

Tomáš Mrkvička, Vladimíra Petrášková,
Michal Houda, Jan Fiala

Probability and Statistics with Worked Examples in MS Excel



SERIA
EDUCATION

Tomáš Mrkvička, Vladimíra Petrášková,
Michal Houda, Jan Fiala

Probability and Statistics with Worked Examples in MS Excel



Probability and Statistics with Worked Examples in MS Excel

Authors: Tomáš Mrkvička
 Vladimíra Petrášková
 Michal Houda
 Jan Fiala

Reviewers: prof. RNDr. Jindřich Klůfa, CSc.
 doc. RNDr. Kateřina Helisová, Ph.D.
 RNDr. Jakub Staněk, Ph.D.

Published by: University of South Bohemia in České Budějovice,
 Faculty of Education

Illustrations: Daniel Kyncl
Cover design: Přemysl Rosa
Typesetting: Přemysl Rosa
 Michal Houda

1st edition

© Tomáš Mrkvička, Vladimíra Petrášková, Michal Houda, Jan Fiala, 2026
© University of South Bohemia in České Budějovice, Faculty of Education,
2026

ISBN 978-80-7694-176-2

Contents

Preface	9
1 Introduction	11
1.1 A Brief Historical Overview	11
1.2 Random Experiment	12
1.3 List of Symbols Used	14
1.4 List of Symbols for Probability Distributions of a Random Variable	15
I Probability Theory	17
2 Random Event	19
2.1 Axiomatic Definition of Probability	19
2.2 Classical Probability Space	25
2.3 Geometric Probability	30
2.4 General Probability Spaces	32
2.5 Exercises	34
3 Conditional probability	41
3.1 Conditional Probability	41
3.2 Independence	46
3.3 Bernoulli process	50
3.4 Exercises	54
4 Total probability, Bayes' formula	61
4.1 Total Probability	61

4.2	Bayes' theorem	63
4.3	Exercises	69
5	Random variable	73
5.1	Introduction of a Random Variable	73
5.2	Distribution of a Random Variable	75
5.3	Cumulative Distribution Function	77
5.4	Discrete Random Variables	81
5.5	Absolutely Continuous Random Variables	84
5.6	Generalization	89
5.7	Exercises	95
6	Characteristics of Random Variables	101
6.1	Expected value of a random variable	101
6.2	Other measures of location	106
6.3	Variance of a Random Variable	110
6.4	Higher-order Moments	113
6.5	Exercises	115
7	Distributions of discrete random variables	119
7.1	Discrete uniform distribution	119
7.2	Degenerate distribution	121
7.3	Bernoulli (Alternative) Distribution	121
7.4	Binomial distribution	123
7.5	Geometric distribution	129
7.6	Negative binomial distribution	133
7.7	Hypergeometric distribution	137
7.8	Poisson distribution	139
7.9	Exercises	146
8	Distributions of continuous random variables	151
8.1	Continuous uniform distribution	151
8.2	Exponential Distribution	153
8.3	Erlang distribution and gamma distribution	159
8.4	General Normal Distribution	164

8.5	Standard (normalized) normal distribution	168
8.6	Lognormal distribution	173
8.7	Pearson Distribution χ^2	176
8.8	Student's t -distribution	178
8.9	Fisher–Snedecor distribution	180
8.10	Exercises	183
9	Random vector	189
9.1	Random Vector	189
9.2	Characteristics of a Random Vector	196
9.3	Some Multivariate Distributions	200
9.3.1	Multinomial distribution	200
9.3.2	General bivariate normal distribution	201
9.4	Independent Random Variables	202
9.5	Exercises	206
10	Function of random variables	213
10.1	Functions of Random Variables	213
10.2	Function of a Random Vector	216
10.3	Convolution	218
10.4	Exercises	225
11	Law of Large Numbers, CLT	229
11.1	Convergence of Random Variables	229
11.2	Law of Large Numbers	231
11.3	Central Limit Theorem	233
11.4	Exercises	239
II	Mathematical Statistics	241
12	Statistical data processing	243
12.1	Basic Concepts	243
12.2	Frequency Distribution and Classification	245
12.3	Characteristics of Location	254
12.4	Measures of Variability (dispersion)	260

12.5 General sample moments	264
12.6 Measures of Symmetry and Concentration	265
12.7 Exercises	268
13 Building Blocks of Statistics	271
13.1 Population and Sample	271
13.2 Distribution of Sample Statistics	275
13.3 Critical Values and Quantiles	276
13.4 Exercises	278
14 Parameter Estimation	281
14.1 Properties of Estimators	282
14.2 Point Estimators of the Mean and Variance	285
14.3 Interval Estimation	285
14.3.1 Confidence Interval for the Mean of a Normal Distribution	286
14.3.2 Determining the Sample Size Required for a Prescribed Width of a Confidence Interval for the Mean	287
14.3.3 Confidence Interval for the Variance σ^2 of a Normal Distribution	288
14.4 Confidence Interval for the Mean Based on the CLT	290
14.5 Exercises	293
15 Parametric Tests	297
15.1 Principles of Hypothesis Testing	297
15.2 One-Sample t -Test	300
15.2.1 Computation of the Two-Sided and One-Sided p -Values of the t -Test	301
15.3 Test for the Variance of a Normal Distribution	303
15.4 Paired t -Test	305
15.5 Two-Sample t -Test	308
15.6 F -Test	310
15.7 Testing the Equality of Means under Unequal Variances	314
15.8 Test for the Mean Based on the CLT	317
15.9 Exercises	319

16 Nonparametric Tests	327
16.1 Sign Test	328
16.2 One-Sample Wilcoxon Test	330
16.3 Two-Sample Wilcoxon Test	333
16.4 Exercises	336
17 Comparison of Multiple Samples	341
17.1 One-Way Analysis of Variance	341
17.1.1 Levene's Test	344
17.2 Kruskal–Wallis Test	348
17.3 Two-Way Analysis of Variance	350
17.4 Friedman's Test	357
17.5 Exercises	361
18 Correlation Analysis	371
18.1 Pearson Sample Correlation Coefficient	371
18.2 Spearman's Rank Correlation Coefficient	374
18.3 Correlation versus Causation	375
18.4 Exercises	377
19 Regression Analysis	381
19.1 Simple Linear Regression	381
19.2 Multiple Linear Regression	388
19.3 Polynomial Regression	395
19.4 Nonlinear Regression	400
19.5 Exercises	403
20 Goodness-of-Fit Tests	409
20.1 Pearson's χ^2 Test	409
20.2 Tests of Normality	411
20.3 Test of the Poisson Distribution	413
20.4 Kolmogorov–Smirnov One-Sample Goodness-of-Fit Test . . .	416
20.5 Exercises	419

21 Contingency Tables	423
21.1 Test of Independence	424
21.2 Test of Homogeneity of Multinomial Distributions	427
21.3 χ^2 Test for 2×2 Contingency Tables	428
21.4 Fisher's Exact Test	429
21.5 McNemar's Test	431
21.6 Test of Symmetry	433
21.7 Exercises	435
 Conclusion	 445
 Index	 447
 Functions and Tools in Microsoft Excel	 455
 Statistical Tables	 463

Preface

Probability-based reasoning, models, and predictions have become a standard component of research practice and find wide-ranging applications across numerous scientific disciplines, including in practical contexts. Probabilistic methods are employed in the analysis of processes influenced by randomness. In situations where the effects of randomness can be taken into account, these methods often enable the derivation of optimal solutions, support rational decision-making, and serve related purposes.

Closely related to probability theory is mathematical statistics, which is used in the evaluation of various tests and experiments. Virtually all modern analysis in both the natural and the social sciences is grounded in statistical inference, that is, in drawing conclusions about a population based on the statistical (scientific) analysis of data subject to variability or uncertainty. In the absence of variability, there is nothing to be analyzed statistically.

The textbook is written in a form accessible to a broader audience, which at times leads to a reduced level of mathematical generality. Individual chapters first present the theoretical exposition, followed by worked examples. Examples solved in the MS Excel environment are indicated by the icon . At the end of each chapter, there are exercises for independent work (always accompanied by final results and, in some cases, partial guidance and recommendations). Chapters marked with an asterisk * contain supplementary material that generalizes and extends the core text.



**This entire book can
fit inside a single
statistician's head**

Authors

Chapter 1

Introduction

1.1 A Brief Historical Overview

Probability theory and mathematical statistics are mathematical disciplines collectively subsumed under the scientific field of *stochastics*¹, which is concerned with the study and modelling of random phenomena. Although an early reference to stochastics can already be found in Plato's work *Philebus*, both disciplines achieved their major development only in the 20th century. The foundations of *probability theory* were laid approximately during the Renaissance, when the most prominent mathematicians of the period addressed, among other topics, problems arising from what was perhaps the most widespread form of entertainment at the time – games of chance involving dice². Girolamo Cardano³ is today regarded as the first Western mathematician to study games of chance. Around 1564, he wrote the book *Liber de Ludo Aleae* (Book on Games of Chance). However, the work was not published until nearly a century later, in 1663, by which time it had already been surpassed by the contributions of Pascal⁴ and Fermat⁵. At that time, a friend of Pascal's, an amateur French



¹From Ancient Greek *stóchos* – to aim at a target, to conjecture.

²In Arabic, dice is *az-zahr*. From there, the term likely entered Spanish as *azar* and French as *hasard*.

³Girolamo Cardano (1501–1576), Italian physicist, mathematician, and gambler.

⁴Blaise Pascal (1623–1662), French mathematician, physicist, inventor, philosopher, writer, and Catholic theologian.

⁵Pierre de Fermat (1601–1665), French mathematician and jurist.

mathematician known among his peers as the Chevalier de Méré⁶, sought to make a fortune using the following pair of dice games.

Example 1.1 (Chevalier de Méré’s Dice Games). The Chevalier de Méré was fond of accepting bets that in four throws of a single die, at least one six would be obtained. He reasoned that his probability of winning in this case was $4 \cdot \frac{1}{6} = \frac{2}{3}$. Although this calculation is incorrect, he was nevertheless able to win in this game when the stakes were even. However, he was not as fortunate in his second dice game. In this case, he threw two dice a total of 24 times and won the pot if at least one double six occurred during these throws. His similarly flawed calculation was as follows: the probability of obtaining a double six is $\frac{1}{36}$; therefore, over 24 throws his probability of winning is $24 \cdot \frac{1}{36} = \frac{2}{3}$, that is, the same as in the first game.

Because in the second game, despite his calculation, he lost a considerable sum of money, he asked the members of the Paris Academy, and above all Pascal, for help in explaining his failure. Reflections on the method of solving this problem (as well as the even more famous, concurrently studied problem of points problem of the division of stakes), embodied in the correspondence between Pascal and Fermat, opened the way to more general considerations of mathematical “expectation”, that is, of probability understood as a scientific discipline. More specific outcomes related to de Méré’s questions include what are now known as Pascal’s triangle, the binomial distribution, and the expected value (or mean value).

1.2 Random Experiment

The central concept of probability theory and mathematical statistics is the *random experiment*. An *experiment* is, in general, any process that generates outcomes (data). We have all encountered experiments during our school years in physics or chemistry classes, for example, rubbing an ebonite rod with a fox tail. Experiments that, upon repetition, lead to one and the same outcome (in this case, the generation of a negative electric charge on the rod) are called *deterministic* experiments. If identical external conditions are maintained, the outcome of a repeated deterministic experiment will always be the same (an electric charge will always be produced on the rod). In chemistry, we may consider another deterministic experiment in which an acid is neutralized by adding a precisely specified amount of a base. This amount is fixed for a given type and concentration of the acid and the

⁶Antoine Gombaud (1607–1684), French writer and salon theorist, who adopted the name Chevalier de Méré after one of his literary characters.

base, and it does not change under repeated trials conducted under otherwise identical conditions.

From a research perspective, experiments whose outcomes vary upon repetition – that is, are subject to a certain degree of uncertainty – are more interesting, even when all other external experimental conditions are kept identical. Such experiments are referred to as *random* experiments. As already mentioned in the introduction, the earliest studied random experiments were dice throws and the elementary components of similar games (coin tosses, card games, lotteries, roulette, etc.). However, modern probability theory does not focus solely on random experiments arising in games of chance (which are, in fact, of relatively limited importance), but provides highly general concepts essential for solving many practical problems and real-life situations. These include, for example, the investigation of the effectiveness of pharmaceutical treatments, the yield of agricultural crops, or quality control in industrial production. In this publication, we will deal exclusively with random experiments.



1.3 List of Symbols Used

$A, B, \dots$	random event
$\{A_i\}$	sequence of events A_i
$\emptyset$	impossible event, empty set
$\bar{A}$	complement of the event A
Ω	certain event, set of all possible outcomes
ω	outcome of a random experiment
$P(\cdot)$	probability of a random event as a variable or a value of a variable
$\mathcal{P}(\Omega)$	power set of the set Ω
$\mathcal{A}$	σ -algebra
f_n	relative frequency
$\mathbb{P}(A)$	probability of the event A in the sense of a probability function
$\mathbb{P}(A B)$	conditional probability of the event A given that the event B has occurred
$X, Y, \dots$	random variables
$p_X(\cdot)$	probability mass function of the discrete random variable X
$\varphi(\cdot)$	probability density of the standard normal distribution
$\Phi(\cdot)$	cumulative distribution function of the standard normal distribution
$f_X(\cdot)$	probability density function of the continuous random variable X
$F_X(\cdot)$	cumulative distribution function of the random variable X
m	moment of a random variable
m_A	number of outcomes favorable to the event A
θ	theoretical population parameter
μ	mean (expected value) as a population parameter
μ^1, μ^2	one-dimensional measure (length), two-dimensional measure (area)
$\mathbb{E}X$	expected value of the random variable X
$\text{var } X$	variance of the random variable X
σ^2	population variance
S^2	sample variance (in the sense of a random variable)
σ	population standard deviation
S	sample standard deviation (in the sense of a random variable)
v	coefficient of variation

$\lfloor x \rfloor$	greatest integer less than or equal to x
$\mathbb{R}$	set of real numbers
$\mathbb{N}$	set of natural numbers
$\mathcal{B}, \mathcal{B}(\mathbb{R})$	Borel σ -algebra (on the set of real numbers)
$\bar{X}$	sample mean (in the sense of a random variable)
$\bar{x}$	arithmetic mean
$\tilde{x}$	median
$\hat{x}$	mode
$\bar{x}_g$	geometric mean
$\bar{x}_h$	harmonic mean
R	range
R	sample correlation coefficient (in the sense of a random variable)
r	computed value of the sample correlation coefficient
d	absolute deviation from the arithmetic mean
$\bar{d}$	mean absolute deviation
$s_{x,y}, \text{cov}(x, y)$	covariance
n	sample size
$\bigcap_{i=1}^n A_i$	intersection of events A_i
$\bigcup_{i=1}^n A_i$	union of events A_i
$\prod_{i=1}^n \mathbb{P}(A_i)$	product of probabilities of events A_i
$\sum_{i=1}^n \mathbb{P}(A_i)$	sum of probabilities of events A_i
$\mathbf{X}$	random vector
$\mathbf{X}^T$	transpose of the vector $\mathbf{X}$ (written as a row/column vector)
B	beta function
Γ	gamma function

1.4 List of Symbols for Probability Distributions of a Random Variable

B	Bernoulli (alternative, zero-one) distribution
Bi	binomial distribution
Dg	degenerate distribution
Erlang	Erlang distribution
Exp	exponential distribution

F	Fisher–Snedecor F -distribution
Γ	gamma distribution
Ge	geometric distribution
HGeom	hypergeometric distribution
χ^2	chi-square distribution
LN	log-normal distribution
N	normal distribution
NB	negative binomial distribution
Po	Poisson distribution
q	studentized range
DU	discrete uniform distribution
t	Student’s t -distribution
U	continuous uniform distribution
Wb	Weibull distribution

Part I

Probability Theory

Chapter 2

Event, random event, probabilities of a random event

2.1 Axiomatic Definition of Probability

In the introductory chapter, we explained what a random experiment represents. To each random experiment, we can assign a set Ω , containing as its elements all possible outcomes of the experiment. For a throw of a die, for example,

$$\Omega = \{\square, \begin{smallmatrix} \square \\ \square \end{smallmatrix}, \begin{smallmatrix} \square & \square \\ \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square & \square \\ \square & \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square & \square & \square \\ \square & \square & \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square & \square & \square & \square \\ \square & \square & \square & \square & \square \end{smallmatrix}\},$$

for a coin toss,

$$\Omega = \{\text{tails}, \text{heads}\},$$

and for two coin tosses,

$$\Omega = \{(\text{tails}, \text{tails}), (\text{heads}, \text{heads}), (\text{heads}, \text{tails}), (\text{tails}, \text{heads})\}.$$

The set Ω may be finite or infinite. However, it only makes sense to consider non-empty sets. A random experiment must be uniquely described by the set Ω , i.e., the set Ω must be exhaustive (exactly one outcome must always occur), and the outcomes must be mutually exclusive (it is impossible for two outcomes to occur simultaneously). The elements $\omega \in \Omega$ are called *elementary events*.

Question 2.1. What do the elementary events look like in both of the Chevalier de Méré's dice games from Example 1.1?

Table 2.1: Notation of basic probability relations and operations

Notation	Probability interpretation
$\omega \in A$	Event A has occurred; the outcome ω of the random experiment is favorable to event A .
$A \subset B$	event A is a subevent of event B ; event B occurs whenever event A occurs.
$B - A$, also $B \setminus A$	The difference of events B and A (in this order) is the event that occurs exactly when event B occurs and at the same time event A does not occur.
$\overline{A} = \Omega - A$	The complement of event A is the event that occurs exactly when event A does not occur.
$A \cup B$	The union of events A and B is the event that occurs whenever at least one of the events A , B occurs.
$A \cap B$	The intersection of events A and B is the event that occurs exactly when both events occur simultaneously.
$A \cap B = \emptyset$	Events A and B are called disjoint; they cannot occur simultaneously.
$A_i \cap A_j = \emptyset, i \neq j,$ $\bigcup_{i=1}^n A_i = C$	The events $A_1, \dots, A_n$ form a partition of event C .

Subsets of Ω are called *random events*. Since the possible outcomes of random experiments in fact form such sets, we employ set-theoretic relations (e.g., subset) and operations (intersection, union, difference, complement) when working with events, supplementing them with probabilistic interpretation and sometimes specific terminology. An overview is given in Table 2.1. The intersection and union of events can also be extended to more events (to a finite or infinite sequence of events). The so-called *certain event* is a random event that occurs in every realization of the experiment. It is denoted by Ω , as it is equivalent to the set of all possible outcomes. The so-called *impossible event* is an event that does not occur in any realization of the experiment. It is denoted by $\emptyset$, since it is equivalent to the empty set.

Not all subsets of Ω must necessarily be random events. Some are not interesting from the point of view of interpretation; others cannot be regarded as events at all because they cannot be “measured”, and it is therefore not possible to build a meaningful theory on them. This led Kolmogorov¹ in 1933 to formulate a general framework of probability based on collections of subsets that are closed under complements and countable unions. These collections are called σ -algebras and contain all subsets of Ω that are meaningful to consider. With a slight simplification, a σ -algebra may be viewed as the collection of all “well-behaved” subsets of Ω .

Definition 2.1 (σ -algebra)

Let $\mathcal{A}$ be a non-empty family (collection) of subsets of a set $\Omega \neq \emptyset$ such that

- a) $\emptyset \in \mathcal{A}$,
- b) if $A \in \mathcal{A}$, then $\bar{A} \in \mathcal{A}$,
- c) if $A_i \in \mathcal{A}$, $i = 1, 2, \dots$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Then $\mathcal{A}$ is called a σ -algebra.

The individual elements of a suitably chosen σ -algebra $\mathcal{A}$ are called *random events*. The pair $(\Omega, \mathcal{A})$ is called an *event space*. Let us now state, without proof, some further properties of random events. Definition 2.1 together with Theorem 2.1 ensure that the results of common set operations on random events are again random events.

Theorem 2.1 (on some properties of random events). *Let Ω be a certain event and let A, B be random events in $\mathcal{A}$.*

- a) *The certain event is a random event: $\Omega \in \mathcal{A}$.*

¹Andrey Nikolaevich Kolmogorov (1903–1987), Russian mathematician, author of the axiomatic definition of probability.

b) *The difference of two random events is a random event:*

if $A, B \in \mathcal{A}$, then $B - A \in \mathcal{A}$.

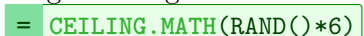
c) *The intersection of countably many random events is a random event:*

if $A_1, A_2, \dots \in \mathcal{A}$, then $\bigcap_{i=1}^{\infty} A_i \in \mathcal{A}$.

Example 2.1. Let $\Omega = \{1, \dots, n\}$. Then the power set $\mathcal{P}(\Omega)$ (i.e. the set of all subsets of Ω) is a σ -algebra and there does not exist a smaller σ -algebra containing all elementary events $\{\omega\}$, $\omega \in \Omega$.

Example 2.2. Let $\Omega = \mathbb{R}$. Then the power set $\mathcal{P}(\mathbb{R})$ is also a σ -algebra, but there also exists a smaller σ -algebra, which we (for reasons mentioned later) prefer. It is mainly the so-called *Borel* σ -algebra, which we define as the smallest σ -algebra containing all open subsets of $\mathbb{R}$. The Borel σ -algebra contains all countable unions of open sets, but also all closed subsets of $\mathbb{R}$. We denote it by $\mathcal{B}(\mathbb{R})$ or simply $\mathcal{B}$.

To each random event, a real number can be assigned representing the “chance” with which the event occurs. This number is called the *probability* of a random event. We have all certainly encountered the concept of probability in everyday life and have used it many times without knowing a precise definition of this concept. Consider, for example, the random experiment of rolling a standard die. In this case, for each outcome $\square, \blacksquare, \dots, \boxtimes$ the probability of occurrence is equal to $\frac{1}{6}$. When tossing a coin, the probability of tails is the same as the probability of heads, i.e. $\frac{1}{2}$. These statements about the probabilities of outcomes of the random experiments of a die roll and a coin toss follow from our experience. We can also look at the concept of probability from a statistical point of view: let us roll a die and examine the relative frequencies $f_n = \frac{m}{n}$ of a chosen event (n is the number of trials, m is the number of trials in which a favorable outcome of the investigated event occurred); with a larger number of trials we find that the relative frequencies of elementary outcomes fluctuate around $\frac{1}{6} = 0.1\bar{6}$. For coin tosses, the relative frequencies fluctuate around $\frac{1}{2} = 0.5$.

Example 2.3 (The relative frequencies of the first dice game of the Chevalier de Méré from Example 1.1 on page 12).  This statistical approach can easily be simulated on a computer. Let us see how the Chevalier de Méré could have dealt with his question regarding the chance of winning in the first dice game if this convenience had already been available in the 17th century. In MS Excel, we use a function for generating random numbers. A single die roll can be simulated using  = `CEILING.MATH(RAND()*6)`

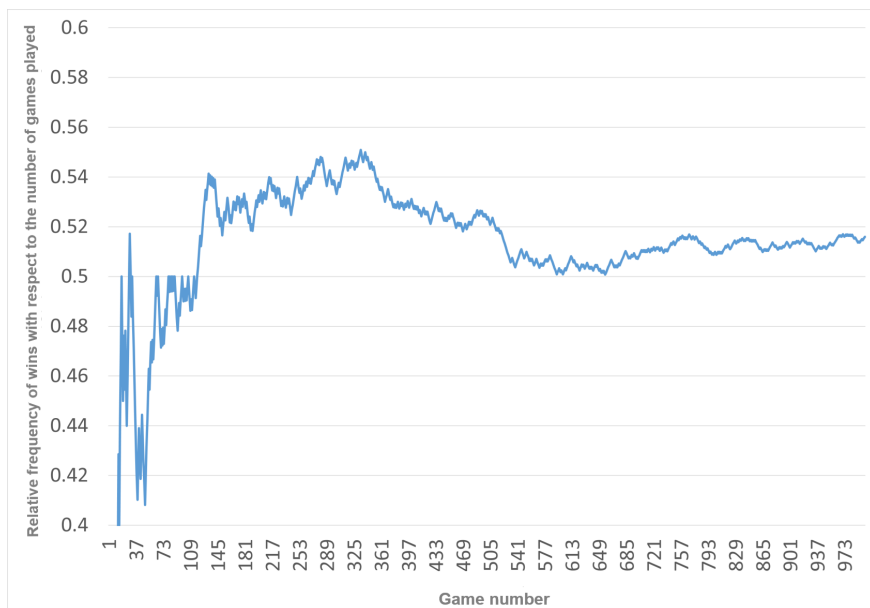
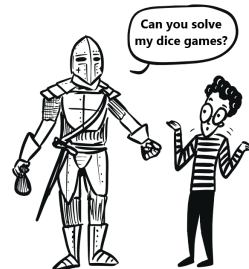


Figure 2.1: The relative frequency of winning in the first dice game of the Chevalier de Méré from Example 2.3

or more simply `= RANDBETWEEN(1,6)`. Each game consists of four such rolls, so we place the formula, for example, into four cells B2 to E2 next to each other. The Chevalier de Méré wins a game and takes the pot if at least one six appears: `= (COUNTIF(B2:E2,6)>0)*1`. Now we copy the entire row as many times as the number of games we want to simulate (for example, one thousand). We sum the number of ones and divide by the number of games, thereby obtaining the desired relative frequency of winning. After a certain number of games (in one simulation illustrated in Figure 2.1, it was after about 430 games), the relative frequency of winning will fluctuate only slightly around the value 0.52.

As we will learn later, the actual probability of winning in this game is 0.518, which is very close to the value we simulated. The Chevalier de Méré actually had the possibility to perform such a simulation even in his time, only carrying out the simulation without the use of a computer (but with a real die) would have taken somewhat longer. Had he realized this, he would certainly have saved a large fortune and would not have attempted to play the second of his games.



Question 2.2. Using simulation, estimate the probability of the Chevalier de Méré's success in his second dice game.

Even from an intuitive perspective, we can deduce that relative frequencies have the following properties:

- 1) $0 \leq f_n(A) \leq 1$ for any random event $A \in \mathcal{A}$,
- 2) if $A_1, A_2, \dots$ a sequence of random events such that for any $i \neq j$, $i, j \in \{1, 2, \dots\}$ it holds that $A_i \cap A_j = \emptyset$, then

$$f_n \left(\bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} f_n(A_i).$$

- 3) $f_n(\emptyset) = 0$ and $f_n(\Omega) = 1$,

If probability is to correspond to the relative frequencies of events, as suggested by example 2.3, then it must exhibit similar properties. This is precisely what Kolmogorov realized, and he derived the following definition of probability:

Definition 2.2 (The probability of a random event)

Let $\Omega \neq \emptyset$ and $\mathcal{A}$ be a σ -algebra defined on Ω , i.e. $(\Omega, \mathcal{A})$ is an event space. Then probability measure (probability, for short) is any real function $\mathbb{P}: \mathcal{A} \rightarrow \mathbb{R}$, which satisfies the following three properties:

- a) (non-negativity) $\mathbb{P}(A) \geq 0$ for any random event $A \in \mathcal{A}$,
- b) (σ -additivity) for any sequence of disjoint events $\{A_n\}_{n=1}^{\infty}$ it holds

$$\mathbb{P} \left(\bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

- c) (normalization) $\mathbb{P}(\Omega) = 1$.

The triple $(\Omega, \mathcal{A}, \mathbb{P})$ is called a probability space. The value $\mathbb{P}(A)$ is called the probability of the event A .

The probability measure defined in this way exhibits a number of further intuitive properties. We summarize them in the following theorem.

Theorem 2.2 (on some properties of the probability of a random event, without proofs). For the probability $\mathbb{P}$ of events $A, B \in \mathcal{A}$ it holds:

- a) $\mathbb{P}(\emptyset) = 0$,



- b) $\mathbb{P}$ is finitely additive, i.e. if $A_1, \dots, A_n \in \mathcal{A}$, $A_i \cap A_j = \emptyset$ for any $i \neq j$,
 $i, j = 1, \dots, n$, then $\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mathbb{P}(A_i)$,
- c) $\mathbb{P}$ is monotone, i.e. if $A, B \in \mathcal{A}$ and $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$,
- d) if $A, B \in \mathcal{A}$ and $A \subset B$, then $\mathbb{P}(B - A) = \mathbb{P}(B) - \mathbb{P}(A)$,
- e) $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$ for any random event $A \in \mathcal{A}$,
- f) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ for any random events $A, B \in \mathcal{A}$,
- g) $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum \mathbb{P}(A_i)$ for any sequence $\{A_i\}$ in $\mathcal{A}$,
- h) if for all n we have $A_n \in \mathcal{A}$, $A_1 \subset A_2 \subset A_3 \dots$, $A = \bigcup_{n=1}^{\infty} A_n$,
then $\mathbb{P}(A) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n)$,
- i) if for all n we have $A_n \in \mathcal{A}$, $A_1 \supset A_2 \supset A_3 \dots$, $A = \bigcap_{n=1}^{\infty} A_n$,
then $\mathbb{P}(A) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n)$.

Property f) can be extended by induction to any finite number of events as follows:

$$\begin{aligned} \mathbb{P}\left(\bigcup_{i=1}^n A_i\right) &= \sum_{i=1}^n \mathbb{P}(A_i) - \sum_{i=1}^{n-1} \sum_{j=i+1}^n \mathbb{P}(A_i \cap A_j) + \\ &+ \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^n \mathbb{P}(A_i \cap A_j \cap A_k) + \dots + (-1)^{n-1} \mathbb{P}\left(\bigcap_{i=1}^n A_i\right). \end{aligned} \quad (2.1)$$

2.2 Classical Probability Space

Kolmogorov's axiomatic definition of probability is non-constructive: it does not specify at all what the event space and the probability should look like; it only states that events must form a σ -algebra according to definition 2.1 and that probability must satisfy the conditions of definition 2.2. Earlier attempts to define probability were mostly constructive in nature – they specified concrete properties of probability, but at the same time lost the necessary generality to be able to describe the random behavior of more complex processes. The oldest studied probability space is the so-called *classical probability*.

Definition 2.3 (Classical probability space²)

A probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is called a classical probability space if

²The given definition is also called the Laplace definition of probability. Its author is Pierre Simon de Laplace (1749–1827), a French mathematician, physicist, astronomer, and politician.

- a) the set Ω is finite with m elements,
 b) for the σ -algebra $\mathcal{A}$ we take the system $\mathcal{P}(\Omega)$ of all subsets of the set Ω ,
 c) all elementary events are equally likely; denoting $p_1, \dots, p_m$ the probabilities of individual elementary events, then

$$p_1 = p_2 = \dots = p_m = \frac{1}{m}.$$

The probability $\mathbb{P}$ of a random event A is then equal to

$$\mathbb{P}(A) = \frac{m_A}{m},$$

where m_A is the number of outcomes favorable to the event A and m is the number of all possible outcomes of the random experiment. The probability defined in this way is called classical probability.

Question 2.3. Verify that classical probability is a probability in the sense of Definition 2.2, i.e. it satisfies all three axioms.

In problems in which classical probability is used, the task is to determine the set of elementary outcomes so that these outcomes are equally likely. Once this is achieved, the computation of probabilities is reduced to the use of combinatorial rules.

Example 2.4. We roll two dice. What is the probability of the event A the sum of the numbers shown by the two dice is less than 5?

Solution. The outcomes of the experiment are ordered pairs of die faces: the first component of the pair corresponds to the result of the roll of the first die and the second component corresponds to the result of the roll of the second die. All possible outcomes are therefore represented by the set

$$\Omega = \{(\square, \square), (\square, \blacksquare), \dots, (\boxtimes, \boxtimes), (\boxtimes, \boxtimes)\}.$$

The number of all possible outcomes is $|\Omega| = m = 36$. The outcomes favorable to the observed event A are $(\square, \square), (\square, \blacksquare), (\blacksquare, \square), (\blacksquare, \blacksquare), (\boxtimes, \square), (\boxtimes, \blacksquare)$, i.e. $m_A = 6$. The sought probability of the event A is therefore equal to

$$\mathbb{P}(A) = \frac{6}{36} = \frac{1}{6}.$$

△

Example 2.5 (The probability of winning in the second dice game of the Chevalier de Méré). What is the probability of winning in the second dice game of the Chevalier de Méré from Example 1.1?

Solution. When rolling two dice simultaneously, we have a total of 36 possible outcomes, as in Example 2.4. Only one of them is $(\boxed{6}, \boxed{6})$, so the probability of rolling two sixes in a single throw is $\frac{1}{36}$. According to the rule for the probability of the complementary event (item e) of Theorem 2.2), the probability of not rolling two sixes is therefore $\frac{35}{36}$. The probability of not rolling two sixes in any of the 24 trials is thus

$$\mathbb{P}(\text{no pair of sixes in 24 trials}) = \left(\frac{35}{36}\right)^{24} \approx 0.5086.$$

The resulting probability that at least one pair of sixes appears in the entire game – in which case the Chevalier de Méré wins and takes the pot – is therefore

$$\mathbb{P}(\text{at least one pair of sixes in 24 trials}) = 1 - \left(\frac{35}{36}\right)^{24} \approx 0.4914.$$

The advantage of the Chevalier de Méré's opponents in this game is therefore about 1.8%. For comparison, the house edge in American roulette (with single and double zero) is 5.26%, and in French roulette (with one green zero) 2.7%. Thus, the Chevalier de Méré did not lose his money as quickly as in a casino, but with a comparable degree of certainty. $\triangle$



Question 2.4. Using the procedure from Example 2.5, compute the probability of the Chevalier de Méré's success in the first dice game, which we estimated by simulation in Example 2.3.

Example 2.6. In a deck there are 32 playing cards, including four aces. We draw one card at random twice in succession, after the first draw, the drawn card either

- is returned to the deck, or
- is not returned to the deck.

Determine the probability of the event A “at least one of the drawn cards is an ace”.



Solution. The outcomes of the experiment are again ordered pairs in both cases: the first component of the pair corresponds to the card drawn in the first draw and the second component to the card drawn in the second draw. The procedure then differs.

- a) In the first draw, a card can be drawn in 32 ways. Since the drawn card is returned to the deck, there are also 32 possibilities in the second draw. The total number of all possible cases is therefore $m = 32^2 = 1024$. The favorable cases correspond to draws with ordered pairs of the type (any card, ace), (ace, any card), (ace, ace). The number of favorable cases is $m_A = 28 \cdot 4 + 4 \cdot 28 + 4 \cdot 4 = 240$. The sought probability is $\mathbb{P}(A) = \frac{240}{1024} = \frac{15}{64} \approx 0.234$.
- b) The number of possible cases is, since after the first draw the card is not returned, $m = 32 \cdot 31 = 992$. The favorable cases again correspond to draws (any card, ace), (ace, any card), (ace, ace). The number of favorable cases is now $m_A = 28 \cdot 4 + 4 \cdot 28 + 4 \cdot 3 = 236$. The sought probability is $\mathbb{P}(A) = \frac{236}{992} = \frac{59}{248} \approx 0.238$.

In this second case, it is also possible to proceed by choosing a representation in which both cards are drawn simultaneously. The outcomes of the experiment are all unordered pairs of possible draws, of which there are $\binom{32}{2}$. The number of cases in which we draw at least one ace can be computed using the complementary event, i.e. using the number of cases in which we draw no ace. The number of such cases is $\binom{28}{2}$.

Altogether, $\mathbb{P}(A) = \frac{\binom{32}{2} - \binom{28}{2}}{\binom{32}{2}} = \frac{59}{248} \approx 0.238$.

△

Example 2.7. We roll six dice once, assigning to each die exactly one of the numbers $1, \dots, 6$. What is the probability that on at least one die the number of dots matches the number with which we labeled the die?



Solution. Let A_i be the event that on the i -th die the number of dots matches the number with which the die is labeled, and let the event A be the event “on at least one die the number of dots matches the number with which the die is labeled”. Then $A = \bigcup_{i=1}^6 A_i$. Since the events $A_1, \dots, A_6$ are not disjoint, according to equation (2.1) on page 25

$$\begin{aligned} \mathbb{P}\left(\bigcup_{i=1}^6 A_i\right) &= \sum_{i=1}^6 \mathbb{P}(A_i) - \sum_{i=1}^5 \sum_{j=i+1}^6 \mathbb{P}(A_i \cap A_j) + \\ &+ \sum_{i=1}^4 \sum_{j=i+1}^5 \sum_{k=j+1}^6 \mathbb{P}(A_i \cap A_j \cap A_k) + \dots + (-1)^5 \mathbb{P}\left(\bigcap_{i=1}^6 A_i\right). \quad (2.2) \end{aligned}$$

Let us determine the probabilities occurring in (2.2). A roll of six dice corresponds to an ordered 6-tuple whose components may repeat. The

number of possible outcomes is therefore $6^6 = 46\,656$. For the event A_i , $i = 1, 2, \dots, 6$, the favorable outcomes are those in which, on the i -th die, the number of dots matches its label, while any result may occur on the other dice. There are $6^5 = 7\,776$ such outcomes. Hence

$$\mathbb{P}(A_i) = \frac{6^5}{6^6} = \frac{1}{6}, \quad i = 1, 2, \dots, 6.$$

Now we will compute the probabilities $\mathbb{P}(A_i \cap A_j)$, where $i, j = 1, 2, \dots, 6$, $i \neq j$. The number of possible outcomes is again $6^6 = 46\,656$. The favorable outcomes are those in which, on the i -th and j -th die, the number of dots matches their labels; any result may occur on the other dice. The number of favorable outcomes is therefore $6^4 = 1\,296$, and

$$\mathbb{P}(A_i \cap A_j) = \frac{6^4}{6^6} = \frac{1}{6^2} \approx 0.028, \quad i, j = 1, 2, \dots, 6, \quad i \neq j.$$

Analogously,

$$\mathbb{P}(A_i \cap A_j \cap A_k) = \frac{6^3}{6^6} = \frac{1}{6^3} \approx 0.0046, \quad i < j < k,$$

$$\vdots$$

$$\mathbb{P}\left(\bigcap_{i=1}^6 A_i\right) = \frac{1}{6^6}.$$

The sought probability of the event A is

$$\mathbb{P}(A) = 6 \cdot \frac{1}{6} - \binom{6}{2} \cdot \frac{1}{6^2} + \dots + (-1)^5 \cdot \frac{1}{6^6} \approx 0.665.$$

The described procedure is rather complicated. In this case, it is simpler to use the complementary event $\overline{A}$ (on no die does the number of dots match its label). The number of favorable outcomes for $\overline{A}$ is 5^6 , hence

$$\mathbb{P}(A) = 1 - \mathbb{P}(\overline{A}) = 1 - \frac{5^6}{6^6} \approx 0.665.$$

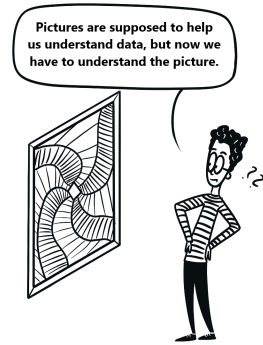
△

Example 2.8. At the beginning of a Poker game, each player receives five cards. What is the probability that a player is dealt two kings and at the same time two aces?

Solution. We consider cards with four different “suits”, from each suit 13 cards with values 2–10, jack, queen, king and ace. The total number of possibilities of how five-card hands can be formed is $\binom{52}{5} = 2\,598\,960$. The number of favorable outcomes for the considered event is $\binom{4}{2} \cdot \binom{4}{2} \cdot \binom{44}{1} = 1\,584$. The resulting probability is $\mathbb{P} = \frac{1\,584}{2\,598\,960} \approx 0.000609$. △

2.3 Geometric Probability

Sometimes, elementary outcomes can be represented using some geometric object (a line segment, rectangle, sphere, etc.), in general by a subset of points from $\mathbb{R}^d$. If all points have the same “probability weight”, i.e. no elementary outcome is preferred over another in terms of chance of occurrence, we speak of the so-called model of geometric probability.



Definition 2.4 (Geometric probability)

A probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is called geometric probability if

- a) $\Omega \subset \mathbb{R}^d$,
- b) $\mathcal{A} = \mathcal{B}(\Omega)$ is the Borel σ -algebra on Ω (i.e. the smallest σ -algebra containing all open subsets of Ω),
- c) the probability of an event $A \in \mathcal{A}$ is directly proportional to the d -dimensional Lebesgue³ measure μ^d of the set A ⁴:

$$\mathbb{P}(A) = \frac{\mu^d(A)}{\mu^d(\Omega)}.$$

Example 2.9. Buses arrive at the stop regularly at 10-minute intervals. A student arrives at the stop at a random time. What is the probability that they will wait longer than 5 minutes?



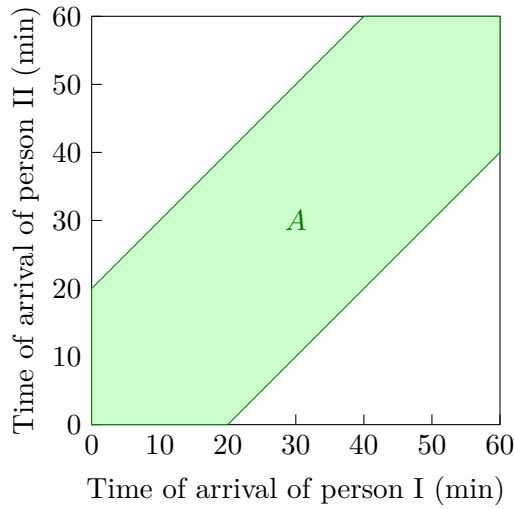
Solution. The outcome of the experiment is the waiting time of the student for the bus. The student can wait from 0 to 10 minutes; set of elementary events is therefore the interval $\Omega = [0, 10)$. The event A (waiting time longer than 5 minutes) is the open interval $(5, 10)$. The sought probability is thus:

$$\mathbb{P}(A) = \frac{\mu^1((5, 10))}{\mu^1([0, 10))} = \frac{5}{10} = \frac{1}{2} = 0.5.$$

△

³Henri Léon Lebesgue (1875–1941), a French mathematician.

⁴For our purposes, it suffices to think of $\mu^1(A)$ as the length of the set A , $\mu^2(A)$ as the area of A , and $\mu^3(A)$ as the volume of A .

Figure 2.2: Meeting region (event A) for Example 2.10

Example 2.10. Two persons (I, II) arrive at a meeting place between 12 and 1 p.m. The times of arrival of both persons are random and independent. Whoever arrives at the meeting place waits for 20 minutes and, if the other does not arrive, leaves. What is the probability that the persons meet?

Solution. The elementary event is here $\omega = (x, y)$, where x denotes the time of arrival of person I and y the time of arrival of person II. The sample space is therefore $\Omega = [0, 60] \times [0, 60]$. Let A denote the event that “the persons meet”, then we have

$$A = \{(x, y) \in \Omega : |x - y| \leq 20\}$$

(see Figure 2.2). The complementary event is

$$\bar{A} = \{(x, y) \in \Omega : y < x - 20 \text{ or } y > x + 20\}.$$

The area of the set $\bar{A}$ consisting of two triangles is $\mu^2(\bar{A}) = 40 \cdot 40 = 1\,600$. The area of Ω is $\mu^2(\Omega) = 3\,600$, which together gives:

$$\mathbb{P}(A) = 1 - \mathbb{P}(\bar{A}) = 1 - \frac{\mu^2(\bar{A})}{\mu^2(\Omega)} = 1 - \frac{4}{9} = \frac{5}{9} = 0.\bar{5}.$$

△

2.4 General Probability Spaces

Classical probability cannot represent systems in which there is an infinite number of elementary events (e.g. random events representing numbers of occurrences). Similarly, neither classical nor geometric probability can handle cases in which elementary events do not have equal weights (e.g. the probability of the birth of a boy is slightly higher than the probability of the birth of a girl). Therefore, we introduce more general probability spaces, in which these requirements are relaxed. These spaces will find use later when introducing the concept of a random variable.

A discrete probability space is a generalization of classical probability, which may have countably many elementary events and in which the probabilities of individual elementary events may differ.

Definition 2.5 (Discrete probability space)

A probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is called a discrete probability space if

- a) $\Omega = \{\omega_1, \omega_2, \dots\}$ has at most countably many elements,
- b) $\mathcal{A} = \mathcal{P}(A)$ is the power set (i.e. the set of all subsets),
- c) to each elementary event $\omega_i \in \Omega$, $i \in \mathbb{N}$ a probability $p_i = \mathbb{P}(\omega_i)$ is assigned such that

$$\sum_{i=1}^{\infty} p_i = 1.$$

The probability of any event $A \in \mathcal{A}$ is then uniquely given by

$$\mathbb{P}(A) = \sum_{\omega_i \in A} p_i. \quad (2.3)$$

Similarly, the continuous case is a generalization of one-dimensional geometric probability. In Chapter 9 on random vectors we introduce multidimensional geometric probability.

Definition 2.6 (One-dimensional continuous probability space)

A probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is called a one-dimensional continuous probability space if

- a) $\Omega = \mathbb{R}$,
- b) $\mathcal{A} = \mathcal{B}(\mathbb{R})$ is the Borel σ -algebra⁵ on $\mathbb{R}$,
- c) there exists a function $f: \mathbb{R} \rightarrow [0, +\infty]$ such that

$$\int_{\mathbb{R}} f(x) dx = 1.$$

⁵See Example 2.2 on page 22.

The probability of any event $A \in \mathcal{A}$ is uniquely given by

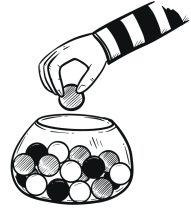
$$\mathbb{P}(A) = \int_A f(x) \, dx. \tag{2.4}$$

2.5 Exercises

2.1 In an urn there are balls of three colors (otherwise identical physical properties). Let events A , B , C successively denote that “a randomly selected ball is black”, “the ball is red”, “the ball is white”.

Determine the meaning of the following events:

- a) $\bar{A} \cap \bar{B}$,
- b) $(A \cup C) \cap B$,
- c) $(A \cap C) \cup B$,
- d) $A \cup B \cup C$.



2.2 We roll a die once. The event A means that “the number obtained is less than 4”, and the event B that “the number obtained is less than 5”. Express the following events using A , B , $\bar{A}$, $\bar{B}$:

- a) “the number obtained is 4”,
- b) “the number obtained is greater than 3”,
- c) “the number obtained is less than or equal to 6”.

2.3 In a biology exam, a randomly selected student is given three questions. Let the event A mean that “the student answers the first question correctly”, event B that “the student answers the second question correctly”, and event C that “the student answers the third question correctly”. Express, using the events A , B , C , $\bar{A}$, $\bar{B}$, $\bar{C}$, that the randomly selected student:

- a) “answers correctly only the first question”,
- b) “answers correctly at least two questions”,
- c) “answers correctly exactly one question”,
- d) “answers correctly at most two questions”.

2.4 We roll three dice. What is the probability that the sum of the numbers of dots on the top faces of these three dice is equal to 5?

2.5 There are 10 tickets in an urn. Six of them have two-digit numbers and four have one-digit numbers. We draw one ticket twice and the tickets are not returned to the urn. The drawn tickets are placed next to each other in the order in which they were drawn. What is the probability that the resulting number is four-digit?

2.6 Among fourteen people, eight have a university education, the others have a secondary education. We randomly select four people. What is the probability that

- a) “all four have secondary education” (event A),
- b) “exactly one has a university education” (event B),
- c) “at least one has a university education” (event C).

2.7 The Smith couple wants to have four children and knows that the probability of the birth of a boy is, according to long-term statistics, about 0.514, and the birth of a girl 0.486. However, Mr. Smith’s mother-in-law claims that with the highest probability the Smiths will have two sons and two daughters. Decide whether the mother-in-law is right. Justify your answer.

2.8 Player I wins if, in six throws of a die, he obtains at least one six. Player II wins if he obtains at least two sixes in twelve throws of a die. Which player has a higher probability of winning?⁶

2.9 Four deer were captured from a population of N deer, marked and released back. To verify that the marked deer are randomly distributed in the population, we captured five deer after a sufficiently long time from the same population. What is the probability that “exactly one marked deer will be recaptured” (event A), if $N = 8$ and $N = 30$? Construct a scatter plot of the dependence of the probability $\mathbb{P}(A)$ on the size of the deer population ($N = 8, 9, \dots, 60$).



2.10 We roll two dice n times in succession. What is the probability that “in at least one throw the sum 12 appears”?

2.11 A student draws three questions from 30 in a mathematics exam. Among the 30 questions, there are 10 from algebra, 15 from mathematical analysis, and 5 from geometry. What is the probability that “the student draws at least two questions from the same discipline”?

⁶This problem was first formulated by Samuel Pepys (1633–1703), a senior official of the English Navy and politician. The problem was solved in 1693 by Sir Isaac Newton (1643–1727).

2.12 From an urn containing n balls, we draw several balls at once. Determine the probability that we have drawn an even number of balls. Assume that all possible draws of balls (distinguishable) are equally likely, i.e. we choose randomly among all possible selections of balls.

2.13 In an urn there are a tickets with even numbers and b tickets with odd numbers. In one draw we draw k tickets. The drawn tickets are placed into a second empty urn, and from it we then draw one ticket. What is the probability that “this ticket has an even number”?

2.14 In a class there are n students.

- a) What is the probability that in the class there is at least one pair who share the same birthday?
- b) What is the probability that in the class there exists at least one classmate who has a birthday on the same day as the class teacher?
- c) What is the smallest n such that the probability that two students have the same birthday is greater than $\frac{1}{2}$?



In the problems, do not consider leap years, the presence of twins, and assume that births are uniformly distributed throughout the year.

2.15 Two steamships that use the same single dock may arrive at any time during 24 hours. Their arrivals are independent. The first steamship occupies the dock for one hour, the second for two hours. What is the probability that neither steamship will have to wait for the dock to become available?

2.16 Let $x, y \in (0, 1)$ be randomly chosen numbers. What is the probability that their sum is less than 1 and their product is greater than 0.09?

2.17 On a line segment of length l , two points X and Y are placed at random, which divide the segment into three parts. With what probability can a triangle be constructed from the three resulting segments?

Solutions to the Exercises

2.1

- a) “the selected ball is not black and at the same time is not red”, i.e. “the selected ball is white”,

- b) “the selected ball is black or white and at the same time red”, an impossible event,
 c) “the selected ball is black and at the same time white, or red”, i.e. “the selected ball is red”,
 d) “the selected ball is black, or red, or white”, a certain event.

2.2 a) $B \cap \bar{A} = B \setminus A$, b) $\bar{A}$, c) $B \cup \bar{B}$ or $A \cup \bar{A}$.

2.3

- a) $A \cap \bar{B} \cap \bar{C}$,
 b) $(A \cap B) \cup (B \cap C) \cup (A \cap C)$,
 c) $(A \cap \bar{B} \cap \bar{C}) \cup (\bar{A} \cap B \cap \bar{C}) \cup (\bar{A} \cap \bar{B} \cap C)$,
 d) $(\bar{A} \cap \bar{B} \cap \bar{C}) \cup (A \cap \bar{B} \cap \bar{C}) \cup (\bar{A} \cap B \cap \bar{C}) \cup (\bar{A} \cap \bar{B} \cap C) \cup (A \cap B \cap \bar{C}) \cup (A \cap \bar{B} \cap C) \cup (\bar{A} \cap B \cap C) = (A \cap B \cap C)$.

2.4 $\frac{1}{36} = 0.02\bar{7}$

2.5 $\frac{1}{3} = 0.\bar{3}$

2.6 a) $\mathbb{P}(A) \approx 0.0150$, b) $\mathbb{P}(B) \approx 0.1598$, c) $\mathbb{P}(C) \approx 0.985$.

2.7 The probability of the birth of two sons and two daughters is, after rounding, 0.374, which is the highest probability that can be achieved among the various combinations of numbers of sons and daughters. The mother-in-law is therefore right.

2.8 The probability of Player I winning is about 0.665, while the probability of Player II winning is about 0.619. Player I has a higher chance of winning.

2.9 For $N = 8$, $\mathbb{P}(A) \approx 0.071$, and for $N = 30$, $\mathbb{P}(A) \approx 0.42$. The scatter plot is shown in Figure 2.3.

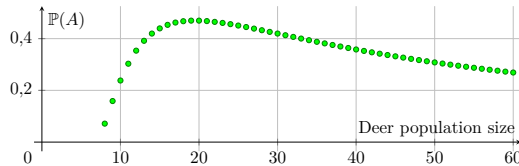


Figure 2.3: Point plot of the dependence of the probability $\mathbb{P}(A)$ on the size of the deer population N

$$\mathbf{2.10} \quad 1 - \left(\frac{35}{36}\right)^n$$

$$\mathbf{2.11} \quad 0.815$$

$$\mathbf{2.12} \quad \frac{2^{n-1} - 1}{2^n - 1}$$

$$\mathbf{2.13} \quad \frac{a}{a+b}$$

2.14

- a) $\mathbb{P}(A) = 1 - 1 \cdot \left(1 - \frac{1}{365}\right) \cdot \left(1 - \frac{2}{365}\right) \cdots \left(1 - \frac{n-1}{365}\right) =$
 $= 1 - \frac{365 \cdot 364 \cdots (365 - n + 1)}{365^n}.$
 b) $\mathbb{P}(B) = 1 - \left(\frac{364}{365}\right)^n.$
 c) $\mathbb{P}(A) > \frac{1}{2} \Leftrightarrow n > 23$, i.e. at least 24 students.

$$\mathbf{2.15} \quad \mathbb{P} \approx 0.878.$$

$$\mathbf{2.16} \quad \mathbb{P}(A) = \frac{\mu^2(A)}{\mu^2(\Omega)} = \frac{\int_{0.1}^{0.9} \left(1 - x - \frac{0.09}{x}\right) dx}{1} \approx 0.202.$$

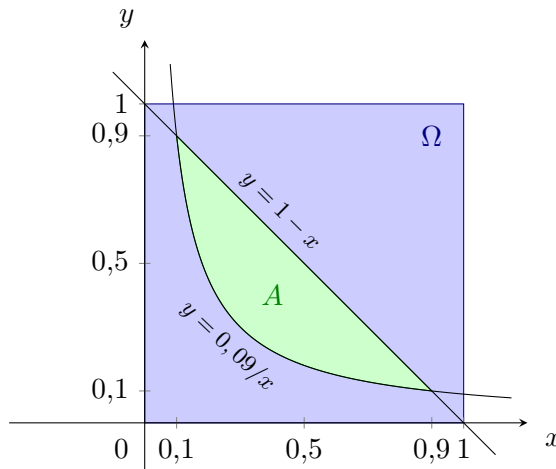


Figure 2.4: The set of all possible outcomes and the outcomes favorable to the event A in Problem 2.16

$$\mathbf{2.17} \quad \mathbb{P}(A) = \frac{\mu^2(A)}{\mu^2(\Omega)} = \frac{l^2/4}{l^2} = \frac{1}{4}.$$

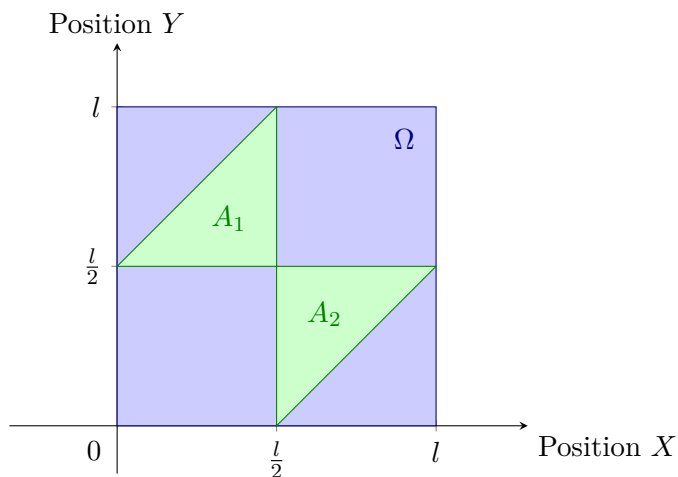


Figure 2.5: The set of all possible outcomes and the outcomes favorable to the event $A = A_1 \cup A_2$ in problem 2.17

Chapter 3

Conditional probability, independence

3.1 Conditional Probability

The probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is a mathematical representation of all possible outcomes of a random experiment together with information about the probability of their occurrence. It is common that, over time, some update of information about this experiment may occur, which may also lead to a change in the probability that a certain event A occurs. Probability theory provides a tool to describe this situation – it is called *conditional probability*.

