the product of the subgroup measures. It is computed anew, from the total number of events and the total number of observations.

18 — a. The categories "lung cancer", "breast cancer", "colorectal cancer" and "other" have no logical ordering, so the scale is nominal.

19 — a. The intercept shows where the line crosses the Y-axis and carries no information about the slope. A positive intercept is compatible with both a positive and a negative slope.

20 — b. The intervals [21 %; 28 %] and [32 %; 42 %] do not overlap, so the difference is statistically significant. The relapse rate is lower under therapy A, which is therefore the more effective one.

21 — a. From smallest to largest random error: typological (stratified) sample, quota sample, cluster sample. Stratification reduces the dispersion within the strata, whereas a cluster sample increases it, because the units within one cluster resemble each other.

22 — a. The rule for rejecting H_0 is $p < 0.05$. The value $p = 0.050$ is not smaller than 0.05, so the null hypothesis is not rejected and the data do not permit a claim of a difference. This is a borderline case: with $p = 0.049$ the conclusion would have been the opposite, which shows why a decision should not be reduced to a single number.

23 — c. The median depends on the **position** of the observations rather than on their values and is therefore the most resistant to extreme values. The arithmetic mean, the standard deviation and the standard error are computed from the values themselves and are strongly affected.

24 — a. $p = 0.006 < 0.05$, so the null hypothesis is rejected. McNemar's test does not apply to independent groups and categorical variables; the t-test used assumes an approximately normal distribution; and the alternative hypothesis states that the means are **different**.

25 — c. Extrapolation does not increase but decreases the validity of the estimate, because it requires the additional and unverifiable assumption that the relationship holds outside the observed range.

26 — a. Robustness is lost with small samples. The absence of outliers and similar sizes and variances are conditions under which the test **retains** robustness.

27 — b. With large samples the t-distribution approaches the normal one; above $df = 120$ the critical values practically coincide with 1.645, 1.960 and 2.576.

28 — b. In the sample studied the rate is known exactly and equals 10 %. The confidence interval estimates the **population** rate, not the rate in the sample.

29 — d. The smallest sample size is required with a homogeneous population (small dispersion) and a low confidence level (smaller critical value).

30 — c. $SE_{\bar{X}} = S/\sqrt{n} = 22/\sqrt{121} = 22/11 = 2$ months.

Preparation Test, Part 3

31 — b. McNemar's test applies to **paired** samples and a dichotomous categorical variable. Here the groups are independent and the variable is quantitative, so the statement is wrong.

32 — c. Critical values from the Z-distribution are used when the population standard deviation is known or has been reliably estimated from a large enough sample.

33 — a. The largest sample size is required with a heterogeneous population (large dispersion) and a high confidence level.

34 — d. Height has a natural zero and permits ratios, so the scale is ratio.

35 — c. The intervals [21 %; 28 %] and [24 %; 38 %] overlap in the region 24–28 %. Overlap alone does not allow the claim that a difference exists. Note that overlap likewise **does not prove** equality; a confidence interval for the difference or an appropriate test is needed for a definitive conclusion.

36 — b. The reverse ordering of question 21: from largest to smallest random error — cluster, quota, typological sample.

37 — b. $p = 0.0051 < 0.05$, so the difference is statistically significant. The lower relapse rate (24 %) is under therapy B, which is therefore the more effective one.

38 — c. A negative regression coefficient means a negative slope: as X increases, the expected value of Y decreases.

39 — d. The ordinal scale orders the categories logically, but it does not guarantee equal distances between them and it is not the most precise scale. Each unit belongs to exactly one category.

40 — a. The standard deviation is computed from the squared deviations and therefore reacts most strongly to extreme values. The interquartile range and the mode are resistant, and the significance level is set by the researcher.

41 — c. The standardized subgroup measures **are summed**, and their sum gives the overall standardized rate.

42 — d. $SE_{\bar{X}} = 9/\sqrt{81} = 9/9 = 1$ month.

43 — c. Robustness is retained with samples of approximately equal size and dispersion.

44 — b. The confidence interval estimates the **population proportion**, not the probability for an individual child. The interval is not a statement about an individual.

45 — b. On an ordinal scale the ordering of the categories carries information. A bar chart preserves it, whereas a pie chart loses it, because a circle has neither a beginning nor an end.

Open Questions

Question 46. Region A keeps its measure of **15 %**, because its structure is taken as the standard. For region B the expected answer is **18 %**.

	Crude	Standard — region A	Standard — region B
Region A	15 %	**15 %**	16 %
Region B	19 %	**18 %**	19 %

The conclusion is the same under both standards: the incidence in region B is higher, and the difference is about 3 percentage points.

Methodological note. The exact value for region B does not follow uniquely from the overall measures alone: its computation requires the subgroup measures and the structure of region A. Two things are certain — region A necessarily keeps 15 %, and the direction of the difference is preserved.

Question 47. The model is $\hat{Y} = 228.2 - 2.61 \cdot X$. The regression coefficient $b = -2.61$ is negative, so the relationship is **inverse**: with an increase of the age at diagnosis by **1 year** the expected survival decreases on average by **2.61 months**.