Let us assume that we are studying the probability of the occurrence of a random event A in the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Suppose that newly available additional information indicates that another event B has occurred with non-zero probability $\mathbb{P}(B) > 0$. We can illustrate the situation with the following example.

Example 3.1. Let us successively roll two dice and consider the random event $A =$ “the sum of the numbers on both dice is at most 5”. The sample space is the same as in Example 2.4:

$$\Omega = \{(\square, \square), (\square, \blacksquare), \dots, (\boxplus, \boxplus), (\boxplus, \boxplus)\},$$

and the probability of the event A before the experiment begins is

$$\mathbb{P}(A) = \frac{10}{36} = 0.2\bar{7},$$

since 10 out of the 36 possible outcomes are favorable to the event A (the reader may enumerate them independently).

Let us now examine the situation after the first of two throws. Suppose that in the first throw the die shows $\blacksquare$. We represent this information over the space $(\Omega, \mathcal{A}, \mathbb{P})$ as the event

$$B = \{(\blacksquare, \blacksquare), (\blacksquare, \blacksquare), \dots, (\blacksquare, \blacksquare)\}.$$

The probability of occurrence of this event is clearly $\mathbb{P}(B) = 6/36$. What is now, in this new situation, the probability of occurrence of the event A ? Out of the ten ordered pairs in the set A , four pairs have $\blacksquare$ as the first element, forming a subset of Ω :

$$\{(\blacksquare, \blacksquare), (\blacksquare, \blacksquare), (\blacksquare, \blacksquare), (\blacksquare, \blacksquare)\}.$$

Note that this set is formally the intersection of the sets A and B , i.e., $A \cap B$, and its original (computed before the first throw) probability is $4/36$. However, since we already know that the event B has occurred, it is desirable to update the probability of the event A so that it takes into account the new information about the realization of the event B . Its new value should be $4/6$ – for the second throw, there are six possible outcomes, four of which are favorable to the event A . Using the original probability information, we can obtain it as the ratio

$$\frac{4/36}{6/36} = \frac{2}{3} = 0.\bar{3},$$

that is, as the ratio of the probability of the intersection of the events $A \cap B$ to the probability of the condition B . We thus see that the probability of the occurrence of the event A has increased after the first of the two throws from the original approximately 28% to about 33%.

In this and similar cases, we do not in fact speak of the “updated” probability of the event A , but rather of the probability of the event A conditional on the event B , also referred to as the conditional probability of the event A given B , which we denote by $\mathbb{P}(A|B)$.

Definition 3.1 (Conditional probability)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let A, B be events such that $\mathbb{P}(B) > 0$. The conditional probability of the event A given that the event B has occurred is defined by the relation

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \quad (3.1)$$

The following theorem states that the conditional probability has all the fundamental properties of (unconditional) probability according to Definition 2.2, and is therefore itself a probability.

Theorem 3.1. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let B be an event such that $\mathbb{P}(B) > 0$. Then, for any event $A \in \mathcal{A}$, we have:

- a) (non-negativity) $\mathbb{P}(A|B) \geq 0$,
- b) (σ -additivity) for every sequence $\{A_n\}_{n=1}^{\infty}$ of disjoint events, it holds that

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n | B\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n | B),$$

- c) (normalization) $\mathbb{P}(\Omega | B) = 1$.

Proof.

- a) $\mathbb{P}(A|B) \geq 0$, because, according to (3.1), it is the ratio of a non-negative and a positive number.
- b) Since $A_1, A_2, \dots$ are disjoint, the sets $A_1 \cap B, A_2 \cap B, \dots$ are also disjoint. From axiom b) of Definition 2.2 and from Definition 3.1, it follows that

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n | B\right) = \frac{\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n \cap B\right)}{\mathbb{P}(B)} = \frac{\sum_{n=1}^{\infty} \mathbb{P}(A_n \cap B)}{\mathbb{P}(B)} = \sum_{n=1}^{\infty} \mathbb{P}(A_n | B).$$

- c) From Definition 3.1, it follows that

$$\mathbb{P}(\Omega | B) = \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1.$$

□

Very often, in application of Definition 3.1, both the conditional and the unconditional probabilities are known, which allows us, using relation (3.1), to derive the probability that two events occur simultaneously:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B) \cdot \mathbb{P}(B). \quad (3.2)$$

Example 3.2 (on defective fuses). Consider a box containing 20 thermal fuses, of which we know that, by mistake, it contains 5 defective ones. We need to repair two devices by replacing their fuses. We therefore take two fuses from the box one after the other (without replacement). What is the probability that both of them are defective?



Solution. Let us denote

A = “the second fuse is defective”

B = “the first fuse is defective”

$A \cap B$ = “both fuses are defective”

We know the conditional probability $\mathbb{P}(A|B)$ and the unconditional probability of the event B ,

$$\mathbb{P}(B) = \frac{5}{20}, \quad \mathbb{P}(A|B) = \frac{4}{19},$$

both of which follow from a simple combinatorial argument. Substituting into the relation (3.2), we obtain

$$\mathbb{P}(A \cap B) = \frac{4}{19} \cdot \frac{5}{20} = \frac{1}{19}.$$

△

This simple rule can be generalized to more events, as formulated in Theorem 3.2.

Theorem 3.2 (on multiplication of probabilities, the chain rule). *If, for any sequence of events $A_1, A_2, \dots, A_n$, the condition $\mathbb{P}(A_1 \cap A_2 \cap \dots \cap A_{n-1}) > 0$ holds, then*

$$\begin{aligned} \mathbb{P}\left(\bigcap_{i=1}^n A_i\right) &= \mathbb{P}(A_1) \cdot \mathbb{P}(A_2|A_1) \cdot \mathbb{P}(A_3|A_1 \cap A_2) \cdots \\ &\quad \cdots \mathbb{P}(A_n|A_1 \cap A_2 \cap \dots \cap A_{n-1}). \end{aligned} \quad (3.3)$$

Proof. By the monotonicity of probability and the assumption of the theorem, we have

$$\mathbb{P}(A_1) \geq \mathbb{P}(A_1 \cap A_2) \geq \dots \geq \mathbb{P}(A_1 \cap \dots \cap A_{n-1}) > 0,$$

and therefore all conditional probabilities appearing in the statement of the theorem are well-defined. By repeatedly applying the formula (3.2), we

obtain

$$\begin{aligned}
 \mathbb{P}\left(\bigcap_{i=1}^{n-1} A_i \cap A_n\right) &= \mathbb{P}\left(\bigcap_{i=1}^{n-1} A_i\right) \mathbb{P}\left(A_n \mid \bigcap_{i=1}^{n-1} A_i\right) = \\
 &= \mathbb{P}\left(\bigcap_{i=1}^{n-2} A_i\right) \mathbb{P}\left(A_{n-1} \mid \bigcap_{i=1}^{n-2} A_i\right) \mathbb{P}\left(A_n \mid \bigcap_{i=1}^{n-1} A_i\right) = \\
 &= \dots = \\
 &= \mathbb{P}(A_1) \mathbb{P}(A_2 \mid A_1) \mathbb{P}(A_3 \mid A_1 \cap A_2) \cdots \mathbb{P}\left(A_n \mid \bigcap_{i=1}^{n-1} A_i\right).
 \end{aligned}$$

□

Example 3.3. From an urn containing n white and n black balls, we randomly draw two balls at a time, n times, without replacing the balls drawn. Determine the probability that in each draw we obtain one white and one black ball.

Solution. Let A_i denote the event “in the i -th draw we obtain one white ball and one black ball”, $i = 1, \dots, n$. We seek the probability of the event

$$A_1 \cap A_2 \cap \dots \cap A_n = \bigcap_{i=1}^n A_i;$$

this is given by formula (3.3) in Theorem 3.2. Let us determine the probabilities appearing in equation (3.3). First, we compute $\mathbb{P}(A_1)$: the number of possible cases is $\binom{2n}{2}$, and the number of favorable cases for the event A_1 is $n \cdot n$, from which it follows that

$$\mathbb{P}(A_1) = \frac{n \cdot n}{\binom{2n}{2}}.$$

When determining the probability $\mathbb{P}(A_2 \mid A_1)$, we proceed from the fact that the urn now contains only $n - 1$ white and $n - 1$ black balls, hence

$$\mathbb{P}(A_2 \mid A_1) = \frac{(n-1)(n-1)}{\binom{2n-2}{2}}.$$

Analogously

$$\begin{aligned}\mathbb{P}(A_3|A_1 \cap A_2) &= \frac{(n-2)(n-2)}{\binom{2n-4}{2}} \\ &\vdots \\ \mathbb{P}\left(A_n \middle| \bigcap_{i=1}^{n-1} A_i\right) &= 1.\end{aligned}$$

The sought probability is therefore equal to

$$\mathbb{P}\left(\bigcap_{i=1}^n A_i\right) = \frac{n \cdot n}{\binom{2n}{2}} \cdot \frac{(n-1)(n-1)}{\binom{2n-2}{2}} \cdot \frac{(n-2)(n-2)}{\binom{2n-4}{2}} \cdots 1 = \frac{n!n!2^n}{(2n)!}.$$

△

3.2 Independence

Let us now consider two events A and B . If the occurrence of the event B does not affect the probability of the event A in any way, that is, if

$$\mathbb{P}(A|B) = \mathbb{P}(A), \tag{3.4}$$

we can say that the event A is independent of the event B . Substituting (3.4) into the formula (3.2), we obtain that in this case it must hold that

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B). \tag{3.5}$$

By the reverse procedure, we can derive from formula (3.5) that relation (3.4) holds symmetrically also for the event B :

$$\mathbb{P}(B|A) = \mathbb{P}(B). \tag{3.6}$$

However, both formulas (3.4) and (3.6) require, due to the definition of conditional probability, that

$$\mathbb{P}(A) > 0 \quad \text{and} \quad \mathbb{P}(B) > 0.$$

The formula (3.5), however, works well even without this requirement and, moreover, reflects well the symmetry of the concept of independence; therefore, it is nowadays preferred when defining the independence of two random events.

Definition 3.2 (Independent events)

Events A and B are (mutually¹) independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B). \quad (3.7)$$

Example 3.4. What is the probability that in two throws of a die an even number appears in the first throw (event A) and a prime number appears in the second throw (event B)?

Solution. The sample space Ω consists of all possible ordered pairs formed successively from the results of the two throws of a die. The probabilities of the events A and B are the same, $\mathbb{P}(A) = \mathbb{P}(B) = \frac{1}{2}$ (the prime numbers are 2, 3 and 5, and the number 1 is not prime by definition). Since the outcome of the second throw is not affected by the outcome of the previous throw, the two events under consideration are independent, and therefore the probability of their simultaneous occurrence is

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = 0.25.$$

△

The concept of independence can also be extended to a collection of events, albeit in several ways.

Definition 3.3 (Pairwise independent events)

Let $A_1, \dots, A_n$ be events. We say that they are pairwise independent if the events A_i, A_j are independent for all $i, j = 1, \dots, n$, $i \neq j$.

Example 3.5. When tossing two coins, we consider the following events: A_1 is the event “heads appears on the first coin”, A_2 is the event “tails appears on the second coin”, and A_3 is the event “the same face appears on both coins”. Determine whether the given events are pairwise independent.



Solution.

$$\begin{array}{lll} \mathbb{P}(A_1) = \frac{1}{2}, & \mathbb{P}(A_2) = \frac{1}{2}, & \mathbb{P}(A_3) = \frac{1}{2}, \\ \mathbb{P}(A_1 \cap A_2) = \frac{1}{4}, & \mathbb{P}(A_1 \cap A_3) = \frac{1}{4}, & \mathbb{P}(A_2 \cap A_3) = \frac{1}{4}, \\ \mathbb{P}(A_1) \cdot \mathbb{P}(A_2) = \frac{1}{4}, & \mathbb{P}(A_1) \cdot \mathbb{P}(A_3) = \frac{1}{4}, & \mathbb{P}(A_2) \cdot \mathbb{P}(A_3) = \frac{1}{4}. \end{array}$$

¹The word *mutually* is often omitted, and we simply speak of independence of events.

Since for all three pairs of events A_1, A_2, A_3 the equality

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i) \cdot \mathbb{P}(A_j), \quad i \neq j,$$

holds, these events are pairwise independent. $\triangle$

Definition 3.4 (Mutually independent events)

Let $A_1, A_2, \dots, A_n$ be events. We say that they are mutually (totally) independent if, for any sequence of indices

$\{k_1, k_2, \dots, k_r\} \subset \{1, \dots, n\}$, $r = 2, \dots, n$ we have

$$\mathbb{P}(A_{k_1} \cap A_{k_2} \cap \dots \cap A_{k_r}) = \mathbb{P}(A_{k_1}) \cdot \mathbb{P}(A_{k_2}) \cdots \mathbb{P}(A_{k_r}). \quad (3.8)$$

Example 3.6 (Continuation of Example 3.5). Determine whether the events from Example 3.5 are mutually independent.

Solution. $\mathbb{P}(A_1 \cap A_2 \cap A_3) = 0$, since $A_1 \cap A_2 \cap A_3$ is an impossible event. Because $\mathbb{P}(A_1) \cdot \mathbb{P}(A_2) \cdot \mathbb{P}(A_3) \neq 0$, the events A_1, A_2, A_3 are not mutually independent, even though they are pairwise independent. $\triangle$

We now present several statements concerning the independence of events.

Theorem 3.3 (on the independence of events).

- a) The impossible event and any event A are independent.
- b) The certain event and any event A are independent.
- c) Let A, B be disjoint events. Then the events A, B are independent if and only if $\mathbb{P}(A) \cdot \mathbb{P}(B) = 0$.

Proof.

$$\text{a) } \left. \begin{array}{l} \mathbb{P}(\emptyset \cap A) = \mathbb{P}(\emptyset) = 0 \\ \mathbb{P}(\emptyset) \cdot \mathbb{P}(A) = 0 \cdot \mathbb{P}(A) = 0 \end{array} \right\} \Rightarrow \mathbb{P}(\emptyset \cap A) = \mathbb{P}(\emptyset) \cdot \mathbb{P}(A).$$

$$\text{b) } \mathbb{P}(\Omega \cap A) = \mathbb{P}(A) = \mathbb{P}(A) \cdot 1 = \mathbb{P}(A) \cdot \mathbb{P}(\Omega).$$

- c) The proof of the equivalence is carried out by verifying the individual implications. Implication from left to right: for disjoint events A, B , we have $A \cap B = \emptyset$. Since we assume that they are independent, it holds that

$$\mathbb{P}(A) \cdot \mathbb{P}(B) = \mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0.$$

Proof of the implication from right to left: since A and B are disjoint, we have $\mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0$. Since, by assumption, $\mathbb{P}(A) \cdot \mathbb{P}(B) = 0$, the events A, B must also be independent. $\square$

Theorem 3.4. *Let A, B be independent events. Then the pairs of events $(A, \bar{B})$, $(\bar{A}, B)$, $(\bar{A}, \bar{B})$ are independent.*

Proof.

$$\begin{aligned}\mathbb{P}(\bar{A} \cap B) &= \mathbb{P}(B - A) = \mathbb{P}(B - (A \cap B)) = \mathbb{P}(B) - \mathbb{P}(A \cap B) = \\ &= \mathbb{P}(B) - \mathbb{P}(B) \cdot \mathbb{P}(A) = \mathbb{P}(B) \cdot (1 - \mathbb{P}(A)) = \mathbb{P}(B) \cdot \mathbb{P}(\bar{A}).\end{aligned}$$

The independence of the events A and $\bar{B}$ can be proved analogously.

If A and B are independent events, then, according to the previous result, the events A and $\bar{B}$ are independent; and from this, again by the previous result, it follows that the events $\bar{A}$ and $\bar{B}$ are also independent. $\square$

Theorem 3.5. *Let $A_1, \dots, A_n$ be mutually (totally) independent events. Then it holds that*

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = 1 - \prod_{i=1}^n [1 - \mathbb{P}(A_i)].$$

Proof. From De Morgan's laws for set operations, it follows that

$$\bigcup_{i=1}^n A_i = \Omega - \bigcap_{i=1}^n \bar{A}_i,$$

hence

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \mathbb{P}(\Omega) - \mathbb{P}\left(\bigcap_{i=1}^n \bar{A}_i\right) = 1 - \mathbb{P}\left(\bigcap_{i=1}^n \bar{A}_i\right).$$

Since $A_1, \dots, A_n$ are independent, the events $\bar{A}_1, \dots, \bar{A}_n$ are also independent, and therefore

$$\mathbb{P}\left(\bigcap_{i=1}^n \bar{A}_i\right) = \prod_{i=1}^n \mathbb{P}(\bar{A}_i) = \prod_{i=1}^n (1 - \mathbb{P}(A_i)).$$

$\square$

Example 3.7. During a given day, ten children were born in a maternity ward. The probability of a boy being born is $p = 0.514$. What is the probability that at least one boy was born in the ward on that day?



Solution. Let A_i , $i = 1, \dots, 10$, denote the event “the i -th child born is a boy”. The event $\bigcup_{i=1}^{10} A_i$ means “at least one of the children born is a boy”. Since $A_1, \dots, A_{10}$ are mutually independent, by Theorem 3.5

$$\mathbb{P}\left(\bigcup_{i=1}^{10} A_i\right) = 1 - \prod_{i=1}^{10} (1 - \mathbb{P}(A_i)) = 1 - 0.486^{10} \approx 0.999265.$$

△

Theorem 3.6. *Events A and B , where $0 < \mathbb{P}(B) < 1$, are independent if and only if*

$$\mathbb{P}(A|B) = \mathbb{P}(A|\bar{B}).$$

Proof. The equality

$$\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A \cap \bar{B})}{\mathbb{P}(\bar{B})}$$

holds if and only if $\mathbb{P}(\bar{B}) \cdot \mathbb{P}(A \cap B) = \mathbb{P}(\bar{A} \cap B) \cdot \mathbb{P}(B)$. Substituting $\mathbb{P}(\bar{B}) = 1 - \mathbb{P}(B)$ yields $\mathbb{P}(A \cap B) - \mathbb{P}(B) \cdot \mathbb{P}(A \cap B) = \mathbb{P}(B) \cdot \mathbb{P}(A \cap \bar{B})$, and after simplification, $\mathbb{P}(A \cap B) = \mathbb{P}(B) \cdot (\mathbb{P}(A \cap B) + \mathbb{P}(A \cap \bar{B})) = \mathbb{P}(B) \cdot \mathbb{P}(A)$. □

3.3 Bernoulli² process

In the introductory chapters, we have discussed random experiments as one-time experiments in which, with a certain probability, we expect exactly one of the possible outcomes to occur. We will now study repeated experiments, i.e., experiments in which the same trial is repeated several times in succession. A special case is the repetition of experiments with two alternative mutually exclusive outcomes, for example 0–1, boy–girl, drawing from an urn with black and white balls, etc. Individual experiments of this type are called *Bernoulli trials*, their basic feature being that the outcomes of the experiment are characterized only by two mutually exclusive attributes. If, in addition, the individual trials are independent of each other (the outcome of one trial is not influenced by the outcomes of previous trials), their sequence is called a *Bernoulli process*. In Chapter 7, we will deal with random variables related to the



²Jacob Bernoulli (1655–1705), Swiss mathematician and physicist.

Bernoulli process (e.g., the Bernoulli distribution in Section 7.3, the binomial distribution in Section 7.4, or the geometric distribution in Section 7.5). The Bernoulli process can be represented by the so-called *Bernoulli scheme*, which we now introduce.

Example 3.8 (Bernoulli scheme). An urn contains b white balls and c black balls. We successively draw one ball n times, returning the ball to the urn after each draw. What is the probability that among the drawn balls there are exactly k white balls and $(n - k)$ black balls?

Solution. In each repeated draw, the urn always contains b white balls and c black balls. Therefore, in each of the n draws, there are $(b + c)$ possibilities for drawing one ball. Since each possibility from the first draw can be combined with each possibility from the second draw, and these possibilities can be combined with each possibility from the third draw, ..., the number of possible outcomes of the experiment is $(b + c)^n$.

We now determine the number of favorable cases. In n draws, we want to draw k white balls, i.e., we distribute k white balls among n positions. The number of such arrangements is $\binom{n}{k}$. If we consider one fixed arrangement, for example, first k white balls are drawn and then $(n - k)$ black balls, we see that such an arrangement can occur in $b^k \cdot c^{n-k}$ ways. Thus, the number of favorable cases is $\binom{n}{k} b^k c^{n-k}$. The sought probability $P_{k,n}$ is therefore equal to

$$P_{k,n} = \binom{n}{k} \frac{b^k \cdot c^{n-k}}{(b + c)^n},$$

which is nothing other than

$$P_{k,n} = \binom{n}{k} \left(\frac{b}{b + c} \right)^k \left(\frac{c}{b + c} \right)^{n-k}. \quad (3.9)$$

Let us denote $p = \frac{b}{b+c}$, $q = \frac{c}{b+c}$; here p represents the probability that in a single draw we obtain a white ball, and q represents the probability that in a single draw we obtain a black ball. Then (3.9) can be written in the form

$$P_{k,n} = \binom{n}{k} p^k q^{n-k}, \quad \text{where } q = 1 - p, \quad p \in (0, 1). \quad (3.10)$$

△

In the case of a repeatedly and independently performed alternative experiment, we compute the probability that in n independent trials the event A occurs k times according to (3.10), where in each trial the event A occurs

with probability p . The probability that in n independent trials the event A occurs at least k times is computed using the formula

$$S_{k,n} = \sum_{i=k}^n P_{i,n}. \quad (3.11)$$

From (3.11), moreover, for $k = 1$, we obtain

$$S_{1,n} = \sum_{i=1}^n P_{i,n} = 1 - P_{0,n} = 1 - q^n, \quad (3.12)$$

which is the probability that in n independent trials the event A occurs at least once.

Let us now return to Example 3.7, which we solved using Theorem 3.5. This example can also be solved using the above Bernoulli scheme if we consider each birth as a Bernoulli trial. In Example 3.7, we thus have ten independent Bernoulli trials; the outcome of each such trial is either the birth of a boy or a girl. The probability of a boy being born in each trial is 0.514; according to (3.12), we therefore obtain

$$S_{1,n} = 1 - (0.486)^{10} \approx 0.999265.$$

We see that the Bernoulli scheme is a relatively efficient method for solving problems of this type.

Example 3.9. We throw a die n times in succession. Compute the probability that in these n throws at least one six appears.

Solution. We again have n independent trials. The outcome of each trial is either a six with probability $p = \frac{1}{6}$, or one of the remaining outcomes with probability $q = \frac{5}{6}$. According to (3.12), we obtain

$$S_{1,n} = 1 - \left(\frac{5}{6}\right)^n,$$

which is the probability that in n throws of a die, a six appears at least once. $\triangle$



We are often more interested in the number of trials n that we must perform in order to be able to assert, with probability at least P , that the

event A has occurred at least once, given that the probability of the event A occurring in each individual trial is p . In other words, we require that

$$P \leq S_{1,n},$$

from which, after simplification, we obtain

$$\begin{aligned} P &\leq 1 - (1 - p)^n, \\ \ln(1 - P) &\geq n \ln(1 - p), \end{aligned}$$

and hence

$$n \geq \frac{\ln(1 - P)}{\ln(1 - p)}. \quad (3.13)$$

Example 3.10. A player submits 30 lottery tickets in a betting office. The probability of winning the 5th prize (guessing three numbers out of 49, filling in only one column) is $p = 0.01765$. Determine the number of tickets the player must submit so that among them there is at least one winning the 5th prize with probability at least 0.75.



Solution. The calculation is carried out according to (3.13), where $p = 0.01765$, $P = 0.75$:

$$n \geq \frac{\ln(1 - 0.75)}{\ln(1 - 0.01765)} = \frac{\ln 0.25}{\ln 0.98235} \approx 77.8.$$

The player must submit 78 lottery tickets in order to win at least one 5th prize with probability at least 0.75. $\triangle$

3.4 Exercises

3.1 There are six balls in an urn. Each ball has a different color (red, white, black, yellow, blue, green). We repeatedly draw one ball independently five times in succession, returning the ball to the urn after each draw. What is the probability that

- a) “in the 1st and the 5th draw we obtain a yellow ball” (A),
- b) “we obtain a yellow ball exactly twice” (B)?

3.2 The machine consists of two components that are connected

- (S) in series,
- (P) in parallel,

and operate independently of each other. The probability of failure of the first component is 0.04, and of the second 0.03. A device with components connected in series functions if both components operate properly simultaneously. A device with components connected in parallel functions if at least one component operates without failure. What is, for each type of connection, the probability that

- a) “the machine operates” (A),
- b) “the machine does not operate” ($\bar{A}$)?



3.3 A machine consists of three components that are all connected in parallel and operate independently of each other. The probability that the first component functions is $\mathbb{P}(A_1)$, for the second $\mathbb{P}(A_2)$, and for the third $\mathbb{P}(A_3)$. The device functions if at least one component operates without failure. Express in general terms the probability $\mathbb{P}(A)$ of the event A that “the machine operates without failure”.

3.4 We roll a die 20 times. What is the probability that

- a) “in all twenty throws, a face with at most 4 dots appears” (event A),
- b) “only in the twentieth throw does a face with at most 4 dots not appear” (event B),
- c) “a face with at most 4 dots appears exactly ten times” (event C).

3.5 We draw one card twice in succession from a deck of 32 cards, without returning the first card to the deck. What is the probability that:

- “in the second draw we draw any king, given that in the first draw we also drew any king”,
- “in the second draw we do not draw a king, given that in the first draw we drew any king”,
- “in the second draw we draw a king, given that in the first draw we did not draw any king”,
- “in the second draw we do not draw a king, given that we also did not draw one in the first draw”.



3.6 Two athletes throw a javelin independently of each other, each having only one attempt. The probability that the first throws 80 meters is 0.8. The probability that the second throws 80 meters is 0.75. What is the probability that neither of them throws 80 meters?



3.7 A professor forgets his umbrella at each visit to a shop with probability $\frac{1}{4}$. If he visited four shops and returned home without the umbrella, what is the probability that he forgot it in the last shop?



3.8 It is given that: $\mathbb{P}(A) = 0.4$, $\mathbb{P}(B) = 0.5$, and $\mathbb{P}(\overline{A} \cap \overline{B}) = 0.3$.

- Are the events A and B independent?
- Compute $\mathbb{P}(A \cup B)$, $\mathbb{P}(A \cap B)$, $\mathbb{P}(\overline{A} \cup \overline{B})$ and $\mathbb{P}(A \cup \overline{B})$.
- Compute $\mathbb{P}(A|\overline{B})$ and $\mathbb{P}(\overline{B}|\overline{A})$.

3.9 Consider two urns. Each contains four tickets. On two tickets there are numbers divisible by two, and on the other two there are numbers divisible by three, with none of the tickets bearing a number divisible by six. Let A be the event that we draw from the first urn a ticket bearing a number divisible by two; let B be the event that we draw from the second urn a ticket

bearing a number divisible by three; and let C be the event that in both draws we draw tickets with numbers divisible by the same number. Determine whether the events A , B , C are pairwise independent and mutually independent.

3.10 A children's picture puzzle consists of twelve cubes, each having on every face a part of one of six pictures. In order to correctly assemble a picture from the partial images on the individual faces of the cubes, each cube must be placed in the correct position, must have the correct face turned upward, and the picture must be oriented in the correct direction. Let the event A mean that "we place each cube in its correct position in the box", the event B that "it will be turned with the correct face upward" (all faces belonging to the same picture), and the event C that "we rotate the correct face into the correct orientation" (for simplicity, we do not consider the upside-down position as correct). Compute $\mathbb{P}(A \cap B \cap C)$, i.e., the probability that by randomly placing the cubes into the box we correctly assemble one of the six pictures.

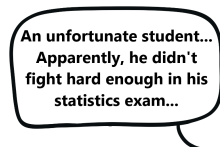
3.11 A gardening company Truhlík a spol. delivers a shipment of one hundred potted plants to a flower shop. When receiving the shipment, the shop assistant performs a random quality check of the goods. She selects five plants at random. If at least one plant is infested with pests, she rejects the shipment. Determine the probability of rejecting the shipment if 2% of the total delivered plants are infested with pests.

3.12 From a deck of 32 cards, among which there are four aces, we draw one card after another without replacement. In total, we draw 7 cards. Compute the probability that an ace appears only as the second and the seventh card.



3.13 In an examination in probability and statistics, 30% of the sixty questions are from probability theory. Once drawn, questions are not returned. So far, ten students have been examined. Determine the probability that

- "exactly four of the drawn questions were from probability theory" (event A),
- "at most three were from probability theory" (event B),



- c) “at least two questions were from probability theory” (event C).

3.14 (Banach’s³ problem) A shop assistant has sweaters on the counter arranged in two equally large piles. Each pile contains n sweaters. When selling a sweater, she randomly selects one of the piles (with equal probability) and removes a sweater from it. Eventually, she selects a pile from which all sweaters have already been sold out. Compute the probability that at that moment k sweaters remain in the other pile.

3.15 (De Montmort’s⁴ matching problem) We have n different letters and n different envelopes. The letters are placed into the envelopes at random, with exactly one letter in each envelope. What is the probability that at least one letter is placed into the correct envelope? Compute the probability for $n = 4$.

3.16 Players A, B, C take turns tossing a coin; A tosses first, B tosses second, and C tosses third. The game ends as soon as one player obtains heads, and that player becomes the winner. Compute the probabilities:

- a) that player A wins,
- b) that player B wins.

3.17 An urn contains six balls, four of which are white. We randomly select four balls without replacement. Let the events be: A “exactly one of the first two drawn balls is white”, B “the fourth drawn ball is white”, C “exactly two balls in the sample are white”. Determine the probabilities of the events A, B, C . Are the events A, B, C mutually independent?

3.18 We roll two dice. Event A means that an odd number appears on the blue die, event B means that an even number appears on the green die, and event C means that the sum of both numbers is odd. Are the events A, B, C independent? Are the events A, B, C pairwise independent?

3.19 There are three rogues in prison: Al Capone, Babinský, and Cimrman.⁵ Two of them are chosen by lot to be executed. Al Capone wants to find out whether he is among those chosen. He knows that the guard would not answer a direct question, so he asks

³Stefan Banach (1892–1945), Polish mathematician.

⁴Pierre Remond de Montmort (1678–1719), French mathematician.

⁵Václav Babinský – a legendary highwayman, Jára Cimrman – a fictional Czech genius.

him to name one of his fellow prisoners who will be executed. The guard tells the truth, and if he has two possibilities, he chooses between them uniformly at random. He names Babinský. Does this answer provide Al Capone with any information about his fate? (Before the conversation with the guard, Al Capone's probability of being executed was equal to $\frac{2}{3}$; is the conditional probability after the guard's answer different?)



Solutions to the Exercises

3.1 a) $\mathbb{P}(A) = \frac{1}{36} \approx 0.028$, b) $\mathbb{P}(B) \approx 0.161$.

3.2

(S) a) $\mathbb{P}(A) = 0.9312$, b) $\mathbb{P}(\overline{A}) = 0.0688$.

(P) a) $\mathbb{P}(A) = 0.9988$, b) $\mathbb{P}(\overline{A}) = 0.0012$.

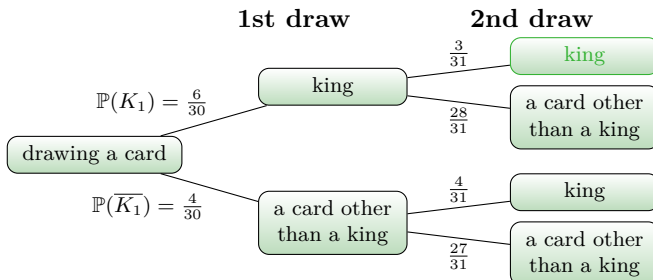
3.3 $\mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) - \mathbb{P}(A_1)\mathbb{P}(A_2) - \mathbb{P}(A_1)\mathbb{P}(A_3) - \mathbb{P}(A_2)\mathbb{P}(A_3) + \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3)$.

3.4 a) $\mathbb{P}(A) \approx 3.01 \cdot 10^{-4}$, b) $\mathbb{P}(B) \approx 1.50 \cdot 10^{-4}$,

c) $\mathbb{P}(C) = \binom{20}{10} \cdot \left(\frac{4}{6}\right)^{10} \cdot \left(\frac{2}{6}\right)^{10} \approx 0.054$.

3.5 a) $\mathbb{P}(K_2|K_1) = \frac{3}{31} \approx 0.097$, b) $\mathbb{P}(\overline{K_2}|K_1) = \frac{28}{31} \approx 0.903$,

c) $\mathbb{P}(K_2|\overline{K_1}) = \frac{4}{31} \approx 0.129$, d) $\mathbb{P}(\overline{K_2}|\overline{K_1}) = \frac{27}{31} \approx 0.871$.



3.6 0.05

3.7 0.154**3.8**

- a) The events A and B are independent, since the events $\bar{A}$ and $\bar{B}$ are also independent:
 $\mathbb{P}(\bar{A}) \cdot \mathbb{P}(\bar{B}) = 0.6 \cdot 0.5 = 0.3 = \mathbb{P}(\bar{A} \cap \bar{B})$.
- b) Given the independence of the events A and B , it is convenient to first compute their intersection: $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B) = 0.4 \cdot 0.5 = 0.2$. Then we determine $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = 0.4 + 0.5 - 0.2 = 0.7$.
 Further, using the same formula: $\mathbb{P}(\bar{A} \cup B) = 0.6 + 0.5 - 0.6 \cdot 0.5 = 0.8$,
 $\mathbb{P}(A \cup \bar{B}) = 0.4 + 0.5 - 0.4 \cdot 0.5 = 0.7$.
- c) From the independence of the events A and B , it follows that $\mathbb{P}(A|\bar{B}) = \mathbb{P}(A) = 0.4$ and $\mathbb{P}(\bar{B}|\bar{A}) = \mathbb{P}(\bar{B}) = 0.5$.

3.9 The events A , B , and C are pairwise independent, but they are not mutually independent.

3.10 $\frac{1}{12!} \cdot \frac{6}{6^{12}} \cdot \frac{1}{4^{12}} \approx 3.4 \cdot 10^{-25}$

3.11 0.098**3.12** 0.00834

3.13 a) $\mathbb{P}(A) \approx 0.213$, b) $\mathbb{P}(B) \approx 0.6575$, c) $\mathbb{P}(C) \approx 0.874$.

3.14 $\binom{2n-k}{n} \cdot \left(\frac{1}{2}\right)^n \cdot \left(\frac{1}{2}\right)^{n-k} = \binom{2n-k}{n} \cdot \left(\frac{1}{2}\right)^{2n-k}$

3.15 $\mathbb{P}_n = 1 - \frac{D_n}{n!}$, where $D_n = n! \left(1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^n}{n!}\right)$. $\mathbb{P}_{n=4} = 1 - \frac{9}{24} = \frac{5}{8} = 0.625$.

3.16 a) $\frac{4}{7} \approx 0.571$, b) $\frac{2}{7} \approx 0.286$.

3.17 $\mathbb{P}(A) = \frac{8}{15}$, $\mathbb{P}(B) = \frac{2}{3}$, $\mathbb{P}(C) = \frac{2}{5}$, The events A , B , C are not mutually independent.

3.18 The events A , B , C are pairwise independent, but they are not mutually independent.

3.19 The probability that Al Capone will be among the two prisoners selected for execution is $\frac{2}{3}$. If the guard truthfully names Babinský as a prisoner to be executed, the conditional probability that the other condemned prisoner is Al Capone is $\frac{1}{6}$. However, Al Capone still remains in three of the remaining possible pairs of two prisoners selected for execution. The guard's information therefore only tells Al Capone that he will be executed together with Babinský with probability $\frac{1}{6}$. The probability that Al Capone will be executed thus remains equal to $\frac{2}{3}$.

Chapter 4

Total probability, Bayes' formula

4.1 Total Probability

The primary purpose of conditional probability, as mentioned in Section 3.1, is to update the probability of the occurrence of an event when additional information related to this event becomes available, specifically when the conditioning event has occurred. However, the concept of conditional probability can also be used to compute the so-called *total (law of total) probability* of a given event in situations where the conditioning event has not yet occurred, but its possible outcomes, their probabilities, and the corresponding conditional probabilities are known. Let us present an example illustrating this situation.

Example 4.1. There are twenty skiers, six cyclists, and four runners in a training group. The probability that a skier meets the training standard is 0.9, for a cyclist it is 0.8, and for a runner 0.75. What is the probability that a randomly selected athlete meets the standard?

Solution. Let us denote A_1 the event “the randomly selected athlete is a skier”, A_2 the event “the randomly selected athlete is a cyclist”, and A_3 the event “the randomly selected athlete is a runner”; further, let B be the event “the randomly selected athlete meets the standard”. Then

$$\mathbb{P}(A_1) = \frac{20}{30} = \frac{2}{3}, \quad \mathbb{P}(A_2) = \frac{6}{30} = \frac{1}{5}, \quad \mathbb{P}(A_3) = \frac{4}{30} = \frac{2}{15}.$$

These probabilities are called *a priori* probabilities – we know in advance the distribution of the occurrence of the individual types of sports within the

population of athletes. Furthermore, we know the *conditional probabilities* of meeting the standard for each category of athletes, namely

$$\mathbb{P}(B|A_1) = 0.9, \quad \mathbb{P}(B|A_2) = 0.8 \quad \text{and} \quad \mathbb{P}(B|A_3) = 0.75.$$

We decompose the event B according to the category of sport into three disjoint subsets (and can therefore sum their probabilities), and repeatedly apply the formula (3.2) to compute the probability of the event B :

$$\begin{aligned} \mathbb{P}(B) &= \mathbb{P}(B \cap A_1) + \mathbb{P}(B \cap A_2) + \mathbb{P}(B \cap A_3) = \\ &= \mathbb{P}(B|A_1) \cdot \mathbb{P}(A_1) + \mathbb{P}(B|A_2) \cdot \mathbb{P}(A_2) + \mathbb{P}(B|A_3) \cdot \mathbb{P}(A_3). \end{aligned} \quad (4.1)$$

Since we have all the necessary information available, we can substitute the current values into formula (4.1), obtaining

$$\mathbb{P}(B) = \frac{2}{3} \cdot 0.9 + \frac{1}{5} \cdot 0.8 + \frac{2}{15} \cdot 0.75 = 0.86.$$

The randomly selected athlete thus meets the standard with probability 0.86. $\triangle$

The procedure from the solution of Example 4.1 can be generalized into Theorem 4.1, including its proof.

Theorem 4.1 (on total probability). *Let $A_1, A_2, \dots$ be events forming a partition of the sample space, i.e.*

$$A_i \cap A_j = \emptyset \text{ for any } i \neq j \quad \text{and} \quad \bigcup_{i=1}^{\infty} A_i = \Omega.$$

Let these events have probabilities $\mathbb{P}(A_1), \mathbb{P}(A_2), \dots$, respectively, where $\mathbb{P}(A_i) > 0$ for all $i = 1, 2, \dots$. Let us consider an arbitrary event B , for which the conditional probabilities $\mathbb{P}(B|A_i)$ are known for all $i = 1, 2, \dots$. Then the so-called law of total probability holds

$$\mathbb{P}(B) = \sum_{i=1}^{\infty} \mathbb{P}(B|A_i) \cdot \mathbb{P}(A_i). \quad (4.2)$$

Proof. Since the events $A_1, \dots, A_n$ form a disjoint partition, it follows that if for any $i \neq j$ we have $(A_i \cap B) \cap (A_j \cap B) = \emptyset$, then

$$\bigcup_{i=1}^{\infty} (A_i \cap B) = B.$$

$$\mathbb{P}(B) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} (A_i \cap B)\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i \cap B) = \sum_{i=1}^{\infty} \mathbb{P}(B|A_i) \cdot \mathbb{P}(A_i).$$

In the proof, we used property b) from Theorem 2.2 on page 25 and formula (3.2). $\square$

The situation of randomly selecting one athlete can also be represented by a tree diagram, as shown in Figure 4.1. The total probability is the sum of the joint probabilities lying on all branches labeled “yes” (met the standard).

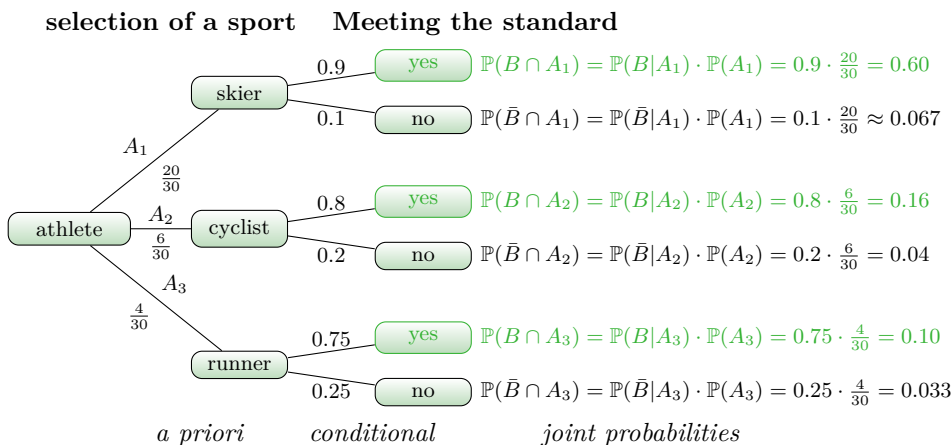


Figure 4.1: Tree diagram of the random selection of an athlete from Problem 4.1

4.2 Bayes' theorem¹

Let us return to Example 4.1 on page 61 and pose a new question.

Example 4.2. In a group of thirty athletes, there are twenty skiers, six cyclists, and four runners. A skier meets the standard with probability 0.9, a cyclist with probability 0.8, and a runner with probability 0.75. It is now known that a randomly selected athlete has met the standard. What is the probability that this athlete is a cyclist?

Solution. Let the events A_1, A_2, A_3, B have the same meaning as in the solution of Example 4.1. In that example, based on the prior and conditional probabilities $\mathbb{P}(A_i)$ and $\mathbb{P}(B|A_i)$, we computed the total probability of the

¹Thomas Bayes (1701–1761), an English clergyman and statistician.

event B . Now we additionally know that event B has occurred, and we wish to “update” the probabilities of the events A_i accordingly (in particular, A_2).

From a probabilistic point of view, we are looking for the probability that a cyclist (event A_2) is selected, given that he or she has met the standard (event B). That is, using formula (3.1) from definition 3.1, we want to compute the value of the conditional probability

$$\mathbb{P}(A_2|B) = \frac{\mathbb{P}(A_2 \cap B)}{\mathbb{P}(B)}.$$

The numerator of the formula is the joint probability

$$\mathbb{P}(A_2 \cap B) = \mathbb{P}(B|A_2) \cdot \mathbb{P}(A_2) = 0.8 \cdot \frac{6}{30} = 0.16,$$

which we have, for instance, represented in the tree diagram 4.1. The denominator $\mathbb{P}(B) = 0.86$ is the value of the total probability that we also computed in Example 4.1. Hence, the resulting probability is

$$\mathbb{P}(A_2|B) = \frac{\mathbb{P}(A_2 \cap B)}{\mathbb{P}(B)} = \frac{0.16}{0.86} \approx 0.186. \quad (4.3)$$

△

The computation in Example 4.2 suggests an outline of the proof of the so-called Bayes' theorem, which we now present in its general form.

Theorem 4.2 (Bayes' theorem). *Let the assumptions of theorem 4.1 be satisfied. Then for any $j = 1, 2, \dots$ we have*

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j) \cdot \mathbb{P}(A_j)}{\sum_{i=1}^{\infty} \mathbb{P}(B|A_i) \cdot \mathbb{P}(A_i)}. \quad (4.4)$$

Proof. According to formula (3.1) for the conditional probability, we have

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(A_j \cap B)}{\mathbb{P}(B)}. \quad (4.5)$$

Substituting (4.2) into (4.5), we obtain

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(A_j \cap B)}{\sum_{i=1}^{\infty} \mathbb{P}(B|A_i) \cdot \mathbb{P}(A_i)}.$$

In the numerator, we additionally use (3.2) and obtain formula (4.4), which completes the proof. □



Formula (4.4) is called Bayes' rule (or first Bayes' theorem), which we apply by formulating, based on the text of the problem, events $A_i, i \in I$, so-called hypotheses, that are mutually exclusive and whose union covers all possibilities – i.e., they form a partition of Ω . We seek the probability of one of these events, say A_j for some $j \in I$, given that event B has occurred. We know in advance the so-called prior probabilities $\mathbb{P}(A_i)$ and the conditional probabilities $\mathbb{P}(B|A_i)$ for all $i \in I$. The probability $\mathbb{P}(A_j|B)$ is called the posterior probability and is computed by substituting the known values into formula (4.4).

Let us now return to Example 4.2 and verify its solution using the first Bayes' theorem, which is in fact exactly the computation (4.3).

Solution. (continuation from page 63) Using the probabilities already known from Example 4.1, we compute $\mathbb{P}(A_2|B)$. Substituting into (4.4), we obtain

$$\mathbb{P}(A_2|B) = \frac{\mathbb{P}(B|A_2) \cdot \mathbb{P}(A_2)}{\sum_{j=1}^3 \mathbb{P}(B|A_j) \cdot \mathbb{P}(A_j)} = \frac{0.8 \cdot \frac{1}{5}}{0.9 \cdot \frac{2}{3} + 0.8 \cdot \frac{1}{5} + 0.75 \cdot \frac{2}{15}} \approx 0.186.$$

The probability that a randomly selected athlete who has met the standard is a cyclist is, after rounding, equal to 0.186. $\triangle$

Question 4.1. Compute the probability that a randomly selected athlete who has met the standard is a skier, and respectively a runner.

Example 4.3. During the examination of a patient, there is a suspicion of three mutually exclusive diseases. The probability of occurrence of the first disease is 0.3, of the second 0.5, and of the third 0.2. A laboratory test yields a positive result for 15% of patients with the first disease, for 30% of patients with the second disease, and for 30% of patients with the third disease. What is the probability of the occurrence of the second disease, given that the result of the laboratory test is positive (i.e., the patient has one of the diseases)?



Solution. Denoting A_1 as the event “the patient has disease 1”, A_2 as the event “the patient has disease 2”, A_3 as the event “the patient has disease

3", and B as the event "the laboratory test yields a positive result", the events A_1, A_2, A_3 form a partition of the sample space, since

$$A_i \cap A_j = \emptyset \text{ for all } i \neq j, \ i, j = 1, 2, 3, \text{ and } \bigcup_{i=1}^3 A_i = \Omega.$$

We know that $\mathbb{P}(A_1) = 0.3$, $\mathbb{P}(A_2) = 0.5$, and $\mathbb{P}(A_3) = 0.2$. We also know the conditional probabilities: $\mathbb{P}(B : A_1) = 0.15$, $\mathbb{P}(B : A_2) = 0.30$, $\mathbb{P}(B : A_3) = 0.30$. We want to determine the value of $\mathbb{P}(A_2 : B)$. We again use Bayes' theorem. Substituting the known probabilities into (4.4), we obtain

$$\mathbb{P}(A_2|B) = \frac{0.5 \cdot 0.3}{0.3 \cdot 0.15 + 0.5 \cdot 0.3 + 0.2 \cdot 0.3} \approx 0.588.$$

After the laboratory test has been performed and the result is positive, the probability of the occurrence of the second disease is, after rounding, equal to 0.588. $\triangle$

Example 4.4. In the entire population, 60% of people are vaccinated, while 40% are unvaccinated. Hospital statistics show that the probability that a patient is vaccinated, respectively unvaccinated, given that they are admitted to an intensive care unit (ICU), is 0.1, respectively 0.9. What is the relative risk that an unvaccinated person will be admitted to the ICU?

Solution. Let us denote the random events: O "The person is vaccinated.", J "The person is admitted to the ICU." From the problem statement, we know: $\mathbb{P}(O) = 0.6$, $\mathbb{P}(\overline{O}) = 0.4$, $\mathbb{P}(O|J) = 0.1$, $\mathbb{P}(\overline{O}|J) = 0.9$.

It holds that

$$\mathbb{P}(J|O) = \frac{\mathbb{P}(O|J) \cdot \mathbb{P}(J)}{\mathbb{P}(O)} = \frac{0.1 \cdot \mathbb{P}(J)}{0.6}$$

and similarly

$$\mathbb{P}(J|\overline{O}) = \frac{\mathbb{P}(\overline{O}|J) \cdot \mathbb{P}(J)}{\mathbb{P}(\overline{O})} = \frac{0.9 \cdot \mathbb{P}(J)}{0.4}.$$

$$\frac{\mathbb{P}(J|O)}{\mathbb{P}(J|\overline{O})} = \frac{\frac{0.1}{0.6}}{\frac{0.9}{0.4}} \approx 0.074.$$

$\triangle$

From the above examples, it is evident that when solving problems using Bayes' theorem, we proceed as follows:

- 1) we determine the random events $A_1, \dots, A_n$ that form a partition of the sample space, i.e., they are mutually disjoint and exhaust all possibilities,
- 2) we determine the event B ,
- 3) we identify the prior probabilities $\mathbb{P}(A_1), \dots, \mathbb{P}(A_n)$ and the conditional probabilities $\mathbb{P}(B : A_1), \dots, \mathbb{P}(B : A_n)$, and
- 4) we substitute into formula (4.4).

Example 4.5 (AIDS). A blood test for the presence of the HIV virus does not always correctly identify the disease. Two types of errors may occur:

- a) the test incorrectly indicates a positive result (presence of the virus), or
- b) the test incorrectly indicates a negative result (absence of the virus).

Statistical observations have shown that this test is highly reliable. More precisely, if an individual is infected, the test will be positive with probability 0.995 (this property is called the sensitivity of the test). That is, $\mathbb{P}(Pos|Inf) = 0.995$, from which we obtain that the probability of error (a false negative result) is $\mathbb{P}(Neg|Inf) = 0.005$. Similarly, a non-infected individual is correctly identified with probability $\mathbb{P}(Neg|NonInf) = 0.996$ (this property is called the specificity of the test), and the probability of error (a false positive result) is $\mathbb{P}(Pos|NonInf) = 0.004$.

Suppose that a law is enacted requiring all people to undergo this “accurate” test in order to identify all infected individuals. If we then randomly select one person with a positive test result, we ask what is the probability that the person actually has HIV. In other words, we are interested in the probability $\mathbb{P}(Inf|Pos)$, which we determine using Bayes’ theorem:

$$\mathbb{P}(Inf|Pos) = \frac{\mathbb{P}(Pos|Inf) \cdot \mathbb{P}(Inf)}{\mathbb{P}(Pos|Inf) \cdot \mathbb{P}(Inf) + \mathbb{P}(Pos|NonInf) \cdot \mathbb{P}(NonInf)}.$$

We still need to determine the prior probabilities $\mathbb{P}(Inf)$ and $\mathbb{P}(NonInf)$. In the Czech Republic, as of July 31, 2021, there were 2 990 registered infected individuals. On the same date, the Czech Republic had a population of 10 524 167. This leads to the estimates

$$\mathbb{P}(Inf) = \frac{2\,990}{10\,524\,167} \approx 0.000284 \quad \text{and} \quad \mathbb{P}(NonInf) \approx 0.999716.$$

In medical diagnostics, $\mathbb{P}(Inf)$ is referred to as the prevalence of the disease². Substituting into (4.4), we obtain

$$\mathbb{P}(Inf|Pos) = \frac{0.995 \cdot 0.000284}{0.995 \cdot 0.000284 + 0.004 \cdot 0.999716} \approx 0.066.$$

²In the analysis of disease spread, prevalence is usually replaced by the so-called incidence rate, i.e., the number of newly diagnosed cases related to the total population size.

Enacting such a law would therefore not be very reasonable, since only 6.6% of positively tested individuals would actually be infected with HIV.

4.3 Exercises

4.1 Two forest nurseries produce spruce seedlings. The first nursery produces 80% of the seedlings, of which ninety out of one hundred are of first quality. The second nursery supplies 20% of the seedlings, of which sixty out of one hundred are of first quality.

- What is the probability that a randomly selected seedling is of first quality?
- What is the probability that a randomly selected first-quality seedling comes from the first forest nursery?

4.2 In a city, there are three commercial companies. The first company operates twenty shops, the second fifteen shops, and the third ten shops. When visiting a shop belonging to the first company, you will be cheated with probability 0.15, in a shop of the second company with probability 0.08, and in a shop of the third company with probability 0.05. What is the probability that you will be cheated when making a purchase in this city?

4.3 Entrance examinations to a certain secondary school take place in four classes. In the first class with fifty students, 80% passed the exam; in the second class with 35 students, 40% passed; in the third class with 40 students, 65% passed the exam; and in the fourth class with 30 students, the success rate was 60%. From all students who took the entrance examinations, one is selected at random. What is the probability that this student successfully passed the exam?



4.4 There are n passengers on a bus. At the next stop, each of them gets off with probability p . In addition, no new passenger boards the bus with probability p_0 , and with probability $1 - p_0$ one new passenger boards. What is the probability that, after the next stop and when the bus departs again, there will be n passengers on board, just as at the beginning?

4.5 There are 36 students in a class, of whom 15 are boys and 21 are girls. Among the boys, five are obese, and among the girls, seven are obese. What is the probability that a randomly selected obese child is a boy?

4.6 Two groups of boys have gathered on a playground. The first group consists of twelve boys, and the second group of ten boys. The boys want to play football. Since each team should have eleven players, one boy is randomly selected from the first group and transferred to the second group. What is the probability that a randomly selected boy from the completed second group is a good football player, given that before the transfer the first group contained eight good and four weaker players, and the second group contained six good and four weaker players?



4.7 There are four urns. Suppose that each urn is equally likely to be selected. Each of the first three urns contains three slips with even numbers and two slips with odd numbers. The fourth urn contains four slips with even numbers and one slip with an odd number. We randomly select an urn, draw one slip from it, and place it into another urn (again selected at random). We then draw one slip from this urn. What is the probability that the last slip drawn has an even number?

4.8 Twelve students of the Mathematics–Physics programme, fourteen students of Mathematics–Biology, and twelve students of Mathematics–Computing are able to solve a problem on Bayes' theorem with probabilities 0.80, 0.65, and 0.95, respectively. What is the probability that Paul is a student of the programme

- a) Mathematics–Physics,
- b) Mathematics–Biology,
- c) Mathematics–Computing,

given that he is unable to solve problems on Bayes' theorem?

4.9 For messages transmitted in Morse code, the following holds: instead of the transmitted signal “dot”, a “dash” is received by mistake in $\frac{2}{5}$ of cases, and instead of the transmitted signal “dash”, a “dot” is mistakenly received in $\frac{1}{3}$ of cases. The transmitted signals dot and dash occur in the ratio 5 : 3. Determine the probability that the transmitted signal was

- a) a dot, given that a dash has been received (event A);
- b) a dot, given that a dot has been received (event B).



4.10 A 100 m race involves three groups of runners. The first group, consisting of ten runners, completes the distance within the time limit with probability 0.90, the second group, consisting of six runners, with probability 0.85, and the third group, also consisting of six runners, with probability 0.75. One runner is chosen at random from all runners, and it is known that this runner completes the 100 m within the time limit. What is the probability that this selected runner comes from the second group?



Solutions to the Exercises

4.1 a) 0.84, b) 0.857.

4.2 0.1044

4.3 0.6323

4.4 $p_0 \cdot (1 - p)^n + (1 - p_0) \cdot n \cdot p \cdot (1 - p)^{n-1}$

4.5 0.417

4.6 0.606

4.7 0.65

4.8 a) 0.304, b) 0.620, c) 0.076.

4.9 a) $\mathbb{P}(A) = 0.5$, b) $\mathbb{P}(B) = 0.75$.

4.10 0.274

Chapter 5

Random variable

Working with random events, that is, with the set of elementary outcomes Ω and a σ -algebra $\mathcal{A}$, can sometimes be difficult, since the construction of both sets may be quite complicated: elementary outcomes may, for instance, be arbitrary objects or their k -tuples, whose number can be very large. Fortunately, many results of random experiments are numerical: either the result is a number by the very nature of the experiment, or it can be expressed as a number. To represent such outcomes, we can therefore use a probability space constructed over the set of real numbers. Mathematical objects representing random experiments with numerical outcomes are called *random variables*.

5.1 Introduction of a Random Variable

Let us consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, where Ω is a non-empty set of all possible outcomes of a random experiment, $\mathcal{A}$ is a σ -algebra constructed on Ω , and $\mathbb{P}: \mathcal{A} \rightarrow [0, 1]$ is a probability measure. If the outcome of the experiment is numerical, we could define the set Ω directly as the set of all possible numerical outcomes of the experiment – for example, when measuring the height of students, we could take $\Omega = (0, \infty)$, choose the Borel σ -algebra $\mathcal{B}([0, \infty))$ as the collection of events, and work directly with this space. However, random variables can also be introduced in situations where the space of outcomes is not numerical, but can be suitably encoded numerically, as will be illustrated in Examples 5.1 and 5.2. For this reason, we define a random variable as a function that assigns to any (random) outcome – whether numerical or abstract – an appropriate numerical representation.

Definition 5.1 (Random variable)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. A real-valued function X defined on Ω is called a random variable if X is a measurable mapping¹ defined as $X: (\Omega, \mathcal{A}) \rightarrow (\mathbb{R}, \mathcal{B})$, that is, for any Borel set $B \in \mathcal{B}$ it holds that

$$\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{A}, \quad (5.1)$$

where $\mathcal{B} = \mathcal{B}(\mathbb{R})$ is the σ -algebra of Borel² subsets of $\mathbb{R}$.

Example 5.1. We roll a single die once. The set of elementary outcomes is $\Omega = \{\square, \blacksquare, \boxdot, \boxtimes, \boxplus, \boxminus\}$, and the σ -algebra (the collection of events) is the power set $\mathcal{A} = \mathcal{P}(\Omega)$. By the random variable X we mean the *numerical outcome of the die roll*, and we define it as a function $X: \Omega \rightarrow \mathbb{R}$ with values

$$X(\square) = 1, X(\blacksquare) = 2, X(\boxdot) = 3, X(\boxtimes) = 4, X(\boxplus) = 5, X(\boxminus) = 6. \quad (5.2)$$

Note that the function defined in this way is indeed a random variable, i.e., a measurable function: for any Borel subset of the real numbers, its preimage belongs to $\mathcal{A}$. For example, for the interval $(\pi, 5]$ the preimage of X is the set $X^{-1}((\pi, 5]) = \{\boxplus, \boxminus\}$, which is an element of $\mathcal{A}$.

Whether a function X is a random variable depends, among other things, on the form of the chosen σ -algebra. If, in Example 5.1, we were to work with a σ -algebra, for instance, of the form

$$\mathcal{A} = \{\emptyset; \{\square, \blacksquare\}; \{\boxdot, \boxtimes, \boxplus, \boxminus\}; \Omega\}, \quad (5.3)$$

the set mentioned in the example, $X^{-1}((\pi, 5]) = \{\boxplus, \boxminus\}$, does not belong to this σ -algebra, which contradicts the measurability condition. Let us recall in this context that, from a practical point of view, the σ -algebra associated with a random experiment represents the information that can be identified as the outcome of the experiment. The set of outcomes in the form (5.3) corresponds to a situation in which, for example, the event “ $\square$ occurs” cannot happen or, more precisely, cannot be identified (for instance, the die was rolled secretly and the statistician is only informed whether the result exceeds two). The construction and use of such specific σ -algebras, however, go beyond the scope of this textbook; for common purposes, it suffices to work with the σ -algebras introduced in Examples 2.1 and 2.2.

¹A pair $(\kappa, \mathcal{A})$ and $(\theta, \mathcal{B})$, where κ and θ are non-empty sets and $\mathcal{A}, \mathcal{B}$ are σ -algebras of their subsets, are called measurable spaces. A mapping $f: \kappa \rightarrow \theta$ is called measurable if for all $B \in \mathcal{B}$ we have $\{x \in \kappa : f(x) \in B\} \in \mathcal{A}$, i.e., the preimages of measurable sets are measurable. A measurable mapping is denoted $f: (\kappa, \mathcal{A}) \rightarrow (\theta, \mathcal{B})$.

²That is, the smallest σ -algebra containing all open subsets of $\mathbb{R}$, see Example 2.2 on page 22.

In the following example, we construct a random variable whose range will be something other than the natural numbers, as we have been accustomed to so far.

Example 5.2. The gender of a newborn child can be represented as a random experiment with two possible outcomes: boy and girl. We represent this experiment by the sample space $\Omega = \{\text{boy}, \text{girl}\}$, and the event space is the power set

$$\mathcal{A} = \{\emptyset, \{\text{boy}\}, \{\text{girl}\}, \Omega\}.$$

A random variable X can in this case be defined, for example, as follows:

$$X(\omega) = \begin{cases} 0, & \text{if } \omega = \text{boy}, \\ 1, & \text{if } \omega = \text{girl}. \end{cases}$$

In this case, the outcomes of the random experiment are of a qualitative nature (boy, girl), and we assign to them a certain, in this case artificially chosen, real number.

Random variables are typically denoted by uppercase letters from the end of the alphabet, such as X, Y, Z . The values $X(\omega)$ that random variables can take are denoted by lowercase letters x, y, z . We also say that x is a *realization* of the random variable X (especially in cases where the value of the random variable is already known because the random experiment has been carried out). We also commonly use a simplified notation: instead of $\{\omega \in \Omega : X(\omega) \in B\}$ we briefly write $\{X \in B\}$, and instead of $\{\omega \in \Omega : X(\omega) < x\}$ we write $\{X < x\}$.

Sums, products, and quotients of random variables are again random variables. Likewise, raising a random variable to a natural power and multiplying a random variable by a scalar results again in a random variable. Proofs can be found, for example, in [7].

5.2 Distribution of a Random Variable

As in Chapter 2, where we introduced the probability measure $\mathbb{P}$ on the abstract event space $(\Omega, \mathcal{A})$ (satisfying the requirements of Definition 2.2), we could similarly introduce a probability measure on the Borel subsets of the space $(\mathbb{R}, \mathcal{B})$, representing the possible numerical outcomes of random variables. However, since we have defined a random variable X on an already established probability space $(\Omega, \mathcal{A}, \mathbb{P})$, the probability of outcomes is determined by the already defined measure $\mathbb{P}$ and by the rule defining the function X .

Example 5.3 (continuation of Example 5.2). The probability of the birth of a boy is 0.514. Let the random variable X be defined as in Example 5.2. If we now want to determine the probability that this random variable takes the value zero, we proceed as follows:

$$\begin{aligned}\mathbb{P}(X = 0) &= \mathbb{P}\{\omega \in \Omega : X(\omega) = 0\} = \mathbb{P}\{\omega \in X^{-1}(0)\} = \\ &= \mathbb{P}\{\omega \in \{\text{boy}\}\} = \mathbb{P}\{\omega = \text{boy}\} = 0.514.\end{aligned}$$

Similarly, we would determine that $\mathbb{P}(X = 1) = \mathbb{P}\{\omega = \text{girl}\} = 0.486$.

We see that in this way we can determine the probability of any Borel set, simply by finding the probability of its preimage³. The probability measure on Borel sets defined in this way (induced by the probability measure $\mathbb{P}$) is called the *probability distribution of a random variable*, or simply the *distribution*⁴. Thus, the probability distribution fully characterizes the probabilistic behaviour of a random variable.

In practical problems, we determine the probabilities of numerical outcomes in the same way as for any other probability space. For example, *discrete random variables*, i.e., variables with a finite or at most countable set of values, are completely characterized by the so-called *probability mass function*, that is, by listing the values of the random variable together with the corresponding probabilities. Such variables include, for example, the outcome of a die roll (see Example 5.1), the numerically encoded gender of a newborn child (Examples 5.2, 5.3), the number of incoming telephone calls (Example 7.16), etc. The situation will also be illustrated by the following example.

Example 5.4 (Statistics exam). In the university study system, it is possible to observe the long-run distribution of grades in the course Statistics. A student may receive one of the grades 1, 2, 3, or 4 on the exam. The probability with which a randomly selected student received a particular grade can be clearly represented by a table of values and their corresponding probabilities (see table 5.1). We define a random variable $X = \text{grade in Statistics}$ for a randomly selected student. Then table 5.1 is also a representation of the probability mass function of the random variable X , and hence of its distribution.

Question 5.1. Using Table 5.1 from Example 5.4, create in Excel a bar chart representing the probability mass function of the random variable X .

³This is precisely why a random variable was defined as a *measurable* function.

⁴The probability distribution can be denoted, for example, by $\mu_X = \mathbb{P} \circ X^{-1}$, since from the viewpoint of formal mathematics it is a composition of the measure $\mathbb{P}$ and the mapping X^{-1} . However, we will not need this notation further.

Table 5.1: Grades and probabilities for a randomly selected student

Grade	1	2	3	4
Probability	0.10	0.25	0.50	0.15

For the computation of probabilities, we use the rules introduced in Chapter 2; for example, in Example 5.4 we have

$$\mathbb{P}(X = 2) = 0.25 \quad \text{or} \quad \mathbb{P}(X \geq 3) = 0.50 + 0.15 = 0.65.$$

Note that the sum of the probabilities in Table 5.1 is equal to 1. A rigorous definition of a discrete random variable will be given in Section 5.4. If the number of possible outcomes of a random variable is uncountable, the probability function is defined using integral calculus and will be introduced in Section 5.5. For this purpose, we will also need another alternative complete characterization of the distribution of a random variable, which we now present.

5.3 Cumulative Distribution Function

Definition 5.2 (Cumulative distribution function)

Let X be a random variable. Its cumulative distribution function is defined as the real function F_X of the real variable x given by

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(\{\omega : X(\omega) \leq x\}). \quad (5.4)$$

The cumulative distribution function represents the probability of sets of the specific form $\{\omega : X(\omega) \leq x\}$, which belong to the σ -algebra $\mathcal{A}$, and are therefore events whose probabilities (i.e., values of the cumulative distribution function) can be determined. For this reason, the cumulative distribution function is defined for all $x \in \mathbb{R}$, and it can be shown that, similarly to the probability mass function, it uniquely determines the distribution of the random variable.

Example 5.5 (Ludo). In the game Ludo, a player enters a piece into play when a six is rolled on a die. We represent the die roll as a random experiment with the sample space

$$\Omega = \{\square, \square, \square, \square, \square, \square\}$$

and take the σ -algebra $\mathcal{A}$ to be the system of all subsets of Ω . Since the probabilities of all elementary outcomes are equal ($\frac{1}{6}$), this is a classical

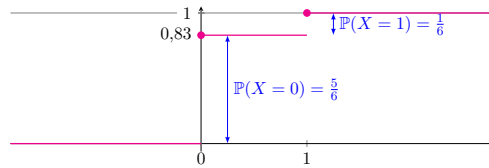


Figure 5.1: Cumulative distribution function of the random variable X from Example 5.5.

probability space, and the probability $\mathbb{P}(A)$ of an event $A \in \mathcal{A}$ is equal to the ratio of the number of favourable outcomes to the total number of outcomes. Now, on this space $(\Omega, \mathcal{A}, \mathbb{P})$, we construct a random variable X representing success in a die roll, which allows a piece to enter the game. We proceed analogously to Example 5.2, using an artificially chosen pair of numbers as follows:

$$X(\omega) = \begin{cases} 0, & \text{if } \omega \in \{\square, \begin{smallmatrix} \square \\ \square \end{smallmatrix}, \begin{smallmatrix} \square & \square \\ \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square & \square \\ \square & \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square & \square & \square \\ \square & \square & \square & \square \end{smallmatrix}\}, \\ 1, & \text{if } \omega = \begin{smallmatrix} \square & \square \\ \square & \square \\ \square & \square \end{smallmatrix}. \end{cases}$$

The probability mass function of the random variable X is given by the values $\mathbb{P}(X = 0) = \frac{5}{6}$ and $\mathbb{P}(X = 1) = \frac{1}{6}$. According to Definition 5.2, the cumulative distribution function F is then defined as follows:

$$F(x) = \begin{cases} 0, & \text{if } x < 0, \\ \frac{5}{6}, & \text{if } 0 \leq x < 1, \\ 1, & \text{if } x \geq 1. \end{cases}$$

The graph of this cumulative distribution function is shown in Figure 5.1.

Note that the cumulative distribution function in Example 5.5 is non-decreasing, since it accumulates probabilities step by step (first $\mathbb{P}(X = 0)$, then $\mathbb{P}(X = 1)$). For this reason, distribution functions are also referred to as *cumulative distribution functions*. The values of a cumulative distribution function lie between zero and one, since they represent probabilities. Finally, the cumulative distribution function of a discrete random variable has a stepwise (jump) character: the jumps occur precisely at the values taken by the random variable (in our case at 0 and 1), and their sizes correspond to the respective probabilities. These properties are characteristic of any cumulative distribution function and will be summarized in Theorem 5.1, to which we will also add information on continuity and improper limits of the cumulative distribution function.

Theorem 5.1 (on properties of the cumulative distribution function). *The cumulative distribution function $F_X(x)$ of a random variable X is*

- a) *non-decreasing, i.e., for any $a, b \in \mathbb{R}$ with $a \leq b$, it holds that $F_X(a) \leq F_X(b)$,*
b) *right-continuous at every point $x \in \mathbb{R}$, i.e., for any $x_0 \in \mathbb{R}$ we have $\lim_{x \rightarrow x_0^+} F_X(x) = F_X(x_0)$,*
c) *$\lim_{x \rightarrow -\infty} F_X(x) = 0$ and $\lim_{x \rightarrow \infty} F_X(x) = 1$,*
d) *has at most countably many points of discontinuity; these points of discontinuity are of the first kind (i.e., jumps), and the size of the jump at a point x_0 is*

$$F_X(x_0) - F_X(x_0^-) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x_0\}),$$

where $F_X(x_0^-) = \lim_{x \rightarrow x_0^-} F_X(x)$.

Proof.

- a) According to Definition 5.2, we have

$$F_X(a) = \mathbb{P}(\{\omega : X(\omega) \leq a\}),$$

$$F_X(b) = \mathbb{P}(\{\omega : X(\omega) \leq b\}).$$

Since $\{\omega : X(\omega) \leq a\} \subset \{\omega : X(\omega) \leq b\}$, by the monotonicity property of probability (Property c)), it follows that

$$\mathbb{P}(\{\omega : X(\omega) \leq a\}) \leq \mathbb{P}(\{\omega : X(\omega) \leq b\}),$$

$$\text{i.e., } F_X(a) \leq F_X(b).$$

- b) Let us choose an arbitrary point $x_0 \in \mathbb{R}$. We need to prove that the function F_X is right-continuous at the point x_0 . Consider an arbitrary sequence $\{x_n\}$ of real numbers such that $x_n \searrow x_0$. If we denote

$$A_n = \{\omega \in \Omega : X(\omega) \leq x_n\}, \quad n = 0, 1, 2, \dots,$$

then

$$F_X(x_n) = \mathbb{P}(A_n), \quad n = 0, 1, 2, \dots,$$

$$A_n \subset A_{n+1}, \quad n = 1, 2, \dots$$

and

$$A_0 = \bigcup_{n=1}^{\infty} A_n.$$

We wish to prove that $\lim_{n \rightarrow \infty} F_X(x_n) = F_X(x_0)$. This can be established by using property h) of probability stated in Theorem 2.2 on

page 25:

$$\begin{aligned}\lim_{n \rightarrow \infty} F_X(x_n) &= \lim_{n \rightarrow \infty} \mathbb{P}(\{\omega : X(\omega) \leq x_n\}) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \\ &= \mathbb{P}(A_0) = F_X(x_0).\end{aligned}$$

$$\text{c) } \lim_{x \rightarrow -\infty} F_X(x) = \lim_{n \rightarrow \infty} F_X(-n) = \lim_{n \rightarrow \infty} \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq -n\}).$$

By property i) of probability on page 25, we have

$$\begin{aligned}\lim_{n \rightarrow \infty} \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq -n\}) &= \\ &= \mathbb{P}\left(\bigcap_{n=1}^{\infty} \{\omega \in \Omega : X(\omega) \leq -n\}\right) = \mathbb{P}(\emptyset) = 0, \\ \lim_{x \rightarrow \infty} F_X(x) &= \lim_{n \rightarrow \infty} F_X(n) = \lim_{n \rightarrow \infty} \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq n\}).\end{aligned}$$

From property h) of probability stated in Theorem 2.2, it follows that

$$\begin{aligned}\lim_{n \rightarrow \infty} \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq n\}) &= \\ &= \mathbb{P}\left(\bigcup_{n=1}^{\infty} \{\omega \in \Omega : X(\omega) \leq n\}\right) = \mathbb{P}(\Omega) = 1.\end{aligned}$$

d) Let us denote by C the set of discontinuity points of the given distribution function, C_n the set of discontinuity points with jumps greater than $\frac{1}{n}$. Then $C = \bigcup_{n=2}^{\infty} C_n$. Since the distribution function $F_X(x)$ takes values between zero and one for every $x \in \mathbb{R}$ and is a non-decreasing function, the set C_n contains at most $(n-1)$ discontinuity points. Therefore, the set C is a union of countably many finite sets, which implies that C is at most countable. Finally, for the jump at x_0 , we compute

$$F_X(x_0) - F_X(x_0^-) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x_0\}).$$

$$\begin{aligned}\mathbb{P}(\{\omega \in \Omega : X(\omega) = x_0\}) &= \\ &= \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x_0\}) - \mathbb{P}(\{\omega \in \Omega : X(\omega) < x_0\}) = \\ &= F_X(x_0) - \mathbb{P}(\{\omega \in \Omega : X(\omega) < x_0\}) = \\ &= F_X(x_0) - \mathbb{P}\left(\bigcup_{n=1}^{\infty} \left\{\omega \in \Omega : X(\omega) \leq x_0 - \frac{1}{n}\right\}\right).\end{aligned}$$

By property i) of probability stated in Theorem 2.2 on page 25, it follows that

$$\begin{aligned} \mathbb{P} \left(\bigcup_{n=1}^{\infty} \left\{ \omega \in \Omega : X(\omega) \leq x_0 - \frac{1}{n} \right\} \right) &= \\ &= \lim_{n \rightarrow \infty} \mathbb{P} \left(\left\{ \omega \in \Omega : X(\omega) \leq x_0 - \frac{1}{n} \right\} \right) = F_X(x_0^-). \end{aligned}$$

□

We now return to the announced formulations of discrete and continuous random variables.

5.4 Discrete Random Variables

Definition 5.3 (Discrete random variable)

A random variable X is called discrete, or a random variable with a discrete probability distribution, if there exists a sequence of real numbers $\{x_n\}$ and a corresponding sequence of non-negative numbers $\{p_n\}$ such that

$$p_n = \mathbb{P}(X = x_n),$$

and it holds that

$$\sum_{n=1}^{\infty} p_n = 1.$$

The probability that a discrete random variable X takes the value x is denoted by $p_X(x)$ or simply $p(x)$; thus $p_X(x) = \mathbb{P}(X = x)$ is the probability mass function of the random variable X , as introduced in Section 5.2. The cumulative distribution function of a discrete random variable is piecewise constant, except at the points x_n , where it has jumps of size $p_n = \mathbb{P}(X = x_n)$. Formally, it has the form

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{\{n: x_n \leq x\}} \mathbb{P}(X = x_n) = \sum_{\{n: x_n \leq x\}} p_n, \quad (5.5)$$

in which the reader can identify the cumulative nature of probabilities. Among other uses, the cumulative distribution function serves to compute the probability of the event $X \in (a; b]$, i.e.

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a) = \sum_{\{n: a < x_n \leq b\}} \mathbb{P}(X = x_n) = \sum_{\{n: a < x_n \leq b\}} p_n$$

for any real numbers a, b such that $a \leq b$.

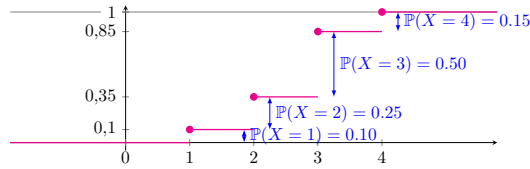


Figure 5.2: Graph of the cumulative distribution function for Example 5.6

Example 5.6 (statistics exam). For the random variable X from Example 5.4, whose probability mass function is given in Table 5.1, construct the cumulative distribution function and compute the probabilities that a randomly selected student

- receives grade 4, i.e., fails the exam;
- passes the exam with a grade from 1 to 3;
- receives grade 3 or 4;
- belongs to the “average of the cohort” with grades 2 or 3.

Solution. The cumulative distribution function F_X of the random variable X is piecewise constant with jumps at the values 1, 2, 3, and 4 of sizes equal to the corresponding probabilities. The graph of the cumulative distribution function is shown in Figure 5.2.

For computing probabilities of general discrete random variables, the function `=PROB` in MS Excel can be used. The arguments of the function are the range of values of the random variable (in our case the assigned grades, which we may assume are in cells A2:A5), the corresponding probabilities (cells B2:B5), and the bounds of the interval for which we compute the probability. Thus, we can easily obtain that

- $\mathbb{P}(X = 4)$ `= PROB(A2:A5,B2:B5,4)` = 0.15;
- $\mathbb{P}(X \leq 3)$ `= PROB(A2:A5,B2:B5, ,3)` = 0.85;
- $\mathbb{P}(X \geq 3)$ `= PROB(A2:A5,B2:B5,3,4)` = 0.65;
- $\mathbb{P}(2 \leq X \leq 4)$ `= PROB(A2:A5,B2:B5,2,4)` = 0.90.

In this simple example, of course, the results can be easily verified by manual computation. Also note that

- $\mathbb{P}(X = 4) = p_X(4)$,
- $\mathbb{P}(X \leq 3) = F_X(3)$,
- $\mathbb{P}(X \geq 3) = 1 - \mathbb{P}(X < 3) = 1 - \mathbb{P}(X \leq 2) = 1 - F_X(2)$,
- $\mathbb{P}(2 \leq X \leq 4) = \mathbb{P}(1 < X \leq 4) = F_X(4) - F_X(1)$.

△

Table 5.2: Values of $p(x)$ and $F(x)$ for Example 5.7

x	0	1	2	3	4	5
$p(x)$	0.03125	0.15625	0.3125	0.3125	0.15625	0.03125
$F(x)$	0.03125	0.1875	0.5	0.8125	0.96875	1

Example 5.7.  Let us consider a discrete random variable X representing the number of heads obtained in five coin tosses. Determine its probability mass function and cumulative distribution function, and construct their graphs.



Solution. Repeated coin tossing is an example of a Bernoulli process (see Section 3.3); the probability that in n tosses exactly k heads occur is therefore given by formula (3.10), where $p = q = \frac{1}{2}$. The probability mass function thus has the form

$$p(x) = \mathbb{P}(X = x) = \begin{cases} \binom{5}{x} \cdot \left(\frac{1}{2}\right)^x \cdot \left(\frac{1}{2}\right)^{5-x} & \text{for } x \in \{0, 1, \dots, 5\}, \\ 0 & \text{otherwise.} \end{cases}$$

The given values can be easily computed in Excel using combinatorial formulas, for example $\mathbb{P}(X = 2)$ as `= COMBIN(5,2)*0.5^2*0.5^(5-2)`, where the value 2 in this formula is in practice replaced by a reference to the corresponding cell of the prepared table of values (the first row of Table 5.2), so that the formula can be copied for all other values of x ⁵. The computed values of the probability mass function are also given in Table 5.2. Finally, we verify that this is indeed a probability distribution:

$$\sum_{k=0}^5 p_k = \sum_{k=0}^5 \binom{5}{k} \cdot \left(\frac{1}{2}\right)^k \cdot \left(\frac{1}{2}\right)^{5-k} = 0.03125 + 0.15625 + \dots + 0.03125 = 1.$$

The cumulative distribution function is then given by the cumulative sum of the above probabilities:

$$F(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{for } x < 0, \\ \sum_{k=0}^{\lfloor x \rfloor} \binom{5}{k} \cdot \left(\frac{1}{2}\right)^k \cdot \left(\frac{1}{2}\right)^{5-k} & \text{for } 0 \leq x \leq 5, \\ 1 & \text{for } x > 5. \end{cases}$$

⁵In Section 7.4 we will present an even simpler formula for computing the values of this probability mass function.

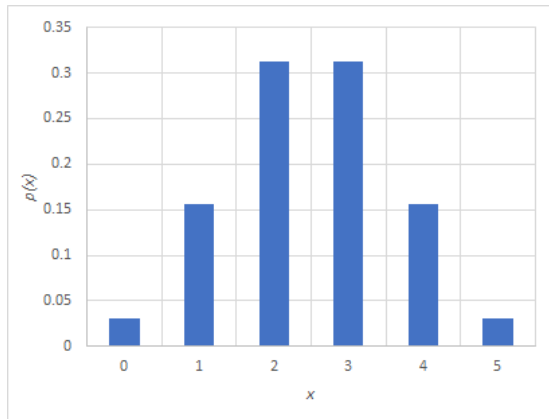


Figure 5.3: Graph of the probability mass function in Example 5.7

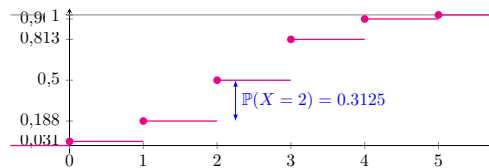


Figure 5.4: Graph of the probability mass function in Example 5.7

In Table 5.2, we have also included the values of the cumulative distribution function for $x = 0, 1, \dots, 5$.

The probability mass function is typically represented by a bar chart, as shown in Figure 5.3. Such a chart can be created, for example, in MS Excel by selecting a column chart from the Insert tab and modifying the horizontal axis labels (in Chart Design, choose Select Data and, for Horizontal (Category) Axis Labels, replace the default values with the range of cells containing the values x). The graph of the cumulative distribution function is shown in Figure 5.4⁶. $\triangle$

5.5 Absolutely Continuous Random Variables

If the number of possible outcomes of a random variable X is uncountable, it is not possible to use the “classical” summation formulas for computing probabilities, since these would diverge to infinity for uncountably many

⁶Graphs of cumulative distribution functions of random variables can be created in MS Excel only approximately and in a relatively complicated way; therefore, we present them to the reader already prepared in another graphical software (specifically PGF/TikZ).

values. Instead of sums, we use a (Lebesgue) integral for the computation, and the role of the probability mass function is taken over by the so-called *probability density function*.

Definition 5.4 (Absolutely continuous random variables)

A random variable X , or equivalently its cumulative distribution function F_X , is called absolutely continuous (a. s.) if there exists a non-negative integrable function f_X such that for any $x \in (-\infty, \infty)$ it holds that

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f_X(t) dt. \quad (5.6)$$

The function f_X is called the probability density function.

Let us briefly recall some properties of the Lebesgue integral. Definition 5.4 ensures that a primitive function exists for the density f_X . In view of (5.6), one such function is the cumulative distribution function (namely the one whose limit at $-\infty$ is zero, as required by Property c) in Theorem 5.1). Furthermore, the probability that $X \in (a, b]$ for some real numbers a, b , is, by Property d) in Theorem 2.2, given by

$$\mathbb{P}(X \in (a, b]) = \mathbb{P}(X \in (-\infty, b]) - \mathbb{P}(X \in (-\infty, a]) = F_X(b) - F_X(a),$$

which is nothing but the rule for computing the definite (Newton) integral of the density f_X over the interval with bounds a, b ,

$$\mathbb{P}(X \in (a, b]) = \int_a^b f_X(x) dx,$$

which can also be visualized (in the Riemann⁷ sense) as the area under the graph of the density f_X between the points a and b (see also Figure 5.5). From this geometric interpretation it follows, among other things, that

$$\mathbb{P}(X = c) = \mathbb{P}(X \in [c, c)) = \int_c^c f_X(x) dx = 0$$

for any real value $c \in \mathbb{R}$ (the “area” under a single point c is, of course, zero).

The fact that the probability of each individual point is zero also implies that the cumulative distribution function has no jump at that point and is therefore continuous there; hence it is continuous over the entire real line $\mathbb{R}$. However, property (5.6) is even stronger, as it ensures differentiability

⁷Georg Friedrich Bernhard Riemann (1826–1866), German mathematician.

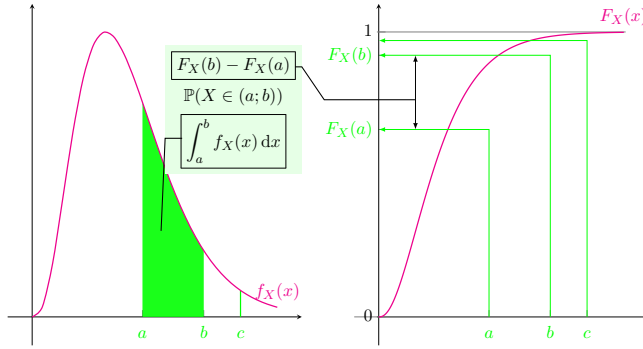


Figure 5.5: Example of a probability density function and cumulative distribution function of an absolutely continuous random variable and the computation of probability

at all points except possibly on an at most countable set of probability zero, at which the first derivative need not be defined. Such functions are called *absolutely continuous*, which also explains the name of the corresponding probability distributions.

As a consequence, when studying probabilities of an absolutely continuous random variable X , we do not need to consider all of its values within a given interval – the corresponding integral remains unchanged if we remove finitely or at most countably many points from the interval. For example, the probabilities

$$\mathbb{P}(X \in (a, b]) = \mathbb{P}(X \in (a, b)) = \mathbb{P}(X \in [a, b)) = \mathbb{P}(X \in [a, b])$$

are all equal for this reason.

In probability theory, special importance is attached to sets of values V (let us briefly call this “property V ”) that a random variable attains with probability 1, i.e.,

$$\mathbb{P}(X \text{ has property } V) = 1.$$

From the above discussion it follows that we may omit (or modify) at most countably many values from the set corresponding to property V without changing this probability of one⁸. In such cases, we say that X has property V “almost surely” (or “almost everywhere”, i.e., except on a set of probability zero), and we use the abbreviation a. s.

The properties of the probability density function derived above can now be summarized in Theorem 5.2.

⁸Consequently, the omitted values have probability zero.

Theorem 5.2 (On the properties of the density). *Let $f_X(x)$ be the probability density function of a random variable X . Then the following holds:*

- a) $f_X(x) = \frac{d}{dx}F_X(x)$ a. s.,
- b) $\int_{-\infty}^{\infty} f_X(x) dx = 1$,
- c) $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x) dx$ for any real numbers a, b such that $a \leq b$.

Proof. All the properties stated follow from Definition 5.4 and from the properties of the cumulative distribution function. $\square$

Example 5.8. Consider a hypothetical population of fish. It is known that the cumulative distribution function of the fish depends on the square of their lifetime and that no fish lives longer than 10 years. That is, $F(x)$ is given by the relation

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0, \\ (c \cdot x)^2 & \text{for } 0 < x \leq 10, \ c \in \mathbb{R}^+, \\ 1 & \text{for } x > 10. \end{cases}$$



Let X be a random variable representing the age of a fish at the time of death.

- a) Determine the constant c such that $F(x)$ is the cumulative distribution function of the random variable X .
- b) Compute the density function $f(x)$ of the lifetime distribution in the fish population.
- c) Calculate the probability that a fish dies between the ages of 3 and 4.

Solution.

- a) It must hold that $F(10) = 1$, hence $(c \cdot 10)^2 = 1$, which implies $c = \frac{1}{10}$, and therefore

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0, \\ \frac{x^2}{100} & \text{for } 0 < x \leq 10, \\ 1 & \text{for } x > 10. \end{cases}$$

- b) $f(x) = \left(\frac{x^2}{100}\right)' = \frac{x}{50}$, hence

$$f(x) = \begin{cases} \frac{x}{50} & \text{for } 0 < x \leq 10, \\ 0 & \text{otherwise.} \end{cases}$$

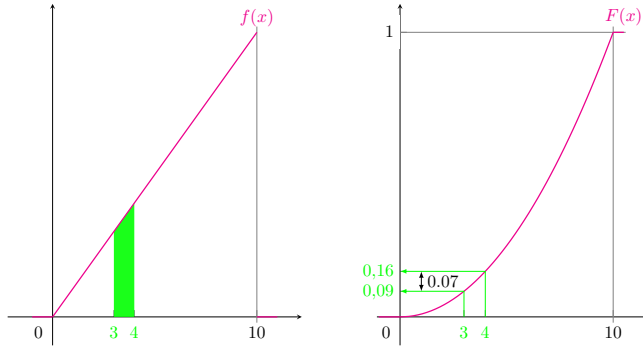


Figure 5.6: Probability density function, cumulative distribution function, and the computation of probability for Example 5.8

c) The probability can be computed either from the density

$$\mathbb{P}(3 \leq X \leq 4) = \int_3^4 f(x) dx = \int_3^4 \frac{1}{50} x dx = \frac{1}{50} \left[\frac{x^2}{2} \right]_3^4 = 0.07,$$

or directly from the cumulative distribution function

$$\mathbb{P}(3 \leq X \leq 4) = F(4) - F(3) = \frac{16}{100} - \frac{9}{100} = 0.07.$$

The computation is illustrated in Figure 5.6.

△

Let us note once again that both approaches to solving the last question of Example 5.8 are equivalent: as mentioned above, the cumulative distribution function is one of the primitive functions of the probability density. Therefore, computing the probability from the density (mathematically, evaluating a Riemann definite integral) is necessarily equivalent to computing it using the cumulative distribution function (mathematically, evaluating a Newton integral), since, due to the properties of the density function, both integrals exist and must therefore be equal.

Let us now present the solution of a similar example, but starting from a given probability density function.

Example 5.9. Determine the constant c such that the function

$$f(x) = \begin{cases} c \cdot x^2 \cdot e^x & \text{for } 0 \leq x \leq 1, \\ 0 & \text{otherwise,} \end{cases}$$

is a probability density function of some random variable X . Compute the cumulative distribution function $F(x)$.

Solution. We need to determine the constant c such that

$$\int_{-\infty}^{\infty} f(x) dx = 1.$$

From the definition of the function $f(x)$ and by repeated use of integration by parts, we obtain

$$\int_{-\infty}^{\infty} f(x) dx = \int_0^1 c \cdot x^2 \cdot e^x dx = c \cdot (e - 2) = 1, \text{ that is, } c = \frac{1}{e - 2}.$$

Thus, the probability density function of the random variable X is

$$f(x) = \begin{cases} \frac{1}{e-2} \cdot x^2 \cdot e^x & \text{for } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

The cumulative distribution function on the interval $[0, 1]$ is

$$\begin{aligned} \int_0^x \frac{1}{e-2} \cdot t^2 \cdot e^t dt &= \frac{1}{e-2} \cdot [t^2 \cdot e^t - 2te^t + 2e^t]_0^x = \\ &= \frac{1}{e-2} \cdot (x^2 e^x - 2x e^x + 2e^x - 2). \end{aligned}$$

The complete expression for the cumulative distribution function is therefore

$$F(x) = \begin{cases} 0 & \text{for } x < 0, \\ \frac{1}{e-2} \cdot (x^2 e^x - 2x e^x + 2e^x - 2) & \text{for } 0 \leq x < 1, \\ 1 & \text{for } x \geq 1. \end{cases}$$

The cumulative distribution function and the probability density function are shown in Figure 5.7. $\triangle$

5.6 Generalization*

For the purpose of measuring the “size” of sets, measure theory introduces the concept of a *set measure*. A measure is a σ -additive set function defined on a space $(\Omega, \mathcal{A})$, i.e.

- a) $\mu: \mathcal{A} \rightarrow [0, \infty]$,
- b) $\mu(\emptyset) = 0$,
- c) if $A_n \in \mathcal{A}$, $n \geq 1$ are pairwise disjoint, then $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$.

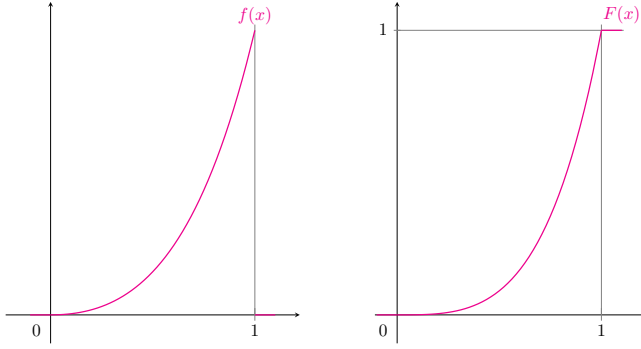


Figure 5.7: Density and cumulative distribution function for Example 5.9

If, in addition, $\mu(\Omega) = 1$, we say that μ is a *probability measure*.

For the purpose of characterizing the distribution of a random variable as a mapping into the space $(\mathbb{R}, \mathcal{B})$, we therefore introduce such a probability measure on this space: to each Borel set $B \in \mathcal{B}$ we can assign a probability measure on $(\mathbb{R}, \mathcal{B})$

$$\mu_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\}),$$

which we call the probability distribution of the random variable X . In the special case $B = (-\infty, x]$, we obtain the cumulative distribution function

$$\mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}) = F_X(x).$$

We see that there is a one-to-one correspondence between the cumulative distribution function and the probability distribution.

If we set $B = (a, b]$ for $-\infty < a \leq b < \infty$, we obtain

$$\mathbb{P}(X \in (a, b]) = F_X(b) - F_X(a) = \mu_F([a, b)).$$

This relation uniquely defines the Lebesgue–Stieltjes measure induced by the cumulative distribution function F , or equivalently by the random variable X . The above relation is defined only for intervals; however, this is sufficient to uniquely determine the measure for all Borel sets. That is,

$$\mathbb{P}(X \in B) = \mu_{F_X}(B) = \int_B d\mu_{F_X} \text{ for any random event } B \in \mathcal{B},$$

and

$$\int_{\Omega} \phi(X(\omega)) d\mathbb{P}(\omega) = \int_{\mathbb{R}} \phi(x) d\mu_{F_X}(x)$$

for any measurable function ϕ for which at least one of the integrals exists.

The integral with respect to the Lebesgue–Stieltjes measure is sometimes written in a simplified form as

$$\int_{\mathbb{R}} \phi(x) \, dF_X(x).$$

Since it is not our intention to present measure theory in this text, we will restrict ourselves to an intuitive explanation of the above definitions.

If the measure λ is defined by the relation

$$\lambda([a, b]) = b - a, \quad -\infty < a \leq b < \infty,$$

then the measure λ is called the *Lebesgue measure*. This measure is not finite and hence is not a probability measure. The integral with respect to the Lebesgue measure, $\int \phi(x) \, d\lambda(x)$, is a generalization of the Riemann integral $\int \phi(x) \, dx$. Therefore, we can work with it in an analogous way.

If we define the measure μ on a bounded interval $[A, B]$ by the relation

$$\mu([a, b]) = \frac{b - a}{B - A}, \quad A \leq a \leq b \leq B,$$

then μ is already a probability measure.

The integral with respect to the measure μ is then computed as

$$\int \phi(x) \, d\mu(x) = \int \phi(x) \frac{1}{B - A} I_{[A, B]}(x) \, dx.$$

In both of these cases, all values x have the same weight. The Lebesgue–Stieltjes integral additionally allows us to assign different weights to different values of x , determined by the increments of the cumulative distribution function F_X . If X is the random variable defined in Example 5.5, then the cumulative distribution function F_X has only two jumps and no other increments. The integral with respect to μ_{F_X} is in fact

$$\int \phi(x) \, d\mu_{F_X}(x) = \frac{5}{6}\phi(0) + \frac{1}{6}\phi(1).$$

Conversely, if F_X has no jumps and is continuous at every point, then F_X has a derivative f at every point, which represents the increment of F_X . The integral with respect to μ_{F_X} is then simply

$$\int \phi(x) \, d\mu_{F_X}(x) = \int \phi(x) f(x) \, dx.$$

The Lebesgue–Stieltjes integral is thus a generalization of both the integral, where the values x have different weights, and sums, where the individual terms also have different weights. These considerations lead us to classify random variables into two basic types: discrete and absolutely continuous random variables. However, combinations of these types may also occur.

Example 5.10. A rocket with a targeting system hits at a distance of 0 to 30 meters from the targeted point, and let us assume that the distances by which it misses the target are all equally likely (so this is a case of geometric probability). We are targeting the center of a building with a diameter of 10 meters. The rocket causes 100% damage if it hits the building in any way, and the damage continuously decreases by 10% for each meter away from the building. What is the probability that after the rocket is fired, the building will be damaged by at least 50%?

Solution. Let us first introduce an auxiliary random variable Y with values in the interval $[0, 30]$, representing the deviation of the rocket's impact point from the targeted point, measured in metres. Next, let us introduce a random variable X representing the damage caused, expressed in percentage points. Clearly, we have:

$$X = \begin{cases} 0 & \text{if } Y \in [15, 30], \\ 10 \cdot (15 - Y) & \text{if } Y \in (5, 15), \\ 100 & \text{if } Y \in [0, 5]. \end{cases}$$

According to the principles of geometric probability (see Section 2.3 on page 30), we have

$$\begin{aligned} \mathbb{P}(X = 0) &= \mathbb{P}(Y \in [15, 30]) = \frac{15}{30} = \frac{1}{2}, \\ \mathbb{P}(X = 100) &= \mathbb{P}(Y \in [0, 5]) = \frac{5}{30} = \frac{1}{6}. \end{aligned}$$



The cumulative distribution function F_X of the random variable X therefore has two jumps, at the points 0 and 100, of sizes $\frac{1}{2}$ and $\frac{1}{6}$, respectively. At all other points, the cumulative distribution function is continuous, and

$$\mathbb{P}(X \in (0, 100)) = \mathbb{P}(Y \in (5, 15)) = 1 - \left(\frac{1}{2} + \frac{1}{6} \right) = \frac{1}{3},$$

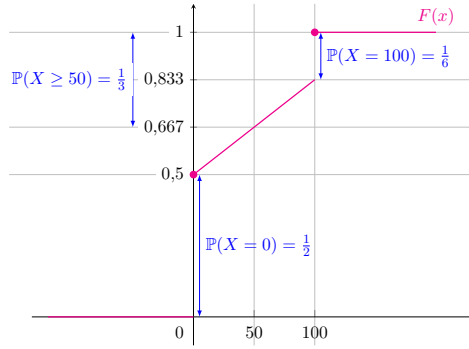


Figure 5.8: Cumulative distribution function for Example 5.10

and hence

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0, \\ \frac{1}{2} + \frac{x}{300} & \text{for } 0 \leq x < 100, \\ 1 & \text{for } x \geq 100. \end{cases}$$

The probability that the building will be damaged by at least 50% is therefore

$$\mathbb{P}(X \geq 50) = \mu_F([50, 100]) = \int_{50-}^{100+} d\mu_F = \int_{50}^{100} \frac{1}{300} dx + \frac{1}{6} = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

The complete computation is illustrated in Figure 5.8. $\triangle$

Let us now also present the existence theorem for the cumulative distribution function, i.e., a theorem that tells us for which real-valued function there exists a probability space and a random variable defined on it such that the given function is its cumulative distribution function.

Theorem 5.3. *Let the function $F: \mathbb{R} \rightarrow \mathbb{R}$ satisfy properties 1), 2), and 3) of Theorem 5.1. Then there exists a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a random variable X defined on it such that*

$$F_X(x) = F(x).$$

Proof. Let us set

$$\Omega = \mathbb{R}, \quad \mathcal{A} = \mathcal{B}, \quad \mathbb{P} = \mu_F,$$

where μ_F is the Lebesgue–Stieltjes measure induced by the function F . Let us also define $X(\omega) = \omega$. We now have

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \leq x) = \mu_F((-\infty, x]) = \lim_{a \rightarrow -\infty} \mu_F((a, x]) \\ &= \lim_{a \rightarrow -\infty} (F(x) - F(a)) = F(x), \quad x \in \mathbb{R}, \end{aligned}$$

where we have used the continuity of the finite measure μ_F and property (b) of the cumulative distribution function. $\square$

5.7 Exercises

5.1 Let the set of elementary outcomes be $\Omega = \{\omega_1, \omega_2\}$, and let the σ -algebra $\mathcal{A}$ be defined on Ω as follows:

$$\mathcal{A} = \{\emptyset, \{\omega_1\}, \{\omega_2\}, \Omega\}.$$

Determine whether the function X given by $X(\omega_1) = 1$, $X(\omega_2) = 0$ is a random variable.

5.2 Let us roll a die once. The set of elementary outcomes is

$$\Omega = \{\square, \begin{smallmatrix} \square \\ \square \end{smallmatrix}, \begin{smallmatrix} \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square & \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square & \square & \square & \square \end{smallmatrix}\}, \quad (5.7)$$

we define the σ -algebra of subsets of the set Ω as follows:

$$\mathcal{A} = \{\emptyset, \{\begin{smallmatrix} \square \\ \square \end{smallmatrix}\}, \{\begin{smallmatrix} \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square & \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square & \square & \square & \square \end{smallmatrix}\}, \Omega\}.$$

Find a function that will be a random variable with respect to $\mathcal{A}$.

5.3 A popular game in the USA is *Chuck-a-luck*, which is played with three standard dice with faces numbered from 1 to 6. Michal plays a game with Honza in which Michal stakes 1 € and chooses a particular outcome, for example that a face with three dots (i.e., the number three) appears on the die. If, after rolling the three dice, this does not occur on any die, Michal loses his stake. If the number three appears on exactly one die, Honza pays Michal 2 €; if it appears on exactly two dice, Michal receives 3 €; and if it appears on all three dice, Michal receives 4 €.

Describe a random variable representing the amount of net gain with respect to the number of occurrences of the number three, and determine the probability distribution of this random variable.⁹

5.4 The random variable X has the cumulative distribution function

$$F_X(x) = \begin{cases} 0 & \text{for } x < 1, \\ (x-1)^2 & \text{for } 1 \leq x < 2, \\ 1 & \text{for } x \geq 2. \end{cases}$$

- a) Determine the probability density function of the random variable X .

⁹According to [6], pp. 281–282.

- b) Represent graphically the density and the cumulative distribution function of the random variable X .
- c) Determine $\mathbb{P}(1.5 < X < 1.75)$.

5.5 The cumulative distribution function is given

$$F(x) = a + b \cdot \arctan \frac{x}{2}, \quad -\infty < x < \infty.$$

Determine:

- a) for which values of a, b the function F is a cumulative distribution function,
- b) the probability density function f ,
- c) $\mathbb{P}(\alpha < X < \beta)$.

5.6 The graph of the probability density function of the random variable X is a broken line formed by three points $A = [0, 0]$, $B = [5, y_B]$, and $C = [10, 0]$. Find the coordinate y_B of point B such that this function is a probability density. Determine the formula and construct the graphs of the probability density function $f(x)$ and the cumulative distribution function $F(x)$ defined by the given broken line.

5.7 The random variable X has the probability density function

$$f(x) = \begin{cases} 0 & \text{for } x < -4, \\ \frac{1}{\pi} \cdot \frac{1}{\sqrt{4^2 - x^2}} & \text{for } -4 \leq x < 4, \\ 0 & \text{for } x \geq 4. \end{cases}$$

Determine the cumulative distribution function F_X .

5.8 The probability density function of the random variable X has the form

$$f(x) = \begin{cases} 0 & \text{for } x \leq -1, \\ \frac{a}{\sqrt{1-x^2}} & \text{for } -1 < x < 1, \\ 0 & \text{for } x \geq 1. \end{cases}$$

Determine:

- a) coefficient a ,
- b) $\mathbb{P}(-\frac{1}{2} < X < \frac{1}{2})$.

5.9 Determine for which values of a, b the function

$$F(x) = \frac{1}{1 + e^{-(a+bx)}}, \quad x \in \mathbb{R}$$

is a cumulative distribution function.

Solutions to the Exercises

5.1 Yes

5.2 For example, $X(\begin{smallmatrix} \blacksquare & \cdot \\ \cdot & \cdot \end{smallmatrix}) = b$, $X(\begin{smallmatrix} \blacksquare & \\ & \cdot \end{smallmatrix}) = X(\begin{smallmatrix} \cdot & \cdot \\ \cdot & \cdot \end{smallmatrix}) = X(\begin{smallmatrix} \cdot & \cdot \\ \cdot & \cdot \end{smallmatrix}) = X(\begin{smallmatrix} \cdot & \cdot \\ \cdot & \cdot \end{smallmatrix}) = X(\begin{smallmatrix} \cdot & \cdot \\ \cdot & \cdot \end{smallmatrix}) = c$, where b, c are arbitrary real constants.

5.3 $X: \Omega \rightarrow \mathbb{R}$, $\Omega = \{(a, b, c) \in \{\begin{smallmatrix} \blacksquare & \cdot \\ \cdot & \cdot \end{smallmatrix}, \begin{smallmatrix} \cdot & \cdot \\ \cdot & \cdot \end{smallmatrix}, \dots, \begin{smallmatrix} \cdot & \cdot \\ \cdot & \cdot \end{smallmatrix}\}^3\}$, $|\Omega| = 6^3 = 216$.

$$X(a, b, c) = \begin{cases} 3 & \text{for } a = b = c = \begin{smallmatrix} \blacksquare & \cdot \\ \cdot & \cdot \end{smallmatrix}, \\ 2 & \text{if exactly two of the } a, b, c \text{ are } \begin{smallmatrix} \blacksquare & \cdot \\ \cdot & \cdot \end{smallmatrix}, \\ 1 & \text{if exactly one of the } a, b, c \text{ is } \begin{smallmatrix} \blacksquare & \cdot \\ \cdot & \cdot \end{smallmatrix}, \\ -1 & \text{otherwise.} \end{cases}$$

$$p(x) = \begin{cases} \frac{1}{216} \approx 0.005 & \text{for } x = 3, \\ \frac{15}{216} \approx 0.069 & \text{for } x = 2, \\ \frac{75}{216} \approx 0.347 & \text{for } x = 1, \\ \frac{125}{216} \approx 0.579 & \text{otherwise.} \end{cases}$$

5.4

a) $f(x) = \begin{cases} 2x - 2 & \text{for } 1 \leq x \leq 2, \\ 0 & \text{otherwise,} \end{cases}$

b) see figure 5.9,

c) $\mathbb{P}(1.5 < X < 1.75) = 0.3125$.

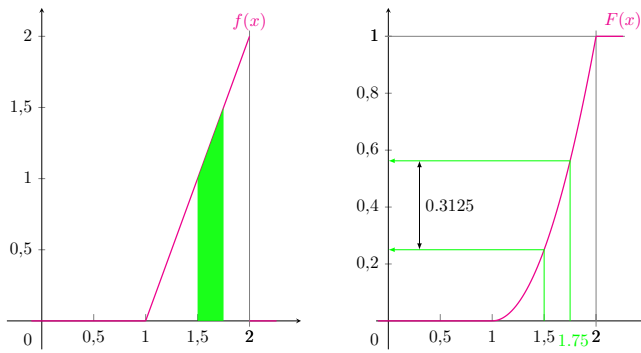


Figure 5.9: Graph of the probability density function and the cumulative distribution function for Problem 5.4

5.5 a) $a = \frac{1}{2}$, $b = \frac{1}{\pi}$, b) $f_X(x) = \frac{2}{\pi} \cdot \left(\frac{1}{4+x^2} \right)$,
 c) $\mathbb{P}(\alpha < X < \beta) = \frac{1}{\pi} \cdot \left(\arctan \frac{\beta}{2} - \arctan \frac{\alpha}{2} \right)$.

This is the so-called Cauchy distribution,¹⁰ which is used, for example, in finance to model extreme investment risks. The graphs of the density and the cumulative distribution function are shown in Figure 5.10.

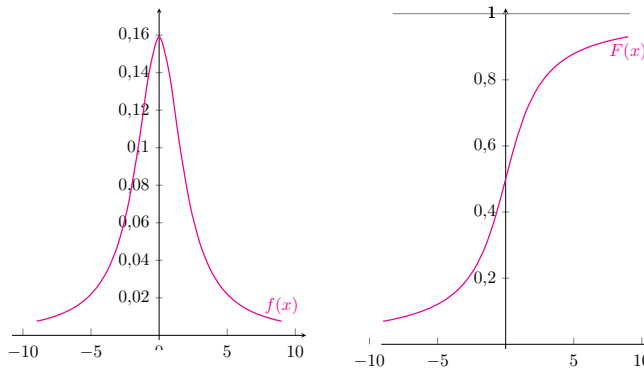


Figure 5.10: Graph of the probability density function and the cumulative distribution function for Problem 5.5

¹⁰Augustin-Louis Cauchy (1789–1857), French mathematician.

5.6 $y_B = 0, 2,$

$$f(x) = \begin{cases} 0.04x & \text{for } 0 \leq x < 5, \\ 0.4 - 0.04x & \text{for } 5 \leq x < 10, \\ 0 & \text{otherwise.} \end{cases}$$

$$F(x) = \begin{cases} 0 & \text{for } x < 0, \\ 0.02x^2 & \text{for } 0 \leq x < 5, \\ 0.4x - 0.02x^2 - 1 & \text{for } 5 \leq x < 10, \\ 1 & \text{for } x \geq 10. \end{cases}$$

This is the so-called triangular distribution. It is used, for example, in simulation studies for variables for which we know only the range (lower and upper bounds) of values they may take, and their most probable value (mode). The graphs of the probability density function and the cumulative distribution function are shown in Figure 5.11.