The coefficient of determination $R^2 = 0.9329$ means that **93.3 %** of the variability of survival is explained by the age at diagnosis; the remaining 6.7 % is due to other factors and to random variation.

The correlation coefficient is $|r| = \sqrt{0.9329} = 0.97$, and the sign is negative, because the slope of the line is negative, that is, $r \approx -0.97$ – a very strong inverse correlation.

Question 48. The main risks of snowball sampling are three. First, **systematic error**: if the initial participants are too similar in place of residence, social status or attitudes, the whole chain inherits their profile and the sample does not represent the population. Second, **high random error**, especially when the initial participants are few. Third, **interruption of the chain**: if a participant refuses or refers no one, recruitment in that branch stops. In addition, the probability of inclusion is unknown, which prevents a correct assessment of the random error.

Question 49. With $df = 21$ the critical value for $p = 0.05$ is exactly **2.080**. The computed $t = 2.080$ coincides with it, that is, $p = 0.05$.

By the rule $p \geq 0.05$ the null hypothesis is not rejected. The result is not statistically significant at $\alpha = 0.05$ and gives no ground to claim a difference in the relapse rate between the two treatments. This does not prove that no difference exists: the value lies exactly on the boundary and a larger sample could lead to a different conclusion.

Question 50. Autocorrelation is the phenomenon in which the residuals of a regression model are not independent but are related to their values at preceding observations. It occurs most often in time series, but also in spatial or systematically ordered data.

The consequence is an underestimation of the standard errors: the confidence intervals look narrower and the p-values smaller than they actually are, which leads to apparently significant results.

It is checked primarily **graphically**: the residuals are plotted in time order or in the order of the observations, and one looks for long waves of consecutive positive and negative values instead of random scatter. Formal tests also exist, for example the Durbin–Watson test.

Question 51. The **binomial distribution** describes the probability of observing exactly k successes in n independent trials in which the probability of success at each trial is the same and equal to p. It serves to model the frequency of dichotomous outcomes of the type “yes / no”, “present / absent”.

The binomial distribution can be approximated by the normal one with a **large enough sample** and a probability p that is not too close to 0 or 1. A practical rule is that the number of cases with the event and the number without it should both be at least 5, that is, $np \geq 5$ and $n(1-p) \geq 5$; the approximation is best for p near 0.5. It is exactly this approximation that underlies the confidence interval for a proportion and the Z-test for two proportions.

Question 52. The body mass index is a quantitative, normally distributed variable measured **twice in the same patients** – before and after the diet. The design is paired, so a **t-test for two paired samples** is applied.

Question 53. Length is normally distributed with a mean of 49 cm and a standard deviation of 1 cm.

- The value 49 cm coincides with the mean, so **50 %** of the newborns, that is 50 children, lie below it.
- The value 51 cm is two standard deviations above the mean, so approximately **2.5 %**, that is about 2 children, lie above it.

The two regions do not overlap, and in total about **52** of the 100 newborns have a length below 49 cm or above 51 cm.

Question 54. One-sided tests are tests in which the alternative hypothesis has a pre-specified direction: the researcher expects one value to be only greater or only smaller than the other.

Their main advantage is **greater statistical power**. The reason is that the whole critical region α is concentrated in one tail of the distribution instead of being split between the two. This lowers the critical value — for example from 1.960 to 1.645 for the normal distribution — and makes the test more sensitive to differences in the expected direction.

The price of this advantage is that an effect in the opposite direction cannot be detected. The direction is therefore declared **before** the data are seen; switching to a one-sided test after the result is known inflates the actual risk of a type I error.

E Glossary of Basic Terms

The terms are arranged in alphabetical order. The chapter in which the term is introduced is given in parentheses.

Absolute value — a named number obtained by direct measurement or counting. (Section 1)

Age-standardized rate — the rate a group would have if it had the age composition of the standard population. (Section 1)

Alternative hypothesis H_1 — the statement opposite to the null hypothesis: the observed difference or association is real. (Section 4)

Case fatality — deaths from a disease relative to the number of diagnosed cases of that disease; a measure of its severity. (Section 1)

Central tendency — a summary value representing the observations as a whole: the arithmetic mean, the median, or the mode. (Section 2)

Chi-square χ^2 — a test statistic for the independence of two categorical variables. (Section 5)

Clinical significance — an assessment of the practical importance of an established effect, independent of statistical significance. (Section 4)

Coefficient of determination R^2 — the proportion of the variance of the dependent variable explained by the model. (Section 7)

Collinearity — a strong correlation between two or more predictors in a multiple regression model. (Section 7)

Confidence interval — an interval of values constructed so that, if the procedure were repeated many times, it would contain the true parameter in a pre-specified proportion of cases. (Section 3)

Confounding factor — a third variable associated with both the factor under study and the outcome, which distorts the observed association. (Section 1)

Contingency table (cross-tabulation) — a table distributing the observations simultaneously by two categorical variables. (Section 5)

Cramér's V **coefficient** — a standardized rate of the strength of the association in a contingency table of any size; it takes values from 0 to 1. (Section 5)

Critical value (Z or t multiplier) — the value by which the standard error is multiplied when a confidence interval is constructed, or against which a test statistic is compared. (Section 3)

Cross-tabulation — see *contingency table*.