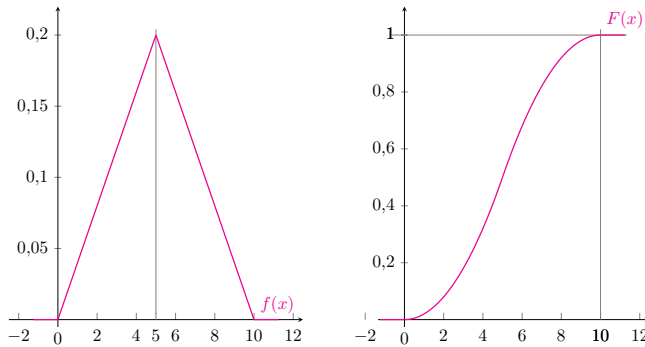


Figure 5.11: Graph of the probability density function and the cumulative distribution function for Problem 5.6

5.7

$$F_X(x) = \begin{cases} 0 & \text{for } x < -4, \\ \frac{2}{\pi} \arcsin \sqrt{\frac{x+4}{8}} & \text{for } -4 \leq x < 4, \\ 1 & \text{for } x \geq 4. \end{cases}$$

This is the so-called arcsine distribution (on a general interval, see Figure 5.12), which is used, for example, in signal processing.

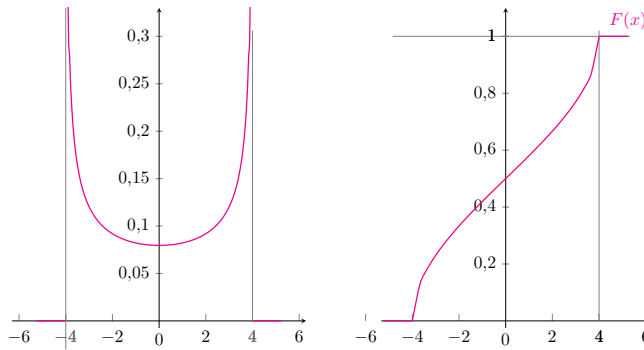


Figure 5.12: Graph of the probability density function and the cumulative distribution function for Problem 5.7

5.8 a) $a = \frac{1}{\pi}$, b) $\frac{1}{3}$. This is also the so-called arcsine distribution.

5.9 $a \in \mathbb{R}$, $b > 0$. This is the so-called logistic distribution, which is used, for example, in modelling random variables representing categorical characteristics.

Chapter 6

Characteristics of Random Variables

The probabilistic behavior of a random variable is completely described by the probability mass function, the probability density function, or the distribution function. A complete characterization of a random variable can, however, sometimes be structurally complicated or, in some cases, even unavailable (e.g., it is poorly or very imprecisely estimated). In practical applications, however, it is often sufficient to summarize the information about the distribution of a random variable into a few suitable numerical characteristics that sufficiently accurately describe the basic properties of this distribution. In such cases, a complete characterization of a random variable by means of the probability mass function or density and the distribution function is not even needed. In this chapter, we will deal only with the so-called characteristics of location (expected value, quantile, median, mode) and characteristics of variability (variance, standard deviation), and with methods for their computation.

6.1 Expected value of a random variable and its properties

Definition 6.1 (Expected value of a random variable)

Let X be a random variable defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and let F_X be its cumulative distribution function. The expected value $\mathbb{E}X$ of the random variable X is defined as the integral

$$\mathbb{E}X = \int_{-\infty}^{\infty} x \, dF_X(x), \tag{6.1}$$

if this integral exists.

The expected value $\mathbb{E}X$ of the random variable X is also often called the expectation and represents a number that characterizes the location of the values of the random variable on the number line with respect to their probabilities. The integral in formula (6.1) is of the Lebesgue–Stieltjes type (see Section 5.6). We now present a special form of Definition 6.1 in the case where the random variable has a discrete or, respectively, an absolutely continuous distribution.

Theorem 6.1.

- a) Let X be a discrete random variable taking values x_i for $i = 1, 2, \dots$ with probabilities $\mathbb{P}(X = x_i) = p_i$. Then the expected value $\mathbb{E}X$ of the random variable X is given by

$$\mathbb{E}X = \sum_{i=1}^{\infty} x_i \cdot p_i, \quad (6.2)$$

if the series in (6.2) converges.

- b) Let X be an absolutely continuous random variable with density f_X . Then the expected value of the random variable X is given by

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f_X(x) dx, \quad (6.3)$$

if the integral converges.

Proof. See Section 5.6. □

Note that the expected value is a weighted sum of the values of the random variable. The role of the weights is played by the probability mass function in the discrete case, and by the probability density function in the continuous case.

Example 6.1. For the discrete random variable X representing the number of heads obtained in five coin tosses (see Example 5.7), compute its expected value $\mathbb{E}X$.

Solution. The expected value is computed according to the formula (6.2)

$$\mathbb{E}X = \sum_{i=1}^5 x_i \cdot p_i = 0 \cdot 0.03125 + 1 \cdot 0.15625 + \dots + 5 \cdot 0.03125 = 2.5.$$

This result can also be easily computed in Excel using the dot product function. For example, if the values of the random variable (the first row of Table 5.2) are stored in cells B1 to G1 and the corresponding probabilities (the second row of Table 5.2) in cells B2 to G2, then the expected value $\mathbb{E}X$ is given by the function `=SUMPRODUCT(B1:G1,B2:G2)`. $\triangle$

Example 6.2. Let us consider an absolutely continuous random variable X with density $f(x) = \ln x$ for $x \in [1, e]$ and 0 otherwise. Compute the expected value $\mathbb{E}X$.

Solution. First, let us verify that $f(x)$ is a density function, i.e., that $\int_{-\infty}^{\infty} f(x) dx = 1$:

$$\int_{-\infty}^{\infty} f(x) dx = \int_1^e \ln x dx = [x \ln x - x]_1^e = 1.$$

The expected value $\mathbb{E}X$ is, according to formula (6.3), equal to the value of the integral

$$\int_1^e x \cdot \ln x dx = \left[\frac{x^2}{2} \cdot \left(\ln x - \frac{1}{2} \right) \right]_1^e = \frac{e^2}{4} + \frac{1}{4} \approx 2.097.$$

$\triangle$

Example 6.3. Consider an absolutely continuous random variable X with density $f(x) = \frac{1}{\pi(1+x^2)}$ (this is the so-called Cauchy distribution). Compute $\mathbb{E}X$.

Solution. First, it is appropriate to verify that the function f is indeed a probability density function, i.e., that

$$\int_{-\infty}^{\infty} f(x) dx = \int_{-\infty}^{\infty} \frac{1}{\pi} \frac{1}{1+x^2} dx = \frac{1}{\pi} [\arctan x]_{-\infty}^{\infty} = 1.$$

We thus see that f is a density. We now attempt to compute $\mathbb{E}X$ according to the formula (6.3):

$$\mathbb{E}X = \int_{-\infty}^{\infty} \frac{1}{\pi} \frac{x}{1+x^2} dx.$$

The integral on the right-hand side of the formula, however, diverges. This probability distribution is therefore an example of a distribution for which the expected value does not exist. $\triangle$

The expected value of a random variable is, by Definition 6.1, a Lebesgue–Stieltjes integral and thus shares all its properties. Therefore, the following statements hold, which we present without proof.

Theorem 6.2 (Properties of the expected value). *Let $X, Y, X_n, n = 1, 2, \dots$ be random variables on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and let a, b be real constants. Then*

a) *the expected value of a constant is the constant:*

$$\mathbb{E}a = a,$$

b) *the expected value is absolutely integrable:*

$$\mathbb{E}X < \infty \quad \text{if and only if} \quad \mathbb{E}|X| < \infty,$$

c) *the expected value is a linear operator:*

$$\mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y,$$

d) *if the expected value of the sum of random variables X_i exists, then it is given by the sum of the expected values of the random variables X_i :*

$$\mathbb{E}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathbb{E}X_i,$$

e) *the expected value is non-negative for non-negative random variables:*

$$\text{if } X \geq 0 \text{ a. s., then } \mathbb{E}X \geq 0,$$

f) *the expected value is monotonic:*

$$\text{if } X_1 \leq X \leq X_2 \text{ a. s., then } \mathbb{E}X_1 \leq \mathbb{E}X \leq \mathbb{E}X_2,$$

g)

$$|\mathbb{E}X| \leq \mathbb{E}|X|,$$

h)

$$\text{if } |X| \leq Y \text{ a. s., } \mathbb{E}Y < \infty, \quad \text{then } \mathbb{E}X < \infty,$$

i) *term-by-term integration:*

$$\text{if } \sum_{n=1}^{\infty} \mathbb{E}|X_n| < \infty, \text{ then } \mathbb{E}\left(\sum_{n=1}^{\infty} X_n\right) = \sum_{n=1}^{\infty} \mathbb{E}X_n,$$

j) *Fatou's¹ lemma holds:*

$$\text{if } X_n \geq 0 \text{ a. s., then } \mathbb{E}(\liminf_{n \rightarrow \infty} X_n) \leq \liminf_{n \rightarrow \infty} \mathbb{E}X_n.$$

¹Pierre Fatou (1878–1929), French mathematician.

Although Properties a)–j) follow from the general properties of the Lebesgue–Stieltjes integral, they can also be proved directly in the case of discrete and absolutely continuous random variables. If X is a discrete random variable, we use the properties of sums. In the case that X is an absolutely continuous random variable, we use the properties of the Riemann or Lebesgue integral.

Example 6.4. According to mortality tables in the USA (1978–1979), the probability that a 32-year-old woman dies within one year is 0.001819. An insurance company offers women of this age that for an annual premium of 100 USD it will pay 25 000 USD to the beneficiaries in the event of the insured person's death. What profit can the insurance company expect if it concludes such insurance contracts with 5000 women of the given age?

Solution. We denote the profit (or loss) of the insurance company in the case of a single contract by a random variable X_i , $i = 1, \dots, 5000$. Its expected value is

$$\mathbb{E}X_i = 100 \cdot 0.998181 - 24\,900 \cdot 0.001819 = 99.8181 - 45.2931 = 54.525.$$

If the insurance company concludes 5 000 such contracts, its expected profit is equal to the expected value of the random variable

$$Y = \sum_{i=1}^{5000} X_i,$$

therefore

$$\mathbb{E}Y = \mathbb{E}\left(\sum_{i=1}^{5000} X_i\right) = 5\,000 \cdot \mathbb{E}X_i = 272\,625.$$

The expected profit of the insurance company is 272 625 USD. $\triangle$

If we define a new random variable Y as a function of an existing random variable X , i.e., $Y = \phi(X)$, then it is possible to use the distribution of X to compute the expected value of Y , as stated in the following theorem. Therefore, in this case it is not necessary to derive the distribution of the random variable Y in order to compute its expected value.

Theorem 6.3. *Let X be a random variable and let $\phi: \mathbb{R} \rightarrow \mathbb{R}$. Then it holds that*

$$\mathbb{E}\phi(X) = \int_{-\infty}^{\infty} \phi(x) \, dF_X(x),$$

if at least one of the integrals converges.

If the random variable X has a discrete distribution $\{x_n, p_n\}_{n \in \mathbb{N}_0}$, then

$$\mathbb{E}\phi(X) = \sum_{n \in \mathbb{N}_0} \phi(x_n)p_n,$$

If the random variable X has an absolutely continuous distribution with density f , then

$$\mathbb{E}\phi(X) = \int_{-\infty}^{\infty} \phi(x)f(x) \, dx,$$

if at least one of the integrals converges.

6.2 Other measures of location of random variables

The expected value of a random variable is classified among the properties of random variables that are collectively referred to as measures of location. Other characteristics of this type are derived from the so-called quantile function.

Definition 6.2 (Quantile function, quantile)

- a) Let F be the cumulative distribution function of a random variable X . We define the function F^{-1} by

$$F^{-1}(u) = \inf\{x : F(x) \geq u\}, \quad 0 < u < 1. \quad (6.4)$$

Then F^{-1} is called the quantile function corresponding to the cumulative distribution function F .

- b) The values of the function $F^{-1}(u)$ are called quantiles. In particular, the α -quantile is the value $F^{-1}(\alpha)$.

From the definition it follows that if F is increasing and continuous, then the quantile function F^{-1} is the inverse function of F . This is also the origin of the notation F^{-1} , which is, however, used in the remaining cases as well². Quantiles are also often expressed in percentages, since α is a number between zero and one. In this context, they are also called *percentiles*.

²If the cumulative distribution function is constant on an interval (a, b) , the inverse function is not defined for the corresponding value, since it cannot be uniquely determined: in principle, any point of this interval may be chosen as the quantile. The infimum formula (6.4) defines the lower bound as the quantile, but a supremum version can also be found in textbooks.

The most important quantile is the so-called *median*; it is denoted by $\tilde{X}$ or $\text{Me}(X)$. The median is the 50% quantile, i.e., $F^{-1}(0.50)$. It divides the values of the random variable into two parts with equal probability 50%:

$$\mathbb{P}(X \leq \tilde{X}) \geq \frac{1}{2} \quad \text{and} \quad \mathbb{P}(X \geq \tilde{X}) \geq \frac{1}{2}. \quad (6.5)$$

The median is used wherever weighting values by probability (i.e., the use of expected values) fails. This typically occurs in the case of distributions with heavy tails, such as the Cauchy distribution from Example 6.3 (in which the expected value does not exist at all), or distributions modelling situations with occurrences of extreme values – if such an extreme value has a non-negligible probability, it pulls the expected value towards itself. A typical example is the average wage of a randomly selected employee in a company: a single high managerial salary significantly distorts the expected value of the wages of the remaining employees towards a level that most employees of the company do not reach. In such cases, the median is a better measure of location of the random variable: in the case of wages, it gives the value that exactly one half of the employees attain. Let us also note that the median is not uniquely determined by condition (6.5) (see Exercise 6.2).

It is, of course, also possible to divide the values of a random variable into parts of unequal size using quantiles: for example, the 25% quantile $F^{-1}(0.25)$ divides the values of the random variable into the lower quarter and the upper three quarters (in terms of probability). This quantile is called the *lower quartile*. Similarly, the *upper quartile* is the 75% quantile $F^{-1}(0.75)$. Smaller or larger quantiles are used in statistics, in particular in the construction of interval estimates (see Chapter 14) and in hypothesis testing (see Chapter 15).

Example 6.5. Let us denote the waiting time of an angler for a catch (in minutes) by the random variable X . Determine the expected value, the lower quartile, the median, and the upper quartile of the waiting time, if the random variable X has the probability density function of the form

$$f(x) = \begin{cases} e^{-x} & \text{for } 0 < x < \infty, \\ 0 & \text{otherwise.} \end{cases} \quad (6.6)$$

Solution. This distribution, whose density is given by formula (6.6), is called the exponential distribution (see Section 8.2 on page 153 for more details). The expected value is determined by integration by parts

$$\mathbb{E}X = \int_0^\infty x \cdot e^{-x} dx = [x \cdot (-e^{-x})]_0^\infty - \int_0^\infty (-e^{-x}) dx = 0 + [-e^{-x}]_0^\infty = 1.$$

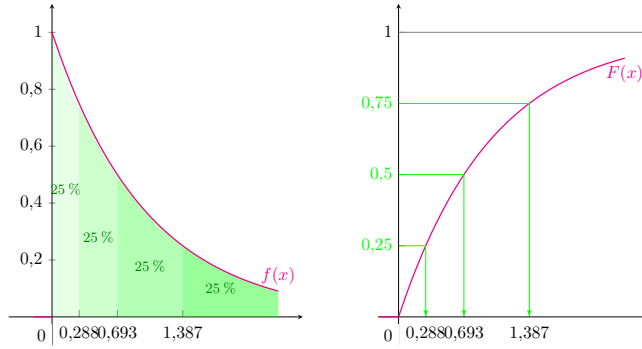


Figure 6.1: First, second, and third quartile of the random variable X from Example 6.5

To compute the lower quartile, median, and upper quartile, we first determine the cumulative distribution function; it suffices for $x > 0$:

$$F(x) = \int_0^x e^{-t} dt = [-e^{-t}]_0^x = 1 - e^{-x}$$

(for $x \leq 0$, the function $F(x)$ is equal to zero). We then derive the quantile function as the inverse function to $F(x)$:

$$F^{-1}(u) = -\ln(1 - u).$$

Substituting the values $u = 0.25$, 0.5 , and 0.75 , we obtain

$$F^{-1}(0.25) \approx 0.288, \quad F^{-1}(0.5) \approx 0.693, \quad F^{-1}(0.75) \approx 1.387.$$

The values are given in minutes; by multiplying the results by 60, we obtain the corresponding times in seconds. The meaning of these quantiles is illustrated in Figure 6.1.

△

The last measure of location of the distribution of a random variable X that we mention here is the so-called *mode*, which is usually denoted by $\hat{X}$, or alternatively by $\text{Mod}(X)$ or $\text{Mo}(X)$. Roughly speaking, the mode represents the “most probable” value that a random variable can take. If the discrete distribution is concentrated at the points $x_1, x_2, \dots$, then the mode $\hat{X}$ is the value for which

$$\mathbb{P}(X = \hat{X}) \geq \mathbb{P}(X = x_i)$$

holds for all $i = 1, 2, \dots$. In other words, it is the value at which the probability mass function $p(x)$ attains its maximum. For an absolutely continuous

random variable, the mode $\hat{X}$ is the value at which the probability density function $f(x)$ attains its maximum, i.e.,

$$f(\hat{X}) \geq f(x)$$

for all $x \in (-\infty, \infty)$. The mode need not be uniquely determined either; a random variable may have two or more modes (see Exercise 6.3).

Example 6.6. Determine the mode $\hat{X}$ of the following random variables:

a) of a discrete random variable X with probability distribution

$$p_n = \begin{cases} \left(\frac{1}{2}\right)^n & \text{for } n = 1, 2, \dots, \\ 0 & \text{otherwise,} \end{cases}$$

b) of a continuous random variable with probability density function

$$f(x) = \begin{cases} \frac{x^2 e^{-x}}{2} & \text{for } x \in (0; \infty), \\ 0 & \text{otherwise.} \end{cases}$$

Solution.

a) We see that

$$p_1 = \frac{1}{2}, \quad p_2 = \frac{1}{4}, \dots$$

Since for increasing n the probability p_n decreases geometrically, the mode is $\hat{X} = 1$.

b) To determine $\hat{X}$ in the case of a continuous distribution, it is necessary to find the maximum of the density. We therefore examine the behaviour of the density function $f(x)$. Since at the endpoints of the interval $(0, \infty)$ the limit of $f(x)$ is equal to zero, we investigate the points at which $f'(x) = 0$.

$$f'(x) = xe^{-x} - \frac{x^2 e^{-x}}{2} = xe^{-x} \left(1 - \frac{x}{2}\right) = 0.$$

The only such point in the interior of the interval $(0, \infty)$ is $x = 2$. Since

$$f''(2) = -e^{-2} < 0,$$

this is indeed a maximum, and hence the mode is $\hat{X} = 2$.

△

6.3 Variance of a Random Variable

The expected value, median, quantiles, and other measures of location provide information about where we can expect the values of a random variable to lie; however, by themselves they say nothing about how far or close these values are with respect to the measure of location. This information is provided by the so-called measures of dispersion, or variability. The basic characteristic of this type is the variance of a random variable.

Definition 6.3 (Variance of a random variable)

The mean squared deviation from the expected value of a random variable, i.e.,

$$\text{var } X = \mathbb{E}(X - \mathbb{E}X)^2,$$

is called the variance of the random variable X .

The variance is the second most important characteristic of a random variable. From its definition, we see that the existence of the expected value is a necessary condition for the existence of the variance; therefore, for example, the Cauchy distribution from Example 6.3 cannot have a variance in addition to lacking an expected value. The variance of a random variable is always non-negative and is equal to zero if and only if $\mathbb{P}(X = c) = 1$, where c is some constant. The variance has a number of useful mathematical properties, which we summarize in the following theorem.

Theorem 6.4 (on properties of the variance). *Let X be a random variable and let a, c be real constants. Then*

a) the so-called computational formula for the variance holds:

$$\text{var } X = \mathbb{E}(X^2) - (\mathbb{E}X)^2, \quad (6.7)$$

b) $\text{var } c = 0$,

c) $\text{var}(aX) = a^2 \text{var } X$,

d) $\text{var}(X + c) = \text{var } X$.

e) If X has a finite expected value and a finite non-zero variance, and

$$Y = \frac{X - \mathbb{E}X}{\sqrt{\text{var } X}},$$

then $\mathbb{E}Y = 0$ and $\text{var } Y = 1$.

Proof.

$$\begin{aligned} \text{a) } \quad \text{var } X &= \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}[X^2 - 2X\mathbb{E}X + (\mathbb{E}X)^2] = \\ &= \mathbb{E}X^2 - 2(\mathbb{E}X)^2 + (\mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2. \end{aligned}$$

- b) Since $\mathbb{E}c = c$, it necessarily follows that $\text{var}(c - \mathbb{E}c)^2 = \text{var}(c - c)^2 = 0$.
 c)
$$\begin{aligned}\text{var}(aX) &= \mathbb{E}(aX - \mathbb{E}aX)^2 = \mathbb{E}[a(X - \mathbb{E}X)]^2 = \\ &= a^2 \mathbb{E}(X - \mathbb{E}X)^2 = a^2 \text{var } X.\end{aligned}$$

Parts d) and e) are left without proof. $\square$

A drawback of the variance is that it expresses the variability of a random variable in the squares of its units, and therefore the numerical values of the variance are difficult to interpret. For this reason, in practical applications the so-called *standard deviation* is used to characterize the variability of random variables; it is the square root of the variance,

$$\sigma = \sqrt{\text{var } X}.$$

The standard deviation measures the variability of a random variable in its original units. However, it does not share all the advantageous properties of the variance, which we will discuss in Chapter 9 (for example, the standard deviation of the sum of independent random variables is *not* equal to the sum of their standard deviations).

Example 6.7 (continuation of Example 6.5). Let us return to the example of the angler waiting for a catch (see page 107). We denoted the waiting time (in minutes) by the random variable X . Compute the variance and the standard deviation of this random variable X .

Solution. Recall that the random variable X has the probability density function

$$f(x) = \begin{cases} e^{-x} & \text{for } 0 < x < \infty, \\ 0 & \text{otherwise.} \end{cases}$$

We have already shown that $\mathbb{E}X = 1$ minute. The variance is computed using formula (6.7) on page 110. First, we determine $\mathbb{E}X^2$.

$$\begin{aligned}\mathbb{E}X^2 &= \int_0^\infty x^2 \cdot e^{-x} \, dx = [x^2 \cdot (-e^{-x})]_0^\infty - \int_0^\infty -2xe^{-x} \, dx = \\ &= 0 + 2 \cdot \int_0^\infty xe^{-x} \, dx = 2 \cdot \mathbb{E}X = 2.\end{aligned}$$

Hence

$$\text{var } X = 2 - 1^2 = 1.$$

The variance of the random variable X is 1 min^2 , and therefore the standard deviation is 1 minute. $\triangle$

Theorem 6.5 (Chebyshev's inequality). *Let X be a random variable with finite variance. Then for any $\varepsilon > 0$ it holds that*

$$\mathbb{P}[|X - \mathbb{E}X| \geq \varepsilon] \leq \frac{\text{var } X}{\varepsilon^2}. \quad (6.8)$$

Formula (6.8) can be rewritten in an alternative form: if we set $\varepsilon = k \cdot \sigma$, where $k > 0$ and $\sigma = \sqrt{\text{var } X}$ is the standard deviation of the random variable X , then

$$\mathbb{P}[|X - \mathbb{E}X| \geq k \cdot \sigma] \leq \frac{1}{k^2}. \quad (6.9)$$

Inequality (6.9) is called the classical Chebyshev's inequality and holds for any $k > 0$. Formula (6.9) provides a rough symmetric interval estimate of probability (often referred to as a confidence interval) that a random variable with finite variance lies within a prescribed distance from its expected value. This distance is expressed as a multiple k of the standard deviation of the random variable. For estimation purposes, it does not make sense to choose $k \leq 1$. The inequality is independent of the distribution of the random variable X , and thus represents a maximally conservative estimate (in fact, an upper bound on probability). Its essential theoretical significance lies in providing a tool for proving Bernoulli's law of large numbers (see Section 11.2 on page 231).

Example 6.8. Suppose we manage an investment portfolio with an average return of 8 000 CZK and a standard deviation of 500 CZK. What is the maximum possible probability that the actual return differs from the expected value by more than $\pm 10\%$ of the mean return?

Solution. Since the distribution of the portfolio return is unknown, we use Chebyshev's inequality for the estimate, as it does not depend on the distribution. Since 10% of 8 000 CZK is 800 CZK, we want to find

$$\mathbb{P}[|X - 8\,000| \geq 800].$$

According to formula (6.9), we estimate

$$\mathbb{P}[|X - 8\,000| \geq 1.6 \cdot 500] \leq \frac{1}{1.6^2} \approx 0.391.$$

However, if we knew the distribution of the random variable X , we could significantly reduce this estimate. $\triangle$

¹Pafnuty Lvovich Chebyshev (1821–1894), Russian mathematician.

6.4 Higher-order Moments

The definition of the expected value and the variance can be generalized by the following definition.

Definition 6.4 (n -th moment)

For a natural number n , the n -th moment of a random variable X is defined as $\mathbb{E}(X^n)$; the n -th absolute moment as $\mathbb{E}(|X|^n)$; and the n -th central moment as $\mathbb{E}[(X - \mathbb{E}X)^n]$.

Using the Lebesgue–Stieltjes measure (see Section 5.6), the moments defined above can be expressed in general by means of integrals.

$$\begin{aligned}\mathbb{E}(X^n) &= \int_{-\infty}^{\infty} x^n dF_X(x), \\ \mathbb{E}(|X|^n) &= \int_{-\infty}^{\infty} |x|^n dF_X(x), \\ \mathbb{E}[(X - \mathbb{E}X)^n] &= \int_{-\infty}^{\infty} (x - \mathbb{E}X)^n dF_X(x).\end{aligned}$$

Formulas for the special cases are obtained by replacing $dF_X(x)$ with the expression $f(x)dx$ in the absolutely continuous case, and by replacing the integral with a weighted sum with weights p_n in the discrete case, similarly as in Theorem 6.1.

From Definition 6.4 we see that the expected value is the first moment. The first central moment is always equal to zero, since

$$\mathbb{E}(X - \mathbb{E}X) = \mathbb{E}X - \mathbb{E}(\mathbb{E}X) = \mathbb{E}X - \mathbb{E}X = 0.$$

The variance of a random variable is the second central moment. In practical applications, the following two characteristics based on the third and fourth central moments are also commonly used.

Definition 6.5

Let $\mu = \mathbb{E}X$ denote the expected value and $\sigma = \sqrt{\text{var } X}$ the standard deviation of the random variable X . Then

a) the skewness of the random variable X is

$$\mathbb{E}\left(\frac{X - \mu}{\sigma}\right)^3 = \frac{\mathbb{E}(X - \mu)^3}{\sigma^3},$$

b) the kurtosis of the random variable X is

$$\mathbb{E}\left(\frac{X - \mu}{\sigma}\right)^4 = \frac{\mathbb{E}(X - \mu)^4}{\sigma^4}.$$

The coefficient of skewness describes the asymmetry of the density of a random variable. Random variables whose density is symmetric with respect to the vertical axis (passing through the median) have zero skewness. Random variables whose mode lies to the right of the median have negative skewness and are called left-skewed. If the mode lies to the left of the median, the skewness is positive and the random variable is said to be right-skewed.

The coefficient of kurtosis characterizes the concentration of the density around the median. The kurtosis of a standardized random variable with the Gaussian density (see Section 8.5 on page 168) equals 3. Random variables whose probability is more concentrated around the median (i.e., whose density near the median is higher than that of the Gaussian curve) have kurtosis greater than 3 and are called *leptokurtic* (*peaked*). Random variables whose probability mass is, compared to the Gaussian curve, more spread into the tails have kurtosis less than 3 and are called *platykurtic* (*flat*). Random variables with kurtosis equal to 3 are referred to as *mesokurtic*.

Sometimes the coefficient of kurtosis is defined with a shift of -3 relative to the above definition (i.e., platykurtic variables then have a negative value of this adjusted coefficient). In English, this characteristic is called *excess kurtosis*, literally “excess peakedness”. For example, Microsoft Excel uses this adjusted definition of kurtosis.

Chebyshev’s inequality can be generalized to any absolute moment; this is formalized in the following lemma. The lemma will also be used for the postponed proof of Chebyshev’s inequality.

Lemma 6.6. *Let the n -th moment of the random variable X exist, where $n > 0$. Then for any $\varepsilon > 0$ it holds that*

$$\mathbb{P}[|X| \geq \varepsilon] \leq \frac{\mathbb{E}|X|^n}{\varepsilon^n}.$$

Proof.

$$\begin{aligned} \mathbb{E}|X|^n &= \int_{-\infty}^{\infty} |x|^n \, dF_X(x) \geq \int_{|x| \geq \varepsilon} |x|^n \, dF_X(x) \geq \\ &\geq \varepsilon^n \int_{|x| \geq \varepsilon} dF_X(x) = \varepsilon^n \mathbb{P}[|X| \geq \varepsilon]. \end{aligned}$$

□

Proof of Theorem 6.5. In the previous lemma, we set $n = 2$ and consider, instead of the random variable X , the random variable $X - \mathbb{E}X$. □

6.5 Exercises

6.1 Property damage in thousands of CZK is represented by a random variable X that follows a distribution with the probability density function

$$f(x) = \begin{cases} \frac{c}{x^3} & \text{for } x > 3, \\ 0 & \text{otherwise.} \end{cases}$$

- Compute the value of c such that the function is a probability density.
- Do the expected value and the variance exist? If so, compute them.
- What is the probability that the damages do not exceed 10 thousand CZK?

6.2 Find a discrete and a continuous distribution for which the median is not uniquely determined.

6.3 Find a discrete and a continuous distribution for which the mode is not uniquely determined.

6.4 Determine for which values of μ the density attains

$$f(x) = \frac{1}{2\sqrt{2\pi}} \left(e^{-\frac{(x-\mu)^2}{2}} e^{-\frac{(x+\mu)^2}{2}} \right)$$

its maximum.

6.5 The tested device consists of five components. The n -th component fails with probability

$$p_n = 0.2 + 0.1(n - 1).$$

Determine the expected value of the number of failed components.

6.6 n devices are tested. The probability of failure is the same for all devices and is equal to p . Determine the expected value of the number of devices that fail during the test.

6.7 In a lottery, there are m_1 prizes of value k_1 , m_2 prizes of value $k_2, \dots, m_n$ prizes of value k_n . In total, there are N tickets. Determine the price of a ticket so that the expected value of the winnings per ticket is equal to one half of its price.

6.8 An urn contains m white balls and n black balls. Balls are drawn until a white ball is drawn. After each draw, the ball is returned to the urn. Determine the expected value and the variance of the number of balls drawn.

6.9 Determine the expected value and the variance of the random variable X with density

$$f(x) = \frac{1}{\pi\sqrt{a^2 - x^2}}, \quad -a \leq x \leq a.$$

Solutions to the Exercises

6.1 a) $c = 18$, b) $\mathbb{E}X = 6$, $\text{var } X$ does not exist, c) $\mathbb{P}(X \leq 10) = 0.91$. This is the so-called Pareto distribution of type I with shape parameter $\alpha = 2$, which is used in finance precisely for modelling the occurrence of extreme events.

6.2 Discrete distribution: any distribution with an even number of possible values. Continuous distribution: a distribution with a density that is symmetric about zero and such that the density is zero on an interval $(-c, c)$ for some constant $c > 0$. Any number from the interval $(-c, c)$ can serve as a median.

6.3 Discrete distribution: a distribution for which two different values have the same probability, which is simultaneously the maximum value of the probability mass function. Continuous distribution: a distribution with a density that has two peaks of equal height, which is simultaneously the maximum value of the density.

6.4 For $\mu = 0$, the maximum is attained at a single point.

6.5 Let $Y = X_1 + \cdots + X_5$ denote the number of failed components, where X_n is a random variable taking the value 1 if the n -th component fails and 0 otherwise.

$$EY = EX_1 + \cdots + EX_5 = \sum_{n=1}^5 [0.2 + 0.1(n-1)] \cdot 1 = 2.$$

6.6 $\mathbb{E}X = np$

6.7

$$\frac{2}{N} (m_1 k_1 + \cdots + m_n k_n)$$

6.8

$$\mathbb{E}X = \frac{m+n}{m}, \quad \text{var } X = \frac{1 - \frac{m}{m+n}}{\left(\frac{m}{n+m}\right)^2} = \frac{n}{m^2}.$$

This is the so-called geometric distribution with success parameter $p = \frac{m}{n+m}$.

6.9 $\mathbb{E}X = 0$, $\text{var } X = \frac{a^2}{2}$. This is the so-called arcsine distribution.

Chapter 7

Selected distributions of discrete random variables

7.1 Discrete uniform distribution

Let D be a set with n elements ($n \in \mathbb{N}$). A random variable X , for which each value from the set D occurs with the same probability p , has a discrete uniform distribution $\text{DU}(n)$ with parameter n , denoted by $X \sim \text{DU}(n)$. The discrete uniform distribution is also known as a “distribution with equally likely outcomes”, and thus represents the model of classical probability in terms of random variables.

For further considerations, we will work with a set of outcomes of a special form¹ $D = \{1, 2, \dots, n\}$, for which we explicitly derive the basic properties of the random variable X . The probability mass function and the cumulative distribution function have the following form:

$$p(x) = \begin{cases} \frac{1}{n} & \text{for } x \in D, \\ 0 & \text{otherwise.} \end{cases} \quad F(x) = \begin{cases} 0 & \text{for } x < 1, \\ \lfloor x \rfloor \cdot \frac{1}{n} & \text{for } 1 \leq x < n \\ 1 & \text{for } x \geq n, \end{cases}$$

where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x . The expected value and the variance are given by

$$\mathbb{E}X = \frac{n+1}{2} \quad \text{and} \quad \text{var } X = \frac{n^2-1}{12},$$

(see exercise 7.1).

¹For other sets D , the same could be done, but the resulting formulas would be more complicated.

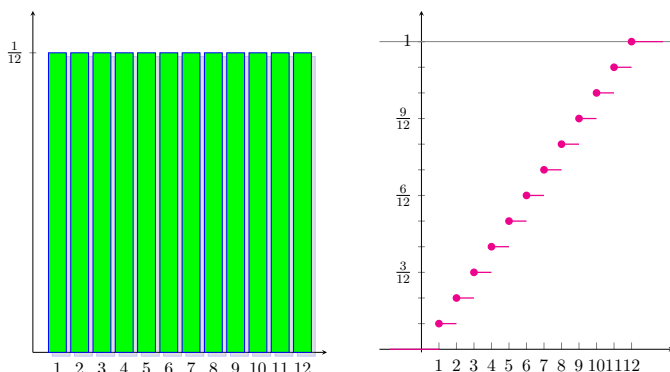


Figure 7.1: Graph of the probability mass function and the cumulative distribution function of the discrete uniform distribution $\text{DU}(12)$

Example 7.1. Consider rolling a regular dodecahedral die whose faces are numbered from 1 to 12.

- Describe the probability distribution of the random variable X representing the outcome (i.e., the number on the top face after the die lands), using the probability mass function and the cumulative distribution function.
- Construct the graphs of the probability mass function and the cumulative distribution function.
- Compute the expected value and the variance.
- Compute the probability that
 - the upper face shows an even number (event A),
 - the upper face shows a prime number (event B).

Solution. All possible outcomes of the die roll are represented by the set $D = \{1, 2, \dots, 12\}$ (thus $n = 12$), and they are all equally likely, each with probability $\frac{1}{12}$.

- The probability mass function and the cumulative distribution function have the form

$$p(x) = \begin{cases} \frac{1}{12} & \text{for } x \in D, \\ 0 & \text{otherwise,} \end{cases} \quad F(x) = \begin{cases} 0 & \text{for } x < 1, \\ \lfloor x \rfloor \cdot \frac{1}{12} & \text{for } 1 \leq x < 12, \\ 1 & \text{for } x \geq 12. \end{cases}$$

- The graphs of the probability mass function and the cumulative distribution function are shown in Figure 7.1.
- $\mathbb{E}X = 6.5$, $\text{var } X = \frac{143}{12}$.

d) The corresponding probabilities are

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(X = 2) + \mathbb{P}(X = 4) + \cdots + \mathbb{P}(X = 12) = 0.5, \\ \mathbb{P}(B) &= \mathbb{P}(X = 2) + \mathbb{P}(X = 3) + \mathbb{P}(X = 5) + \mathbb{P}(X = 7) + \mathbb{P}(X = 11) = \\ &= \frac{5}{12}.\end{aligned}$$

△

7.2 Degenerate distribution

A random variable X that takes a constant value c with probability 1, i.e., $\mathbb{P}(X = c) = 1$, and any other value with probability 0, i.e., $\mathbb{P}(X \neq c) = 0$, has a degenerate distribution $\text{Dg}(c)$ with parameter c , denoted by $X \sim \text{Dg}(c)$. The degenerate distribution is also known as the Dirac² measure (centered) at the point c and is denoted by δ_c . Its significance lies mainly in the fact that it allows one to treat a non-random quantity – whose value is known in advance – as if it were a random variable. For example, a degenerate distribution arises for a random variable representing the number of pips obtained when throwing a die whose faces all have six pips.

7.3 Bernoulli (Alternative) Distribution

In Section 3.3 we considered a special type of random experiment called a Bernoulli trial. A Bernoulli trial has only two possible outcomes, usually referred to as “success” and “failure”. Examples include a success in a coin toss, success (rolling a six) in the game Ludo, detecting a defective product during quality control, the birth of a boy (or a girl, depending on the reader’s preference), or whether a randomly selected household has access to high-speed internet, etc.



In practice, it is convenient to describe this experiment in terms of random variables using the so-called *Bernoulli distribution*, also known as the *alternative distribution* or the *0–1 distribution*³.

²Paul Dirac (1902–1984), British mathematician and physicist.

³The term Bernoulli distribution is internationally widespread, as it directly relates to the Bernoulli trial. In Poland and some other countries, the term 0–1 distribution is used, reflecting the values of the random variable. The term alternative distribution is traditional in Czech, Slovak, and German literature.

Assume that the probability of success in a Bernoulli trial is given by a parameter p , where we assume $0 < p < 1$ ⁴. We introduce a random variable X with two possible values 0 and 1, representing failure and success in the Bernoulli trial as follows:

$$X = \begin{cases} 0 & \text{if failure occurs,} \\ 1 & \text{if success occurs in the Bernoulli trial.} \end{cases}$$

It is easy to show that the probability mass function and the cumulative distribution function of this distribution have the form

$$p(x) = \begin{cases} 1-p & \text{for } x = 0, \\ p & \text{for } x = 1, \\ 0 & \text{otherwise,} \end{cases} \quad F(x) = \begin{cases} 0 & \text{for } x < 0, \\ 1-p & \text{for } 0 \leq x < 1, \\ 1 & \text{for } x \geq 1. \end{cases}$$

In Exercise 7.3, the reader can verify that the expected value and the variance are given by

$$\mathbb{E}X = p \quad \text{and} \quad \text{var } X = p(1-p).$$

The Bernoulli distribution will be denoted by $B(p)$ ⁵.

Example 7.2. An urn contains one white ball and two black balls. Find the probability mass function and the distribution function of the random variable X , which takes the value 1 if, upon a single draw of exactly one ball from the urn, we drew a white ball, and the value 0 if we drew a black ball. What is the probability that we draw a white ball? Compute the expected value and the variance.

Solution. $X \sim B(p)$, where $p = \frac{1}{3}$ is the probability of drawing a white ball. Therefore,

$$p(x) = \begin{cases} \frac{2}{3} & \text{for } x = 0, \\ \frac{1}{3} & \text{for } x = 1, \\ 0 & \text{otherwise,} \end{cases} \quad F(x) = \begin{cases} 0 & \text{for } x < 0, \\ \frac{2}{3} & \text{for } 0 \leq x < 1, \\ 1 & \text{for } x \geq 1. \end{cases}$$

The graphs of $p(x)$ and $F(x)$ for $X \sim B(\frac{1}{3})$ are shown in Figure 7.2. The expected value and the variance are

$$\mathbb{E}X = \frac{1}{3}, \quad \text{var } X = \frac{1}{3} \cdot \left(1 - \frac{1}{3}\right) = \frac{2}{9}.$$

△

⁴If $p = 0$ or $p = 1$, the distribution would be degenerate.

⁵It is also sometimes denoted by $\text{Alt}(p)$ or $A(p)$ in Czech, Slovak, and German-language publications.

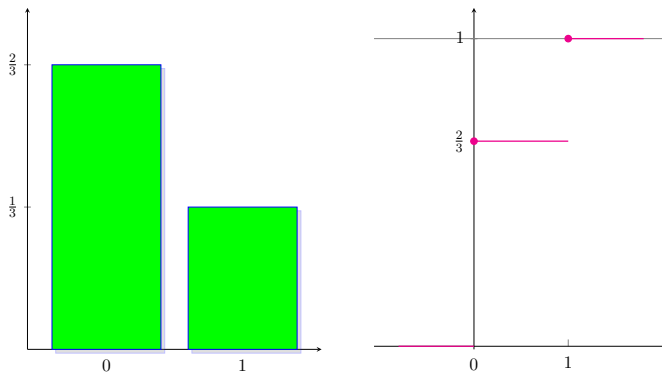


Figure 7.2: Graph of the probability mass function and the cumulative distribution function of the Bernoulli (alternative) distribution $B(1/3)$

7.4 Binomial distribution

We consider a sequence of n independent Bernoulli trials with constant probability of success $p \in (0, 1)$. These can be represented by a sequence of random variables $X_1, \dots, X_n$ with identical Bernoulli (alternative) distributions $X_i \sim B(p)$, $i = 1, \dots, n$. The distribution of the random variable X representing the number of successes in this sequence of independent Bernoulli trials is called the *binomial distribution*. From the nature of this definition, it follows that the random variable X can be represented as the sum

$$X = \sum_{i=1}^n X_i, \quad (7.1)$$

and can take values from the set $\{0, 1, \dots, n\}$. The binomial distribution will be denoted by $\text{Bi}(n, p)$, where n is the number of Bernoulli trials and p is the probability of success.

The probability mass function of the binomial distribution has already been derived in Example 3.8 as formula (3.10), and the cumulative distribution function is given by the cumulative sum of these probabilities. The probability that exactly k successes occur in a sequence of n Bernoulli trials is therefore

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, \dots, n. \quad (7.2)$$

The cumulative distribution function $F(x)$ has the form

$$F(x) = \begin{cases} 0 & \text{for } x < 0, \\ \sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} p^k (1-p)^{n-k} & \text{for } 0 \leq x < n, \\ 1 & \text{for } x \geq n. \end{cases}$$

To compute the values of the probability mass function and the cumulative distribution function, we can use the Excel function `=BINOM.DIST`, as shown in Example 7.3. The last parameter of this function specifies the type of computation: use the value `FALSE` for computing the probability mass function, and `TRUE` for the values of the cumulative distribution function. The first three parameters of the function correspond to the values of the parameters k , n , and p , respectively.

Example 7.3. Construct the graph of the probability mass function of the random variable X with a binomial distribution.

- a) $\text{Bi}(16, 0.1)$,
- b) $\text{Bi}(16, 0.5)$,
- c) $\text{Bi}(16, 0.8)$

and the corresponding graphs of the cumulative distribution functions. For case c), determine the probability that the random variable X takes the value

- 1) exactly 12,
- 2) at most 12,
- 3) less than 12,
- 4) at least 12,
- 5) greater than 12,
- 6) at least 12 and at most 15,
- 7) greater than 12 and less than 15.

Solution. The graphs of the probability mass functions and the cumulative distribution functions are shown in Figure 7.3. The probability calculations follow.

$$1) \mathbb{P}(X = 12) = p(12) = \binom{16}{12} 0.8^{12} 0.2^4 \approx 0.2,$$

$$= \text{BINOM.DIST}(12, 16, 0.8, \text{FALSE})$$

$$2) \mathbb{P}(X \leq 12) = p(0) + \cdots + p(12) = F(12) = \\ = \binom{16}{0} 0.8^0 0.2^{16} + \cdots + \binom{16}{12} 0.8^{12} 0.2^4 \approx 0.402,$$

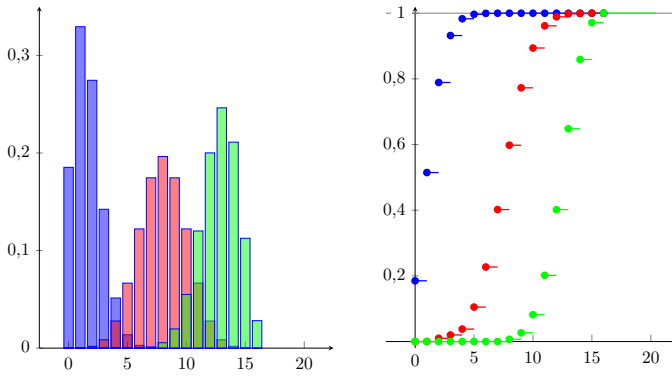


Figure 7.3: Graphs of the probability mass functions and cumulative distribution functions of the binomial distributions $\text{Bi}(16, 0.1)$ (blue), $\text{Bi}(16, 0.5)$ (red), $\text{Bi}(16, 0.8)$ (green)

`= BINOM.DIST(12,16,0.8,TRUE)`

$$\begin{aligned} 3) \quad \mathbb{P}(X < 12) &= p(0) + \cdots + p(11) = F(11) = \\ &= \binom{16}{0} 0.8^0 0.2^{16} + \cdots + \binom{16}{11} 0.8^{11} 0.2^5 \approx 0.202, \end{aligned}$$

`= BINOM.DIST(11,16,0.8,TRUE)`

$$\begin{aligned} 4) \quad \mathbb{P}(X \geq 12) &= p(12) + \cdots + p(16) = 1 - F(11) = \\ &= 1 - \left[\binom{16}{0} 0.8^0 0.2^{16} + \cdots + \binom{16}{11} 0.8^{11} 0.2^5 \right] \approx 0.798, \end{aligned}$$

`= 1-BINOM.DIST(11,16,0.8,TRUE)`

$$\begin{aligned} 5) \quad \mathbb{P}(X > 12) &= p(13) + \cdots + p(16) = 1 - F(12) = \\ &= 1 - \left[\binom{16}{0} 0.8^0 0.2^{16} + \cdots + \binom{16}{12} 0.8^{12} 0.2^4 \right] \approx 0.598, \end{aligned}$$

`= 1-BINOM.DIST(12,16,0.8,TRUE)`

$$\begin{aligned} 6) \quad \mathbb{P}(12 \leq X \leq 15) &= p(12) + \cdots + p(15) = F(15) - F(11) = \\ &= \binom{16}{12} 0.8^{12} 0.2^4 + \cdots + \binom{16}{15} 0.8^{15} 0.2^1 \approx 0.770, \end{aligned}$$

`= BINOM.DIST(15,16,0.8,TRUE)-BINOM.DIST(11,16,0.8,TRUE)`

$$\begin{aligned} 7) \quad \mathbb{P}(12 < X < 15) &= p(13) + p(14) = F(14) - F(12) = \\ &= \binom{16}{13} 0.8^{13} 0.2^3 + \binom{16}{14} 0.8^{14} 0.2^2 \approx 0.457, \end{aligned}$$

`= BINOM.DIST(14,16,0.8,TRUE)-BINOM.DIST(12,16,0.8,TRUE)`

or also as the sum

```
= BINOM.DIST(14,16,0.8,FALSE) and
= BINOM.DIST(13,16,0.8,FALSE).
```

△

The name binomial distribution arises from the fact that the probabilities $p_k = \mathbb{P}(X = k)$ are the terms of the binomial expansion of the expression

$$1^n = [p + (1 - p)]^n = \sum_{k=0}^n \binom{n}{k} p^k (1 - p)^{n-k} = \sum_{k=0}^n p_k. \quad (7.3)$$

From formula (7.3), it is immediately clear that the probabilities p_k satisfy the conditions for a probability distribution, namely

a) non-negativity $p_k \geq 0$ for any k , and

b) normalization $\sum_{k=0}^n p_k = 1$.

A random variable X with distribution $\text{Bi}(n, p)$ has expected value np and variance $np(1 - p)$. This can be derived directly from the definitions:

$$\begin{aligned} \mathbb{E}X &= \sum_{k=0}^n k p_k = \sum_{k=0}^n k \binom{n}{k} p^k (1 - p)^{n-k} = \\ &= np \sum_{k=1}^n \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} = \\ &= np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{n-1-j} = \\ &= np(p + (1-p))^{n-1} = np. \end{aligned}$$

To compute the variance, we use the computational formula (6.7), from which

$$\text{var } X = \mathbb{E}X(X - 1) + \mathbb{E}X - (\mathbb{E}X)^2.$$

From the definition, we obtain the required value

$$\begin{aligned} \mathbb{E}X(X - 1) &= \sum_{k=0}^n k(k-1) \binom{n}{k} p^k (1-p)^{n-k} = \\ &= n(n-1)p^2 \sum_{k=2}^n \frac{(n-2)!}{(k-2)!(n-k)!} p^{k-2} (1-p)^{n-k} = \\ &= n(n-1)p^2 \sum_{j=0}^{n-2} \binom{n-2}{j} p^j (1-p)^{n-2-j} = \\ &= n(n-1)p^2. \end{aligned}$$

Substituting, we finally obtain

$$\text{var } X = n(n-1)p^2 + np - (np)^2 = np(1-p).$$

In deriving the expected value, one can also make use of property d) of Theorem 6.2 on the sum of expected values, together with the fact that, according to (7.1), the random variable X can be represented as such a sum. A similar property holds for the variance (Theorem 9.2); however, it is based on the independence of random variables, which we will discuss later in Section 9.4.

Example 7.4. If the probability of the birth of a boy is 0.515, what is the probability that among four consecutively born children there will be

- a) the first two boys and the next two girls,
- b) exactly two boys?
- c) Determine how many children must be born so that the probability that at least one of them is a boy is greater than or equal to 0.99.

For solving all parts, we assume the simplifying condition that the sexes of consecutively born children are independent events⁶.

Solution.

- a) Since the order of births of boys and girls is fixed, the required probability is

$$0.515^2 \cdot 0.485^2 \approx 0.062.$$

- b) In this case, the order of birth does not matter; we must consider all possibilities, of which there are $\binom{4}{2}$. The resulting probability is

$$p_2 = \binom{4}{2} \cdot 0.515^2 \cdot 0.485^2 \approx 0.374.$$

This probability can also be obtained by introducing a random variable X representing the number of boys born; it follows a binomial distribution $\text{Bi}(4, 0.515)$. The probability $p_2 = \mathbb{P}(X = 2) \approx 0.374$ can be computed in MS Excel.

`= BINOM.DIST(2,4,0.515,FALSE)`

- c) The event complementary to the required event “at least one boy is born” is “no boy is born”, i.e., “ n girls are born”, whose probability is

$$\mathbb{P}(n \text{ girls are born}) = (1 - 0.515)^n = 0.485^n.$$

⁶This assumption is usually not valid in the case of twins.

Therefore, the number of children must, according to the requirements of the problem, satisfy the inequality

$$1 - (0.485)^n \geq 0.99,$$

from which we easily obtain

$$n \geq \log_{0.485} 0.01 = \frac{\log 0.01}{\log 0.485} \approx 6.47.$$

Hence, at least seven children are needed.

△

Example 7.5. Suppose that the probability of growing a healthy seedling from a seed is 0.62. Let the random variable X denote the number of healthy seedlings grown from 27 seeds. Determine

- a) the most probable number of healthy plants and its probability,
- b) the expected value and the variance of the random variable X .

Solution. Assume that the seeds germinate and the plants grow independently of each other. Then from the problem statement it follows that the random variable X has the distribution $\text{Bi}(27, 0.62)$.

- a) We are to determine the most probable value of the random variable X , i.e., the mode $\hat{X}$. From the definition of $\hat{X}$ it follows that

$$\frac{\mathbb{P}(X = \hat{X} - 1)}{\mathbb{P}(X = \hat{X})} \leq 1 \quad \text{and} \quad \frac{\mathbb{P}(X = \hat{X} + 1)}{\mathbb{P}(X = \hat{X})} \leq 1.$$

Since X has the distribution $\text{Bi}(27, 0.62)$, it follows that

$$\begin{aligned} \frac{\binom{27}{\hat{X}-1} 0.62^{\hat{X}-1} (1-0.62)^{27-\hat{X}+1}}{\binom{27}{\hat{X}} 0.62^{\hat{X}} (1-0.62)^{27-\hat{X}}} &\leq 1, \\ \frac{\binom{27}{\hat{X}+1} 0.62^{\hat{X}+1} (1-0.62)^{27-\hat{X}-1}}{\binom{27}{\hat{X}} 0.62^{\hat{X}} (1-0.62)^{27-\hat{X}}} &\leq 1, \end{aligned}$$

that is

$$\frac{\hat{X}}{27 - \hat{X} + 1} \cdot \frac{1 - 0.62}{0.62} \leq 1 \quad \text{and} \quad \frac{27 - \hat{X}}{\hat{X} + 1} \cdot \frac{0.62}{1 - 0.62} \leq 1.$$

From both inequalities, we express $\hat{X}$ and obtain

$$16.36 \leq \hat{X} \leq 17.36.$$

The most probable number of seedlings that can be grown from 27 seeds is 17. An alternative way to determine the mode of the distribution is to compute the values of the probability mass function in MS Excel using `= BINOM.DIST` for all arguments $k = 0, 1, \dots, 27$ and to find among them the one with the highest value of $\mathbb{P}(X = k)$. This is the probability that exactly 17 seedlings are grown,

$$\mathbb{P}(X = 17) = \binom{27}{17} 0.62^{17} 0.38^{10} \approx 0.1565.$$

`= BINOM.DIST(17, 27, 0.62, FALSE)`

- b) Using the formulas for the expected value and variance of the binomial distribution, we obtain

$$\begin{aligned}\mathbb{E}X &= 27 \cdot 0.62 = 16.74, \\ \text{var } X &= 27 \cdot 0.62 \cdot 0.38 = 6.3612.\end{aligned}$$

△

7.5 Geometric distribution

A random variable X has the (asymmetric) geometric distribution $\text{Ge}(p)$, denoted by $X \sim \text{Ge}(p)$, if X represents the number of failures before the first success in a Bernoulli process. It takes values $r = 0, 1, 2, \dots$ with probabilities

$$P(X = r) = p_r = p(1 - p)^r,$$

where the parameter $p \in (0, 1)$ is the probability of success in a single Bernoulli trial and is the same for all independent trials. Note that these probabilities form a geometric sequence (they decrease geometrically with increasing r), which is the origin of the name of this distribution. It is clear that $p_r \geq 0$ for all r and $\sum_{r=0}^{\infty} p_r = 1$, since

$$\sum_{r=0}^{\infty} p_r = \sum_{r=0}^{\infty} p(1 - p)^r = p \sum_{r=0}^{\infty} (1 - p)^r = p \frac{1}{1 - (1 - p)} = 1.$$

The probability mass function of the geometric distribution is therefore of the form

$$p(x) = \begin{cases} p(1 - p)^x & \text{for } x = 0, 1, \dots, \\ 0 & \text{otherwise.} \end{cases}$$

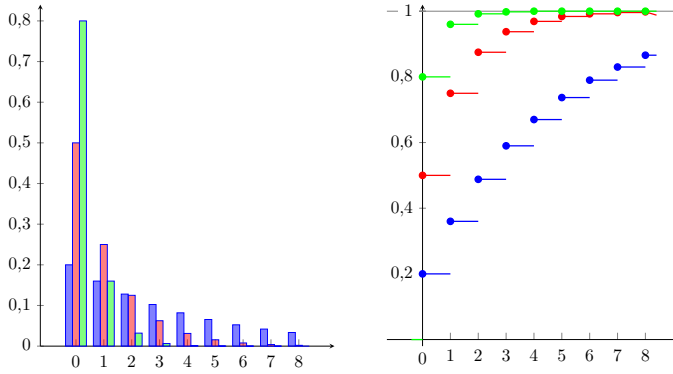


Figure 7.4: Graphs of the probability mass functions and cumulative distribution functions of the geometric distributions $\text{Ge}(0.2)$ (blue), $\text{Ge}(0.5)$ (red), $\text{Ge}(0.8)$ (green)

The cumulative distribution function of the geometric distribution has the form

$$F(x) = \begin{cases} 0 & \text{for } x < 0, \\ \sum_{r=0}^{\lfloor x \rfloor} p(1-p)^r & \text{for } x \geq 0. \end{cases}$$

Example 7.6. A random variable X represents the number of failures before the first success in a sequence of independent trials. The probability of success is 0.2. Compute the probability that before the first success there occur

- exactly four failures,
- at most four failures,
- at least four failures,
- more than two and at most four failures.