Crude rate — a rate computed for a whole population without regard to its composition. (Section 1)

Degrees of freedom df — a parameter of the distribution that determines which row of the statistical table is used. (Section 4)

Descriptive statistics — the description of the data in the sample. (Section 2)

Expected frequency E — the frequency that would be observed if the null hypothesis of independence were true. (Section 5)

Extrapolation — applying an established relationship outside the range of the data on which it was built. (Section 6)

Extreme value (outlier) — an observation outside the limits $Q_1 - 1.5 \cdot IQR$ and $Q_3 + 1.5 \cdot IQR$. (Section 2)

Frequency distribution — an ordered presentation of the data together with their frequencies; it may be ungrouped (by individual values) or grouped (by class intervals). (Section 2)

Incidence — the number of new cases arising in the population at risk during a stated period. (Section 1)

Independence, test of — a test checking whether the distribution of one variable is the same across all categories of the other. (Section 5)

Interquartile range IQR — the difference between the third and the first quartile; a measure of dispersion around the median. (Section 2)

Measurement scale — nominal, ordinal, interval, or ratio; it determines which methods are permissible. (Section 2)

Median Me — the value dividing the data, arranged in ascending order, into two equal parts. (Section 2)

Mode M_0 — the most frequently occurring value in the statistical series. (Section 2)

Normal distribution — a symmetric, bell-shaped distribution in which the mean, the median, and the mode coincide. (Section 2)

Null hypothesis H_0 — the statement that no difference or association exists and that the observed discrepancy is due to chance. (Section 4)

Outlier — see *extreme value*.

p-value — the probability, under a true null hypothesis, of obtaining a result as extreme as, or more extreme than, the observed one. (Section 4)

Parameter — a measure of the population; usually unknown. (Section 3)

Pearson correlation coefficient r — a measure of the strength and direction of the linear relationship between two quantitative variables; it takes values from -1 to $+1$. (Section 6)

Person-time — the accumulated time during which the persons under observation were at risk; the denominator of a true rate. (Section 1)

Phi coefficient φ — a measure of the strength and direction of the association in a 2 × 2 table; it is interpreted like a correlation coefficient. (Section 5)

Pooled proportion — the overall proportion calculated from the two groups together and used in the denominator of the Z-test. (Section 4)

Population — all units to which the conclusion of the study refers. (Section 2)

Power — the probability of detecting a difference that truly exists; it equals $1 - \beta$. (Section 4)

Predicted value $\hat{Y}$ — the mean expected value of the dependent variable at a given value of the predictor. (Section 7)

Prevalence — the number of existing cases, new and old, at a given moment. (Section 1)

Proportion — a structural measure showing how a whole is distributed among categories; the sum is 100 %. (Section 1)

Quartiles Q_1, Q_2, Q_3 — the values dividing the ordered statistical series into four equal parts. (Section 2)

Range R — the difference between the largest and the smallest value. (Section 2)

Rate — a frequency measure showing how often an event occurs in the population at risk. (Section 1)

Regression coefficient b — the average change in the dependent variable when the predictor increases by one unit. (Section 7)

Relative frequency (proportion) $\hat{p}$ — the share of observations with a given characteristic, expressed as a percentage. (Section 2)

Relative value — the result of dividing two absolute values. (Section 1)

Residual — the difference between the observed and the predicted value of the dependent variable. (Section 7)

Sample — the subset of units that was actually studied. (Section 2)

Significance level α — the pre-selected acceptable risk of a type I error, usually 0.05. (Section 4)

Standard deviation S — a measure of the dispersion of the individual observations around the arithmetic mean. (Section 2)

Standard error SE — a measure of the precision of a sample estimate; it describes the dispersion of the estimate from sample to sample. (Section 3)

Standard population — the composition adopted as common in standardization; the sum of its weights is 100 %. (Section 1)

Standardization (direct method) — a procedure applied to age-specific rates that removes the influence of differences in the composition of the compared groups and yields age-standardized rates. (Section 1)

Statistic — a measure calculated from the sample. (Section 3)

Statistical significance — a result in which the p-value is below the chosen level α. (Section 4)

Statistical variable — a distinguishing feature, a characteristic by which the units of observation differ. (Section 2)

Test statistic — the number calculated from the data (Z, t, χ^2), whose value is compared with the critical values. (Section 4)

Three-sigma rule — under a normal distribution, approximately 68 %, 95 %, and 99.7 % of the observations fall within $\bar{X} \pm 1S$, $\bar{X} \pm 2S$, and $\bar{X} \pm 3S$, respectively. (Section 2)

Type I error α — rejecting a true null hypothesis, that is, declaring a discovery of something that does not exist. (Section 4)

Type II error β — failing to reject a false null hypothesis, that is, missing a difference that truly exists. (Section 4)

Y-intercept a — the value of the predicted dependent variable at $X = 0$; the intercept on the Y-axis. (Section 7)