Example 7.7. Construct the graphs of the probability mass function and the cumulative distribution function of the random variable X with geometric distribution

- $\text{Ge}(0.2)$,
- $\text{Ge}(0.5)$,
- $\text{Ge}(0.8)$.

Before computing the expected value and the variance, we recall the formulas for the sums of series

$$\sum_{r=1}^{\infty} r q^{r-1} = \frac{1}{(1-q)^2} \quad \text{and} \quad \sum_{r=2}^{\infty} r(r-1) q^{r-2} = \frac{2}{(1-q)^3},$$

which are obtained by differentiating with respect to q the formula for the sum of a geometric series

$$\sum_{r=0}^{\infty} q^r = \frac{1}{1-q}.$$

Using these formulas, we obtain

$$\begin{aligned}\mathbb{E}X &= \sum_{r=0}^{\infty} rp(1-p)^r = p(1-p) \sum_{r=0}^{\infty} r(1-p)^{r-1} = \\ &= p(1-p) \frac{1}{(1-(1-p))^2} = \frac{1-p}{p}. \\ \mathbb{E}X(X-1) &= \sum_{r=2}^{\infty} r(r-1)p(1-p)^r = \\ &= p(1-p)^2 \sum_{r=2}^{\infty} r(r-1)(1-p)^{r-2} = \\ &= p(1-p)^2 \cdot \frac{2}{(1-(1-p))^3} = \frac{2(1-p)^2}{p^2}.\end{aligned}$$

It follows that

$$\begin{aligned}\text{var } X &= \mathbb{E}X(X-1) + \mathbb{E}X - (\mathbb{E}X)^2 = \\ &= \frac{2(1-p)^2}{p^2} + \frac{1-p}{p} - \frac{(1-p)^2}{p^2} = \frac{1-p}{p^2}.\end{aligned}$$

The geometric distribution is used as a simple probabilistic model, for example, for the discrete “lifetime” (durability, reliability) of a product measured in days, months, or years before it fails, the waiting time until a certain event occurs (e.g., until a customer reports an insurance claim), or the duration of sick leave until a person is able to return to work.

In some publications, the term geometric distribution is also used for distributions of related random variables, namely

- the number of trials $X' = X + 1$ needed to achieve the first success, with probability mass function

$$\mathbb{P}(X' = n) = \mathbb{P}(X = n-1) = p(1-p)^{n-1}, \text{ or}$$

(also referred to as the *shifted geometric distribution*),

- the number of successes X'' before the first failure, with probability mass function

$$\mathbb{P}(X'' = k) = p^k(1-p),$$

(here the roles of the parameters are reversed).

In this respect, the terminology and notation are not standardized, and one should always verify which parameterization has been chosen by the author for the given random variable.

Example 7.8. With what probability does a six occur when rolling a die

- a) only on the eighth roll,
- b) at the latest on the third roll,
- c) on rolls other than the second and third?

Solution. Let the random variable X represent the number of rolls on which a six does not occur. The probability of success, i.e., that a six occurs in a single roll, is $p = \frac{1}{6}$.

- a) The event “a six occurs only on the eighth roll” means that the successful eighth roll is preceded by seven unsuccessful rolls, and thus

$$\mathbb{P}(X = 7) = \frac{1}{6} \left(1 - \frac{1}{6}\right)^7 \approx 0.0465.$$

- b) The event “a six occurs at the latest on the third roll” means that it occurs either on the first roll (i.e., no failure precedes it), on the second roll (preceded by one unsuccessful roll), or on the third roll (preceded by two unsuccessful rolls). Therefore

$$\begin{aligned} \mathbb{P}(X \leq 2) &= p(0) + p(1) + p(2) = F(2) = \sum_{r=0}^2 p(1-p)^r = \\ &= \frac{1}{6} + \frac{1}{6} \cdot \frac{5}{6} + \frac{1}{6} \left(\frac{5}{6}\right)^2 \approx 0.421. \end{aligned}$$

- c) $\mathbb{P}(X \neq 2, X \neq 3) = 1 - p(1) - p(2) \approx 0.7454.$

△

Example 7.9. Assume that a predatory fish survives if it catches a fish at least once within two days. During these two days, it engages in eight attempts, with probability of catching a fish $p = 0.25$. What is the probability that the predatory fish does not die?

Solution. The predatory fish survives if it achieves its first success within eight attempts. The sought probability is therefore the sum

$$\sum_{n=1}^8 p_{n-1},$$

where p_{n-1} is the probability that the predatory fish achieves its first success in the n -th attempt (so $n - 1$ is the number of preceding failures). These probabilities follow a geometric distribution with parameter p :

$$\begin{aligned} \sum_{n=1}^8 p_{n-1} &= \sum_{r=0}^7 p_r = \sum_{r=0}^7 p(1-p)^r = \sum_{r=0}^7 0.25 \cdot 0.75^r = \\ &= \frac{1}{4} + \frac{3}{16} + \frac{9}{64} + \frac{27}{256} + \frac{81}{1024} + \frac{243}{4096} + \frac{729}{16384} + \frac{2187}{65536} \approx 0.90. \end{aligned}$$

Thus, the predatory fish has a 10% chance of dying before it manages to catch its prey.

For comparison, let us also consider the expected value $\mathbb{E}X$ of the number of unsuccessful encounters for $X \sim \text{Ge}(0.25)$ and the most probable number of unsuccessful encounters, i.e., the mode $\hat{X}$:

$$\mathbb{E}X = \frac{1-p}{p} = \frac{0.75}{0.25} = 3, \quad \hat{X} = 0,$$

since $p_0 = \frac{1}{4} > p_r = \frac{1}{4} \cdot \left(\frac{3}{4}\right)^r$ for any $r = 1, 2, \dots$. We see that the expected value of a geometric random variable can be considerably far from the most probable value (the more so, the smaller the probability of success is). $\triangle$

7.6 Negative binomial distribution

A random variable X with the negative binomial distribution represents the number $r \in \mathbb{N}$ of unsuccessful trials that precede the k -th successful trial ($k \in \mathbb{N}$) in a sequence of $n = k + r$ independent Bernoulli trials, in which success occurs with probability $p \in (0, 1)$. This distribution is denoted by $\text{NB}(k, p)$ or $\text{NeBi}(k, p)$. It is sometimes also referred to as the Pascal or Pólya⁷ distribution.

The probability mass function of the negative binomial distribution, under the above parameterization, has the form

$$p(r) = \mathbb{P}(X = r) = \begin{cases} \binom{k+r-1}{r} p^k (1-p)^r & \text{for } r = 0, 1, \dots, \\ 0 & \text{otherwise,} \end{cases}$$

where p is the probability of success. The cumulative distribution function has the form

$$F(x) = \begin{cases} 0 & \text{for } x < 0, \\ \sum_{r=0}^{\lfloor x \rfloor} \binom{k+r-1}{r} p^k (1-p)^r & \text{for } x \geq 0. \end{cases}$$

⁷György Pólya (1887–1985), Hungarian mathematician.

The formulas for the expected value and the variance are given by

$$\mathbb{E}X = \frac{k(1-p)}{p}, \quad \text{var } X = \frac{k(1-p)}{p^2}.$$

The negative binomial distribution is, for $k = 1$, a special case of the geometric distribution (see Section 7.5). Therefore, for computing geometric probabilities in Examples 7.8 and 7.9, besides direct calculation, it is also possible to use the Excel function `= NEGBINOM.DIST`, which is intended for computing the probability mass function and the cumulative distribution function of the negative binomial distribution. For example, the probability in Example 7.9 can be obtained more easily as the value of the function `= NEGBINOM.DIST(7,1,0.25,TRUE)`.

Example 7.10. Let the random variable $X \sim \text{NB}(2, 0.8)$ represent the number of failures before the second success in a sequence of independent Bernoulli trials. Compute the probability that

- no failure occurs,
- four failures occur,
- at most four failures occur,
- more than five and at most eight failures occur.

Solution. In independent trials, the probability of success is 0.8.

- a) $X = 0$ means that no failure occurs before the second success. Thus

$$\mathbb{P}(X = 0) = p(0) = F(0) = \binom{0+2-1}{0} 0.8^2 (1-0.8)^0 = 0.8^2 = 0.64,$$

$$= \text{NEGBINOM.DIST}(0, 2, 0.8, \text{FALSE})$$

- b) $\mathbb{P}(X = 4) = p(4) = \binom{4+2-1}{4} 0.8^2 (1-0.8)^4 = 0.00512,$

$$= \text{NEGBINOM.DIST}(4, 2, 0.8, \text{FALSE})$$

- c) $\mathbb{P}(X \leq 4) = p(0) + \cdots + p(4) = F(4) =$

$$= \binom{0+2-1}{0} 0.8^2 (1-0.8)^0 + \cdots + \binom{4+2-1}{4} 0.8^2 (1-0.8)^4 = 0.9984,$$

$$= \text{NEGBINOM.DIST}(4, 2, 0.8, \text{TRUE})$$

- d) $\mathbb{P}(5 < X \leq 8) = p(6) + p(7) + p(8) = F(8) - F(5) =$

$$= \binom{6+2-1}{6} 0.8^2 (1-0.8)^6 + \binom{7+2-1}{7} 0.8^2 (1-0.8)^7 + \binom{8+2-1}{8} 0.8^2 (1-0.8)^8 \approx 0.000367.$$

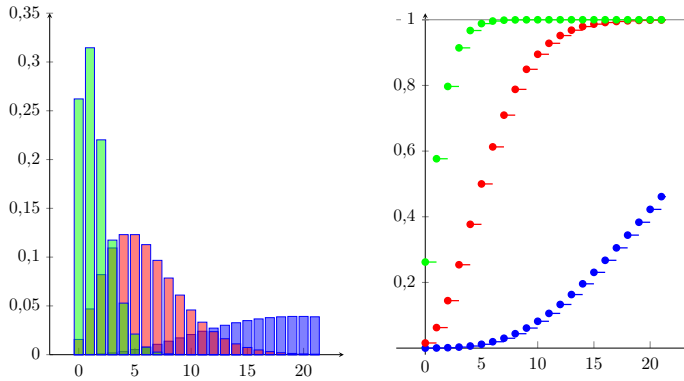


Figure 7.5: Graphs of the probability mass functions and cumulative distribution functions of the negative binomial distributions $\text{NB}(6, 0.2)$ (blue), $\text{NB}(6, 0.5)$ (red), $\text{NB}(6, 0.8)$ (green)

= `NEGBINOM.DIST(8,2,0.8,TRUE)-NEGBINOM.DIST(5,2,0.8,TRUE)`

△

Example 7.11. Construct the graphs of the probability mass functions and the cumulative distribution functions of the random variable X with the negative binomial distribution $\text{NB}(6, p)$ for

- a) $p = 0.2$,
- b) $p = 0.5$,
- c) $p = 0.8$.

Solution. The graphs of the probability mass functions and the cumulative distribution functions are shown in Figure 7.5. △

The random variable X with the negative binomial distribution (representing the number of failures before the k -th success) carries the same information as the random variable W representing the waiting time for the k -th success, i.e., the number of trials needed to achieve the k -th success. In this case,

$$\mathbb{P}(W = n) = \mathbb{P}(X = n - k) = \binom{n-1}{k-1} p^k (1-p)^{n-k}.$$

Alternatively, the situation can be described by the random variable Y representing the number of successes before the r -th failure, in which case

$$\mathbb{P}(Y = k) = \binom{k+r-1}{k} p^k (1-p)^r.$$

The distributions of all three random variables X , W , Y are often collectively referred to in the literature as negative binomial distributions, although they are mathematically distinct distributions.

Example 7.12. What is the probability that “to obtain two sixes when rolling a die, we will need to repeat the roll”

- a) exactly 10 times,
- b) more than 20 times,
- c) at least 12 times and at most 17 times,
- d) 100 times or more?

Solution. Assume that the die is fair, i.e., all outcomes of a die roll occur with the same probability $p = \frac{1}{6}$. We consider as the random experiment the occurrence of a six in a single roll, with two possible outcomes: “a six occurs” (success) and “a six does not occur” (failure). Let W denote the number of rolls we must perform to obtain two sixes, i.e., $k = 2$. Then for $2 \leq n < \infty$ it holds that

$$\mathbb{P}(W = n) = \binom{n-1}{2-1} \left(\frac{1}{6}\right)^2 \left(1 - \frac{1}{6}\right)^{n-2} = \binom{n-1}{1} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^{n-2}.$$

$$\text{a) } \mathbb{P}(W = 10) = \binom{10-1}{1} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^{10-2} = \binom{9}{1} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^8 \approx 0.058.$$

For computing the probability using Excel, we need to reparameterize the random variable as the number of failures before the second success, i.e., as $X = W - 2$ with distribution $X \sim \text{NB}\left(2, \frac{1}{6}\right)$. In that case, $\mathbb{P}(W = 10) = \mathbb{P}(X = 8)$, and therefore we can compute

$$= \text{NEGBINOM.DIST}(8, 2, 1/6, \text{FALSE})$$

$$\text{b) } \mathbb{P}(W > 20) = p(21) + p(22) + p(23) + \cdots = 1 - \mathbb{P}(W \leq 20) =$$

$$= 1 - \sum_{n=2}^{20} \binom{n-1}{1} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^{n-2} \approx 0.13042,$$

$$= 1 - \text{NEGBINOM.DIST}(18, 2, 1/6, \text{FALSE})$$

$$\text{c) } \mathbb{P}(12 \leq W \leq 17) = p(12) + p(13) + p(14) + p(15) + p(16) + p(17) =$$

$$= \sum_{n=12}^{17} \binom{n-1}{2-1} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^{n-2} \approx 0.232,$$

$$\text{X} \quad \text{1} \quad \text{NEGBINOM.DIST}(10; 2; 1/6; \text{FALSE}) + \dots + \text{NEGBINOM.DIST}(15, 2, 1/6, \text{FALSE})$$

Alternatively: $\mathbb{P}(12 \leq W \leq 17) = F(17) - F(11) \approx 0.232$,

$$= \text{NEGBINOM.DIST}(15, 2, 1/6, \text{TRUE}) - \text{NEGBINOM.DIST}(9, 2, 1/6, \text{TRUE})$$

$$\begin{aligned}
 \text{d) } \mathbb{P}(W \geq 100) &= p(100) + p(101) + p(102) + \cdots = \\
 &= 1 - \mathbb{P}(W < 100) = 1 - \mathbb{P}(X \leq 99) = 1 - F(99) = \\
 &= \sum_{n=2}^{99} \binom{n-1}{1} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^{99-2} \approx 3.014 \cdot 10^{-7}.
 \end{aligned}$$

$$= 1 - \text{NEGBINOM.DIST}(97, 2, 1/6, \text{TRUE})$$

△

7.7 Hypergeometric distribution

Discrete distributions introduced in the preceding paragraphs were based on the Bernoulli process, which can be illustratively represented as a repeated experiment of draws with replacement, where the probability of success p remains constant for all draws. If, however, the drawn element is not returned to the population, the probability of success will change depending on the outcomes of the individual draws. Consider a random variable X representing the number of successes in such a sequence of n draws without replacement from an urn containing a total of N elements, of which M elements represent success⁸. This random variable has the so-called *hypergeometric distribution* and we will denote it by $\text{HGeom}(n, M, N)$ ⁹.

The probability mass function of the hypergeometric distribution can be easily derived using combinatorial rules (compare with Example 2.6.b))

$$p(x) = \begin{cases} \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}} & \text{for } x = \max\{0, M - N + n\}, \dots, \min\{n, M\}, \\ 0 & \text{otherwise.} \end{cases}$$

The distribution function is then of the form

$$F(x) = \begin{cases} 0 & \text{for } x < [M - N + n]^+, \\ \sum_{k=[M-N+n]^+}^{\lfloor x \rfloor} \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} & \text{for } [M - N + n]^+ \leq x \leq \min\{n, M\}, \end{cases}$$

where $[M - N + n]^+ = \max\{0, M - N + n\}$ denotes the so-called positive part of the number $M - N + n$. The formulas for the mean and variance are

$$\mathbb{E}X = n \frac{M}{N}, \quad \text{var } X = n \frac{M}{N} \left(1 - \frac{M}{N}\right) \frac{N - n}{N - 1}.$$

⁸In the case of an experiment with replacement, this variable would have a binomial distribution with parameter $p = \frac{M}{N}$.

⁹It is also denoted by $\text{Hg}(n, M, N)$ and sometimes with a different ordering of the individual parameters.

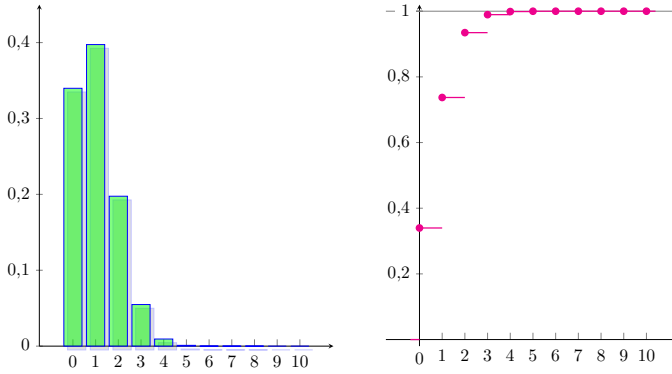


Figure 7.6: Graph of the probability mass function and the distribution function of the hypergeometric distribution $\text{HGeom}(10, 180, 200)$

Example 7.13. A package contains 200 items. Upon receipt of the goods, ten randomly selected items from the entire package are inspected, since the goods may be damaged during transport, which regularly happens to 10% of all products. The goods are accepted if there are at most two broken products in the inspected sample. Determine the probability mass function and the distribution function of the distribution of the random variable that gives the number of broken items, and construct their graphs. With what probability will the package of products be accepted?

Solution. The total number of items in the package is $N = 200$, the number of broken items is $M = 20$ (10% of all items), the number of non-broken items is $N - M = 200 - 20 = 180$, and the number of randomly selected items is $n = 10$. The number of broken items is a random variable with a hypergeometric distribution, i.e. $X \sim \text{HGeom}(10, 180, 200)$. The probability mass function has, for $k = 0, 1, \dots, 10$, the form

$$p(k) = \frac{\binom{20}{k} \binom{180}{10-k}}{\binom{200}{10}}.$$

The graphs of the probability mass function and the distribution function are shown in Figure 7.6. The probability that there are at most two broken items in the package is

$$\begin{aligned} \mathbb{P}(X \leq 2) &= F(2) = p(0) + p(1) + p(2) = \\ &= \frac{\binom{20}{0} \binom{180}{10-0}}{\binom{200}{10}} + \frac{\binom{20}{1} \binom{180}{10-1}}{\binom{200}{10}} + \frac{\binom{20}{2} \binom{180}{10-2}}{\binom{200}{10}} \approx 0.935. \end{aligned}$$

```

1 (COMBIN(180,10)+20*COMBIN(180,9)
2 +COMBIN(20,2)*COMBIN(180,8))/COMBIN(200,10)

```

or, more simply

```
= HYPGEOM.DIST(2,10,20,200,TRUE).
```

The goods will be accepted with a probability of approximately 93.5%. $\triangle$

Note that the binomial distribution is a limiting case of the hypergeometric distribution: if $N \rightarrow \infty$ and $\frac{M}{N} \rightarrow p \in (0, 1)$, then for $k = 0, 1, \dots, n$ we have

$$\lim_{N \rightarrow \infty} \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} = \binom{n}{k} p^k (1-p)^{n-k}.$$

Thus, if the numbers of elements N and M are significantly larger than the number of draws n , the probabilities in the model with replacement and without replacement will not differ substantially. In Example 7.13, for instance, the probability of accepting the goods computed approximately using the binomial distribution would be 0.930.

```
= BINOM.DIST(2,10,20/200,TRUE)
```

7.8 Poisson distribution

In the Bernoulli process, we observed the outcome of a random experiment at predetermined discrete time instants: each such instant corresponds to one Bernoulli trial. If in a Bernoulli trial we observe a success, let us say that an *event* has occurred. In this context, the Bernoulli process is also alternatively called a *discrete-time event process*. Now we relax the condition on discrete time instants: consider a time interval, for simplicity initially of length one time unit¹⁰, and on this interval observe the occurrence of a certain event. If individual events arrive randomly, one by one, and independently of each other (we will explain precisely what this means below), then we speak of a so-called *continuous-time event process*, or also of a *Poisson process*. The Poisson process is characterized by the following axioms:

- 1) the numbers of events in any two non-overlapping time intervals are mutually independent events;
- 2) the probability of the occurrence of one event in a short time interval $(t, t+h]$ (for $t \in [0, 1)$ and small $h > 0$) is directly proportional to the

¹⁰In Section 8.2 we will generalize this assumption.

length of this interval, with a constant coefficient of proportionality λ . Mathematically,

$$\lim_{h \rightarrow 0} \frac{1}{h} \mathbb{P}(\text{one event in the interval } (t, t+h]) = \lambda;$$

- 3) the probability of the occurrence of two or more events at the same time instant is negligible. Mathematically,

$$\lim_{h \rightarrow 0} \frac{\mathbb{P}(\text{two or more events in the interval } (t, t+h])}{\mathbb{P}(\text{one event in the interval } (t, t+h])} = 0.$$

The parameter $\lambda > 0$ is called the *intensity of events* of the Poisson process and represents the average number of observed events per chosen time unit. Now consider a random variable X representing the number of events per chosen time unit. This random variable taking values $k = 0, 1, 2, \dots$ has the so-called *Poisson distribution* with intensity parameter λ , which we denote by $\text{Po}(\lambda)$. The probability mass function can be derived by solving a differential equation based on the axioms of the Poisson process; for $k = 0, 1, 2, \dots$ it has the form

$$\mathbb{P}(X = k) = p_k = \frac{1}{k!} \lambda^k e^{-\lambda}.$$

Here we only verify the conditions on the probabilities p_k :

- a) $p_k \geq 0$ for any $k = 0, 1, 2, \dots$,
b) $\sum_{k=0}^{\infty} p_k = e^{-\lambda} \sum_{k=0}^{\infty} \frac{1}{k!} \lambda^k = e^{-\lambda} \cdot e^{\lambda} = 1$.¹¹

The distribution function is of the form

$$F(x) = \begin{cases} 0 & \text{for } x < 0, \\ \sum_{k=0}^{\lfloor x \rfloor} \frac{1}{k!} \lambda^k e^{-\lambda} & \text{for } x \geq 0. \end{cases}$$

The expected value $\mathbb{E}X$ of a Poisson distribution is equal to the parameter λ , since

$$\begin{aligned} \mathbb{E}X &= \sum_{k=0}^{\infty} k p_k = \sum_{k=1}^{\infty} k \frac{1}{k!} \lambda^k e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{1}{(k-1)!} \lambda^{k-1} = \\ &= \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{1}{j!} \lambda^j = \lambda e^{-\lambda} e^{\lambda} = \lambda. \end{aligned}$$

¹¹The sum was obtained using the Taylor series $e^{\lambda} = \sum_{k=0}^{\infty} \frac{1}{k!} \lambda^k$.

We observe that this is consistent with interpreting the parameter λ as the intensity, i.e., the average number of events per unit time. For the calculation of the variance, we use the relation

$$\text{var } X = \mathbb{E}X(X-1) + \mathbb{E}X - (\mathbb{E}X)^2,$$

where

$$\begin{aligned}\mathbb{E}X(X-1) &= \sum_{k=0}^{\infty} k(k-1) \frac{1}{k!} \lambda^k e^{-\lambda} = e^{-\lambda} \lambda^2 \sum_{k=2}^{\infty} \frac{1}{(k-2)!} \lambda^{k-2} = \\ &= e^{-\lambda} \lambda^2 \sum_{j=0}^{\infty} \frac{1}{j!} \lambda^j = e^{-\lambda} \lambda^2 e^{\lambda} = \lambda^2.\end{aligned}$$

Hence, it follows immediately that

$$\text{var } X = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

We see that for the Poisson distribution the variance is also equal to λ .

Example 7.14. A random variable $X \sim \text{Po}(5)$ is given. Compute:

- a) $\mathbb{P}(X = 6)$,
- b) $\mathbb{P}(X \leq 6)$,
- c) $\mathbb{P}(X < 6)$,
- d) $\mathbb{P}(3 < X \leq 5)$,
- e) $\mathbb{P}(X > 7)$.

Solution.

a) $\mathbb{P}(X = 6) = p(6) = \frac{1}{6!} 5^6 e^{-5} \approx 0.146,$

= POISSON.DIST(6,5,FALSE)

b) $\mathbb{P}(X \leq 6) = p(0) + \cdots + p(6) = F(6) = \sum_{k=0}^6 \frac{1}{k!} 5^k e^{-5} \approx 0.762,$

= POISSON.DIST(6,5,TRUE)

c) $\mathbb{P}(X < 6) = p(0) + \cdots + p(5) = F(5) = \sum_{k=0}^5 \frac{1}{k!} 5^k e^{-5} \approx 0.616,$

= POISSON.DIST(5,5,FALSE)

d) $\mathbb{P}(3 < X \leq 5) = p(4) + p(5) = F(5) - F(3) = \sum_{k=4}^5 \frac{1}{k!} 5^k e^{-5} \approx 0.351,$

= POISSON.DIST(4,5,FALSE)+POISSON.DIST(5,5,FALSE)

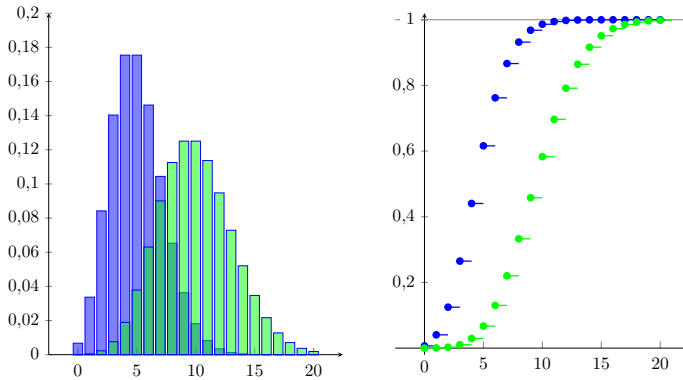


Figure 7.7: Graph of the probability mass function and the cumulative distribution function of the Poisson distribution $\text{Po}(5)$ (blue), $\text{Po}(10)$ (green)

or equivalently

$$= \text{POISSON.DIST}(5, 5, \text{TRUE}) - \text{POISSON.DIST}(3, 5, \text{TRUE})$$

e) $\mathbb{P}(X > 7) = 1 - \mathbb{P}(X \leq 7) = 1 - F(7) \approx 0.133.$

$$= 1 - \text{POISSON.DIST}(7, 5, \text{TRUE})$$

△

Example 7.15. Construct the graphs of the probability mass functions and the cumulative distribution functions of random variables with Poisson distributions $\text{Po}(5)$ and $\text{Po}(10)$.

Solution. The graphs of the probability mass functions and cumulative distribution functions of the Poisson distribution are shown in Figure 7.7. △

A random variable that follows a Poisson distribution represents the number of occurrences of a given event in a specified time interval. We assume that the event may occur at any point in time and that the number of occurrences during the interval depends only on its length, and not on its starting point nor on how many times the event occurred before the interval began. Typical examples of a Poisson-distributed random variable include the number of defective products in a large batch (when the probability of a defect is very small), the number of customers entering a shop during a given time interval, the number of telephone calls received within a certain period, the number of reported insurance claims over the course of a year under a given policy, the number of cases of a rare disease within one month, or the number of customers who sign a credit agreement within a month. The Poisson distribution is often referred to as the “law of rare events”.

Table 7.1: Theoretical probabilities of observing a given number of telephone calls in one minute for Example 7.16

x	0	1	2	3	4	5	6	7	8
$p(x)$	0.135	0.271	0.271	0.180	0.090	0.036	0.012	0.003	0.001

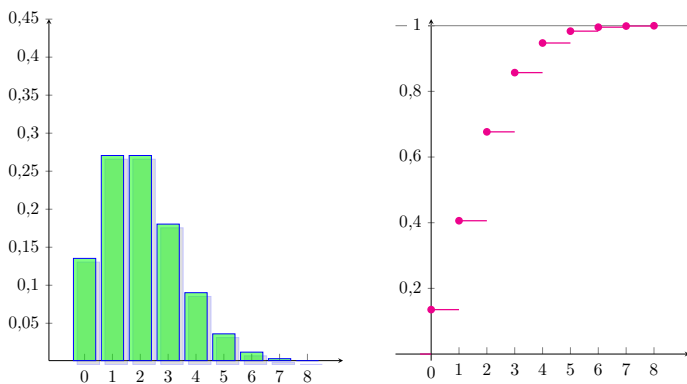


Figure 7.8: Graph of the probability mass function and the cumulative distribution function for Example 7.16

Example 7.16. Consider a random variable X denoting the number of telephone calls per 1 minute, with intensity corresponding to an average of two calls in this time interval. Construct the graph of the probability mass function $p(x)$ and the cumulative distribution function $F(x)$ of the distribution of the random variable X .

Solution. A random variable X follows a Poisson distribution $\text{Po}(2)$. Table 7.1 contains a partial list of values and probabilities of the random variable X . In fact, the variable X can take countably many values, i.e., during one minute an arbitrarily large (natural) number of calls may occur, although with very small probabilities. The graph of the probability mass function $p(x)$ and the cumulative distribution function $F(x)$ is shown in Figure 7.8. $\triangle$

Example 7.17. Ticket vending machines are to be installed at a railway station. After the appropriate coin has been inserted, each machine issues the requested ticket within ten seconds. Suppose that, during the peak passenger period, on average six people per minute will want to use a machine. How



many machines must be installed so that, with probability greater than 0.95, every customer is served immediately during the peak period?

Solution. Let the random variable X denote the number of customers per time unit of ten seconds during the period of peak passenger traffic. During this period, the service demand intensity is 6 persons per minute. Clearly, the random variable X follows a Poisson distribution with parameter $\lambda = \frac{10}{60} \cdot 6 = 1$ passenger per 10 seconds. We now seek the smallest value k such that the random variable X does not exceed k with probability greater than 0.95, i.e., the smallest value k for which $\mathbb{P}(X \leq k) > 0.95$. Since

$$\begin{aligned}\mathbb{P}(X \leq 0) &= e^{-1} \approx 0.368 \\ \mathbb{P}(X \leq 1) &= e^{-1} + e^{-1} \approx 0.736, \\ \mathbb{P}(X \leq 2) &= e^{-1} + e^{-1} + \frac{e^{-1}}{2} \approx 0.920, \\ \mathbb{P}(X \leq 3) &= e^{-1} + e^{-1} + \frac{e^{-1}}{2} + \frac{e^{-1}}{6} \approx 0.981,\end{aligned}$$

we must install at least three service machines so that, with probability greater than 0.95, every customer is served immediately. $\triangle$

For sufficiently large n (e.g., at least $n > 30$) and sufficiently small p (e.g., $p < 0.1$), the binomial distribution can be approximated by the Poisson distribution with parameter $\lambda = np$. In fact, the Poisson distribution $\text{Po}(\lambda)$ is a limiting case of the binomial distribution $\text{Bi}(n, p)$ as $n \rightarrow \infty$, $p \rightarrow 0$, and $np \rightarrow \lambda$. Let us examine the k -th term of the binomial distribution for the special case $n \rightarrow \infty$ and $p \rightarrow 0$ with $np = \lambda$:

$$\begin{aligned}& \lim_{n \rightarrow \infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = \\ &= \lim_{n \rightarrow \infty} \frac{n(n-1) \cdots (n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} = \\ &= \frac{\lambda^k}{k!} \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) \left(1 - \frac{\lambda}{n}\right)^{-k} \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n.\end{aligned}$$

The first limit equals 1, the second limit equals $e^{-\lambda}$, which directly yields the formula for the probability mass function of the Poisson distribution. This fact is illustrated in Example 7.18.

Example 7.18. Suppose that an insurance company records 1,000 insured events over a certain period. The probability that an insured event occurs within one minute is 0.006. What is the probability that ten insured events

occur within one minute? Solve the problem using both the Poisson and the binomial distributions. Round the results to six decimal places and compare them.

Solution. $\lambda = np = 1\,000 \cdot 0.006 = 6$, thus, on average, we expect six insured events per minute. Hence, using the Poisson distribution,

$$\mathbb{P}(X = 10) = e^{-6} \frac{6^{10}}{10!} \approx 0.041303.$$

= `POISSON.DIST(10,6,FALSE)`

Using the binomial distribution, we obtain

$$\mathbb{P}(X = 10) = \binom{1000}{10} 0.006^{10} 0.994^{990} \approx 0.041178.$$

= `BINOM.DIST(10,1000,0.006,FALSE)`

Under the chosen approximation, the results differ only in the fourth decimal place. △

7.9 Exercises

7.1 Prove that a random variable X following the discrete uniform distribution on the set $D = \{1, 2, \dots, n\}$ has expectation $\mathbb{E}X$ equal to $\frac{n+1}{2}$ and variance $\text{var } X = \frac{n^2-1}{12}$.

7.2 Construct the graphs of the probability mass function and the cumulative distribution function of the random variable X with the degenerate distribution $\text{Dg}(1)$.

7.3 Prove that a random variable X following the Bernoulli distribution has expectation $\mathbb{E}X$ equal to p and variance $\text{var } X = p(1-p)$.

7.4 A manufacturer supplies products in packages of 15 items. For each package in which at least one product is defective, the customer will be refunded the price of the entire package. It is known that the probability of producing a non-defective product is 0.95 and that the cost of manufacturing one whole package is 2 \$. What price per product must the manufacturer set in order to expect a profit of 23%? By profit we mean the difference between the revenue from the sale of products and the costs of all production, including products for which a refund is claimed.

7.5 Let X be a discrete random variable with support $\{1, 2, 3, 4, 5\}$ and probabilities $\mathbb{P}(X = i) = \frac{c}{i}$ for $i = 1, 2, \dots, 5$.

- a) Determine the constant c .
- b) Determine $\mathbb{E}X$ and $\text{var } X$.

7.6 Consider the eight well-known blood groups A+, A−, B+, B−, AB+, AB−, 0+, and 0−, and suppose that they occur with equal probability. What is the probability that, in order to find three donors with blood type A+, we need to examine:

- a) exactly ten individuals who do not know their blood type (A),
- b) more than nine individuals who do not know their blood type (B),
- c) more than seven and fewer than twelve individuals who do not know their blood type (C)?

7.7 Among one hundred eggs, 5% are cracked. What is the probability that if we select 20 eggs, there will be:

- a) “exactly two cracked” (A),
- b) “at least two cracked” (B),
- c) “at most two cracked” (C)?

7.8 A telephone exchange connects, on average, 15 calls per hour. What is the probability that during four minutes the exchange will connect

- a) “exactly one call” (event A),
- b) “at least two calls” (event B),
- c) “at least two and at most five calls” (event C)?

7.9 A participant calls a telephone exchange during peak load, when the probability that the line is not busy is 0.25. The individual attempts to establish a connection repeatedly at intervals of several minutes until the connection is made. Determine

- a) the probability that “the connection is established only on the fifth attempt” (event A),
- b) the expectation and variance of the number of unsuccessful attempts up to the moment the connection is established.

7.10 A player rolls a die until three sixes occur. What is the probability that the player will have to roll the die ten times?

7.11 The expected number of failures of a transmitter during 10,000 hours of operation is 10. Assume that individual occurrences of failures over time are independent. Determine the probability of a failure of the device during 100 hours of operation.

7.12 A telephone exchange is connected to 326 subscribers. Each of them calls the exchange during one hour with probability 0.01. Assume that the call times of individual subscribers are mutually independent. What is the probability that during one hour

- a) 4 subscribers will call,
- b) at least 6 subscribers will call,
- c) at most 2 subscribers will call?

7.13 On 500 pages of a book, there are 500 misprints. Assume that the occurrences of individual errors are mutually independent. Determine the probability that on a randomly selected page there are

- a) “at least three errors”,

- b) “exactly one error”,
- c) “at most two errors”.

7.14 On average, 60 customers arrive at a pub per hour. What is the probability that during half a minute, while no one is being served, no customer will arrive?

7.15 Determine the probability that among 136 products there are

- a) exactly seven defective products,
- b) at least three defective products,
- c) at most ten defective products,

if it is known that defective products constitute on average 2.6% of the production.

7.16 In order to save money on expensive blood tests when testing for whooping cough during the Second World War, the army proposed the following plan: instead of testing each soldier’s blood sample separately, they mixed the blood of ten soldiers and tested the mixture. If the test was negative, they knew that all soldiers in the group were negative. (Assume that mixing several blood samples has no effect on the identification of the infection.) If the test was positive, they had to test each soldier in the given group separately. It is clear that if the number of infected soldiers among all soldiers is high, this approach is not worthwhile. Under what conditions does this procedure save money compared with immediately testing each soldier separately? That is, determine the highest possible probability with which a tested soldier may be positive for testing ten soldiers simultaneously to be worthwhile.

7.17 A restaurant provides, with each meal, a card depicting one player from the local team. On each visit to the restaurant, you receive one card at random.

- a) If the cards depict basketball players (five players in the starting lineup), how many times must you visit the restaurant on average in order to obtain at least one card of each player?
- b) If the cards depict baseball players (nine players in the starting lineup), how many times must you visit the restaurant on average in order to obtain at least one card of each player?

Solutions to the Exercises

7.1

$$\begin{aligned}
 \mathbb{E}X &= \sum_{i=1}^n i \cdot \frac{1}{n} = \frac{1}{n} \cdot 1 + \frac{1}{n} \cdot 2 + \cdots + \frac{1}{n} \cdot n = \\
 &= \frac{1}{n} \cdot (1 + \cdots + n) = \frac{1}{n} \cdot \frac{n}{2} \cdot (n+1) = \frac{n+1}{2}, \\
 \mathbb{E}X^2 &= \frac{1}{n} \cdot 1^2 + \frac{1}{n} \cdot 2^2 + \cdots + \frac{1}{n} \cdot n^2 = \frac{1}{n} \cdot (1^2 + 2^2 + \cdots + n^2) = \\
 &= \frac{1}{n} \cdot \frac{n \cdot (n+1) \cdot (2n+1)}{6} = \frac{(n+1) \cdot (2n+1)}{6}, \\
 \text{var } X &= \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{(n+1) \cdot (2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n^2 - 1}{12}.
 \end{aligned}$$

7.2

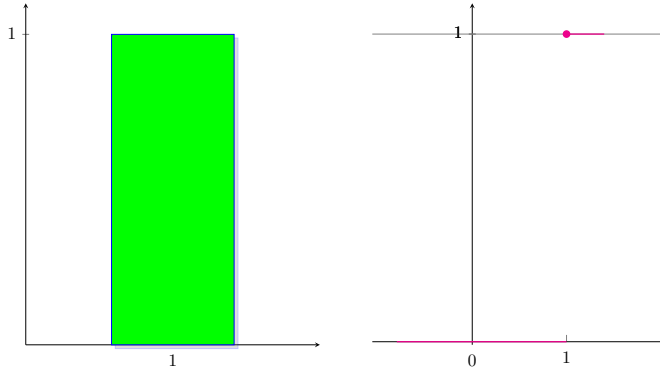


Figure 7.9: Graph of the probability mass function and the cumulative distribution function for Problem 7.2

7.3

$$\begin{aligned}
 \mathbb{E}X &= 0 \cdot (1-p) + 1 \cdot p = p, \\
 \mathbb{E}X^2 &= 0^2(1-p) + 1^2p = p, \\
 \text{var } X &= \mathbb{E}X^2 - (\mathbb{E}X)^2 = p - p^2 = p(1-p).
 \end{aligned}$$

7.4 The probability that a package contains no defective item and thus yields a profit is $0.95^{15} \approx 0.463291$. For example, for 100 packages the total cost is 200 \$, and the revenue must be 246 \$ in order to achieve a profit of 23%. Therefore, the price per product must be $246 : 46.3291 = 5.31$ \$.

7.5

$$\text{a) } 1 = \sum_{i=1}^5 \mathbb{P}(X = i) = c \cdot \sum_{i=1}^5 \frac{1}{i} = c \cdot \left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5}\right) = c \cdot \frac{137}{60},$$

$$\Rightarrow c = \frac{60}{137}.$$

$$\text{b) } \mathbb{E}X = \sum_{i=1}^5 i \cdot \mathbb{P}(X = i) = \frac{60}{137} \cdot \sum_{i=1}^5 1 = \frac{300}{137} \approx 2.190,$$

$$\mathbb{E}(X^2) = \sum_{i=1}^5 i^2 \cdot \mathbb{P}(X = i) = \frac{60}{137} \cdot \sum_{i=1}^5 i = \frac{60}{137} \cdot \frac{5 \cdot 6}{2} = \frac{900}{137},$$

$$\text{var } X = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \frac{900}{137} - \left(\frac{300}{137}\right)^2 \approx 1.774.$$

$$\text{7.6 a) } \mathbb{P}(A) \approx 0.028, \text{ b) } \mathbb{P}(B) \approx 0.908, \text{ c) } \mathbb{P}(C) \approx 0.103.$$

$$\text{7.7 a) } \mathbb{P}(A) \approx 0.207, \text{ b) } \mathbb{P}(B) \approx 0.261, \text{ c) } \mathbb{P}(C) \approx 0.947.$$

$$\text{7.8 a) } \mathbb{P}(A) \approx 0.3679, \text{ b) } \mathbb{P}(B) \approx 0.2642, \text{ c) } \mathbb{P}(C) \approx 0.2636.$$

$$\text{7.9 a) } \mathbb{P}(A) \approx 0.079, \text{ b) } 3, 12.$$

$$\text{7.10 } 0.0465$$

$$\text{7.11 } 0.095$$

$$\text{7.12 } \lambda = 3.26; \text{ a) } 0.1807, \text{ b) } 0.1124, \text{ c) } 0.3675.$$

$$\text{7.13 a) } 0.080, \text{ b) } 0.368, \text{ c) } 0.920.$$

$$\text{7.14 } 0.607$$

$$\text{7.15 a) } 0.039, \text{ b) } 0.689, \text{ c) } 0.999091.$$

$$\text{7.16 } p < 0.004087$$

$$\text{7.17 a) } 11.41, \text{ b) } 24.46.$$

Chapter 8

Selected distributions of continuous random variables

8.1 Continuous uniform distribution

Consider the interval $[a, b]$ of real values, in which none of the values is preferred in terms of occurrence. In such a case, we may use the model of geometric probability discussed in Section 2.3 to characterize the probabilistic behaviour. Alternatively, we may define a random variable X on the probability space with values in the interval $[a, b]$, which we say has the so-called *continuous uniform distribution* $U(a, b)$ ¹. Since this is a continuous distribution, the probabilistic behaviour is characterized by a density that is constant over the given interval $[a, b]$. Furthermore, the cumulative distribution function can easily be derived by integration – the cumulative distribution function is a primitive function of the density, and the density is the derivative of the cumulative distribution function (see also Exercise 8.2). Thus, the density of the uniform distribution and its cumulative distribution function have the form

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{otherwise,} \end{cases} \quad F(x) = \begin{cases} 0 & \text{for } x < a, \\ \frac{x-a}{b-a} & \text{for } a \leq x \leq b, \\ 1 & \text{for } x > b. \end{cases}$$

¹The notation U is derived from the English word *uniform*, meaning consistent, even, homogeneous, identical, constant, etc., which are among the properties of the graph of the probability density function of this distribution; it is a line segment parallel to the x -axis. In Czech literature, the continuous uniform distribution is denoted by $Rs(a, b)$ or simply $R(a, b)$.

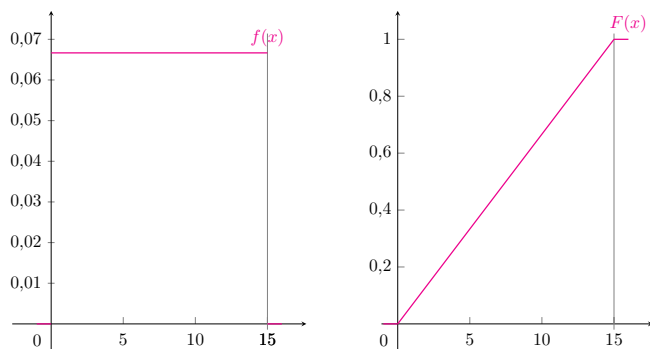


Figure 8.1: The probability density function and the cumulative distribution function for Example 8.1

The density is symmetric about the midpoint of the interval $[a, b]$; therefore, this midpoint is both the expectation and the median of the random variable X . The variance of the random variable X can be computed using the computational formula (6.7) (see Exercise 8.1), from which we obtain

$$\mathbb{E}X = \tilde{X} = \frac{a+b}{2}, \quad \text{var } X = \frac{1}{12}(b-a)^2. \quad (8.1)$$

Example 8.1. Construct the graph of the probability density function and the graph of the cumulative distribution function of the continuous uniform distribution of the random variable X on the interval $[0, 15]$. Compute the expectation and the variance of the random variable X .

Solution. $f(x) = \frac{1}{15}$, $F(x) = \frac{x}{15}$, $x \in [0, 15]$. The graph of the probability density function and the cumulative distribution function is shown in Figure 8.1. The expectation and the variance have the values

$$\mathbb{E}X = \frac{0+15}{2} = 7.5, \quad \text{var } X = \frac{1}{12}(15-0)^2 = 18.75.$$

△

The continuous uniform distribution $U(a, b)$ arises, for example, in the analysis of rounding errors in numerical computations. If the numbers entering the computations are infinite decimal fractions that are rounded to k decimal places, then the rounding error can be regarded as a random variable with the uniform distribution on the interval $[-5 \cdot 10^{-k-1}, 5 \cdot 10^{-k-1}]$.

For the continuous uniform distribution, there is a very illustrative visualization of probability calculations, since we compute the area of a rectangle under the probability density function (see Example 8.2 and Figure 8.2).

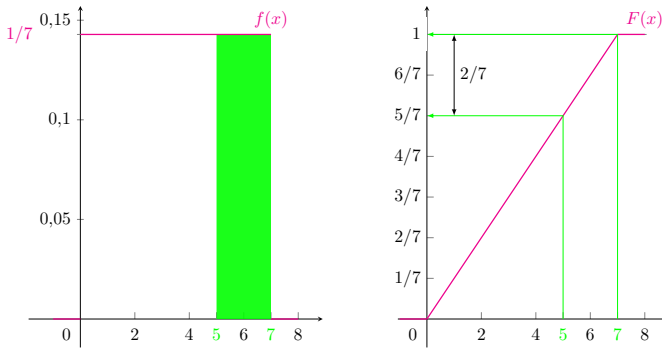


Figure 8.2: Probability density function, cumulative distribution function, and probability calculation for Example 8.2

Example 8.2. Urban buses depart from a stop at intervals of seven minutes. A passenger may arrive at the stop at any moment. What is the probability that a randomly arriving passenger will have to wait more than 5 minutes? What are the expectation and the variance of the waiting time for the departure of the bus from the stop?

Solution. The waiting time for the departure of the bus is a random variable with a continuous uniform distribution on the interval $[0, 7]$, hence

$$f(x) = \frac{1}{7}$$

on the interval $[0, 7]$. The expectation and the variance are given, according to formula (8.1), by

$$\mathbb{E}X = \frac{0+7}{2} = 3.5 \text{ min} \quad \text{a} \quad \text{var } X = \frac{(7-0)^2}{12} = \frac{49}{12} = 4.08\bar{3} \text{ min}^2.$$

The probability that a randomly arriving passenger will have to wait at the stop for more than five minutes is

$$\mathbb{P}(X \geq 5) = \int_5^7 \frac{1}{7} dx = \frac{1}{7}(7-5) = \frac{2}{7} \approx 0.286.$$

The computation of this probability is illustrated in Figure 8.2. $\triangle$

8.2 Exponential Distribution

In Section 7.8 we introduced the Poisson process, i.e., a process of events in continuous time with intensity parameter λ events per prescribed time

unit. Consider the random variable T representing the time elapsed between two successive events of the Poisson process. We now derive the probability distribution of this random variable. First, define an auxiliary random variable X representing the number of events occurring during a prescribed time interval $t > 0$ since the last event. This variable has the Poisson distribution with parameter λt . The probability that the waiting time for the next event exceeds time t is therefore equal to the probability that no event occurs during time t , i.e.

$$\mathbb{P}(T > t) = \mathbb{P}(X = 0) = \frac{1}{0!}(\lambda t)^0 e^{-\lambda t} = e^{-\lambda t} \quad (8.2)$$

for $t > 0$. From this, we directly obtain the formula for the cumulative distribution function of the random variable T ,

$$F_T(t) = \begin{cases} 0 & \text{for } t \leq 0, \\ 1 - \mathbb{P}(T > t) = 1 - e^{-\lambda t} & \text{for } t > 0 \end{cases}$$

and by differentiating $F_T(t)$, we also obtain the formula for the probability density function of the random variable T ,

$$f_T(t) = \begin{cases} 0 & \text{for } t \leq 0, \\ \lambda e^{-\lambda t} & \text{for } t > 0. \end{cases}$$

The distribution of the random variable T is called the *exponential distribution* and is denoted by $\text{Exp}(\lambda)$ ². It is easy to verify that the function $f(t)$ is indeed a density. First, $f(t) \geq 0$ for all $t \in \mathbb{R}$, and moreover

$$\int_{-\infty}^{\infty} f(t) dt = \int_0^{\infty} \lambda e^{-\lambda t} dt = 1,$$

according to the rule for integrating the exponential function

$$\int e^{ct} dt = \frac{1}{c} e^{ct} \quad \text{for any } c \in \mathbb{R}.$$

By repeated integration, we also derive the expectation and the variance of the random variable T , which are equal to

$$\mathbb{E}X = \frac{1}{\lambda} \quad \text{and} \quad \text{var } X = \frac{1}{\lambda^2}.$$

²It is sometimes denoted by $\text{Ex}(\lambda)$ or simply $\text{E}(\lambda)$. Note that the parameterization of the exponential distribution is not entirely uniform in the literature: in some publications, instead of the intensity parameter λ , the mean parameter $\beta = 1/\lambda$ is used, for example $\text{Exp}(\beta)$. Moreover, the symbol λ is sometimes also used for the mean parameter. Therefore, the reader should be careful when consulting the literature to determine which form of parameterization the author adopts. In this book, we consistently use the intensity parameter denoted by λ , and reserve the symbol β for the mean parameter.

The exponential distribution $\text{Exp}(\lambda)$ is an appropriate model for the probabilistic characterization of the “waiting time for an event” (the time until the occurrence of a given event), for example, the lifetime of a certain device. In the following theorem 8.1, we show that the probability distribution of the remaining waiting time does not depend on how long we have already been waiting. This property is called the “memoryless property” of the exponential distribution.

Theorem 8.1. *If the random variable T has the exponential distribution, then*

$$\mathbb{P}(T > t_0 + t | T > t_0) = \mathbb{P}(T > t) \text{ for any } t_0 > 0, t > 0.$$

Proof. Let $T \sim \text{Exp}(\lambda)$. Then, according to formula (8.2) and by the definition of conditional probability, we have

$$\mathbb{P}(T > t_0 + t | T > t_0) = \frac{\mathbb{P}(T > t_0 + t)}{\mathbb{P}(T > t_0)} = \frac{e^{-\lambda(t_0+t)}}{e^{-\lambda t_0}} = e^{-\lambda t}.$$

□

To compute the values of the probability density function and the cumulative distribution function, we may use the Excel function `=EXPON.DIST`, as shown in Example 8.3. The last parameter of this function specifies the type of calculation: we use the value `FALSE` to compute the values of the probability density function and `TRUE` to compute the values of the cumulative distribution function. The first two parameters of the function are, respectively, the variable t (the argument of the density or the cumulative distribution function) and the value of the parameter λ .

Example 8.3. Let the random variable T have the exponential distribution $\text{Exp}(\frac{1}{3})$. Compute:

- $\mathbb{P}(T = 1)$,
- $\mathbb{P}(T \leq 2)$,
- $\mathbb{P}(T > 5)$,
- $\mathbb{P}(T < 3)$,
- $\mathbb{P}(T \geq 5)$,
- $\mathbb{P}(2 < T \leq 4)$.

Solution.

- $\mathbb{P}(T = 1) = 0$, since this is a continuous distribution,
- $\mathbb{P}(T \leq 2) = F(2) = 1 - e^{-\frac{1}{3} \cdot 2} \approx 0.487$,

$$= \text{EXPON.DIST}(2, 1/3, \text{TRUE})$$

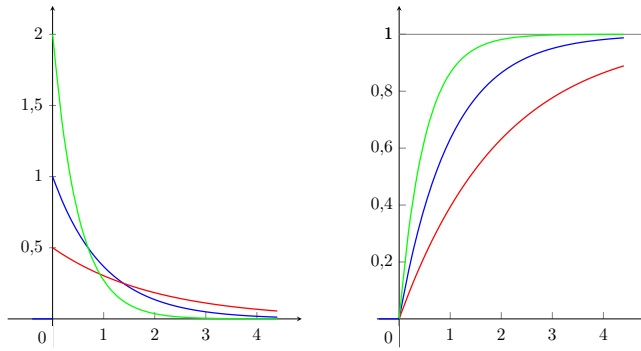


Figure 8.3: Graphs of the probability density functions and cumulative distribution functions of the exponential distributions: $\text{Exp}(\frac{1}{2})$ (red), $\text{Exp}(1)$ (blue), $\text{Exp}(2)$ (green)

c) $\mathbb{P}(T > 5) = 1 - \mathbb{P}(T \leq 5) = 1 - F(5) = 1 - (1 - e^{-\frac{1}{3} \cdot 5}) \approx 0.189$.

`= 1-EXPON.DIST(5,1/3,TRUE)`

d) $\mathbb{P}(T < 3) = \mathbb{P}(T \leq 3) = 1 - e^{-\frac{1}{3} \cdot 3} \approx 0.632$,

`= EXPON.DIST(3,1/3,TRUE)`

e) $\mathbb{P}(T \geq 5) = \mathbb{P}(T > 5) \approx 0.189$, see c),

f) $\mathbb{P}(2 < T \leq 4) = F(4) - F(2) = 1 - e^{-\frac{1}{3} \cdot 4} - [1 - e^{-\frac{1}{3} \cdot 2}] \approx 0.250$.

`= EXPON.DIST(4,1/3,TRUE)-EXPON.DIST(2,1/3,TRUE)`

△

Example 8.4. Construct the graphs of the probability density functions and the cumulative distribution functions of the random variable X with the exponential distribution

- a) $\text{Exp}(\frac{1}{2})$,
- b) $\text{Exp}(1)$,
- c) $\text{Exp}(2)$.

Solution. The graphs of the probability density functions and the cumulative distribution functions are shown in Figure 8.3. Observe that the graph of the density always starts at the value of the intensity parameter λ on the vertical axis and then decreases exponentially. △

Example 8.5. The time between the arrivals of two successive patients follows an exponential distribution with mean 15 minutes. Compute the probability that the reception will wait for the next patient:

- a) at most 15 minutes,
- b) at least 15 minutes,
- c) exactly 15 minutes,
- d) between 5 and 20 minutes.

Solution. Let T denote the time between the arrivals of two successive patients. It is given that $\mathbb{E}T = 15 = \frac{1}{\lambda}$; hence $\lambda = \frac{1}{15}$ and $T \sim \text{Exp}(\frac{1}{15})$.

$$\begin{aligned} \text{a) } \mathbb{P}(T \leq 15) &= F(15) = \int_0^{15} \frac{1}{15} e^{-\frac{t}{15}} dt = \frac{1}{15} \int_0^{15} e^{-\frac{t}{15}} dt = \\ &= \frac{1}{15} \cdot (-15) \left[e^{-\frac{t}{15}} \right]_0^{15} = -e^{-1} + 1 \approx 0.632, \\ &= \text{EXPON.DIST}(15, 1/15, \text{TRUE}) \end{aligned}$$

$$\begin{aligned} \text{b) } \mathbb{P}(T \geq 15) &= 1 - \mathbb{P}(T < 15) = 1 - F(15) \approx 1 - 0.632 = 0.368, \\ &= 1 - \text{EXPON.DIST}(15, 1/15, \text{TRUE}) \end{aligned}$$

or directly $\mathbb{P}(T \geq 15) = e^{-\frac{1}{15} \cdot 15} = e^{-1} \approx 0.368$.

c) $\mathbb{P}(T = 15) = 0$, since this is a continuous distribution,

$$\begin{aligned} \text{d) } \mathbb{P}(5 < T < 20) &= \int_5^{20} \frac{1}{15} e^{-\frac{t}{15}} dt = \frac{1}{15} \int_5^{20} e^{-\frac{t}{15}} dt = \\ &= \frac{1}{15} \cdot (-15) \left[e^{-\frac{t}{15}} \right]_5^{20} = - \left(e^{-\frac{4}{3}} - e^{-\frac{1}{3}} \right) \approx \\ &\approx 0.7165 - 0.2636 \approx 0.453, \end{aligned}$$

however, also

$$\begin{aligned} \mathbb{P}(5 < T < 20) &= F(20) - F(5) \approx 0.7364 - 0.2835 \approx 0.453. \\ &= \text{EXPON.DIST}(20, 1/15, \text{TRUE}) - \text{EXPON.DIST}(5, 1/15, \text{TRUE}) \end{aligned}$$

The reception will wait for the next patient at most 15 minutes with probability approximately 63%, at least 15 minutes with probability approximately 37%, exactly 15 minutes with zero probability, and between 5 and 20 minutes with probability approximately 45%. $\triangle$

In Microsoft Excel, there is no built-in function for computing the values of the quantile function for a random variable with an exponential distribution. Therefore, it is necessary to perform a manual calculation when needed, as shown in the following example 8.6.

Example 8.6. From long-term observations, it is known that a laptop experiences a failure on average once every 10,000 hours. Suppose that the “waiting time until failure” is a random variable T with an exponential distribution. Determine the value t such that the probability that the laptop operates for longer than t is 0.99.

Solution. In this example, we first derive the formula for the quantile function for a general intensity λ and a prescribed probability α . Let

$$\alpha = F(t) = 1 - e^{-\lambda t},$$

and hence

$$1 - \alpha = e^{-\lambda t},$$

from which we obtain

$$t = F^{-1}(\alpha) = -\frac{1}{\lambda} \ln(1 - \alpha). \quad (8.3)$$

By substituting $\lambda = \frac{1}{10\,000}$ and $\alpha = 0.01$ (since $\mathbb{P}(T > t) = 1 - F(t) = 1 - \alpha = 0.99$), we obtain

$$t = -10\,000 \cdot \ln 0.99 \approx 100.5.$$

With probability 99%, the laptop will operate for longer than 100.5 hours. $\triangle$

Example 8.7. On average, 30 customers arrive at a shop per hour. What is the probability that no customer will arrive during $\frac{1}{2}$ minute while the salesperson is absent?

Solution. The problem can be solved using both the exponential and the Poisson distributions.

- a) Let the random variable T denote the waiting time for the arrival of the next customer in minutes. The variable T follows an exponential distribution with intensity parameter $\lambda = \frac{1}{2}$ customers per minute (since the mean waiting time for one customer is $\mathbb{E}T = \frac{1}{\lambda} = 2$ minutes). The required probability is then, according to the formula (8.2)

$$\mathbb{P}\left(T > \frac{1}{2}\right) = e^{-\frac{1}{2} \cdot \frac{1}{2}} = e^{-\frac{1}{4}} \approx 0.7788.$$

= 1-EXPON.DIST(0.5,0.5,TRUE)

- b) Let the random variable X denote the number of customers who arrive at the shop during $\frac{1}{2}$ minute. The variable X follows a Poisson distribution with parameter $\lambda' = \frac{1}{4}$. Note that the parameter of the distribution differs from the previous approach, because the time unit has changed – we have

$$\lambda' = \lambda t = \frac{1}{2} \cdot \frac{1}{2}.$$

The required probability is then

$$\mathbb{P}(X = 0) = \frac{1}{0!} \left(\frac{1}{4}\right)^0 e^{-\frac{1}{4}} = e^{-\frac{1}{4}} \approx 0.7788.$$

`= POISSON.DIST(0,0.25,FALSE)`

The probability that no customer will arrive during the salesperson's absence lasting half a minute is approximately 78%. $\triangle$

8.3 Erlang³ distribution and gamma distribution

Now consider the random variable W_k , which represents the waiting time until the occurrence of the k -th event in a Poisson process with intensity $\lambda > 0$ events per unit time. This means that in the time interval $(0, t)$, $(k - 1)$ events occur with probability

$$\frac{1}{(k - 1)!} (\lambda t)^{k-1} e^{-\lambda t},$$

which follows the Poisson distribution, and the last k -th event occurs in a short time interval $(t, t + dt)$ with probability λdt , as follows from the second axiom of the Poisson process (see Section 7.8). Since the two time intervals do not overlap, the corresponding events are independent by the first axiom of the Poisson process. Therefore, the probability that they occur simultaneously is their product,

$$\frac{1}{(k - 1)!} (\lambda t)^{k-1} e^{-\lambda t} \lambda dt.$$

By taking the limit as $dt \rightarrow 0$, we obtain the probability density function of the random variable W_k , which has the form

$$f_{W_k}(t) = \begin{cases} \frac{1}{(k-1)!} \lambda^k t^{k-1} e^{-\lambda t} & \text{for } t > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (8.4)$$

This distribution is called the *Erlang distribution*. In this section, we denote it by $\text{Erlang}(k, \lambda)$, where the parameter k is the number of required events and the parameter λ is the intensity of their occurrence. It can be shown that the Erlang distribution is the distribution of the sum of k independent

³Agner Krarup Erlang (1878–1929), Danish mathematician, statistician, and engineer.

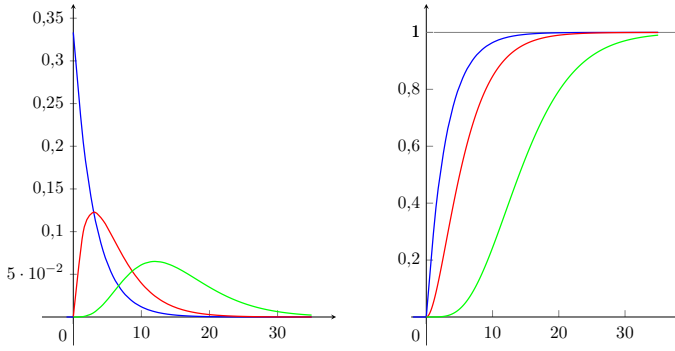


Figure 8.4: Graphs of the probability density functions and cumulative distribution functions of a random variable with the Erlang distribution: $\text{Erlang}(1, \frac{1}{3})$ (blue), $\text{Erlang}(2, \frac{1}{3})$ (red), $\text{Erlang}(5, \frac{1}{3})$ (green)

random variables with exponential distributions having the same parameter λ . In particular, $\text{Exp}(\lambda) = \text{Erlang}(1, \lambda)$. The cumulative distribution function of this distribution has the form

$$F(t) = \begin{cases} 1 - e^{-\lambda t} \cdot \sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!} & \text{for } t > 0, \\ 0 & \text{otherwise} \end{cases}$$

and the expectation and the variance are given by

$$\mathbb{E}W_k = \frac{k}{\lambda} \quad \text{and} \quad \text{var } W_k = \frac{k}{\lambda^2}.$$

Example 8.8. Construct the graphs of the probability density functions and the cumulative distribution functions of the random variable $X \sim \text{Erlang}(k, \frac{1}{3})$, and investigate the effect of the parameter k on the shape of the graphs. Choose the first parameter successively as

- a) $k = 1$,
- b) $k = 2$,
- c) $k = 5$.

Solution. The graphs of the probability density functions and the cumulative distribution functions are shown in Figure 8.4. $\triangle$

If we look more closely at Figure 8.4, we observe that the parameter k determines the shape of the density and, consequently, the cumulative distribution function. If we abandon the concept of a discrete number of

events in a Poisson process, it is possible (in other contexts) to work with densities involving a parameter k that need not be an integer. The density of such a distribution will have a form similar to that in formula (8.4), where we only need to replace the factorial $(k-1)!$ in the denominator with some continuous analogue. Let us denote this, for the time being unknown, function by $\Gamma(\alpha)$. If the argument α is an integer, say k , we require this function to have the same properties as the factorial $(k-1)!$. In particular, we require⁴ which reads

$$\begin{aligned}\Gamma(k) &= (k-1)!, \\ \Gamma(\alpha) &= (\alpha-1)\Gamma(\alpha-1), \\ \Gamma(1) &= 1.\end{aligned}$$

Once such a function is found, the density of the Erlang distribution in the form

$$f(t) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha t^{\alpha-1} e^{-\lambda t} \quad (8.5)$$

will be well defined also for non-integer values of α . However, in order for it to represent a valid density, it must necessarily satisfy the condition

$$\int_0^{+\infty} \frac{1}{\Gamma(\alpha)} \lambda^\alpha t^{\alpha-1} e^{-\lambda t} dt = 1. \quad (8.6)$$

In particular, (8.6) must hold for the value $\lambda = 1$, from which we obtain the formula

$$\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} e^{-t} dt. \quad (8.7)$$

For a general value $\lambda > 0$, we obtain, by a simple substitution,

$$\int_0^{+\infty} t^{\alpha-1} e^{-\lambda t} dt = \frac{1}{\lambda^\alpha} \int_0^{+\infty} t^{\alpha-1} e^{-t} dt = \frac{\Gamma(\alpha)}{\lambda^\alpha},$$

thus, the required identity (8.6) also holds for a general value $\lambda > 0$.

The function Γ defined by formula (8.7) is called the *gamma function*. It is generally defined for all complex numbers except zero and negative integers, and the integral in (8.7) converges whenever $\alpha > 0$. The gamma function is also often referred to as the *Euler integral of the second kind* or the *Euler function*, and in probability theory it serves as a normalizing factor in the formula (8.5), ensuring that the integral of the probability

⁴Recall that for the factorial of any natural number k , the recursive relation $(k-1)! = (k-1)(k-2)!$ holds with the initial condition $0! = 1$.

density function equals one, as demonstrated above. A distribution with density of the form

$$f(t) = \begin{cases} \frac{1}{\Gamma(\alpha)} \lambda^\alpha t^{\alpha-1} e^{-\lambda t} & \text{for } t > 0, \\ 0 & \text{for } t \leq 0, \end{cases}$$

is called the *gamma distribution* and is denoted by $\Gamma(\alpha, \lambda)$ ⁵, where $\alpha > 0$ is the shape parameter and $\lambda > 0$ is the intensity parameter of the distribution. The cumulative distribution function of this distribution for $x > 0$ is given by the integral⁶

$$F(x) = \int_0^x f(t) dt = \frac{1}{\Gamma(\alpha)} \int_0^{\lambda x} t^{\alpha-1} e^{-t} dt. \quad (8.8)$$

The expectation and the variance of a random variable X with the gamma distribution have, analogously to the Erlang distribution, the form

$$\mathbb{E}X = \frac{\alpha}{\lambda}, \quad \text{var } X = \frac{\alpha}{\lambda^2}.$$

Similarly to the exponential distribution, the gamma distribution can also be parameterized using the parameter $\beta = 1/\lambda$ (see the footnote 2 on page 154), which in the case of the gamma distribution is called the scale parameter. This parameterization is used, for example, by Microsoft Excel in the function `= GAMMA.DIST`, which is employed to compute probabilities for the Erlang distribution and, more generally, the gamma distribution. We will demonstrate this in the following examples.

Example 8.9. Let $X \sim \text{Erlang}(2, \frac{1}{3})$. Using an appropriate function in MS Excel, compute:

- a) $\mathbb{P}(X \leq 10)$,
- b) $\mathbb{P}(X \geq 20)$,
- c) $\mathbb{P}(10 < X < 20)$.

Solution. For the solution, we use the function `= GAMMA.DIST`, into which we substitute the parameters $\alpha = k = 2$ and $\beta = \frac{1}{\lambda} = 3$.

- a) $\mathbb{P}(X \leq 10) = F(10) \approx 0.845$,

$$= \text{GAMMA.DIST}(10, 2, 3; \text{TRUE})$$

⁵Sometimes the notation Ga is also used to distinguish it from the symbol Γ for the gamma function.

⁶The integral on the right-hand side of formula (8.8) is called the *lower incomplete gamma function*.

b) $\mathbb{P}(X \geq 20) = 1 - \mathbb{P}(X < 20) \approx 0.00975$,

$$= 1 - \text{GAMMA.DIST}(20, 2, 3, \text{TRUE})$$

c) $\mathbb{P}(10 < X < 20) = F(20) - F(10) \approx 0.145$.

$$= \text{GAMMA.DIST}(20, 2, 3, \text{TRUE}) - \text{GAMMA.DIST}(10, 2, 3, \text{TRUE})$$

△

Example 8.10. Let $X \sim \Gamma(2.5, \frac{1}{3})$. Using an appropriate function in MS Excel, compute:

a) $\mathbb{P}(X \leq 10)$,

b) $\mathbb{P}(X \geq 20)$,

c) $\mathbb{P}(10 < X < 20)$.

Solution. For the solution, we use the function $= \text{GAMMA.DIST}$, into which we substitute the parameters $\alpha = 2.5$ and $\beta = \frac{1}{\lambda} = 3$.

a) $\mathbb{P}(X \leq 10) = F(10) \approx 0.753$,

$$= \text{GAMMA.DIST}(10, 2.5, 3, \text{TRUE})$$

b) $\mathbb{P}(X \geq 20) = 1 - \mathbb{P}(X < 20) = 1 - F(20) \approx 0.02$,

$$= 1 - \text{GAMMA.DIST}(20, 2.5, 3, \text{TRUE})$$

c) $\mathbb{P}(10 < X < 20) = F(20) - F(10) \approx 0.226$.

$$= \text{GAMMA.DIST}(20, 2.5, 3, \text{TRUE}) - \text{GAMMA.DIST}(10, 2.5, 3, \text{TRUE})$$

△

Example 8.11. The mean lifetime of an electric shaver is two years. What is the probability that the average lifetime of 150 randomly selected shavers will exceed 27 months?

Solution. Let $X_1, X_2, \dots, X_{150}$ be random variables representing the lifetimes of the 1st, 2nd, ..., 150th shaver. Each of these random variables follows an exponential distribution with an identical intensity parameter: $\lambda = \frac{1}{24}$ and $X_i \sim \text{Exp}(\frac{1}{24})$, $i = 1, \dots, 150$, if we choose one month as the time unit.

Define the random variable

$$Y = \sum_{i=1}^{150} X_i.$$

Assuming that the lifetimes of the individual shavers are independent random variables, it follows that

$$Y \sim \text{Erlang}\left(150, \frac{1}{24}\right).$$

We seek the probability that “the average lifetime of the 150 randomly selected shavers will exceed 27 months”, i.e.

$$\mathbb{P}\left(\frac{1}{150}Y > 27\right) = \mathbb{P}(Y > 27 \cdot 150) = \mathbb{P}(Y > 4050).$$

To compute $\mathbb{P}(Y > 4050)$, we use the function

`= 1-GAMMA.DIST(4050,150,24,TRUE)`

$\mathbb{P}(Y > 4050) = 1 - \mathbb{P}(Y \leq 4050) \approx 0.067$, i.e. 6.7%. △

The Erlang distribution is used primarily in reliability theory: if the exponential distribution models the time to failure of a component, then the operating time until the k -th failure is described by the Erlang distribution. The previous example 8.11 also falls into this category. The gamma distribution with a non-integer shape parameter α appears in specialized statistical methods⁷ or as a model for the magnitude of an insurance claim. A special case of the gamma distribution is also the so-called Pearson χ^2 distribution, which will be discussed in more detail in Section 8.7 on page 176.

8.4 General Normal Distribution

General normal distribution, also referred to as the *Gaussian distribution* is defined by the density of the form

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty, \quad (8.9)$$

where the parameters of the distribution are a real number μ representing the mean of the distribution and a positive real number σ^2 representing the variance of the distribution. The probability density function $f(x)$ has a characteristic bell-shaped form; its graph is symmetric about the vertical line passing through the mean μ , and it attains its maximum at the point μ , so that μ is also the mode of the distribution.

The cumulative distribution function has the form

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt, \quad -\infty < x < \infty. \quad (8.10)$$

The cumulative distribution function of the normal distribution is one of the well-known functions for which no closed-form analytical expression exists,

⁷For example, it arises in the estimation of the parameter λ of a Poisson process using Bayesian statistics.

and it is available only in the integral form (8.10). The expectation and the variance have the forms

$$\mathbb{E}X = \mu, \quad \text{var } X = \sigma^2.$$

Due to symmetry, the mean μ is also the median of the distribution.

It can be shown that the normal distribution is uniquely determined by its mean μ and variance σ^2 . It is customary to denote it by $N(\mu, \sigma^2)$ ⁸. A special role is played by the so-called normalized (standardized) normal distribution $N(0, 1)$, which is discussed in detail in a separate section 8.5.

Example 8.12. For the random variable $X \sim N(0, 2)$, compute the probability that the random variable X is:

- a) exactly equal to 0,
- b) less than or equal to -2 ,
- c) greater than 4,
- d) at least -1 and at most 2.

Solution. For the computation, we use the function `= NORM.DIST` in Microsoft Excel. The distribution is parameterized in terms of the standard deviation, therefore we substitute the value $\sigma = \sqrt{2}$ as the third parameter.

a) $\mathbb{P}(X = 0) = 0,$

b) $\mathbb{P}(X \leq -2) = \frac{1}{\sqrt{2\pi}\sqrt{2}} \int_{-\infty}^{-2} e^{-\frac{t^2}{4}} dt = F(-2) \approx 0.07865,$

`= NORM.DIST(-2,0,SQRT(2),TRUE)`

c) $\mathbb{P}(X > 4) = \frac{1}{\sqrt{2\pi}\sqrt{2}} \int_4^{\infty} e^{-\frac{t^2}{4}} dt$
 $= 1 - \mathbb{P}(X \leq 4) = 1 - F(4) \approx 0.0023,$

`= 1-NORM.DIST(4,0,SQRT(2),TRUE)`

d) $\mathbb{P}(-1 \leq X \leq 2) = \frac{1}{\sqrt{2\pi}\sqrt{2}} \int_{-1}^2 e^{-\frac{t^2}{4}} dt = F(2) - F(-1) \approx 0.6816.$

`= NORM.DIST(2,0,SQRT(2),TRUE)-NORM.DIST(-1,0,SQRT(2),TRUE)`

△

Example 8.13. Construct the graphs of the probability density functions and the cumulative distribution functions of random variables that follow the distribution

⁸This convention is again not entirely standardized in the literature; some authors use the standard deviation σ as the second parameter and write $N(\mu, \sigma)$. In this textbook, however, we adhere to the parametrization in terms of the variance σ^2 .

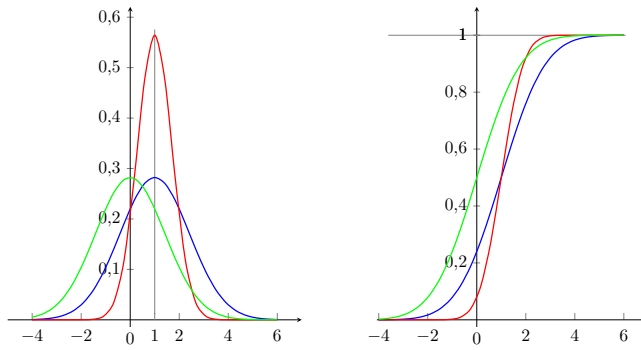


Figure 8.5: Graphs of the probability density functions and the cumulative distribution functions of a random variable with the normal distribution: $N(1, 2)$ (blue), $N(1, \frac{1}{2})$ (red), $N(0, 2)$ (green)

- a) $N(1, 2)$,
- b) $N(1, \frac{1}{2})$,
- c) $N(0, 2)$.

Assess the effect of the parameter σ^2 on the shape of the graphs.

Solution. The graphs are shown in Figure 8.5. The variance parameter σ^2 affects the “spread of the density” (more precisely, its peakedness): the values of the random variable occur predominantly in the interval $(\mu - 3\sigma, \mu + 3\sigma)$; for example, in case (a), they lie in the interval $(-5, 7)$. A more detailed explanation can be found on page 173 in Section 8.5 on the standard normal distribution. $\triangle$

Remark 8.1. *It can be shown that the expectation $\mathbb{E}X = \mu$, the variance $\text{var } X = \sigma^2$, and that all odd central moments (except for the first) are zero, i.e.*

$$\mathbb{E}(X - \mathbb{E}X)^{2k-1} = 0, \quad k = 2, 3, \dots,$$

and that the even central moments are of the form

$$\mathbb{E}(X - \mathbb{E}X)^{2k} = 1 \cdot 3 \cdot 5 \cdots (2k - 1) \sigma^{2k}, \quad k = 1, 2, \dots$$

The normal distribution is of fundamental importance in probability theory and mathematical statistics. In Section 11.3, it will be shown that the sum of a large number of random variables, under very general assumptions about their distributions, is approximately normally distributed. This explains the key role of this distribution in probability theory and mathematical statistics. Random variables encountered in the real world can very

often be regarded as the result of the combined effect of a large number of small random influences with different distributions. Owing to the above result, one may expect that the normal distribution provides a suitable model for such random variables. The most common type of normally distributed random variables are random errors, for example measurement errors caused by a large number of unknown and mutually independent factors. The normal distribution is also an appropriate model for many physical, technical, and biological quantities, such as the height of individuals in a homogeneous population, the annual amount paid by an insurance company for claims, etc.

Example 8.14. From biological statistics, it is known that a population of a certain species of flowers attains height X (in cm) according to the normal distribution $N(20, 16)$. Compute the probability that a randomly selected flower has height (in centimeters)

- a) less than 16,
- b) greater than 20,
- c) between 12 and 28,
- d) less than 12 or greater than 28,
- e) equal to 22.

Solution.

a) $\mathbb{P}(X < 16) = F(16) \approx 0.159,$

$= \text{NORM.DIST}(16, 20, 4, \text{TRUE})$

b) $\mathbb{P}(X > 20) = 1 - \mathbb{P}(X \leq 20) = 1 - F(20) = 0.5.$

$= 1 - \text{NORM.DIST}(20, 20, 4, \text{TRUE})$

Note that in this case we did not need to compute the probability, since it follows directly from the fact that 20 is the median of the distribution.

c) $\mathbb{P}(12 \leq X \leq 28) = F(28) - F(12) \approx 0.955,$

$= \text{NORM.DIST}(28, 20, 4, \text{TRUE}) - \text{NORM.DIST}(12, 20, 4, \text{TRUE})$

d) $\mathbb{P}(X < 12) + \mathbb{P}(X > 28) = F(12) + 1 - F(28) \approx 0.046,$

$= \text{NORM.DIST}(12, 20, 4, \text{TRUE}) + 1 - \text{NORM.DIST}(28, 20, 4, \text{TRUE})$

e) $\mathbb{P}(X = 22) = 0.$

△

Example 8.15. In banking regulation, a measure called the *value at risk* is used to assess the riskiness of an investment portfolio. It represents the magnitude of the portfolio loss that will be exceeded only with a prescribed small probability; here, the loss of a portfolio is typically understood as the negative of its return.

Suppose that the daily loss X of a certain portfolio (in percentage points) during a financial crisis follows a normal distribution with mean 0.015 and variance 0.01. Compute the 1% and 5% value at risk of this portfolio.

Solution. The 1%, respectively 5%, value at risk of the portfolio corresponds, from a probabilistic perspective, to the 99%, respectively 95%, quantile of the loss distribution $N(0.015, 0.01)$ (the risk is represented by the right tail⁹ of the distribution, which represents large losses). For its computation, we use the function `= NORM.INV`, into which we substitute the standard deviation $\sigma = \sqrt{0.01} = 0.1$ as the third parameter.

$$= \text{NORM.INV}(0.99, 0.015, 0.1), \quad = \text{NORM.INV}(0.95, 0.015, 0.1)$$

The values at risk of the portfolio are $F^{-1}(0.99) \approx 0.248$, and $F^{-1}(0.95) \approx 0.179$. These values represent the portfolio loss in percentage points whose exceedance is only weakly probable. $\triangle$

8.5 Standard normal distribution

The so-called *standard* (also standardized or standard) *normal distribution* is a special case of the normal distribution for which $\mu = 0$ and $\sigma^2 = 1$. It arises as the result of the so-called standardization process, described mathematically by the following theorem, which we state without proof.

Theorem 8.2 (Standardization of a normally distributed random variable).
Let the random variable $Z \sim N(0, 1)$ and

$$X = \mu + \sigma Z.$$

Then $X \sim N(\mu, \sigma^2)$. Conversely, if $X \sim N(\mu, \sigma^2)$, then the random variable

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1). \quad (8.11)$$

The transformation (8.11) simplifies the formulation of many theorems in probability theory and statistics, since it essentially states that a general normal distribution and the standard normal distribution can be easily transformed into one another. The practical use of this transformation is illustrated in Example 8.16.

Example 8.16. Return to the formulation of Example 8.14, compute the corresponding probabilities using the transformation to the standard normal distribution, and compare them with the original results.

⁹A tail is the part of the graph of a probability density function that deviates only slightly from the x -axis compared to the rest of the density.

Solution. We transform the random variable X to the standard normal variable

$$Z = \frac{X - 20}{4}$$

and apply the same transformation to the other side of the inequality; this is called the z -score. The z -score can be computed using the `= STANDARDIZE` function in MS Excel if the values of μ and σ of the non-standardized random variable are known. In our example, for part a), we have $\mu = 20$, $\sigma = 4$, and thus

$$\mathbb{P}(X < 16) = \mathbb{P}\left(\frac{X - 20}{4} < \frac{16 - 20}{4}\right) = \mathbb{P}(Z < -1),$$

where the right-hand side of the inequality is obtained as the result

$$= \text{STANDARDIZE}(16, 20, 4)$$

The remaining cases are computed in a similar way. The cumulative distribution function of the standard normal distribution is denoted in the following examples by the traditional symbol Φ .

a) $\mathbb{P}(X < 16) = \mathbb{P}(Z < -1) = \Phi(-1) \approx 0.1586,$

$$= \text{NORM.S.DIST}(-1, \text{TRUE})$$

b) $\mathbb{P}(X > 20) = \mathbb{P}\left(Z > \frac{20 - 20}{4}\right) = \mathbb{P}(Z > 0) = 1 - \Phi(0) = 0.5.$

$$= \text{STANDARDIZE}(20, 20, 4)$$

$$= 1 - \text{NORM.S.DIST}(0, \text{TRUE})^{10}$$

In this case, however, it is not necessary to compute the value $\Phi(0)$ using any software: we can simply observe that 0 is also the median of the distribution $\mathcal{N}(0, 1)$, from which the answer follows directly.

c) $\mathbb{P}(12 \leq X \leq 28) = \mathbb{P}\left(\frac{12 - 20}{4} \leq Z \leq \frac{28 - 20}{4}\right) =$

$$= \mathbb{P}(-2 \leq Z \leq 2) = \Phi(2) - \Phi(-2) \approx 0.9545.$$

x	
1	<code>STANDARDIZE(12, 20, 4) STANDARDIZE(28, 20, 4)</code>
2	<code>NORM.S.DIST(2, TRUE) - NORM.S.DIST(-2, TRUE)</code>

In this case as well, the computation can be simplified by exploiting the symmetry of the distribution. Indeed, $\Phi(-2) = 1 - \Phi(2)$ (that is, the left tail of the distribution up to -2 has the same size as the right tail

¹⁰The function `= STANDARDIZE` can be inserted as the first argument of the function `= NORM.S.DIST`, yielding

$$= \text{NORM.S.DIST}(\text{STANDARDIZE}(20, 20, 4), \text{TRUE}).$$

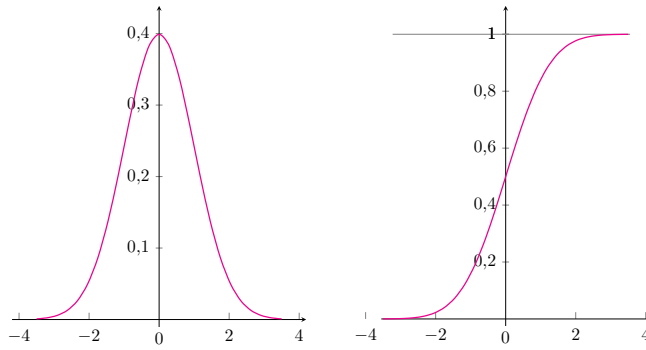


Figure 8.6: Graphs of the probability density function and the cumulative distribution function of a random variable with the standard normal distribution

from 2), hence $\Phi(2) - \Phi(-2) = 2\Phi(2) - 1$. From the above values of the cumulative distribution function, it is therefore sufficient in practice to compute only one of them.

$$= 2 * \text{NORM.S.DIST}(2, \text{TRUE}) - 1$$

d) $\mathbb{P}(X < 12) + \mathbb{P}(X > 28) = 1 - \mathbb{P}(12 \leq X \leq 28) \approx 0.04550$,

where we make use of the previously computed result of Problem c).

△

Substituting the parameters $\mu = 0$ and $\sigma^2 = 1$ into (8.9), we obtain the formula for the probability density function of the standard normal distribution in the form

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty.$$

In Figure 8.6 (left), the graph of the probability density function $\varphi(z)$ is shown; it has the shape of a bell curve, known as the Gaussian¹¹ curve. This curve is symmetric with respect to the vertical axis passing through 0, perpendicular to the horizontal axis, and it attains its maximum $\frac{1}{\sqrt{2\pi}} \approx 0.39894$ at this point.

The cumulative distribution function of the standard normal distribution again has no closed-form analytical expression and is therefore given only in its integral form

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-t^2/2} dt, \quad -\infty < z < \infty.$$

¹¹Johann Carl Friedrich Gauss (1777–1855), German mathematician and physicist.

The graph of this cumulative distribution function is shown in Figure 8.6 on the right. The cumulative distribution functions of the general normal distribution and the standard normal distribution can also be transformed into each other: as a consequence of Theorem 8.2, we have the relation

$$F_X(x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

for $X \sim \mathbf{N}(\mu, \sigma^2)$.

For the notation of the standard normal distribution, the symbol $\mathbf{U}$ (without specifying parameters) is sometimes used, and the corresponding random variable is denoted by U (written as $U \sim \mathbf{U}$). This notation must be distinguished from that used for the uniform distribution. We therefore adhere to the notation $\mathbf{N}(0, 1)$ and write $Z \sim \mathbf{N}(0, 1)$. Note also that the notation φ for the density and Φ for the cumulative distribution function of the standard normal distribution is a fairly well-established convention, which highlights the importance of this distribution in probability theory.

First, let us verify that $\varphi(z)$ is indeed a density. For the calculations, we will use the so-called Gaussian integral (the integral of the Gaussian function)

$$\int_{-\infty}^{\infty} e^{-y^2} dy = \sqrt{\pi}. \quad (8.12)$$

We first observe that $\varphi(z) \geq 0$ for all $z \in \mathbb{R}$. Let us further verify

$$\int_{-\infty}^{\infty} \varphi(z) dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sqrt{2} e^{-y^2} dy = 1,$$

and hence $\varphi(z)$ is a valid density. The expectation and all odd moments are equal to zero, i.e.

$$\mathbb{E}Z = \mathbb{E}Z^{2k-1} = 0, \quad k = 1, 2, 3, \dots,$$

since the function

$$\frac{1}{\sqrt{2\pi}} z^{2k-1} e^{-\frac{z^2}{2}}$$

is an odd function and is integrable over $\mathbb{R}$. Below, we will show that the even moments are equal to

$$\mathbb{E}Z^{2k} = 1 \cdot 3 \cdot 5 \cdots (2k-1) = \frac{(2k)!}{2^k k!}, \quad k = 1, 2, 3, \dots \quad (8.13)$$

Since $\mathbb{E}Z = 0$, by substituting the value $k = 1$ we obtain

$$\text{var } Z = \mathbb{E}Z^2 = 1.$$

Note also the fourth absolute moment $\mathbb{E}Z^4$, i.e. the kurtosis of the distribution¹², which is equal to the constant value 3. If we perform in (8.12) the substitution $y = z\sqrt{\frac{t}{2}}$, we obtain

$$\int_{-\infty}^{\infty} e^{-\frac{tz^2}{2}} dz = \sqrt{\frac{2\pi}{t}}. \quad (8.14)$$

Differentiate both sides of (8.14) with respect to t . We obtain

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{tz^2}{2}} dz = t^{-\frac{3}{2}}.$$

In general, after k repeated differentiations, we obtain

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^{2k} e^{-\frac{tz^2}{2}} dz = 1 \cdot 3 \cdots (2k-1) \cdot t^{-\frac{2k+1}{2}}. \quad (8.15)$$

The result (8.13) then follows from (8.15) by setting $t = 1$.

The values of the probability density function $\varphi(z)$ of the standard normal distribution for a given value of z can be determined in MS Excel using the function `=PHI`, which is equivalent to the function `=NORM.S.DIST(z,FALSE)`. For example, on page 170 we determined the maximum value of the standard normal distribution, which the density attains at 0: here `=PHI(0)`, i.e. $\varphi(0) \approx 0.398942$.

Example 8.17. Let the random variable Z have the standard normal distribution. Compute the probability that the random variable Z is

- a) less than -1 ,
- b) greater than 2 ,
- c) between -1 and 1 .

Solution.

$$\text{a) } \mathbb{P}(Z < -1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-1} e^{-\frac{t^2}{2}} dt = \Phi(-1) \approx 0.159,$$

$$= \text{NORM.S.DIST}(-1, \text{TRUE})$$

$$\text{b) } \mathbb{P}(X > 2) = \frac{1}{\sqrt{2\pi}} \int_2^{\infty} e^{-\frac{t^2}{2}} dt = 1 - \Phi(2) \approx 0.023,$$

$$= 1 - \text{NORM.S.DIST}(2, \text{TRUE})$$

¹²See Definition 6.5 on page 113.

$$c) \mathbb{P}(-1 < X < 1) = \frac{1}{\sqrt{2\pi}} \int_{-1}^1 e^{-\frac{t^2}{2}} dt = \Phi(1) - \Phi(-1) \approx 0.683.$$

= `NORM.S.DIST(1,TRUE)-NORM.S.DIST(-1,TRUE)`

△

Probabilities of the standard normal distribution are also used for the interpretation of the standard deviation. For the standard normal distribution, we have

- 68.3% of the population lies in the interval $(-1, 1)$ ¹³,
- 95% of the population lies in the interval $(-1.96, 1.96)$,
- 95.5% of the population lies in the interval $(-2, 2)$,
- 99.7% of the population lies in the interval $(-3, 3)$,
- 99.9% of the population lies in the interval $(-3.29, 3.29)$.

In terms of the quantile function, we may write

$$\Phi^{-1}\left(\frac{1}{2} + \frac{0.683}{2}\right) = \Phi^{-1}(0.841) \approx 1,$$

and similarly, for example, $\Phi^{-1}(0.975) \approx 1.96$ and $\Phi^{-1}(0.9987) \approx 3$. Applying Theorem 8.2, we obtain the analogous statements for a general normally distributed random variable:

- 68.3% of the population lies in the interval $(\mu - \sigma, \mu + \sigma)$,
- 95% of the population lies in the interval $(\mu - 1.96\sigma, \mu + 1.96\sigma)$,
- 95.5% of the population lies in the interval $(\mu - 2\sigma, \mu + 2\sigma)$,
- 99.7% of the population lies in the interval $(\mu - 3\sigma, \mu + 3\sigma)$,
- 99.9% of the population lies in the interval $(\mu - 3.29\sigma, \mu + 3.29\sigma)$.

8.6 Lognormal distribution

A random variable X has a *lognormal* (informally also lognormal) *distribution* $\text{LN}(\mu, \sigma^2)$ with parameters μ and σ^2 if the random variable $\ln X$ has the normal distribution $\text{N}(\mu, \sigma^2)$. The probability density function and the cumulative distribution function of the lognormal distribution are of the form

$$f(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(\ln x - \mu)^2}{\sigma^2}} & x \in (0, \infty), \\ 0 & \text{otherwise,} \end{cases} \quad F(x) = \begin{cases} 0 & \text{for } x \leq 0, \\ \Phi\left(\frac{\ln x - \mu}{\sigma}\right) & \text{for } x > 0. \end{cases}$$

¹³We mean that for $Z \sim \text{N}(0, 1)$, $\mathbb{P}(-1 \leq Z \leq 1) = 0.683$.

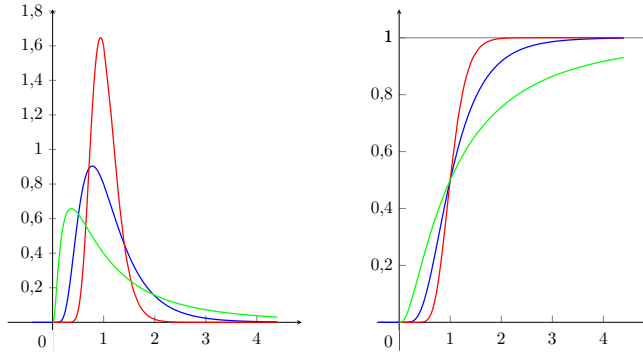


Figure 8.7: Graphs of the probability density functions and cumulative distribution functions of a random variable with the lognormal distribution: $\text{LN}(0, 0.25)$ (blue), $\text{LN}(0, 0.0625)$ (red), $\text{LN}(0, 1)$ (green)

The density of the lognormal distribution is asymmetric, skewed to the right. The expectation and variance of $X \sim \text{LN}(\mu, \sigma^2)$ are given by

$$\mathbb{E}X = e^{\mu + \frac{\sigma^2}{2}}, \quad \text{var } X = e^{2\mu + \sigma^2} (e^{\sigma^2} - 1).$$

Example 8.18. Construct the graph of the probability density function and the cumulative distribution function of a random variable with the lognormal distribution

- $\text{LN}(0, 0.25)$,
- $\text{LN}(0, 0.0625)$,
- $\text{LN}(0, 1)$.

Compute the values of the 17% quantile for all the distributions mentioned above.

Solution. The graphs of the probability density functions and the distribution functions are shown in figure 8.7. We compute the values of the quantile as follows:

a) $F^{-1}(0.17) \approx 0.6205$,

$$= \text{LOGNORM.INV}(0.17, 0, 0.5)$$

b) $F^{-1}(0.17) \approx 0.7878$,

$$= \text{LOGNORM.INV}(0.17, 0, 0.25)$$

c) $F^{-1}(0.17) \approx 0.3851$.

$$= \text{LOGNORM.INV}(0.17, 0, 1)$$

△

The lognormal distribution is used for random variables that take only positive values, where a large number of low values with occasional occurrence of higher values is typical. Such random variables arise, for example, in economics, insurance, and reliability theory. We present an example of the use of the lognormal distribution in insurance for computing the probability that a so-called insurance claim will be filed: if the loss in an insured event is greater than the value r , which represents the increase in the insurance premium in the following insurance period provided that the claim is filed, then the client has an incentive to file the claim. At the beginning, after concluding the contract with the insurance company, the client is assigned to the zero bonus class, where the premium is equal to 100% of the list price. Introduce the following random variables: let Y denote the number of insured events, which follows the Poisson distribution with parameter λ : $Y \sim \text{Po}(\lambda)$. Further, let Z denote the amount of loss, which we assume follows the lognormal distribution with parameters μ and σ^2 , i.e. $Z \sim \text{LN}(\mu, \sigma^2)^{14}$. Finally, let A denote the event “the client files an insurance claim”. Then the probability of filing an insurance claim is given by the relation

$$\mathbb{P}(A) = \mathbb{P}(Y > 0) \cdot \mathbb{P}(Z > r) = (1 - e^{-\lambda}) \left[1 - \Phi\left(\frac{\ln r - \mu}{\sigma}\right) \right]$$

under the assumption that the events of the occurrence of an insured event ($Y > 0$) and the exceedance of the loss amount ($Z > r$) are mutually independent.

Example 8.19. A client of an insurance company is in the first bonus class, where they pay 60% of the premium for the agreed car insurance, which amounts to 2 500 CZK. The parameter λ of the Poisson distribution for the number of insured events is 0.15, and the loss amount follows a lognormal distribution with parameters $\mu = 6$ and $\sigma^2 = 2$. Determine the probability that the client will file an insurance claim.

Solution. The client will file an insurance claim if the loss exceeds the increase in the premium r , where in this case $r = 2\,500 - 0.6 \cdot 2\,500 = 1\,000$ CZK. Then

$$\mathbb{P}(A) = (1 - e^{-0.15}) \left[1 - \Phi\left(\frac{\ln 1\,000 - 6}{\sqrt{2}}\right) \right] \approx 0.036.$$

```
= (1-POISSON.DIST(0,0.15,TRUE))*
```

```
= *(1-LOGNORM.DIST(1000,6,SQRT(2),TRUE))
```

¹⁴This area is indeed characterized by a large number of events with relatively low losses, while occasionally an event with a substantial loss also occurs.

The client will file an insurance claim with probability approximately 3.6%. $\triangle$

8.7 Pearson Distribution χ^2

From the standard normal distribution introduced in Section 8.5, a family of three distributions (Pearson, Student, and Fisher–Snedecor) is derived, which find direct applications in statistics, as we shall see in the second part of this book.

Let the random variables $Z_1, Z_2, \dots, Z_k$ be independent¹⁵ and have the standard normal distribution $N(0, 1)$. Then the variable

$$Q_k = \sum_{i=1}^k Z_i^2 \quad (8.16)$$

has the so-called *Pearson*¹⁶ χ^2 *distribution* (read chi-square) with k degrees of freedom. The number of degrees of freedom is a parameter of this distribution and is equal to the number of summands in equation (8.16). The distribution may be denoted by $\chi^2(k)$, but the notation with a subscript, χ_k^2 , is more common, and we adopt it in this text.

It can be shown that the Pearson distribution χ_k^2 is a special case of the gamma distribution (see Section 8.3 on page 159) with parameters $\alpha = \frac{k}{2}$ and $\lambda = \frac{1}{2}$. From this we obtain the formula for the probability density function

$$f(z) = \frac{1}{\Gamma\left(\frac{k}{2}\right) \cdot 2^{\frac{k}{2}}} \cdot z^{\frac{k}{2}-1} \cdot e^{-\frac{z}{2}}, \quad z > 0,$$

for the cumulative distribution function

$$F(x) = \frac{1}{\Gamma\left(\frac{k}{2}\right)} \int_0^{\frac{x}{2}} t^{\frac{k}{2}-1} e^{-t} dt,$$

and for the expectation and variance

$$\mathbb{E}\chi_k^2 = k, \quad \text{var } \chi_k^2 = 2k.$$

Example 8.20. Let the random variable X have the distribution χ_5^2 . Compute the probability that the random variable X :

a) is less than or equal to 5,

¹⁵For more on the independence of random variables, see Section 9.4 on page 202.

¹⁶Karl Pearson (1857–1936), British mathematician.

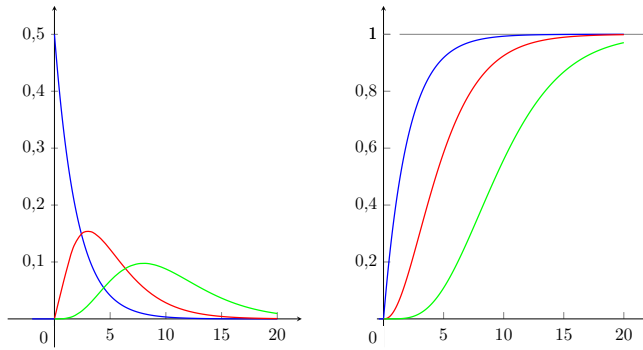


Figure 8.8: Graphs of the probability density functions and the cumulative distribution functions of a random variable with the Pearson distribution: χ^2_2 (blue), χ^2_5 (red), χ^2_{10} (green)

- b) is not greater than 10,
- c) takes values between 5 and 10.

Solution.

- a) $\mathbb{P}(X \leq 5) = F(5) \approx 0.584$,
 $\text{= CHISQ.DIST}(5, 5, \text{TRUE})$
- b) $\mathbb{P}(X > 10) = 1 - \mathbb{P}(X \leq 10) = 1 - F(10) \approx 0.07$,
 $\text{= } 1 - \text{CHISQ.DIST}(10, 5, \text{TRUE})$
- c) $\mathbb{P}(5 < X < 10) = F(10) - F(5) \approx 0.341$.

$$\text{= CHISQ.DIST}(10, 5, \text{TRUE}) - \text{CHISQ.DIST}(5, 5, \text{TRUE})$$

△

Example 8.21. Construct the graphs of the probability density function and the cumulative distribution function of the Pearson distribution χ^2_k for parameter values $k = 2$, $k = 5$, and $k = 10$.

Solution. Figure 8.8 shows the probability density functions and cumulative distribution functions of the Pearson distribution χ^2_k for $k = 2$, $k = 5$, and $k = 10$. Observe that for small values of k , the probability density function is highly asymmetric. For larger values of k , the density begins to resemble that of the normal distribution $N(k, 2k)$. △

For the computation of quantiles, traditionally denoted by $\chi^2_k(p)$, tables were previously used (typically for $k = 1, \dots, 100$). Nowadays, software is employed; for example, in MS Excel one uses the function = CHISQ.INV .

The Pearson χ^2 distribution is applied primarily in statistics, for instance in confidence intervals for the variance (see Section 14.3.3), in tests concerning the variance (Section 15.3), or in goodness-of-fit tests (Chapter 20).

8.8 Student's¹⁷ t -distribution

Let us consider two independent¹⁸ random variables: a random variable Z with the standard normal distribution $N(0, 1)$ and a random variable Q with the χ^2 distribution with k degrees of freedom. It can be shown that the random variable

$$T = \frac{Z}{\sqrt{Q}} \cdot \sqrt{k} = \frac{Z}{\sqrt{Q/k}}$$

has a distribution whose density can be written in the form

$$f(t) = \frac{1}{B(\frac{1}{2}, \frac{k}{2}) \cdot \sqrt{k}} \cdot \left(1 + \frac{t^2}{k}\right)^{-\frac{(k+1)}{2}}, \quad t \in \mathbb{R}, \quad (8.17)$$

where B is the so-called *beta function*, also known as the *Euler function* or the *Euler integral of the first kind*,

$$B(\alpha, \beta) = \frac{\Gamma(\alpha) \cdot \Gamma(\beta)}{\Gamma(\alpha + \beta)} \mathrm{d}x = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} \mathrm{d}x = \int_0^\infty \frac{x^{\alpha-1}}{(1+x)^{\alpha+\beta}}, \quad (8.18)$$

where $k > 0$, and Γ is the gamma function introduced in Section 8.3 by formula (8.7). The distribution with density (8.17) is called the *Student's distribution* with k degrees of freedom and is denoted by t_k ¹⁹. For the random variable $T \sim t_k$, the expectation exists for $k > 1$, and for $k > 2$ the variance is given by

$$\mathbb{E}T = 0, \quad \mathrm{var} T = \frac{k}{k-2}.$$

The density of the Student's distribution is symmetric with respect to the vertical axis; hence, for the quantiles, typically denoted by $t_k(p)$, we have

$$t_k(p) = -t_k(1-p).$$

This relation is used primarily to simplify formulas in statistics, as we shall see, for example, in Section 14.3.1.

¹⁷William Sealy Gosset (1876–1937), an English chemist and pioneer of statistics, who developed the t -distribution and published under the pseudonym *Student*.

¹⁸For more on the independence of random variables, see Section 9.4 on page 202.

¹⁹Also denoted $t(k)$ in the literature.

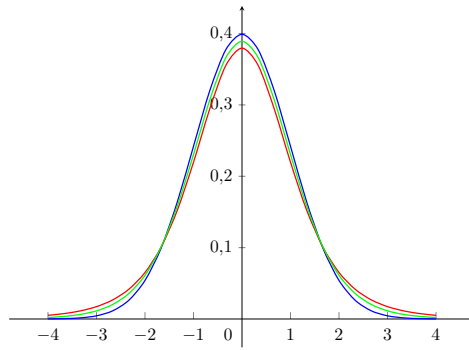


Figure 8.9: Approximation of the probability density function of the standard normal distribution using the Student's distribution: $N(0, 1)$ (blue), t_5 (red), t_{10} (green)

The Student's distribution with a small number of degrees of freedom k is applied in statistics (see, for example, Section 15.2) as a substitute for the normal distribution in cases where the probability of extreme outcomes is higher (it belongs to the class of distributions with so-called heavy tails). For larger values of k , it approaches the standard normal distribution. In fact, if $T_k \sim t_k$, it can be shown that

$$F_{T_k} \xrightarrow[k \rightarrow \infty]{} \Phi,$$

and a similar relation holds also for the densities of the distributions. This will be illustrated in Example 8.22.

Example 8.22. Using the graphs of the probability density functions, illustrate the closeness of the standard normal distribution and the Student's distribution. Choose t_5 and t_{10} . Determine the probability that the random variable $T \sim t_5$ takes values

- less than -1 ,
- greater than or equal to 2 ,
- between -1 and 1 .

Solution. The graphs of the probability density functions of random variables with the standard normal and the Student's distributions with five and ten degrees of freedom are shown in Figure 8.9.

- $\mathbb{P}(T < -1) = F(-1) \approx 0.1816$,

$$= \text{T.DIST}(-1, 5, \text{TRUE})$$

b) $\mathbb{P}(T \geq 2) = 1 - F(2) \approx 0.051,$

$= 1 - \text{T.DIST}(2, 5, \text{TRUE})$

c) $\mathbb{P}(-1 < T < 1) = F(1) - F(-1) \approx 0.6368.$

$= \text{T.DIST}(1, 5, \text{TRUE}) - \text{T.DIST}(-1, 5, \text{TRUE})$

△

Example 8.23. Let the random variable T have a Student's distribution with ten degrees of freedom. Compute:

a) the 1% quantile,

b) the 5% quantile,

c) the 95% quantile.

Solution. a) $t_{10}(0.01) = F_T^{-1}(0.01) \approx -2.76,$

$= \text{T.INV}(0.01, 10)$

b) $t_{10}(0.05) = F_T^{-1}(0.05) \approx -1.81,$

$= \text{T.INV}(0.05, 10)$

c) $t_{10}(0.95) = F_T^{-1}(0.95) \approx 1.81.$

$= \text{T.INV}(0.95, 10)$

△

A special case of the Student's distribution is the Student's distribution with one degree of freedom, which is called the Cauchy distribution and is used, for example, in finance to model extreme investment risks. Its cumulative distribution function and probability density function were studied in Exercise 5.5 and in Example 6.3. As we saw in Example 6.3, the Cauchy distribution has neither expectation nor variance (the corresponding integrals diverge).

8.9 Fisher–Snedecor²⁰ distribution

Let two independent²¹ random variables follow the χ^2 distribution, namely Q_1 with ν_1 degrees of freedom and Q_2 with ν_2 degrees of freedom. It can be shown that the random variable

$$F = \frac{Q_1/\nu_1}{Q_2/\nu_2} \quad (8.19)$$

²⁰Ronald Aylmer Fisher (1890–1962), a British mathematician, statistician, biologist, and geneticist. George Waddel Snedecor (1881–1974), an American mathematician and statistician

²¹For more on the independence of random variables, see Section 9.4 on page 202.

has a distribution whose density can be written in the form

$$f(x) = \frac{1}{B(\frac{\nu_1}{2}, \frac{\nu_2}{2})} \cdot \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \cdot \frac{x^{\frac{\nu_1-2}{2}}}{\left(1 + x \cdot \frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1+\nu_2}{2}}}, \quad x > 0, \quad (8.20)$$

where $B(\cdot, \cdot)$ is the beta function defined by formula (8.18) on page 178. The distribution with density (8.20) is called the *Fisher–Snedecor distribution*, or simply the *F distribution*, denoted by $F(\nu_1, \nu_2)$. The parameters ν_1 and ν_2 are called the degrees of freedom (in this order). The expectation (for $n > 2$) and the variance (for $n > 4$) are given by

$$\mathbb{E}F = \frac{\nu_2}{\nu_2 - 2}, \quad \text{var } F = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}.$$

The quantiles of the distribution are traditionally denoted by $F_F^{-1}(p) = F_{\nu_1, \nu_2}(p)$. The Fisher–Snedecor distribution is widely used in statistical hypothesis testing, for example in the analysis of variance (see Section 17.1) or in regression analysis (see Section 19.2).

Example 8.24. Let the random variable F follow the Fisher–Snedecor distribution with parameters $\nu_1 = 5$ and $\nu_2 = 10$. Compute the probability that the random variable F takes the value

- a) less than or equal to 1,
- b) greater than 3,
- c) between 1 and 2.

Solution.

a) $\mathbb{P}(F \leq 1) = F(1) \approx 0.535,$

$= \text{F.DIST}(1, 5, 10, \text{TRUE})$

b) $\mathbb{P}(F > 3) = 1 - F(3) \approx 0.066,$

$= 1 - \text{F.DIST}(3, 5, 10, \text{TRUE})$

c) $\mathbb{P}(1 < F < 2) = F(2) - F(1) \approx 0.301.$

$= \text{F.DIST}(2, 5, 10, \text{TRUE}) - \text{F.DIST}(1, 5, 10, \text{TRUE})$

△

Example 8.25. Let the random variable F have the F distribution with 7 and 9 degrees of freedom. Compute the value

- a) the 0.1% quantile,
- b) the 1% quantile,
- c) the 5% quantile,

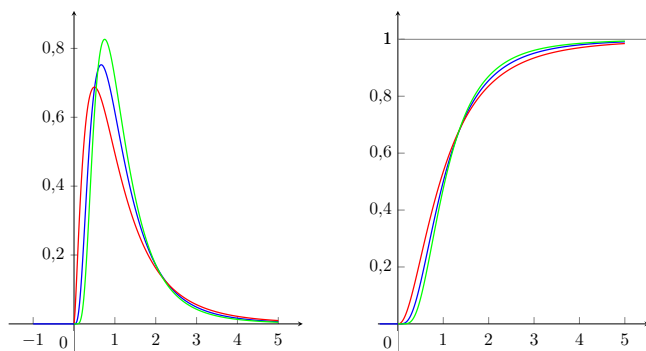


Figure 8.10: Graphs of the probability density functions and the cumulative distribution functions of a random variable with the Fisher–Snedecor distribution: $F(10, 10)$ (blue), $F(5, 10)$ (red), $F(20, 10)$ (green)

d) the 99% quantile.

Solution. a) $F_F^{-1}(0.001) = F_{7,9}(0.001) \approx 0.0698$,

$$= \text{F.INV}(0.001, 7, 9)$$

b) $F_F^{-1}(0.01) = F_{7,9}(0.01) \approx 0.1488$,

$$= \text{F.INV}(0.01, 7, 9)$$

c) $F_F^{-1}(0.05) = F_{7,9}(0.05) \approx 0.2720$,

$$= \text{F.INV}(0.05, 7, 9)$$

d) $F_F^{-1}(0.99) = F_{7,9}(0.99) \approx 5.6129$.

$$= \text{F.INV}(0.99, 7, 9)$$

△

Example 8.26. Create graphs of probability density functions and graphs of distribution functions of random variables with distribution

- a) $F(5, 10)$,
- b) $F(10, 10)$,
- c) $F(20, 10)$.

Solution. The graphs are shown in Figure 8.10.

△

8.10 Exercises

8.1 Verify the general validity of the formula for computing the expectation and the variance of a random variable X that follows the continuous uniform distribution.

8.2 Prove the correctness of the formula for the cumulative distribution function of the continuous uniform distribution.

8.3 A red light at a traffic signal lasts 90 seconds. A cyclist who arrives at the signal and does not know how long the red light has already been on wants to determine the probability that they will have to wait

- a) less than ten seconds,
- b) more than sixty seconds,
- c) between 20 and 40 seconds.

8.4 Charles travels to work by tram, which arrives at the stop every 8 minutes. What is the probability that if Charles arrives at the tram stop at a random time,

- a) he will not have to wait more than 3 minutes,
- b) he will have to wait more than 8 minutes,
- c) he will not have to wait at all and will board the tram immediately upon arrival,
- d) he will have to wait between 4 and 7 minutes?



Sketch the graph of the probability density function and the cumulative distribution function of the random variable considered in the problem.

8.5 The lifetime of a certain product follows an exponential distribution with mean 3 years. What warranty period can the manufacturer provide to customers if they require that the relative frequency of products that cease to function during the warranty period be approximately 10%?

8.6 A store expects a delivery of goods on a given day sometime between 12 and 16 hours. According to the supplier, the delivery is equally likely

to occur at any moment during this time interval. What is the probability that the goods will be delivered between one o'clock and half past one?

8.7 Construct the graph of the probability density function and the cumulative distribution function of the random variable T with the exponential distribution $\text{Exp}(\frac{1}{3})$. Determine and indicate on the graphs:

- a) $\mathbb{P}(T \leq 1)$,
- b) $\mathbb{P}(T = 1.4)$,
- c) $\mathbb{P}(T > 5)$,
- d) $\mathbb{P}(2 \leq T \leq 4)$.

8.8 Determine the mean service time in a shop, given that the probability of completing a service in less than 4 minutes is 0.25. Assume that the service time has an exponential distribution.

8.9 A stolen car is found within time t with probability

$$p(t) = 1 - e^{-\gamma t}, \quad \gamma > 0.$$

Determine the expected value of the search time T required to find the car.

8.10 What is the probability that a random variable X with the normal distribution $N(3, 1)$ takes the value

- a) less than 2,
- b) less than or equal to 2,
- c) greater than 4.5,
- d) equal to 1,
- e) between 2.5 and 3.5?



8.11 Construct the graph of the probability density function and the cumulative distribution function of a continuous random variable X with the standard normal distribution.

8.12 What are the values of μ and σ^2 for a normally distributed random variable X , given that $\mathbb{P}(X < 85) = 0.9$ and $\mathbb{P}(X < 95) = 0.95$?

8.13 What is the probability that a random variable with the standard normal distribution takes the value

- a) less than 1.64,
- b) greater than -1.64 ,

- c) between -1.96 and 1.96 ,
- d) greater than 2.33 ,
- e) less than -2.33 ,
- f) exactly 0.5 ?

8.14 A random variable X , representing a measurement error, has the distribution $N(0.2, 0.64)$. Determine

- a) the probability that the absolute value of X is less than 1 ,
- b) the value of the measurement error that is exceeded with probability 0.05 only.

8.15 In a quality inspection, all products whose length exceeds 77 cm are accepted. It has been found that the mean length of the product (the random variable X) μ is 75 cm and the standard deviation (the square root of the variance) σ is 5 cm. Assuming that the random variable under consideration is approximately normally distributed, determine

- a) the probability that a product that has passed the inspection is longer than 80 cm,
- b) how many products longer than 80 cm can be expected if $2\,261$ items are accepted during the inspection?

8.16 Construct the graph of the probability density function and the cumulative distribution function of a random variable X with the lognormal distribution, i.e. $X \sim \text{LN}(0, 1)$, on the interval $(0, 10]$. Compute the 23% quantile of the distribution $\text{LN}(0, 1)$.

8.17 Consider the time a student needs to solve a problem as a random variable with a gamma distribution, for which the mean is 1.5 and the variance is 0.75 . What is the probability that the time required to solve the problem exceeds 2 hours?

Solutions to the Exercises

8.1

$$\begin{aligned}\mathbb{E}X &= \int_{-\infty}^{\infty} xf(x) \, dx = \int_a^b x \frac{1}{b-a} \, dx = \frac{1}{b-a} \left[\frac{x^2}{2} \right]_a^b = \frac{a+b}{2}, \\ \text{var } X &= \mathbb{E}X^2 - (\mathbb{E}X)^2 = \int_a^b x^2 \frac{1}{b-a} \, dx - \left(\frac{a+b}{2} \right)^2 = \\ &= \frac{1}{b-a} \left[\frac{x^3}{3} \right]_a^b - \frac{(a+b)^2}{4} = \frac{a^2 + ab + b^2}{3} - \frac{a^2 + 2ab + b^2}{4} = \\ &= \frac{(a-b)^2}{12}.\end{aligned}$$

8.2 For $x \in [a, b]$ we have

$$F(x) = \int_a^x \frac{1}{b-a} \, dt = \frac{1}{b-a} [t]_a^b = \frac{x-a}{b-a}.$$

8.3 $X \sim \text{U}(0, 90)$, a) $1/9$, b) $1/3$, c) $2/9$.

8.4 $X \sim \text{U}(0, 8)$

- a) $\mathbb{P}(X \leq 3) = F(3) = 0.375$,
- b) $\mathbb{P}(X \geq 7) = 1 - \mathbb{P}(X \leq 7) = 1 - 0.875 = 0.125$,
- c) $\mathbb{P}(X = 0) = 0$,
- d) $\mathbb{P}(4 \leq X \leq 7) = F(7) - F(3) = 0.375$.

8.5 0.32 years, i.e. just under 4 months

8.6 $X \sim \text{U}(12, 16)$, 0.125.

8.7 a) 0.283 (area in green), b) 0 (“area” in cyan), c) 0.189 (area in blue), d) 0.250 (area in red).

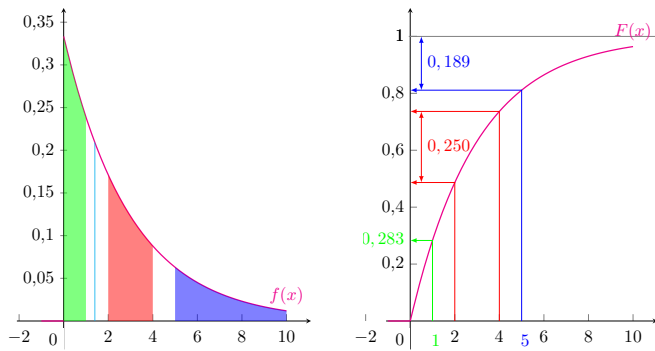


Figure 8.11: Graph of the probability density function and the cumulative distribution function from Problem 8.7

8.8 13 minutes and 54.25 seconds

8.9 $\mathbb{E}T = \frac{1}{\gamma}$

8.10 a) $\mathbb{P}(X < 2) \approx 0.159$, b) $\mathbb{P}(X \leq 2) \approx 0.159$, c) $\mathbb{P}(X > 4.5) \approx 0.068$, d) $\mathbb{P}(X = 1) = 0$, e) $\mathbb{P}(2.5 \leq X \leq 3.5) \approx 0.383$.

8.11 The solution is shown in Figure 8.6.

8.12 $\mu \approx 49.7$; $\sigma^2 \approx 757.64$.

8.13 a) 0.9495, b) 0.9495, c) 0.9500, d) 0.0099, e) 0.0099, f) 0.

8.14 a) 0.7745, b) 1.516.

8.15

- a) $X \sim N(75, 25)$, $\mathbb{P}(X > 80) = 1 - \mathbb{P}(X \leq 80) \approx 0.1587$,
 b) $6562 \cdot 0.1587 \approx 1041$.

8.16 The graph of the density and the cumulative distribution function is shown in Figure 8.7 on page 174. $F^{-1}(0.23) \approx 0.478$.

8.17 $X \sim \Gamma(3, 2)$, $\mathbb{P}(X > 2) = 1 - \mathbb{P}(X \leq 2) \approx 0.2381$.

Chapter 9

Random vector, independence of random variables

9.1 Random Vector

If we need to work simultaneously with several random variables, it is appropriate to introduce the concept of a random vector as a collection of random variables and to define the necessary probabilistic notions on it. In this chapter, we will do so and, in addition, introduce the concept of correlation and extend the notion of independence to random variables.

Definition 9.1 (Random vector)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space on which random variables $X_1, X_2, \dots, X_n$ are defined. Then the ordered n -tuple $\mathbf{X} = (X_1, \dots, X_n)^T$ is called an n -dimensional random vector¹.



A random vector is thus a mapping from Ω into $\mathbb{R}^n$. The values of a random vector can be interpreted geometrically as a point in the n -dimensional space. In particular, for $n = 2$ we write $\mathbf{X} = (X_1, X_2)^T$ or simply $\mathbf{X} = (X, Y)^T$, we speak of a two-dimensional random vector, and its values are geometrically interpreted as points in the plane with coordinates $[x, y]$.

¹Also sometimes referred to as an n -dimensional random variable.

Similarly as for a random variable, the behaviour of the random vector $\mathbf{X} = (X_1, \dots, X_n)^T$ is described by its distribution function, which in this case is n -dimensional.

Definition 9.2 (Distribution function of a random vector)

Let $\mathbf{X} = (X_1, \dots, X_n)^T$ be a random vector defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The joint distribution function $F_{\mathbf{X}}$ of the random vector $\mathbf{X}$ is defined as a real-valued function of n variables on $\mathbb{R}^n$ by the relation

$$\begin{aligned} F_{\mathbf{X}}(x_1, \dots, x_n) &= \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n) = \\ &= \mathbb{P}\left(\bigcap_{i=1}^n \{\omega : X_i(\omega) \leq x_i\}\right), \quad -\infty < x_i < \infty, \quad i = 1, \dots, n. \end{aligned}$$

A two-dimensional random vector $\mathbf{X} = (X, Y)^T$ is thus characterized by the distribution function $F_{\mathbf{X}}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$ of two real variables x, y . For the purposes of the following theorem, let us introduce the notion of an n -dimensional interval $(\mathbf{a}, \mathbf{b}]$, where $\mathbf{a} = (a_1, \dots, a_n)$, $\mathbf{b} = (b_1, \dots, b_n)$, as

$$(\mathbf{a}, \mathbf{b}] = \bigtimes_{i=1}^n (a_i, b_i].$$

In the two-dimensional case, this set is a rectangle; in the three-dimensional case, a rectangular parallelepiped (cuboid). Next, denote by $\Delta_{k,n}$ the set of all $\binom{n}{k}$ n -tuples $(z_1, \dots, z_n)$ such that each z_i is equal to either a_i or b_i , with the former occurring exactly k times (and the latter thus $(n - k)$ times). The union

$$\Delta = \bigcup_{k=0}^n \Delta_{k,n}$$

therefore represents the set of all 2^n vertices of the interval $(\mathbf{a}, \mathbf{b}]$.

Theorem 9.1 (on properties of the distribution function). *Let $F_{\mathbf{X}}$ be the distribution function of the random vector $\mathbf{X} = (X_1, \dots, X_n)^T$. Then $F_{\mathbf{X}}$ has the following properties:*

- a) $F_{\mathbf{X}}(x_1, \dots, x_n)$ is non-decreasing in each of its variables when the remaining variables are held fixed.
- b) $F_{\mathbf{X}}(x_1, \dots, x_n)$ is right-continuous in each variable.
- c) $\lim_{x_i \rightarrow -\infty} F_{\mathbf{X}}(x_1, \dots, x_n) = 0$, $i = 1, \dots, n$, where the values x_j are held fixed for all $j \neq i$, $j = 1, \dots, n$.
- d) $\lim_{\substack{x_1 \rightarrow \infty \\ x_2 \rightarrow \infty \\ \vdots \\ x_n \rightarrow \infty}} F_{\mathbf{X}}(x_1, \dots, x_n) = 1$.

e) For $(\mathbf{a}, \mathbf{b}]$, it holds that

$$\sum_{k=0}^n (-1)^k \sum_{\boldsymbol{\delta} \in \Delta_{k,n}} F_{\mathbf{X}}(\boldsymbol{\delta}) \geq 0.$$

Conversely, every real-valued function of n variables defined on $\mathbb{R}^n$ that satisfies properties a)–e) is the distribution function of some random vector.

Proof. Properties a), b), c), d) are proved analogously to Theorem 5.1; property e) is proved, for example, in [8]. $\square$

The random variables that constitute the components of a random vector may be of either discrete or continuous type. A somewhat simpler case is the following situation, in which these components are of the same type.

Definition 9.3 (Discrete distribution of a random vector)

A random vector $\mathbf{X}$ has a discrete distribution if there exists a sequence of value vectors $\{\mathbf{x}_k\}_{k=1}^{\infty}$, $\mathbf{x}_k \in \mathbb{R}^n$, and a corresponding sequence of positive numbers $\{p_k\}_{k=1}^{\infty}$ such that

$$\sum_{k=1}^{\infty} p_k = 1,$$

where

$$p_k = \mathbb{P}(\mathbf{X} = \mathbf{x}_k) = \mathbb{P}(\{\omega \in \Omega : \mathbf{X}(\omega) = \mathbf{x}_k\}).$$

The distribution function of a random vector $\mathbf{X}$ of discrete type is, for any $\mathbf{x} \in \mathbb{R}^n$, of the form

$$F_{\mathbf{X}}(\mathbf{x}) = \sum_{\{k : \mathbf{x}_k \leq \mathbf{x}\}} p_k,$$

where $\mathbf{x}_k \leq \mathbf{x}$ denotes $\mathbf{x}_k \in (-\infty, \mathbf{x}]$.

Example 9.1. Consider two urns with balls: the first urn contains four balls labeled 1 through 4, while the second urn contains five balls labeled 1 through 5. One ball is drawn from each urn. Let the random variable X denote the number on the ball drawn from the first urn, and let Y denote the number on the ball drawn from the second urn. Determine the joint probability mass function and construct its graph.

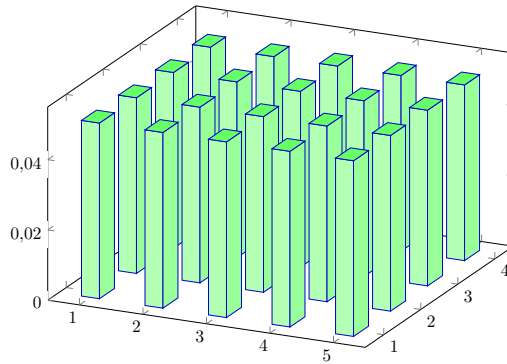


Figure 9.1: Graph of the joint probability mass function $p(x, y)$ in Example 9.1

Table 9.1: Joint and marginal probabilities for Example 9.2

(x, y)	2	3	4	5	6	7	8	9	10	11	12	$\mathbb{P}(X = x)$
0	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	0	0	$\frac{25}{36}$
1	0	0	0	0	0	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	0	$\frac{10}{36}$
2	0	0	0	0	0	0	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$
$\mathbb{P}(Y = y)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	1

Solution. The random variables X and Y form a discrete random vector $\mathbf{X} = (X, Y)^T$, which is completely characterized by its joint probability mass function $p(x, y)$ with domain $(x, y) : x = 1, 2, 3, 4, y = 1, 2, 3, 4, 5$. For each pair (x, y) , the corresponding probability in this case is identical, namely

$$p(x, y) = P(X = x, Y = y) = \frac{1}{20}.$$

The graph of the joint probability mass function consists of twenty points $(x, y, p(x, y))$ in a three-dimensional coordinate system; however, for better visualization, these are often represented by bars, see Figure 9.1. $\triangle$

Example 9.2. Construct a random vector $\mathbf{X} = (X, Y)^T$, where the random variable X denotes the number of sixes obtained when rolling two dice, and the random variable Y denotes the sum of the numbers appearing on the upper faces of both dice. Find the joint probability mass function and construct its graph.

Solution. When rolling two dice, the set of all possible outcomes is

$$\Omega = \{(\square, \square), (\square, \blacksquare), \dots, (\blacksquare, \blacksquare), (\blacksquare, \blacksquare)\}.$$

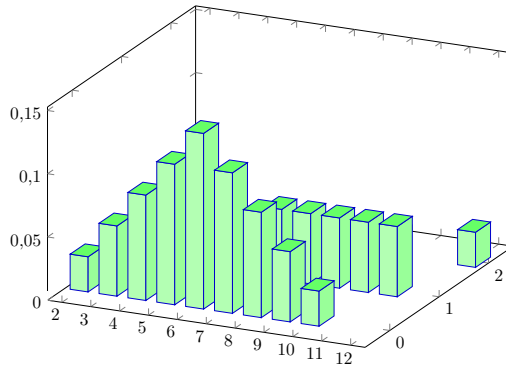


Figure 9.2: Graph of the joint probability mass function $p(x, y)$ in Example 9.2

The cardinality of the sample space Ω is 36. According to the number of sixes obtained and the sum of the face values on both dice, each outcome is associated with a pair of values of the random vector $\mathbf{X}$ from the set

$$\{(x, y) : x = 0, 1, 2, y = 2, 3, \dots, 12\}.$$

For example, the outcome $(\boxed{1}, \boxed{8})$ corresponds to the situation $X = 1, Y = 8$, that is, to the pair $(1, 8)$. For each pair (x, y) from the given set, we compute the corresponding probability $p(x, y) = \mathbb{P}(X = x, Y = y)$ (see Table 9.1).

Since $\sum_{(x,y)} p(x, y) = 1$, the list of probabilities constitutes a distribution, and $p(x, y)$ is its joint probability mass function. The graph of the joint probability mass function consists of 16 points in the coordinate system $(x, y, p(x, y))$; however, these are usually represented by bars, as in Figure 9.2. Graphs similar to Figures 9.2 can also be constructed in Excel using the 3D column chart tool on the Insert tab, based on data such as those presented in Table 9.1. $\triangle$

Definition 9.4 (Absolutely continuous distribution of a random vector)

A random vector $\mathbf{X} = (X_1, \dots, X_n)^T$ has an absolutely continuous distribution, if there exists a non-negative function $f_{\mathbf{X}}$ of n real variables such that

$$F_{\mathbf{X}}(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{\mathbf{X}}(t_1, \dots, t_n) dt_1 \dots dt_n,$$

where the function $f_{\mathbf{X}}$ is called the probability density function of the random vector $\mathbf{X}$, or equivalently the joint density of the random variables $X_1, \dots, X_n$.

For the density $f_{\mathbf{X}}$, it holds that

$$f_{\mathbf{X}}(x_1, \dots, x_n) = \frac{\partial^n F_{\mathbf{X}}(X_1, \dots, X_n)}{\partial x_1 \partial x_2 \dots \partial x_n}$$

at all points $(x_1, \dots, x_n)$ at which the derivative exists, and

$$\int_{\mathbb{R}^n} f_{\mathbf{X}}(\mathbf{x}) \, d\mathbf{x} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, \dots, x_n) \, dx_1 \dots dx_n = 1.$$

Example 9.3. Consider a function f of continuous random variables X and Y , which, for $c \in \mathbb{R}$, is of the form

$$f(x, y) = \begin{cases} c \cdot y \cdot \sin x & \text{for } 0 \leq x \leq \pi, \, 0 \leq y \leq \pi, \\ 0 & \text{otherwise.} \end{cases}$$

Determine c such that $f(x, y)$ is a probability density function of the random vector $\mathbf{X} = (X, Y)^T$. Determine the joint distribution function.

Solution.

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, dx \, dy = \int_0^{\pi} \int_0^{\pi} cy \sin x \, dx \, dy = 2c \int_0^{\pi} y \, dy = c\pi^2 = 1,$$

from which it follows that $c = \frac{1}{\pi^2}$. The joint probability density function is therefore of the form

$$f(x, y) = \begin{cases} \frac{1}{\pi^2} \cdot y \cdot \sin x & \text{for } 0 \leq x \leq \pi, \, 0 \leq y \leq \pi, \\ 0 & \text{otherwise,} \end{cases}$$

and the joint distribution function is of the form

$$F(x, y) = \begin{cases} 0 & \text{pro } x < 0 \text{ or } y < 0, \\ \int_0^y \int_0^x \frac{1}{\pi^2} t \sin s \, ds \, dt = \frac{1}{\pi^2} \int_0^y t(-\cos x + 1) \, dt = \\ = \frac{y^2}{2\pi^2} (1 - \cos x) & \text{for } 0 \leq x \leq \pi, \, 0 \leq y \leq \pi, \\ \frac{1}{2} (1 - \cos x) & \text{for } 0 \leq x \leq \pi, \, y > \pi, \\ \frac{y^2}{\pi^2} & \text{for } x > \pi, \, 0 \leq y \leq \pi, \\ 1 & \text{for } x > \pi, \, y > \pi. \end{cases}$$

△

The distribution function, respectively the density, of a component of the random vector $\mathbf{X}$ is called the *marginal* distribution function, respectively density. As an example, let us present the marginal distribution function of the vector $(X_1, \dots, X_k)$, for $k = 1, \dots, n - 1$:

$$F_{(X_1, \dots, X_k)}(x_1, \dots, x_k) = \lim_{\substack{x_{k+1} \rightarrow \infty \\ x_{k+2} \rightarrow \infty \\ \vdots \\ x_n \rightarrow \infty}} F_{\mathbf{X}}(x_1, \dots, x_k, x_{k+1}, \dots, x_n).$$

The *marginal density* of the vector $(X_1, \dots, X_k)^T$, which is a component of the random vector $\mathbf{X} = (X_1, \dots, X_n)^T$ with an absolutely continuous distribution, is obtained by integrating out the redundant variables:

$$f_{\mathbf{X}}(x_1, \dots, x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, \dots, x_n) dx_{k+1} dx_{k+2} \cdots dx_n.$$

The marginal distribution function of the vector $(X_1, \dots, X_k)$ is then

$$\begin{aligned} F_{(X_1, X_2, \dots, X_k)}(x_1, x_2, \dots, x_k) &= \\ &= \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_k} f_{(X_1, \dots, X_k)}(t_1, \dots, t_k) dt_k dt_{k-1} \cdots dt_1, \end{aligned}$$

Analogous formulas hold for all subvectors of the vector $(X_1, \dots, X_k)^T$. If, moreover, the joint distribution function or the joint density exists, then all marginal distribution functions and marginal densities exist. The converse statement does not hold in general.

Example 9.4. In Example 9.3, determine the explicit form of the marginal density functions of the random variables X and Y .

Solution. The marginal probability density functions of the random variables X and Y are

$$\begin{aligned} f_X(x) &= \int_0^\pi \frac{1}{\pi^2} \cdot y \cdot \sin x \, dy = \frac{\sin x}{2} \quad \text{for } 0 \leq x \leq \pi, \\ f_Y(y) &= \int_0^\pi \frac{1}{\pi^2} \cdot y \cdot \sin x \, dx = \frac{2y}{\pi^2} \quad \text{for } 0 \leq y \leq \pi. \end{aligned}$$

Outside the specified intervals, the density is zero, since the joint density is also zero there, and integration with respect to the other variable does not change this.

△

9.2 Characteristics of a Random Vector

The most important characteristic of a random vector, similarly as in the one-dimensional case, is its expectation.

Definition 9.5

The expectation of a random vector $\mathbf{X} = (X_1, \dots, X_n)^T$ is the vector of expectations of the marginal distributions

$$\mathbb{E}\mathbf{X} = (\mathbb{E}X_1, \dots, \mathbb{E}X_n)^T.$$

To characterize the relationship between the components of a random vector, we introduce the notions of covariance and correlation of two random variables.

Definition 9.6 (Covariance of random variables, correlation coefficient)

Let X and Y be random variables such that $\mathbb{E}X^2 < \infty$, $\mathbb{E}Y^2 < \infty$. The covariance $\text{cov}(X, Y)$ of the random variables X and Y is defined by the relation

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]. \quad (9.1)$$

The correlation coefficient $\text{corr}(X, Y)$ is defined by the relation

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var } X} \sqrt{\text{var } Y}} \quad (9.2)$$

for $\text{var } X > 0$ and $\text{var } Y > 0$.

Covariance of two random variables expresses the relationship between the deviations from the expected values of these variables. If, for example, low (below-average) values of X are typically associated with low (below-average) values of Y , and at the same time high (above-average) values of X are typically associated with high values of Y , then the product $Z = (X - \mathbb{E}X)(Y - \mathbb{E}Y)$ will typically take positive values, and hence the covariance (which is the expectation of the variable Z) and consequently the correlation will be positive (we speak of positively correlated variables). Analogously, negative covariance and correlation can be interpreted.

The covariance of random variables depends on their scale; therefore, the interpretation of numerical values of covariance is quite difficult. The Cauchy–Schwarz–Bunyakovsky inequality² provides an upper bound for the

²Hermann Amandus Schwarz (1843–1921), German mathematician. Viktor Yakovlevich Bunyakovsky (1804–1889), Russian mathematician and physicist.

absolute value of covariance by the product of standard deviations. It follows that the correlation coefficient can take only values between -1 and 1 . The correlation coefficient equals 1 if the random variables are linearly dependent with a positive proportionality coefficient. The correlation coefficient equals -1 if the random variables are linearly dependent with a negative proportionality coefficient. In these cases, we speak of perfect positive or negative linear dependence.

Values of correlation close to -1 and 1 indicate a strong relationship between random variables. Conversely, if covariance and correlation are zero, the random variables are not linearly related at all. In such cases, we speak of *uncorrelated* random variables.

For the actual computation of covariance, instead of the defining formula (9.1), one may use the so-called computational formula

$$\text{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y. \quad (9.3)$$

Its derivation is left to the reader as an exercise.

Question 9.1. Prove relation (9.3).

Example 9.5. Compute the correlation coefficient of the random variables X and Y from Example 9.2 on page 192.

Solution. By summing the rows of Table 9.1 on page 192, we obtain the marginal distribution of the variable X , and the column sums of the table yield the marginal distribution of the variable Y . These are subsequently used to compute the expectations:

$$\begin{aligned} \mathbb{E}X &= 0 \cdot \frac{25}{36} + 1 \cdot \frac{10}{36} + 2 \cdot \frac{1}{36} = \frac{12}{36} = \frac{1}{3}, \\ \mathbb{E}Y &= 2 \cdot \frac{1}{36} + \cdots + 12 \cdot \frac{1}{36} = \frac{252}{36} = 7. \end{aligned}$$

The joint probabilities from the same table are used to compute

$$\mathbb{E}(XY) = 0 \cdot 2 \cdot \mathbb{P}(X = 0, Y = 2) + \cdots + 2 \cdot 12 \cdot \mathbb{P}(X = 2, Y = 12) = \frac{114}{36}$$

and then for the computation of the covariance using the formula (9.3):

$$\text{cov}(X, Y) = \frac{114}{36} - \frac{1}{3} \cdot 7 = \frac{30}{36} = \frac{5}{6} \approx 0.83.$$

Next, we compute the variances of the random variables X and Y :

$$\begin{aligned}\text{var } X &= \mathbb{E}X^2 - (\mathbb{E}X)^2 = 0 \cdot \frac{25}{36} + 1^2 \cdot \frac{10}{36} + 2^2 \cdot \frac{1}{36} - \left(\frac{1}{3}\right)^2 = \frac{10}{36}, \\ \text{var } Y &= \mathbb{E}Y^2 - (\mathbb{E}Y)^2 = 2^2 \cdot \frac{1}{36} + \cdots + 12^2 \cdot \frac{1}{36} - 7^2 = \frac{210}{36}.\end{aligned}$$

Finally, we compute the correlation using the formula (9.2):

$$\text{corr}(X, Y) = \frac{\frac{5}{6}}{\sqrt{\frac{10}{36}} \cdot \sqrt{\frac{210}{36}}} \approx 0.65.$$

A value of the correlation coefficient 0.65 indicates a moderately strong positive relationship between the two random variables. $\triangle$

Definition 9.7 (Variance–covariance matrix and correlation matrix of a random vector)

Let $\mathbf{X} = (X_1, \dots, X_n)^T$ be a random vector whose components have finite second moments. The variance–covariance matrix $\text{var } \mathbf{X}$ of this random vector is defined as an $n \times n$ matrix with entries

$$\text{cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)], \quad 1 \leq i, j \leq n.$$

The correlation matrix $\text{corr } \mathbf{X}$ is a matrix with entries

$$\text{corr}(X_i, X_j) = \frac{\text{cov}(X_i, X_j)}{\sqrt{\text{var } X_i} \sqrt{\text{var } X_j}}, \quad 1 \leq i, j \leq n.$$

It is evident that for $i = 1, \dots, n$ it holds that

$$\text{cov}(X_i, X_i) = \text{var } X_i \quad \text{a} \quad \text{corr}(X_i, X_i) = 1, \quad 1 \leq i \leq n.$$

For example, the variance–covariance matrix and the correlation matrix of the two-dimensional random vector $\mathbf{X} = (X, Y)^T$ have the form

$$\text{var } \mathbf{X} = \begin{pmatrix} \text{var } X & \text{cov}(X, Y) \\ \text{cov}(X, Y) & \text{var } Y \end{pmatrix}, \quad \text{corr } \mathbf{X} = \begin{pmatrix} 1 & \text{corr}(X, Y) \\ \text{corr}(X, Y) & 1 \end{pmatrix}.$$

In Theorem 6.2, we stated property d) concerning the expectation of the sum of random variables. An analogous rule also holds for the variance of the sum of random variables, as formulated in the following theorem, which we present without proof.

Theorem 9.2. *If the variance of the sum of random variables X_i exists, then it is equal to*

$$\text{var} \left(\sum_{i=1}^n X_i \right) = \sum_{i=1}^n \text{var} X_i + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{cov}(X_i, X_j).$$

If the random variables X_i are, moreover, uncorrelated, then it holds that

$$\text{var} \left(\sum_{i=1}^n X_i \right) = \sum_{i=1}^n \text{var} X_i. \quad (9.4)$$

Example 9.6. Diversification of a portfolio consists in selecting assets whose returns are as weakly correlated as possible (or even negatively correlated). In this way, the overall variance (risk) of the portfolio is reduced, since negatively correlated components offset each other. With negative correlation between the components, an investor can significantly reduce risk without necessarily reducing the expected return.

Consider two random vectors X and Y representing two components of a given (e.g. financial) portfolio. Suppose that both components have the same expectation μ and the same variance σ^2 . Compute the variance of the sum of these portfolio components for different values of the correlation coefficient:

- a) $\text{corr}(X, Y) = 0.5$,
- b) $\text{corr}(X, Y) = 0$,
- c) $\text{corr}(X, Y) = -0.5$.

Solution.

$$\text{var}(X + Y) = \text{var} X + \text{var} Y + 2 \cdot \text{cov}(X, Y),$$

where $\text{var} X = \text{var} Y = \sigma^2$.

$$\text{cov}(X, Y) = \text{corr}(X, Y) \cdot \sigma_X \cdot \sigma_Y = \text{corr}(X, Y) \cdot \sigma^2.$$

Upon substitution, we obtain

$$\text{var}(X + Y) = \sigma^2 + \sigma^2 + 2 \cdot \text{corr}(X, Y) \cdot \sigma^2 = 2\sigma^2 \cdot (1 + \text{corr}(X, Y)).$$

For individual values of the correlation coefficient, we obtain:

- a) $\text{var}(X + Y) = 3\sigma^2$,
- b) $\text{var}(X + Y) = 2\sigma^2$,
- c) $\text{var}(X + Y) = \sigma^2$.

For positive correlation $\text{corr}(X, Y) = 0.5$, the variance of the sum of the components is $3\sigma^2$, which means that the components reinforce each other and diversification is not effective. A high correlation between the components leads to greater overall variability of the portfolio and thus to increased risk. For zero correlation $\text{corr}(X, Y) = 0$, the variance of the sum of the components is $2\sigma^2$, which is lower than in the case of positive correlation and results in a weaker form of diversification, but not in maximal diversification. Only for $\text{corr}(X, Y) = -0.5$ is the variance of the sum of the components equal to σ^2 , which is the lowest value among these three cases. Negative correlation between the components means that when one component increases, the other decreases, which results in strong diversification of the portfolio and leads to a substantial reduction of its overall risk. $\triangle$

9.3 Some Multivariate Distributions

9.3.1 Multinomial distribution

The multinomial distribution is a generalization of the binomial distribution and is arguably the most important discrete multivariate distribution. Consider an urn containing balls of k different colors. Let the probability of drawing a ball of the i -th color be p_i , $i = 1, 2, \dots, k$, where

$$0 < p_i < 1, \quad p_1 + p_2 + \dots + p_k = 1.$$

from this urn we draw n times independently one ball at a time (after each draw, the ball is returned to the urn). Let X_i denote the number of balls of the i -th color drawn. From combinatorial arguments, it follows that the random vector $\mathbf{X} = (X_1, X_2, \dots, X_k)^T$ has the joint probability mass function

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = \frac{n!}{x_1! x_2! \dots x_k!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$$

for $x_i = 0, 1, \dots, n$, $i = 1, 2, \dots, k$ and $x_1 + x_2 + \dots + x_k = n$.

The parameters of the multinomial distribution are $(n, p_1, p_2, \dots, p_k)$. The marginal distribution of X_i is binomial with parameters (n, p_i) for $i = 1, 2, \dots, k$. The expectation is $\mathbb{E}X_i = np_i$ for $i = 1, 2, \dots, k$. The entries



of the variance–covariance matrix are

$$\begin{aligned}\text{var } X_i &= np_i(1 - p_i) \quad \text{for } i = 1, 2, \dots, k, \\ \text{cov}(X_i, X_j) &= -np_i p_j \quad \text{for } i \neq j.\end{aligned}$$

Example 9.7. From long-term statistics in the company, it is known that a project is accepted with probability 0.35, rejected with probability 0.5, and postponed for further evaluation with probability 0.15. Determine the probability that out of 15 projects, the company accepts five, rejects seven, and postpones three for further evaluation.

Solution. The random variable X_1 denotes the number of accepted projects, the random variable X_2 denotes the number of rejected projects, and the random variable X_3 denotes the number of projects postponed for further evaluation, with $p_1 = 0.35$, $p_2 = 0.5$, $p_3 = 0.15$.

$$\mathbb{P}(X_1 = 5, X_2 = 7, X_3 = 3) = \frac{15!}{5! \cdot 7! \cdot 3!} \cdot 0.25^4 \cdot 0.5^7 \cdot 0.15^3 \approx 0.037.$$

The first factor can be computed as $\text{MULTINOMIAL}(5, 7, 3)$.

△

9.3.2 General bivariate normal distribution

A normal distribution can also be multivariate; for simplicity, we present only the two-dimensional case, i.e., the bivariate normal distribution.

$$N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

is the distribution of the random vector $(X, Y)^T$ with density

$$\begin{aligned}f(x, y) &= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}} \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left\{\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{x-\mu_1}{\sigma_1}\frac{y-\mu_2}{\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right\}\right\}\end{aligned}$$

for $(x, y) \in \mathbb{R}^2$, where $\mu_i \in \mathbb{R}$, $\sigma_i^2 > 0$, $i = 1, 2$ and $\rho \in (-1, 1)$ are parameters. The expression in the exponent can also be written as

$$-\frac{1}{2}\begin{pmatrix} x - \mu_1 \\ y - \mu_2 \end{pmatrix}^T \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} x - \mu_1 \\ y - \mu_2 \end{pmatrix}.$$

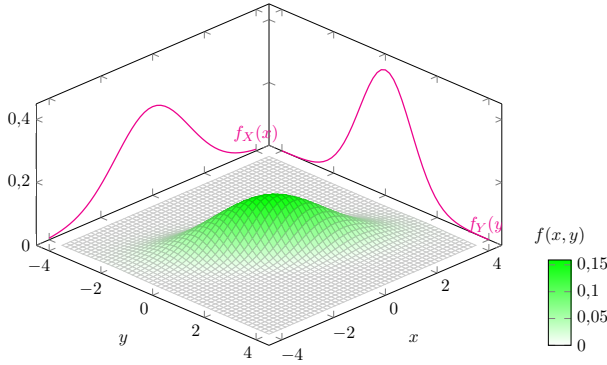


Figure 9.3: Graph of the probability density function of the bivariate normal distribution $N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\right)$

The marginal distributions of X and Y are $N(\mu_1; \sigma_1^2)$ and $N(\mu_2; \sigma_2^2)$, and

$$\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

are the entries of the variance–covariance matrix of the vector $(X, Y)^T$; ρ is the correlation coefficient $\text{corr}(X, Y)$. If we standardize X and Y , then the pair of random variables $W = \frac{(X - \mu_1)}{\sigma_1}$ and $Z = \frac{(Y - \mu_2)}{\sigma_2}$ has a standardized bivariate normal distribution. $N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$.

9.4 Independent Random Variables

In Section 3.2, we defined the independence of random events. We now extend this notion to the independence of random variables $X_1, \dots, X_n$.

Definition 9.8 (Mutually independent random variables)

Random variables $X_1, X_2, \dots, X_n$ are mutually independent if

$$\mathbb{P}\left(\bigcap_{j=1}^r \{\omega : X_{i_j} < x_{i_j}\}\right) = \prod_{j=1}^r \mathbb{P}(\{\omega : X_{i_j} < x_{i_j}\}) \quad (9.5)$$

for all r -tuples $i_1, i_2, \dots, i_r \subset 1, 2, \dots, n$, $1 \leq r \leq n$, and for all $(x_1, \dots, x_n) \in \mathbb{R}^n$.

Similarly as for random events, we can also define pairwise independence of random variables $X_1, X_2, \dots, X_n$. This weaker notion of independence

would be obtained from Definition 9.8 if we restricted the value of r only to $r = 2$.

Verifying independence of random variables according to Definition 9.8 is rather demanding; therefore, we provide criteria that are used in practice to verify independence of random variables.

Theorem 9.3. *Let the random vector $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$ have the joint distribution function $F_{\mathbf{X}}(x_1, x_2, \dots, x_n)$. Let $F_{X_i}(x_i)$ be the marginal distribution function of the random variable X_i , $i = 1, 2, \dots, n$. Then the random variables $X_1, X_2, \dots, X_n$ are independent if and only if the following holds*

$$F_{\mathbf{X}}(x_1, x_2, \dots, x_n) = F_{X_1}(x_1) \cdot F_{X_2}(x_2) \cdots F_{X_n}(x_n), \quad (9.6)$$

for all $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$.

Proof. If $X_1, X_2, \dots, X_n$ are independent, then according to Definition 9.8, relation (9.5) holds for all subsets $i_1, i_2, \dots, i_r \subset 1, 2, \dots, n$, $1 \leq r \leq n$, and for any $x \in \mathbb{R}^n$. Therefore, by the definition of the distribution function, relation (9.6) also holds.

Now assume that (9.6) holds. Consider any subset $i_1, i_2, \dots, i_r$ of the set $1, 2, \dots, n$. Then the assertion of the theorem follows from the definition of the marginal distribution function. $\square$

If we consider a random vector $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$ of absolutely continuous type, then the independence of its components is verified according to the following criterion.

Theorem 9.4. *Let $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$ be a random vector of absolutely continuous type. The random variables $X_1, X_2, \dots, X_n$ are mutually independent if and only if it holds that*

$$f_{\mathbf{X}}(x_1, x_2, \dots, x_n) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdots f_{X_n}(x_n),$$

for all $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$.

Proof. It follows from Theorem 9.3. $\square$

For verifying the independence of the components of a random vector $\mathbf{X}$ of discrete type, the following criterion holds:

Theorem 9.5. Let $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$ be a random vector of discrete type. The random variables $X_1, X_2, \dots, X_n$ are mutually independent if and only if it holds that

$$\mathbb{P}(X_1 = x_1^{(i)}, \dots, X_n = x_n^{(i)}) = \prod_{j=1}^n \mathbb{P}(X_j = x_j^{(i)}),$$

where $\mathbf{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)})$, $i = 1, 2, \dots$ are all possible outcomes of the random vector $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$.

Proof. It follows from Theorem 9.3. □

Example 9.8. The lifetime (in years) of a pair of components is described by the density

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{2}e^{-x-\frac{y}{2}} & x > 0, y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

- Determine the explicit form of the distribution function $F_{\mathbf{X}}$ of the random vector $\mathbf{X} = (X, Y)^T$.
- What is the probability that “the second component outlives the first” (event A)?
- What is the probability that the second component will live at least twice as long as the first?
- Compute the marginal densities.
- Determine whether the components X and Y are independent.
- Determine $\mathbb{E}\mathbf{X}$ and the variance–covariance matrix $\text{var } \mathbf{X}$.

Solution.

- a) $\int_0^y \int_0^x \frac{1}{2}e^{-s-\frac{t}{2}} ds dt = e^{-x-\frac{y}{2}} + 2e^{-x} - e^{-\frac{y}{2}} + 1$, hence

$$F_X(x, y) = \begin{cases} e^{-x-\frac{y}{2}} + 2e^{-x} - e^{-\frac{y}{2}} + 1 & x > 0, y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

- b) The required probability is given by the integral

$$\begin{aligned} \mathbb{P}(A) &= \int_0^\infty \int_x^\infty \frac{1}{2}e^{-x-\frac{y}{2}} dy dx = \\ &= \int_0^\infty \frac{1}{2}e^{-x} \left(\int_x^\infty e^{-\frac{y}{2}} dy \right) dx = \int_0^\infty \frac{1}{2}e^{-x} \left[-2 \cdot e^{-\frac{y}{2}} \right]_x^\infty dx = \\ &= \int_0^\infty \frac{1}{2}e^{-x} \cdot 2e^{-\frac{x}{2}} dx = \int_0^\infty e^{-\frac{3}{2}x} dx = \left[-\frac{2}{3}e^{-\frac{3}{2}x} \right]_0^\infty = \frac{2}{3}. \end{aligned}$$

- c) The probability is given by the integral

$$\int_0^\infty \int_{2x}^\infty \frac{1}{2} e^{-x-\frac{y}{2}} dy dx = \frac{1}{2},$$

which we compute in the same way as in the previous case.

- d) The marginal densities are:

$$f_X(x) = \int_0^\infty \frac{1}{2} e^{-x-\frac{y}{2}} dy = \frac{1}{2} e^{-x} \cdot [-2e^{-\frac{y}{2}}]_0^\infty = e^{-x} \text{ for } x > 0,$$

$$f_Y(y) = \int_0^\infty \frac{1}{2} e^{-x-\frac{y}{2}} dx = \frac{1}{2} e^{-\frac{y}{2}} \cdot [-e^{-x}]_0^\infty = \frac{1}{2} e^{-\frac{y}{2}} \text{ for } y > 0,$$

zero in all other cases.

- e) The components are independent if and only if $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$ for all x, y . For $x > 0$ and $y > 0$, this equality holds according to d). In the remaining cases, both the joint density and at least one of the marginal densities are zero, and hence the equality also holds.
- f) To determine $\mathbb{E}(X, Y)$ and the variance–covariance matrix, we first compute:

$$\mathbb{E}X = \int_0^\infty x f_X(x) dx = \int_0^\infty x \cdot e^{-x} dx = 1$$

$$\mathbb{E}Y = \int_0^\infty y f_Y(y) dy = \int_0^\infty y \cdot \frac{1}{2} e^{-\frac{y}{2}} dy = 2$$

$$\text{var}X = \mathbb{E}X^2 - (\mathbb{E}X)^2, \text{var}Y = \mathbb{E}Y^2 - (\mathbb{E}Y)^2$$

$$\mathbb{E}X^2 = \int_0^\infty x^2 \cdot e^{-x} dx = 2$$

$$\mathbb{E}Y^2 = \int_0^\infty y^2 \cdot \frac{1}{2} e^{-\frac{y}{2}} dy = 8$$

$$\text{var}X = 2 - 1^2 = 1$$

$$\text{var}Y = 8 - 2^2 = 4$$

$$\begin{aligned} \mathbb{E}XY &= \int_0^\infty \int_0^\infty x \cdot y \frac{1}{2} e^{-x-\frac{y}{2}} dx dy = \int_0^\infty y \frac{1}{2} e^{-\frac{y}{2}} \cdot \int_0^\infty x \frac{1}{2} e^{-x} dx dy = \\ &= \int_0^\infty \frac{1}{2} e^{-\frac{y}{2}} dy \cdot \int_0^\infty x \frac{1}{2} e^{-x} dx = (\mathbb{E}Y) \cdot (\mathbb{E}X), \end{aligned}$$

from which it follows that $\text{cov}(X, Y) = \mathbb{E}XY - (\mathbb{E}X) \cdot (\mathbb{E}Y) = 0$. The variance–covariance matrix is therefore equal to

$$\text{var}(X, Y) = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}.$$

△

9.5 Exercises

9.1 Let a function be given

$$p(x, y) = \begin{cases} \frac{1}{11}(y - x) & x = 1, 2, 3, \ y = x^2 + 1, \\ 0 & \text{elsewhere.} \end{cases}$$

- Prove that the given function $p(x, y)$ is a distribution of some random vector $\mathbf{X}$.
- Plot (represent graphically) the function $p(x, y)$.
- Compute the expectation $\mathbb{E}\mathbf{X}$.
- Determine the variance–covariance matrix.

9.2 A function of two variables x and y is given by the relation

$$f(x, y) = \begin{cases} c \cdot x \cdot y \cdot e^{-x^2-y^2} & \text{for } (x, y) \in (0, \infty) \times (0, \infty), \\ 0 & \text{elsewhere.} \end{cases}$$

- Determine the constant c such that the function $f(x, y)$ is the density of some random vector $(X, Y)^T$.
- Determine the distribution function and the marginal distribution functions.
- Determine the marginal densities.
- Determine the expectations.
- Compute the variance–covariance matrix.
- Verify whether the random variables are independent.

9.3 Let the joint density of the random vector $(R, \Phi)^T$ be

$$g(r, \varphi) = \begin{cases} \frac{1}{2\pi\sigma^2} \cdot e^{-\frac{r^2}{2\sigma^2}} \cdot r & r \in (0, \infty), \ \varphi \in (0, 2\pi), \\ 0 & \text{elsewhere,} \end{cases}$$

where r and φ are independent polar coordinates of a point in the Gaussian plane; the random variable R has a uniform distribution and the variable Φ has a normal distribution. (The function $g(r, \varphi)$ is the density of the so-called Rayleigh³ continuous distribution, which is used, for example, in meteorology.)

- Show that $g(r, \varphi)$ is a density of the vector $(R, \Phi)^T$.

³John William Strutt, 3rd Baron Rayleigh (1842–1919), English physicist.

- b) Determine the distribution function $G(r, \varphi)$.
- c) Find the marginal densities of the random variables R and Φ .
- d) Determine whether the random variables R and Φ are independent.

9.4 Let a function be given

$$f(x, y) = \begin{cases} c \cdot (x^2 + y) \cdot e^x & \text{for } (x, y) \in [0, 1] \times [0, 1], \\ 0 & \text{elsewhere.} \end{cases}$$

- a) Find the constant c such that $f(x, y)$ is the density of some random vector $(X, Y)^T$.
- b) Compute the distribution function of the random vector $\mathbf{X} = (X, Y)^T$.
- c) Determine the marginal densities $f_X(x)$, $f_Y(y)$.
- d) Determine the expectation $\mathbb{E}\mathbf{X}$.
- e) Determine the variance–covariance matrix and decide whether the random variables X and Y are independent.

9.5 The tested device consists of five components. The n -th component fails with probability

$$p_n = 0.2 + 0.1(n - 1).$$

Failures of individual components are independent. Determine the variance of the number of failed components.

9.6 We roll a die ten times in succession. Let the random variable X denote the number of sixes obtained in the 10 throws, and let the random variable Y denote the number of ones.

- a) Find the probability distribution (the joint probability mass function) of the random vector $(X, Y)^T$ and construct its graph.
- b) Compute the expectation and the variance–covariance matrix.

9.7 Let the random vector $(X, Y)^T$ have the density

$$f(x, y) = \begin{cases} \frac{1}{(b_1 - a_1)(b_2 - a_2)} & a_1 \leq x \leq b_1, \ a_2 \leq y \leq b_2, \\ 0 & \text{elsewhere.} \end{cases}$$

(The random vector $(X, Y)^T$ with this density has the so-called bivariate uniform distribution.)

- a) Verify that $f(x, y)$ is a probability density function.
- b) Find the distribution function of the random vector $(X, Y)^T$.
- c) Compute the expectation and the variance–covariance matrix.

9.8 Determine the constant c such that the function

$$f(x, y, z) = \begin{cases} c \cdot z & 0 \leq x \leq 2, \\ & 0 \leq y \leq \sqrt{4 - x^2}, \\ & 0 \leq z \leq \sqrt{4 - x^2 - y^2}, \\ 0 & \text{elsewhere.} \end{cases}$$

so that it is the density of a random vector $(X, Y, Z)^T$.

9.9 Let the random vector $(X, Y, Z)^T$ have the joint density

$$f(x, y, z) = \begin{cases} \frac{1}{8}(1 + x \cdot y \cdot z) & -1 \leq x \leq 1, -1 \leq y \leq 1, -1 \leq z \leq 1, \\ 0 & \text{elsewhere.} \end{cases}$$

- Verify that $f(x, y, z)$ is a probability density function.
- Determine the marginal densities of the random variables X, Y, Z .
- Determine the marginal densities of the random vectors $(X, Y)^T$, $(X, Z)^T$ and $(Y, Z)^T$.
- Determine whether the random variables X, Y, Z are pairwise independent and whether they are mutually independent.

9.10 n employees of a certain company have lunch in one of three restaurants, each employee choosing a restaurant at random. The capacities of the restaurants are n_1, n_2 , and n_3 , where $n_i \geq n$, $i = 1, 2, 3$. Derive the distribution of $(X_1, X_2, X_3)^T$, where X_i is the number of employees having lunch in the i -th restaurant. Derive the expected number of unoccupied seats in each restaurant.

9.11 From an urn containing two white and two black balls, two balls are drawn successively with replacement. Define the random variables X_1, X_2 as follows:

$$X_1 = \begin{cases} 1 & \text{if the first drawn ball is white,} \\ 0 & \text{otherwise,} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if the second drawn ball is white,} \\ 0 & \text{otherwise.} \end{cases}$$

Determine the distribution function of the vector $(X_1, X_2)^T$ and determine whether the random variables X_1 and X_2 are independent.

9.12 Let the random variables U, V have a discrete distribution given by Table 9.2. Determine the marginal distributions of each of the two random variables, their expectations, variances, and covariance.

Table 9.2: Distribution of the random variables for Exercise 9.12

U / V	1	2	3
1	0.1	0.2	0.3
2	0.2	0.1	0.1

Solutions to the Exercises

9.1

- a) $p(x, y)$ is a discrete distribution of a random vector $\mathbf{X}$, since $\mathbb{P}[\mathbf{X} = (1; 2)^T] = p_1 = \frac{1}{11}$, $\mathbb{P}[\mathbf{X} = (2; 5)^T] = p_2 = \frac{3}{11}$ and $\mathbb{P}[\mathbf{X} = (3; 10)^T] = p_3 = \frac{7}{11}$, hence $p_1 + p_2 + p_3 = 1$.

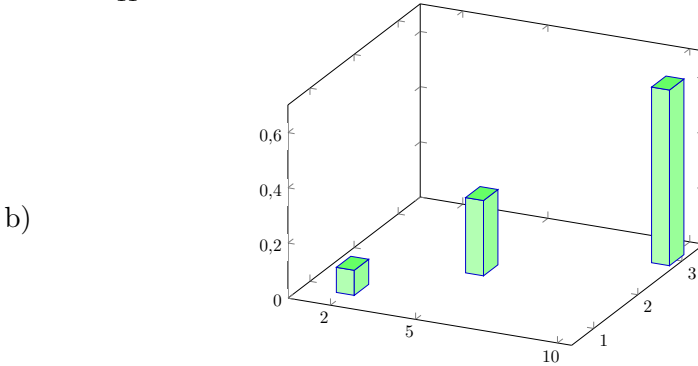


Figure 9.4: Graph of the probability distribution function $p(x, y)$ in Exercise 9.1

- c) $\mathbb{E}X = \frac{28}{11}$, $\mathbb{E}Y = \frac{87}{11}$, i.e. $\mathbb{E}\mathbf{X} = \left(\frac{28}{11}, \frac{87}{11}\right)^T$.
d) $\text{var } X = \frac{52}{121}$, $\text{var } Y = \frac{1000}{121}$, $\text{cov}(X, Y) = \frac{226}{121}$,
i.e. $\text{var } \mathbf{X} = \begin{pmatrix} \frac{52}{121} & \frac{226}{121} \\ \frac{226}{121} & \frac{1000}{121} \end{pmatrix}$.

9.2

- a) $\int_0^\infty \int_0^\infty c \cdot x \cdot y \cdot e^{-x^2} \cdot e^{-y^2} dx dy = \frac{c}{4} = 1$, whence $c = 4$, using the formula $\int x \cdot e^{kx^2} dx = \frac{1}{2k} \cdot e^{kx^2}$, here with $k = -1$.
b) $F(X, Y) = (1 - e^{-x^2}) \cdot (1 - e^{-y^2})$, $F_X(x) = 1 - e^{-x^2}$, $x \geq 0$,
 $F_Y(y) = 1 - e^{-y^2}$, $y \geq 0$.

- c) $f_X(x) = 2 \cdot x \cdot e^{-x^2}$, $f_Y(y) = 2 \cdot y \cdot e^{-y^2}$.
 d) $\mathbb{E}X = 1$, $\mathbb{E}Y = 1$.
 e) $\text{var } \mathbf{X} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.
 f) $f_X(x) \cdot f_Y(y) = 4xye^{-x^2}e^{-y^2} = f(x, y)$ pro $x > 0, y > 0$ (otherwise 0). The random variables X and Y are independent.

9.3

- a) $\int_0^{2\pi} \int_0^\infty \frac{1}{2\pi\rho^2} \cdot e^{-\frac{r^2}{2\rho^2}} \cdot r \, dr \, d\varphi = 1$,
 b) distribution function $G(r, \varphi) = \frac{\varphi}{2\pi} \cdot \left(1 - e^{-\frac{r^2}{2\sigma^2}}\right)$,
 c) marginal densities $g_R(r) = \frac{r}{\rho^2} \cdot e^{-\frac{r^2}{2\rho^2}}$, $g_\Phi(\varphi) = \frac{1}{2\pi}$,
 d) because $g(r, \varphi) = g_R(r) \cdot g_\Phi(\varphi)$, the random variables R and Φ are independent.

9.4

- a) $c = \frac{2}{3e-5}$, the density is of the form $f(x, y) = \frac{2}{3e-5} \cdot (x^2 + y) \cdot e^x$ for $(x, y) \in [0, 1] \times [0, 1]$,
 b) $F(x, y) = \frac{2e^x y}{3e-5} \left(x^2 + \frac{y}{2} - 2x + 2\right)$ for $(x, y) \in [0, 1] \times [0, 1]$,
 c) $f_X(x) = c \cdot e^x \cdot \left(x^2 + \frac{1}{2}\right)$ for $x \in [0, 1]$, $f_Y(y) = c \cdot (2e - ye + y - 3)$ for $y \in [0, 1]$.
 d) Expectation:

$$\mathbb{E}X = \frac{-4e + 13}{3e - 5}, \quad \mathbb{E}Y = \frac{4e - 7}{3 \cdot (3e - 5)},$$

$$\mathbb{E}\mathbf{X} = \left(\frac{-4e + 13}{3e - 5}, \frac{4e - 7}{3 \cdot (3e - 5)} \right)^T,$$

$$\mathbb{E}X^2 = \frac{-30e}{3e - 5}, \quad \text{var } X = \frac{-106e^2 + 254e - 169}{(3e - 5)^2},$$

$$\mathbb{E}Y^2 = \frac{5e - 9}{6 \cdot (3e - 5)}, \quad \text{var } Y = \frac{13e^2 - 24e + 37}{18 \cdot (3e - 5)},$$

$$\text{cov}(X, Y) = \frac{-56x^2 + 345e - 284}{3 \cdot (3e - 5)^2},$$

$$\text{var } \mathbf{X} = \begin{pmatrix} \frac{-106e^2 + 254e - 169}{(3e - 5)^2} & \frac{-56x^2 + 345e - 284}{3 \cdot (3e - 5)^2} \\ \frac{-56x^2 + 345e - 284}{3 \cdot (3e - 5)^2} & \frac{13e^2 - 24e + 37}{18 \cdot (3e - 5)} \end{pmatrix}.$$

- e) The random variables X and Y are not independent; they have non-zero covariance.

9.5 $\text{var } X = 1.1$.

9.6

- a) X takes values in $0, 1, \dots, 10$, Y takes values in $0, 1, \dots, 10$, while $x + y \leq 10$.

$$p(x, y) = \binom{10}{x} \binom{10-x}{y} \left(\frac{1}{6}\right)^{x+y} \left(\frac{5}{6}\right)^{10-x-y}$$

pro $0 \leq x \leq 10$, $0 \leq y \leq 10 - x$, is zero in all other cases.

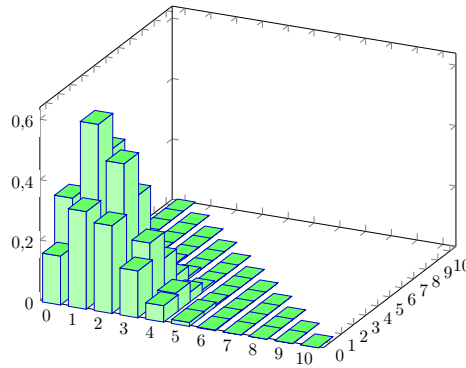


Figure 9.5: Graph of the joint probability mass function for Exercise 9.6

- b) $\mathbb{E}X = \frac{5}{3}$, $\mathbb{E}Y = \frac{5}{3}$, $\text{var } X = \frac{25}{18}$, $\text{var } Y = \frac{25}{18}$, $\text{cov}(X, Y) = 0$,
 $\text{var}(X, Y)^T = \begin{pmatrix} \frac{25}{18} & 0 \\ 0 & \frac{25}{18} \end{pmatrix}$.

9.7

- a) $\int_{a_2}^{b_2} \int_{a_1}^{b_1} \frac{1}{(b_1 - a_1)(b_2 - a_2)} dx dy = 1$.
- b)
$$F(x, y) = \begin{cases} 1, & x > b_1, y > b_2, \\ \frac{(x-a_1)(y-a_2)}{(b_1-a_1)(b_2-a_2)}, & a_1 \leq x \leq b_1, a_2 \leq y \leq b_2, \\ 0, & \text{elsewhere.} \end{cases}$$
- c) $\mathbb{E}X = \frac{a_1+b_1}{2}$, $\mathbb{E}Y = \frac{a_2+b_2}{2}$, $\text{var } X = \frac{(b_1-a_1)^2}{12}$, $\text{var } Y = \frac{(b_2-a_2)^2}{12}$,
 $\text{cov}(X, Y) = 0$, $\text{var}(X, Y)^T = \begin{pmatrix} \frac{(b_1-a_1)^2}{12} & 0 \\ 0 & \frac{(b_2-a_2)^2}{12} \end{pmatrix}$.

9.8 $c = \frac{4}{\pi}$

9.9 a) $\int_{-1}^1 \int_{-1}^1 \int_{-1}^1 \frac{1}{8}(1+x \cdot y \cdot z) \, dx \, dy \, dz = 1$. b) $f_1(x) = f_2(y) = f_3(z) = \frac{1}{2}$.
c) $f_{12}(x, y) = f_{13}(x, z) = f_{23}(y, z) = \frac{1}{4}$. d) X, Y, Z are pairwise independent but are not mutually independent.

9.10 The random vector $(X_1, X_2, X_3)^T$ has a multinomial distribution with parameters n and (p_1, p_2, p_3) :

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{n!}{x_1!x_2!x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3},$$

where x_1, x_2, x_3 are non-negative integers satisfying $x_1 + x_2 + x_3 = n$. The expected number of unoccupied seats in the i -th restaurant is $\mathbb{E}Y_i = n_i - np_i$.

9.11 $(X_1, X_2) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. $\mathbf{X} = (X_1, X_2)^T$ with domain of values $\{(j, k) : j = 0, 1, k = 0, 1\}$, $\mathbb{P}[\mathbf{X} = (1, 1)^T] = \mathbb{P}[\mathbf{X} = (1, 0)^T] = \mathbb{P}[\mathbf{X} = (0, 1)^T] = \mathbb{P}[\mathbf{X} = (0, 0)^T] = \frac{1}{4}$. Distribution function of a random vector

$$F_{X_1, X_2}(x_1, x_2) = \begin{cases} 0 & \text{for } (x_1, x_2) \in (-\infty, 0) \times (-\infty, 0) \\ 0 & \text{for } (x_1, x_2) \in [0, 0] \times (-\infty, 0) \\ 0 & \text{for } (x_1, x_2) \in (-\infty, 0) \times [0, 0] \\ \frac{1}{4} & \text{for } (x_1, x_2) \in [0, 0] \times [0, 0] \\ \frac{1}{2} & \text{for } (x_1, x_2) \in [0, 0] \times [0, 1] \\ \frac{1}{2} & \text{for } (x_1, x_2) \in [0, 1] \times [0, 0] \\ 1 & \text{for } (x_1, x_2) \in [0, 1] \times [0, 1] \end{cases}.$$

The random variables X_1 and X_2 are independent.

9.12 $\mathbb{P}(U = 1) = 0.6$, $\mathbb{P}(U = 2) = 0.4$, $\mathbb{P}(V = 1) = 0.3$, $\mathbb{P}(V = 2) = 0.3$, $\mathbb{P}(V = 3) = 0.4$, $\mathbb{E}U = 1.4$, $\mathbb{E}V = 2.1$, $\text{var } U = 0.24$, $\text{var } V = 0.69$, $\text{cov}(U, V) = -0.14$.

Chapter 10

Function of random variables

10.1 Functions of Random Variables

In solving certain probability problems, we encounter situations where the distribution of a random variable X is known and we seek the distribution of a random variable Y that is a function of X , i.e., $Y = \phi(X)$. The following theorems provide the form of the distribution functions and densities of such variables.

Theorem 10.1. *Let X be a random variable with distribution function F , and let $\phi: \mathbb{R} \rightarrow \mathbb{R}$. Denote $Y = \phi(X)$ and let G be its distribution function. Then for $y \in \mathbb{R}$*

$$G(y) = \int_{\{x: \phi(x) \leq y\}} dF(x). \quad (10.1)$$

Specifically, if F is discrete with values x_n, p_n , then

$$G(y) = \sum_{\{x_n: \phi(x_n) \leq y\}} p_n,$$

and if it is absolutely continuous with density f , then

$$G(y) = \int_{\{x: \phi(x) \leq y\}} f(x) dx.$$

Proof. Let us denote $B_y = \{\omega : X(\omega) \leq y\}$; we have the following equalities:

$$\begin{aligned} G(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(\phi(X) \leq y) = \mathbb{P}(X \in B_y) = \\ &= \int_{B_y} d\mu_F = \int_{\{x: \phi(x) \leq y\}} dF(x). \end{aligned}$$

□

Theorem 10.2. *Let X be an absolutely continuous random variable with distribution function $F(x)$ and density $f(x)$. Let t be a strictly monotonic real function that has a derivative everywhere. Define $Y = t(X)$. Let t^{-1} denote the inverse function of t . Then Y has, for $y \in \mathbb{R}$, the density*

$$g(y) = f(t^{-1}(y)) |t^{-1}(y)|'.$$

Proof. Let $G(y)$ denote the distribution function of Y . Assume that t is increasing. Then it holds that

$$G(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(t(X) \leq y) = \mathbb{P}(X \leq t^{-1}(y)) = F(t^{-1}(y)).$$

We see that G is continuous and has a derivative everywhere, except possibly at finitely many points. Hence it holds that

$$G'(y) = f(t^{-1}(y))(t^{-1}(y))' = g(y),$$

where g is the density of the random variable Y . In the case that t is decreasing, the proof is analogous. $\square$

Example 10.1.

- a) Let X be a random variable with distribution function F , and define $Y = a + bX$ for $b \neq 0$, let G be the distribution function of Y . If $b > 0$, then $a + bX < Y$ is equivalent to $X < \frac{Y-a}{b}$, and from (10.1) it follows that

$$G(y) = \int_{\{x: x < \frac{y-a}{b}\}} dF(x) = F\left(\frac{y-a}{b}\right).$$

Conversely, if $b < 0$, then $a + bX < Y$ is equivalent to $X > \frac{Y-a}{b}$, and we obtain that

$$G(y) = \int_{\{x: x > \frac{y-a}{b}\}} dF(x) = 1 - F\left(\frac{y-a}{b}\right).$$

If F has density f , then G has density g , which is obtained by differentiating the previous relations. We obtain (for $b > 0$ as well as $b < 0$) the relation

$$g(y) = \frac{1}{|b|} f\left(\frac{y-a}{b}\right).$$

- b) If X has a standard normal distribution and $Y = \mu + \sigma X$, then from a) it follows that Y has a normal distribution with parameters μ and σ^2 .

- c) If X has a normal distribution with parameters μ, σ^2 and $Y = a + bX$, then from a) it follows that Y also has a normal distribution with parameters $a + b\mu$ and $b^2\sigma^2$, since

$$g(y) = \frac{1}{\sqrt{2\pi}|b|\sigma} e^{-\frac{\left(\frac{y-a}{b}-\mu\right)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}|b|\sigma} e^{-\frac{(y-(a+b\mu))^2}{2\sigma^2 b^2}}.$$

- d) If the density f of the random variable X is an even function, i.e., if $f(x) = f(-x)$ for every x , then the random variables X and $-X$ have the same distribution.

Let $Y = -X$, i.e., $a = 0$ and $b = -1$; from (10.1) it follows that

$$g(y) = f(-y), \quad \text{i.e.} \quad g(y) = f(y),$$

since f is an even function.

- e) If f is the density of a random variable X and is an even function, then the distribution function F satisfies

$$F(x) = 1 - F(-x) \text{ for any } x \in \mathbb{R},$$

because

$$F(-x) = \int_{-\infty}^{-x} f(t) dt = \int_x^{\infty} f(t) dt = 1 - F(x).$$

- f) Let X have a standard normal distribution. Let $Y = X^2$, then Y has the density

$$g(y) = \begin{cases} 0 & \text{for } y < 0, \\ \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}} & \text{for } y \geq 0. \end{cases}$$

Since, according to (10.1), it holds that

$$\begin{aligned} G(y) &= \int_{\{x: \phi(x) < y\}} f(x) dx = \frac{1}{\sqrt{2\pi}} \int_{\{x: x^2 < y\}} e^{-\frac{x^2}{2}} dx \\ &= \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{y}} e^{-\frac{x^2}{2}} dx, \end{aligned}$$

let us perform the substitution $x^2 = t$, and we obtain

$$G(y) = \frac{1}{\sqrt{2\pi}} \int_0^y t^{-\frac{1}{2}} e^{-\frac{t}{2}} dt.$$

It follows that the density $g(y)$ has the above form.

10.2 Function of a Random Vector

Similarly to the one-dimensional case (see Section 5.6), let us define the Lebesgue–Stieltjes measure μ_F induced by the distribution function F also for the multivariate case (i.e., on the space $\mathbb{R}^n$) by the formula

$$\mu_F((\mathbf{a}, \mathbf{b}]) = \sum_{k=0}^n (-1)^k \sum_{\boldsymbol{\delta} \in \Delta_{k,n}} F(\boldsymbol{\delta})$$

for semi-closed intervals $(\mathbf{a}, \mathbf{b}] = \times_{i=1}^n (a_i, b_i]$. This definition can be uniquely extended to all Borel sets $\mathcal{B}^n$. The integral with respect to the Lebesgue–Stieltjes measure of a function $\phi(\mathbf{x})$ will be denoted, as in the one-dimensional case, as

$$\int_{\mathbb{R}^n} \phi(\mathbf{x}) \, dF_{\mathbf{X}}(\mathbf{x}).$$

Example 10.2 (distribution of the minimum and maximum). Let X_i be, for $1 \leq i \leq n$, independent and identically distributed random variables with continuous distribution function F . Determine the distribution functions of the random variables

$$U = \max_{1 \leq i \leq n} X_i \quad \text{and} \quad V = \min_{1 \leq i \leq n} X_i.$$

Solution. From the independence of $X_1, \dots, X_n$, we obtain, according to Theorem 9.3

$$\begin{aligned} F_U(u) &= \mathbb{P}(U \leq u) = \mathbb{P}(X_1 \leq u, X_2 \leq u, \dots, X_n \leq u) = \\ &= \prod_{i=1}^n \mathbb{P}(X_i \leq u) = [F(u)]^n. \end{aligned}$$

$$\begin{aligned} F_V(v) &= \mathbb{P}(V \leq v) = 1 - \mathbb{P}(V > v) = \\ &= 1 - \mathbb{P}(X_1 > v, X_2 > v, \dots, X_n > v) = \\ &= 1 - \prod_{i=1}^n \mathbb{P}(X_i > v) = 1 - [1 - F(v)]^n. \end{aligned}$$

△

In cases where, for the characterization of the random variable $Z = \phi(\mathbf{X})$, it suffices to compute its expectation, it is not necessary to derive the distribution of Z ; it is enough (analogously to the one-dimensional case in Theorem 6.3) to know the distribution of the random vector $\mathbf{X}$.

Theorem 10.3. *Let $\mathbf{X}$ be a random vector and let $\phi: \mathbb{R}^n \rightarrow \mathbb{R}$ be a real-valued function. Then it holds that*

$$\mathbb{E}\phi(\mathbf{X}) = \int_{\mathbb{R}^n} \phi(\mathbf{x}) dF_{\mathbf{X}}(\mathbf{x}),$$

provided that at least one of the integrals exists. In particular

$$\mathbb{E}\phi(\mathbf{X}) = \sum_{m=1}^{\infty} \phi(\mathbf{x}_m) p_m$$

for the discrete distribution and

$$\mathbb{E}\phi(\mathbf{X}) = \int_{\mathbb{R}^n} \phi(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

for the absolutely continuous distribution.

In the case of independence, it is even sufficient to know the marginal distributions. For simplicity, we formulate the following theorem only for a pair of random variables.

Theorem 10.4. *Let X, Y be independent random variables with distribution functions F_X, F_Y and with Lebesgue–Stieltjes measures μ_{F_X}, μ_{F_Y} . Let $\phi: \mathbb{R}^2 \rightarrow \mathbb{R}$ be a real-valued function; then for the random variable $Z = \phi(X, Y)$ it holds that*

$$\mathbb{E}Z = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \phi(x, y) d\mu_{F_Y}(y) \right) d\mu_{F_X}(x) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \phi(x, y) dF_Y(y) \right) dF_X(x),$$

if $\mathbb{E}|Z| < \infty$.

Proof. From Theorem 9.3, it follows that for the Lebesgue–Stieltjes measures of independent random variables it holds that

$$\mu_{F_{X,Y}} = \mu_{F_X} \cdot \mu_{F_Y}.$$

From this, we obtain

$$\mathbb{E}Z = \int_{\mathbb{R}^2} \phi(x, y) d\mu_{F_{X,Y}}(x, y) = \int_{\mathbb{R}^2} \phi(x, y) d\mu_{F_Y}(y) d\mu_{F_X}(x)$$

and, by Fubini's theorem, we obtain the assertion of Theorem 10.4. □

Corollary 10.5. *The distribution function of the random variable $Z = \phi(X, Y)$ is*

$$G(z) = \int_{-\infty}^{\infty} \left(\int_{\{y: \phi(x, y) \leq z\}} d\mu_{F_Y}(y) \right) d\mu_{F_X}(x).$$

Proof. If in Theorem 10.4 we replace the function $\phi(x, y)$ by the indicator function $I_{(x, y) \in \mathbb{R}^2 : \phi(x, y) \leq z}(x, y)$, the statement follows immediately. $\square$

Corollary 10.6. *If X and Y are independent random variables with finite expectations, then*

a) $\mathbb{E}XY = (\mathbb{E}X)(\mathbb{E}Y).$

If, moreover, $\mathbb{E}X^2 < \infty$ and $\mathbb{E}Y^2 < \infty$, then

b) $\text{cov}(X, Y) = 0,$

c) $\text{var}(aX + bY) = a^2 \text{var } X + b^2 \text{var } Y$ for any $a, b \in \mathbb{R}.$

Proof.

a) Considering $|x|I_{[|x| \leq n]}|y|I_{[|y| \leq n]}$, we obtain by Lévy's theorem

$$\mathbb{E}|XY| = \mathbb{E}|X| \cdot \mathbb{E}|Y| < \infty.$$

We can therefore apply Theorem 10.4, in which we substitute $\phi(x, y) = x \cdot y$.

b) $(X - \mathbb{E}X), (Y - \mathbb{E}Y)$ are also independent random variables. Hence, by (a), we obtain

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}(X - \mathbb{E}X)\mathbb{E}(Y - \mathbb{E}Y) = 0.$$

c) According to (b), we obtain

$$\begin{aligned} \text{var}(aX + bY) &= \mathbb{E}[a(X - \mathbb{E}X) + b(Y - \mathbb{E}Y)]^2 = \\ &= a^2\mathbb{E}(X - \mathbb{E}X)^2 + b^2\mathbb{E}(Y - \mathbb{E}Y)^2 + 2ab\mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y) = \\ &= a^2 \text{var } X + b^2 \text{var } Y + 0. \end{aligned}$$

$\square$

Statement b) postulates that independent random variables are always uncorrelated. However, the converse does not hold; uncorrelated variables need not be independent¹. An exception is given by normally distributed random variables, for which equivalence between uncorrelatedness and independence holds.

10.3 Convolution

Let X and Y be two independent random variables with distribution functions $F(x)$ and $G(y)$. We are interested in the distribution of the sum

¹Uncorrelatedness corresponds to independence only “from a linear perspective”.

$Z = X + Y$. Let the distribution function of the random variable Z be denoted by $H(z)$. Then it holds that

$$\begin{aligned} H(z) &= \int \int_{x+y \leq z} dF(x) dG(y) = \int_{-\infty}^{\infty} F(z-y) dG(y) = \\ &= \int_{-\infty}^{\infty} G(z-x) dF(x). \end{aligned} \quad (10.2)$$

The distribution with distribution function $H(z)$ is called the *convolution of the distributions* with distribution functions $F(x)$ and $G(y)$. The function H is called the *convolution of the distribution functions* F and G . The convolution operation will be denoted by $H = F * G$.

The convolution operation is clearly commutative and associative, since for independent random variables X_1, X_2, X_3 it holds

$$X_1 + X_2 = X_2 + X_1$$

and

$$X_1 + (X_2 + X_3) = (X_1 + X_2) + X_3.$$

Thus, for the distribution functions, the following relations hold

$$F_1 * F_2 = F_2 * F_1$$

and

$$F_1 * (F_2 * F_3) = (F_1 * F_2) * F_3.$$

Theorem 10.7. *Let X and Y be independent random variables with absolutely continuous distribution functions $F(x)$ and $G(y)$ and densities $f(x)$ and $g(y)$. Then the convolution $H = F * G$ is also absolutely continuous, and its density $h(z)$ (i.e., the density of the random variable $Z = X + Y$) satisfies*

$$h(z) = \int_{-\infty}^{\infty} f(x)g(z-x) dx = \int_{-\infty}^{\infty} f(z-x)g(y) dy. \quad (10.3)$$

Proof. Formula (10.2) is, under our assumptions, identical to

$$H(z) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^z f(x-y) dx \right) dG(y) = \int_{-\infty}^z \left(\int_{-\infty}^{\infty} f(x-y) dG(y) \right) dx.$$

By differentiation, we obtain

$$h(z) = \int_{-\infty}^{\infty} f(z-y) dG(y),$$

and from this it follows (10.3). □

From the proof, it is also evident that, for the absolute continuity of the distribution of the random variable $Z = X + Y$, it actually suffices that the distribution of one of the random variables X and Y is absolutely continuous, without any assumptions on the distribution of the other random variable. The function $h(z)$ defined by relation (10.3) is called the *convolution of densities* $f(x)$ and $g(y)$, and we denote it by $h = f * g$. It is indeed a density, since from (10.3) it follows that $h(z) \geq 0$ and

$$\begin{aligned} \int_{-\infty}^{\infty} h(z) dz &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x-y)g(y) dy dx = \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(x-y) dx \right) g(y) dy = \\ &= \int_{-\infty}^{\infty} 1 \cdot g(y) dy = 1. \end{aligned}$$

The discrete analogue of relation (10.3) is given by the following theorem.

Theorem 10.8. *Let F, G be discrete distribution functions with jumps at natural numbers of sizes p_n, q_n , i.e.*

$$F(x) = \sum_{0 \leq n < x} p_n, \quad G(y) = \sum_{0 \leq n < y} q_n.$$

*Let $H = F * G$. Then H is a discrete distribution function with jumps at the natural numbers, and it holds that*

$$H(z) = \sum_{0 \leq n < z} h_n, \quad \text{where } h_n = \sum_{k=0}^n p_k q_{n-k}.$$

Proof. The theorem follows directly by applying the law of total probability. $\square$

We now present some examples of convolution of distributions.

1) **Convolution of uniform distributions** Let

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b, \\ 0 & \text{otherwise} \end{cases}$$

and

$$g(y) = \begin{cases} \frac{1}{d-c} & c \leq y \leq d, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that $d - c \geq b - a$. Then, for the convolution $h(z)$ of the densities of random variables with densities $f(x)$ and $g(y)$, it holds that

$$h(z) = \begin{cases} 0 & z \leq a + c \text{ or } b + d \leq z, \\ \frac{z - (a+c)}{(b-a)(d-c)} & a + c \leq z \leq b + c, \\ \frac{1}{d-c} & b + c \leq z \leq a + d, \\ \frac{(b+d) - z}{(b-a)(d-c)} & a + d \leq z \leq b + d. \end{cases}$$

The graph is a trapezoid with its base on the x -axis. We see that the density $h(z)$ is continuous everywhere, although $f(x)$ and $g(y)$ have points of discontinuity (the convolution “smooths out” discontinuities). In the special case where both random variables X and Y have the same distribution (i.e., $a = c$, $b = d$), the density $h(x)$ has the shape of a triangle; this distribution is called the *Simpson distribution*.

2) Convolution of binomial distributions

Let X_1, X_2 be independent random variables with distributions $\text{Bi}(n_1, p)$ and $\text{Bi}(n_2, p)$. Then the random variable $Y = X_1 + X_2$ has a binomial distribution with parameters $(n_1 + n_2, p)$.

3) Convolution of Poisson distributions

Let X_1 have $\text{Po}(\lambda_1)$ and X_2 have $\text{Po}(\lambda_2)$. Assume moreover that X_1 and X_2 are independent. Then the random variable $Y = X_1 + X_2$ has a Poisson distribution with parameter $\lambda_1 + \lambda_2$.

4) Convolution of normal distributions

Let X_1, X_2 be independent random variables, where $X_1 \sim \text{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \text{N}(\mu_2, \sigma_2^2)$. Then $Y = X_1 + X_2$ has the distribution $\text{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

5) Convolution of exponential distributions

If X_1, X_2 are independent random variables with the same exponential distribution with parameter $\lambda > 0$, then the random variable $Y = X_1 + X_2$ has the distribution $\text{Erlang}(2, \lambda)$.

Example 10.3 (Distribution of the sample mean $\bar{X}$). Consider a hypothetical population of meadow flowers whose height follows a normal distribution $\text{N}(20, 25)$. Compute the probability that the average height calculated from 20 randomly selected flowers differs from 20 by more than 3 cm.

Solution. Let $X_1, \dots, X_{20}$ be independent and identically distributed random variables with distribution $X_i \sim \text{N}(20, 25)$. The convolution of normal

distributions then yields

$$\sum_{i=1}^{20} X_i \sim \mathbf{N}(20 \cdot 20, 20 \cdot 25).$$

and, after a simple linear transformation $\bar{X} = \frac{1}{20} \cdot \sum_{i=1}^{20} X_i$, we obtain that

$$\bar{X} \sim \mathbf{N}\left(20, \frac{25}{20}\right).$$

The desired probability is therefore $\mathbb{P}(\bar{X} < 17) + \mathbb{P}(\bar{X} > 23)$. If we standardize $\bar{X}$ to a variable Z with a standard normal distribution, we obtain

$$\begin{aligned} & \mathbb{P}\left(\frac{\bar{X} - 20}{\sqrt{\frac{5}{4}}} < \frac{17 - 20}{\sqrt{\frac{5}{4}}}\right) + \mathbb{P}\left(\frac{\bar{X} - 20}{\sqrt{\frac{5}{4}}} > \frac{23 - 20}{\sqrt{\frac{5}{4}}}\right) = \\ &= \mathbb{P}\left(Z < \frac{-3}{\sqrt{\frac{5}{4}}}\right) + \mathbb{P}\left(Z > \frac{3}{\sqrt{\frac{5}{4}}}\right) = \Phi\left(\frac{-3}{\sqrt{\frac{5}{4}}}\right) + 1 - \Phi\left(\frac{3}{\sqrt{\frac{5}{4}}}\right) = \\ &= 2 \cdot \left(1 - \Phi\left(\frac{3}{\sqrt{\frac{5}{4}}}\right)\right) \approx 2 \cdot (1 - 0.996) \approx 0.008. \end{aligned}$$

△

Example 10.4. A female of a certain insect lays r eggs with probability $p_r = \frac{e^{-\lambda} \lambda^r}{r!}$. Each egg hatches into a live individual with probability p , which is the same for all eggs. The fates of the individual eggs are independent. We are interested in the probability with which the female gives life to exactly k new individuals.

Solution. Let $N \sim \mathbf{Po}(\lambda)$ be the random variable denoting the number of eggs, and let X be the random variable denoting the number of hatched individuals. We know that, conditional on knowing N , we have $X \sim \mathbf{Bi}(N, p)$. Thus, $\mathbb{P}(X = k)$ is computed by the law of total probability

$$\begin{aligned}
\mathbb{P}(X = k) &= \sum_{n=k}^{\infty} \mathbb{P}(X = k | N = n) \mathbb{P}(N = n) = \\
&= \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} \cdot e^{-\lambda} \frac{\lambda^n}{n!} = \\
&= e^{-\lambda} p^k \cdot \sum_{n=k}^{\infty} \frac{n!}{k!(n-k)!} \cdot (1-p)^{n-k} \frac{\lambda^n}{n!} = \\
&= \frac{e^{-\lambda} p^k \cdot \lambda^k}{k!} \cdot \sum_{n=k}^{\infty} \frac{(1-p)^{n-k} \lambda^{n-k}}{(n-k)!} = \\
&= \frac{e^{-\lambda} p^k \cdot \lambda^k}{k!} \cdot \sum_{n=0}^{\infty} \frac{((1-p) \cdot \lambda)^n}{n!} = \\
&= \frac{e^{-\lambda} p^k \cdot \lambda^k}{k!} \cdot e^{(1-p) \cdot \lambda} = e^{-\lambda p} \cdot \frac{(\lambda p)^k}{k!}.
\end{aligned}$$

In other words, $X \sim \text{Po}(\lambda p)$. △

Example 10.5. Compute the covariance of the random variables X and $Y = X^2$, where X has a uniform distribution on the interval $[-1, 1]$. Are these random variables independent?

Solution.

$$\begin{aligned}
\text{cov}(X, Y) &= \mathbb{E}XY - \mathbb{E}X \cdot \mathbb{E}Y = \mathbb{E}X^3 - \mathbb{E}X \cdot \mathbb{E}X^2 \\
X &\sim \text{U}[-1, 1] \Rightarrow \mathbb{E}X = 0 \\
\mathbb{E}X^3 &= \int_{-1}^1 x^3 \cdot \frac{1}{2} dx = 0.
\end{aligned}$$

Hence, we obtain that $\text{cov}(X, Y) = 0$, but the random variables X and $Y = X^2$ are clearly not independent. △

Example 10.6. The random variable Y is a function of an absolutely continuous random variable X . What is the density $g(y)$ if

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad -\infty < x < \infty \quad \text{and} \quad Y = X^2?$$

Solution.

$$t(x) = x^2, \quad -\infty < x < \infty.$$

In this interval, the function $t(x)$ is not monotonic. The distribution function of the random variable Y is, in general, of the form

$$\begin{aligned} G(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(X^2 \leq y) = \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) = \\ &= \mathbb{P}(X \leq \sqrt{y}) - \mathbb{P}(X \leq -\sqrt{y}) = F(\sqrt{y}) - F(-\sqrt{y}). \end{aligned}$$

Hence, for the density $g(y)$ we obtain

$$g(y) = f(\sqrt{y}) \frac{1}{2\sqrt{y}} - f(-\sqrt{y}) \left(-\frac{1}{2\sqrt{y}}\right) = \frac{1}{2\sqrt{y}} [f(\sqrt{y}) + f(-\sqrt{y})].$$

Specifically, for $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, it holds that

$$g(y) = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{2\sqrt{y}} \left(e^{-\frac{y}{2}} + e^{-\frac{y}{2}} \right) = \begin{cases} \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}} & y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

We have obtained the density of the χ^2 distribution with one degree of freedom, which is used in mathematical statistics [13]. $\triangle$

10.4 Exercises

10.1 The random variable Y is a function of an absolutely continuous random variable X . What is the probability density function $g(y)$ if

- a) $Y = 8X^3$, and $f(x) = \begin{cases} 2x & 0 < x < 1, \\ 0 & \text{otherwise,} \end{cases}$
- b) $Y = -2 \ln X$, and $f(x) = \begin{cases} 1 & 0 < x < 1, \\ 0 & \text{otherwise,} \end{cases}$
- c) $Y = X^3$, and $f(x) = \begin{cases} \frac{x^2}{9} & 0 < x < 3, \\ 0 & \text{otherwise,} \end{cases}$
- d) $Y = X^2$, and $f(x) = \begin{cases} 2xe^{-x^2} & x > 0, \\ 0 & \text{otherwise,} \end{cases}$
- e) $Y = e^X$ and $f(x) = \begin{cases} 1 & 0 < x < 1, \\ 0 & \text{otherwise,} \end{cases}$
- f) $Y = |X|$ and $f(x)$ is arbitrary $-\infty < x < \infty$,
- g) $Y = \sin X$, and $f(x) = \begin{cases} \frac{1}{\pi} & -\frac{\pi}{2} < x < \frac{\pi}{2}, \\ 0 & \text{otherwise,} \end{cases}$
- h) $Y = |1 - X|$, and $f(x) = \begin{cases} \frac{1}{2} & 0 < x < 2, \\ 0 & \text{otherwise.} \end{cases}$

10.2 Determine the distribution of the random variable Y if

- a) $Y = 2X + 1$ and

$$p_n = \begin{cases} \frac{1}{3} & n = 1, 2, 3, \\ 0 & \text{otherwise,} \end{cases}$$

- b) $Y = X^3$

$$p_n = \begin{cases} \left(\frac{1}{2}\right)^n & n = 1, 2, 3, \dots, \\ 0 & \text{otherwise.} \end{cases}$$

- c) Let X, Y be independent random variables, where X has an exponential distribution with parameter $\lambda > 0$, and Y has a uniform distribution on $(0, \theta)$, $\theta > 0$.

- (a) Determine the distribution of $X + Y$.
- (b) Determine $\mathbb{E}(X + Y)$.

(c) Determine $\text{var}(X + Y)$.

10.3 Let X_i , $1 \leq i \leq n$, be independent random variables taking values ± 1 with probabilities $\mathbb{P}(X_i = 1) = p$, $\mathbb{P}(X_i = -1) = 1 - p$. Determine the distribution of the random variable $S_n = \sum_{i=1}^n X_i$.

10.4 We roll three one-crown coins and four five-crown coins. Let X denote the total value of those one-crown coins that land heads, and let Y denote the total value of those five-crown coins that land heads. Define $W = X + Y$. Compute the covariance of the random variables X and W .



10.5 Let $X \sim U[1, 2]$. Compute the covariance of the random variables X and $\frac{1}{X}$.

10.6 Let $(X_1, \dots, X_n)^T$ be a sequence of independent random variables with a uniform distribution on $(0, 1)$. Denote

$$U = \max_{1 \leq i \leq n} X_i, \quad V = \min_{1 \leq i \leq n} X_i.$$

- a) Determine the distribution functions and densities of the random variables U and V .
- b) Determine $\mathbb{E}U$, $\text{var } U$, $\mathbb{E}V$, $\text{var } V$.

10.7 Let X and Y be independent random variables, both having the same exponential distribution.

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0, \\ 0 & \text{for } x < 0. \end{cases}$$

Find the distribution function of the random variable $Z = \max(X, Y)$ and its expectation. Are the random variables $\max(X, Y)$ and $\min(X, Y)$ independent?

10.8 Let the random vector $Y = (X_1, X_2)^T$ have the joint density

$$f_Y(x_1, x_2) = \frac{1}{2\pi\sigma^2} \cdot e^{-\frac{(x_1^2 + x_2^2)}{2\sigma^2}}.$$

Determine the probability density function of the random variable $Z = X_1 + X_2$.

Solutions to the Exercises

10.1

$$\begin{aligned}
\text{a) } g(y) &= \begin{cases} \frac{1}{6}y^{-\frac{1}{3}} & 0 < y < 8, \\ 0 & \text{otherwise,} \end{cases} \\
\text{b) } g(y) &= \begin{cases} \frac{1}{2}e^{-\frac{y}{2}} & y > 0, \\ 0 & \text{otherwise,} \end{cases} \\
\text{c) } g(y) &= \begin{cases} \frac{1}{27} & 0 < y < 27, \\ 0 & \text{otherwise,} \end{cases} \\
\text{d) } g(y) &= \begin{cases} e^{-y} & y \geq 0, \\ 0 & \text{otherwise,} \end{cases} \\
\text{e) } g(y) &= \begin{cases} \frac{1}{y} & 1 < y < e, \\ 0 & \text{otherwise,} \end{cases} \\
\text{f) } g(y) &= \begin{cases} f(y) + f(-y) & y \geq 0, \\ 0 & y < 0, \end{cases} \\
\text{g) } g(y) &= \begin{cases} \frac{1}{\pi\sqrt{1-y^2}} & -1 < y < 1, \\ 0 & \text{otherwise,} \end{cases} \\
\text{h) } g(y) &= \begin{cases} 1 & 0 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

10.2

$$\begin{aligned}
\text{a) } \mathbb{P}(Y = y) &= \begin{cases} \frac{1}{3} & y = 3, 5, 7, \\ 0 & \text{otherwise.} \end{cases} \\
\text{b) } \mathbb{P}(Y = y) &= \begin{cases} \left(\frac{1}{2}\right)^n & y = n^3, n = 1, 2, 3, \dots, \\ 0 & \text{otherwise.} \end{cases} \\
\text{c) } f_Z(z) &= \frac{1}{\theta} \cdot e^{-z\lambda} \cdot (e^{\theta\lambda} - 1), \quad z \geq 0, \\
\mathbb{E}(X + Y) &= \frac{1}{\lambda} + \frac{\theta}{2}, \\
\text{var}(X + Y) &= \frac{1}{\lambda^2} + \frac{\theta^2}{12}.
\end{aligned}$$

10.3 The random variable S_n has a distribution defined by the probabilities

$$\mathbb{P}(S_n = s) = \binom{n}{\frac{s+n}{2}} p^{\frac{s+n}{2}} (1-p)^{n-\frac{s+n}{2}},$$

where $s = \{-n, -n+2, -n+4, \dots, n\}$ and $\frac{s+n}{2}$ are integers.

10.4 $\text{cov}(X, W) = 0.75$

10.5 $\text{cov}(X, \frac{1}{X}) = 1 - \frac{3}{2} \ln 2$

10.6

$$F_U(u) = \begin{cases} 0, & u \leq 0, \\ u^n, & u \in (0, 1), \\ 1, & u \geq 1, \end{cases} \quad f_U(u) = \begin{cases} nu^{n-1}, & u \in (0, 1), \\ 0, & u \notin (0, 1), \end{cases}$$

$$F_V(v) = \begin{cases} 0, & v \leq 0, \\ 1 - (1 - v)^n, & v \in (0, 1), \\ 1, & v \geq 1, \end{cases} \quad f_V(v) = \begin{cases} n(1 - v)^{n-1}, & v \in (0, 1), \\ 0, & v \notin (0, 1). \end{cases}$$

$$\mathbb{E}U = \frac{n}{n+1}, \mathbb{E}V = \frac{1}{n+1}, \text{var } U = \frac{n}{n+2} - \left(\frac{n}{n+1}\right)^2, \text{var } V = \frac{1}{(n+1)(n+2)}.$$

10.7 $F_Z(z) = \mathbb{P}(Z \leq z) = \mathbb{P}(\max(X, Y) \leq z) = \begin{cases} 0 & \text{for } z < 0, \\ (1 - e^{-\lambda z})^2 & \text{for } z \geq 0. \end{cases}$

$\mathbb{E}Z = \frac{3}{2\lambda}$. The random variables $\max(X, Y)$ and $\min(X, Y)$ are independent.

10.8 $f(z) = \frac{1}{2\sqrt{\pi\sigma^2}} \cdot e^{-\frac{z^2}{4\sigma^2}}, z \in \mathbb{R}.$

Chapter 11

Law of Large Numbers, Central Limit Theorem

11.1 Convergence of Random Variables

So far, we have considered a random variable and its characteristics under the assumption that these characteristics are known exactly. However, this is not a very typical situation; more often, we describe some property of a population by means of a random variable, but its characteristics are unknown – we then speak of the theoretical properties of the population, or of the random variable (theoretical distribution, theoretical expectation, etc.). If, however, we are able to repeat independently a random experiment associated with a theoretical random variable, we can construct from the observed values the distribution of relative frequencies and summarize the information about this distribution again by means of its characteristics. This distribution and its characteristics are called the empirical distribution and empirical characteristics. For example, we estimate the expectation of a theoretical distribution by taking a random sample from this distribution – that is, by performing a sequence of independent random experiments whose outcomes are independent random variables having the same distribution as the theoretical random variable – and we use as an estimate of the theoretical expectation the empirical characteristic, namely the arithmetic mean.

Under suitable conditions, we may expect that the empirical distribution (respectively, its characteristics) will approach the theoretical distribution (respectively, the theoretical characteristics), and this approximation improves as the number of performed experiments increases. This is generally expressed by the so-called law of large numbers. However, it should be noted that the convergence of empirical values to theoretical values is not

of the usual type of mathematical convergence, but rather of a probabilistic nature. By probabilistic convergence, we mean that as the number of experiments increases, the probability of large deviations of empirical values from the theoretical ones decreases. There are, in fact, several types of probabilistic convergence, differing in strength. The most commonly used types will be introduced in the following definition. Finally, let us remark that some statements in this chapter will be presented without proof; the proofs can be found, for example, in [8].

Definition 11.1 (Convergence of random variables)

Let $X_1, X_2, X_3, \dots$ be a sequence of random variables and let X be a random variable. Assume that all these variables are defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

a) We say that X_n converges to X almost surely if

$$\mathbb{P}\{\omega : X_n(\omega) \xrightarrow[n \rightarrow \infty]{} X(\omega)\} = 1.$$

b) If for every $\varepsilon > 0$ it holds that

$$\mathbb{P}\{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\} \xrightarrow[n \rightarrow \infty]{} 0,$$

then we say that X_n converges to X in probability.

c) If

$$\mathbb{E}X_n^2 < \infty, \quad n = 1, 2, 3, \dots,$$

and if

$$\mathbb{E}(X_n - X)^2 \xrightarrow[n \rightarrow \infty]{} 0,$$

then we say that X_n converges to X in mean square (of order 2).

The following theorem shows the relationship between the different modes of convergence.

Theorem 11.1 (on relationships between modes of convergence). *Consider convergence of a sequence of random variables almost surely and convergence in probability. Then the following statements hold:*

a) *Almost sure convergence implies convergence in probability.*

b) *Convergence in mean square (of order 2) implies convergence in probability.*

The statements of Theorem 11.1 cannot, without additional assumptions, be strengthened. None of the implications can be reversed.

Lemma 11.2. *For any non-negative random variable X with expectation $\mathbb{E}X$, it holds that*

$$\mathbb{P}(X \geq \varepsilon) \leq \frac{\mathbb{E}X}{\varepsilon}.$$

Proof.

$$\mathbb{P}(X \geq \varepsilon) = \int_{x \geq \varepsilon} dF_X(x) \leq \int_{x \geq \varepsilon} \frac{x}{\varepsilon} dF_X(x) \leq \frac{1}{\varepsilon} \int x dF_X(x) = \frac{1}{\varepsilon} \mathbb{E}X.$$

□

Lemma 11.2 is also meaningful for larger values of ε , not only for very small ε , as is customary in mathematical analysis.

Example 11.1. Let X be a random variable such that $\mathbb{E}X = 3$ and $\mathbb{E}X^2 = 13$. Estimate the probability that the random variable X takes values in the interval $(-2, 8)$.

Solution. The expectation is $\mathbb{E}X = 3$. We determine the variance of the random variable X :

$$\text{var } X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = 13 - 3^2 = 4.$$

The required probability is estimated using Chebyshev's inequality:

$$\mathbb{P}(-2 < X < 8) = \mathbb{P}(|X - \mathbb{E}X| < 5) \geq 1 - \frac{4}{25} \approx 0.84.$$

△

11.2 Law of Large Numbers

Classical results of probability theory include the behaviour of the arithmetic mean when the number of summands in its definition tends to infinity. These properties are described by the laws of large numbers.

Theorem 11.3 (Weak Law of Large Numbers). *Let $X_1, X_2, \dots$ be independent random variables with identical expectations μ and identical variances $\sigma^2 < \infty$. Then, as $n \rightarrow \infty$,*

$$\frac{1}{n}(X_1 + X_2 + \dots + X_n) \rightarrow \mu$$

in probability.

Proof. Let us denote

$$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n).$$

Let us compute the expectation and the variance of the random variable $\overline{X}_n$:

$$\begin{aligned}\mathbb{E}\overline{X}_n &= \mathbb{E}\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] = \frac{1}{n}[\mathbb{E}X_1 + \mathbb{E}X_2 + \cdots + \mathbb{E}X_n] = \\ &= \frac{1}{n}n\mu = \mu, \\ \text{var } \overline{X}_n &= \text{var}\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] = \\ &= \frac{1}{n^2}[\text{var } X_1 + \text{var } X_2 + \cdots + \text{var } X_n] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.\end{aligned}$$

Now we use Chebyshev's inequality (Theorem 6.5) for the random variable $\overline{X}_n$, and we obtain that for every $\varepsilon > 0$ it holds that

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

whence, as $n \rightarrow \infty$, we have $\frac{\sigma^2}{n\varepsilon^2} \rightarrow 0$, and therefore also

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \varepsilon) \rightarrow 0.$$

□

This version of the law of large numbers belongs among the historically earliest results. In essence, it states that if we are interested in estimating the theoretical expectation μ of a random variable, we can approximate it arbitrarily closely by the arithmetic mean of a sufficiently large number of random variables with the same expectation and the same variance. The statement of Theorem 11.3 can be further strengthened. One such result is given in the following theorem. Further generalizations, for example to random vectors with different expectations, can be found in [8].

Theorem 11.4 (Strong Law of Large Numbers). *Let $\{X_n\}_{n=1}^\infty$ be a sequence of independent and identically distributed random variables with expectation $\mathbb{E}X_1 = \mu$, and assume that $\mathbb{E}|X_1| < \infty$. Then*

$$\lim_{n \rightarrow \infty} \overline{X}_n = \mu$$

both in probability and almost surely.

11.3 Central Limit Theorem

The essence of the Central Limit Theorem (hereinafter CLT) is the statement that a random variable Y_n , formed as the sum of a large number of mutually independent random variables $X_1, X_2, \dots, X_n$, has approximately a normal distribution under very general conditions. We say that a random variable Y_n whose limiting distribution is normal has an *asymptotically normal distribution*.

The oldest and simplest case of the CLT is the *de Moivre–Laplace theorem^{*1}. The random variable Y_n is the sum of n mutually independent random variables, each having a Bernoulli (alternative) distribution with parameter p . Then the random variable Y_n has the distribution $\text{Bi}(n, p)$ with expectation $\mathbb{E}Y_n = np$ and variance $\text{var } Y_n = np(1 - p)$. The de Moivre–Laplace theorem states that for the standardized random variable

$$Z_n = \frac{Y_n - np}{\sqrt{np(1 - p)}}$$

it holds that the following limit relation is satisfied

$$\lim_{n \rightarrow \infty} \mathbb{P}(Z_n \leq x) = \Phi(x), \quad -\infty < x < \infty,$$

where $\Phi(x)$ is the distribution function of the standard normal distribution $N(0, 1)$. Thus, the de Moivre–Laplace theorem states that, for a sufficiently large number of independent trials, the binomial distribution can be approximated by a normal distribution.

Theorem 11.5 (De Moivre–Laplace theorem). *Let, for any $n \geq 1$, Y_n be a random variable with distribution $\text{Bi}(n, p)$, $0 < p < 1$. Define*

$$Z_n = \frac{Y_n - np}{\sqrt{np(1 - p)}}$$

and let $F_n(x)$ denote the distribution function of the random variable Z_n . Then

$$\lim_{n \rightarrow \infty} F_n(x) = \Phi(x), \quad -\infty < x < \infty. \quad (11.1)$$

Example 11.2. What is the probability that at least 14 sixes occur in 120 throws of a fair die?

¹Abraham de Moivre (1667–1754), French mathematician.

Solution. Let $X_i \sim B\left(\frac{1}{6}\right)$ be the random variable indicating whether a six occurs in the i -th throw of a die. For X_i we have $\mathbb{E}X_i = \frac{1}{6}$ and $\text{var } X_i = \frac{5}{36}$. We therefore wish to compute

$$\mathbb{P}\left(\sum_{i=1}^{120} X_i \geq 14\right).$$

This example can first be computed exactly. The random variable $Y_{120} = \sum_{i=1}^{120} X_i$ has the binomial distribution $\text{Bi}\left(120, \frac{1}{6}\right)$. Using the formula (7.2) for the probability mass function of the binomial distribution (or by using the Excel function for the binomial distribution), we obtain that

$$\mathbb{P}(Y_{120} \geq 14) = 1 - \sum_{k=0}^{13} \mathbb{P}(Y_{120} = k) = 1 - \sum_{k=0}^{13} \binom{120}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{120-k} \approx 0.950.$$

$$= 1 - \text{BINOM.DIST}(13, 120, 1/6, \text{TRUE})$$

Now let us compute this probability approximately using the CLT approximation. The use of the CLT consists in transforming the random variable Y_{120} into the random variable Z_{120} according to Theorem 11.5. Thus, we write

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^{120} X_i \geq 14\right) &= \mathbb{P}\left(\frac{\sum_{i=1}^{120} X_i - 120 \cdot \frac{1}{6}}{\sqrt{120 \cdot \frac{1}{6} \cdot \frac{5}{6}}} \geq \frac{14 - 120 \cdot \frac{1}{6}}{\sqrt{120 \cdot \frac{1}{6} \cdot \frac{5}{6}}}\right) \approx \\ &\approx \mathbb{P}(Z_{120} \geq -1.470) = 1 - \mathbb{P}(Z_{120} \leq -1.470). \end{aligned}$$

The random variable Z_{120} has, by the CLT, an asymptotically standard normal distribution, and thus we may finally write that

$$\mathbb{P}\left(\sum_{i=1}^{120} X_i \geq 14\right) \approx 1 - \Phi(-1.470) \approx 0.929.$$

$$= 1 - \text{NORM.S.DIST}(-1.470, \text{TRUE})$$

The required probability is approximately (after applying the CLT approximation and rounding) 0.929. We see that the deviation from the exactly computed probability is only about two percentage points. $\triangle$

Example 11.3. How many times must we roll a die so that the probability of obtaining at least ten sixes is greater than or equal to 0.95?

Solution. Analogously to Example 11.2, let $X_i \sim B(\frac{1}{6})$ be the random variable indicating whether or not a six occurs in the i -th throw of the die. The problem can be rewritten as the inequality

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq 10\right) \geq 0.95, \quad (11.2)$$

where the unknown is n , i.e., the number of die throws. This problem can be solved by brute force: for increasing values of n , one computes the binomial probabilities on the left-hand side of inequality (11.2) until one of them exceeds the value 0.95. A more elegant and faster approach is to estimate the value of n using the CLT. This again consists in transforming the inequality into a form involving a random variable that asymptotically approaches a normally distributed random variable:

$$\mathbb{P}\left(\frac{\sum_{i=1}^n X_i - \frac{n}{6}}{\sqrt{\frac{5n}{36}}} \geq \frac{10 - \frac{n}{6}}{\sqrt{\frac{5n}{36}}}\right) \geq 0.95. \quad (11.3)$$

Thus, the value $\frac{10 - \frac{n}{6}}{\sqrt{\frac{5n}{36}}}$ is also determined as the 5% quantile² of the random variable $Z_n = \frac{\sum_{i=1}^n X_i - \frac{n}{6}}{\sqrt{\frac{5n}{36}}}$. By the CLT, this random variable has an asymptotically standard normal distribution, whose 5% quantile we can compute by $\text{= NORM.S. INV}(0.05) = \Phi^{-1}(0.05) \approx -1.64$. Hence

$$\frac{10 - \frac{n}{6}}{\sqrt{\frac{5n}{36}}} \approx -1.64.$$

From this equation, we easily obtain the result $n = 96$. In other words, the die must be rolled at least 96 times in order to achieve (approximately) a 95% probability that at least ten sixes occur. $\triangle$

Example 11.4. According to mortality tables, the probability of death of a 32-year-old woman is 0.001819. In the event of death, the insurance company pays the family 1 million CZK. The insurance company insures 5 000 32-year-old women. Determine the premium that the insurance company should charge so that the probability that the company earns at least 2 million CZK is equal to 0.95.

²Recall that the quantile of a random variable, as introduced in Section 6.2, specifies the location of the left tail of the random variable, whereas inequality (11.3) specifies the location of the right tail.

Solution. Let x denote the insurance premium. The number of women who die follows a binomial distribution, i.e., $Y_n \sim \text{Bi}(n, p)$, where $n = 5,000$ and $p = 0.001819$. From the problem statement, we can formulate the equation

$$\mathbb{P}(x \cdot 5\,000 - 1\,000\,000 \cdot Y_n > 2\,000\,000) = 0.95,$$

$$\mathbb{P}\left(Y_n < \frac{5\,000x - 2\,000\,000}{1\,000\,000}\right) = 0.95.$$

Now we transform the inequality so that we can apply the de Moivre–Laplace version of the CLT.

$$\begin{aligned}\mathbb{P}\left(\frac{Y_n - 5\,000p}{\sqrt{5\,000p(1-p)}} < \frac{\frac{1}{200}x - 2 - 5\,000p}{\sqrt{5\,000p(1-p)}}\right) &= 0.95, \\ \mathbb{P}\left(Z_n < \frac{\frac{1}{200}x - 11.095}{3.013}\right) &= 0.95, \\ \frac{\frac{1}{200}x - 11.095}{3.013} &\approx \Phi^{-1}(0.95) \approx 1.64.\end{aligned}$$

$$= \text{NORM.S.INV}(0.95)$$

Hence, $x \approx 3\,207.28$. $\triangle$

The Lévy–Lindeberg theorem³ is a strengthening of the de Moivre–Laplace theorem and expresses convergence to the normal distribution under more general conditions. In this case, the random variable Y_n is the sum of n independent random variables $X_1, X_2, \dots, X_n$, which are identically distributed with finite expectation $\mathbb{E}X_i = \mu$ and finite variance $\text{var } X_i = \sigma^2$ for $i = 1, 2, \dots, n$. Then, according to the Lévy–Lindeberg theorem, for the standardized random variable

$$Z_n = \frac{Y_n - n\mu}{\sqrt{n\sigma^2}}$$

it also satisfies the corresponding limit relation

$$\lim_{n \rightarrow \infty} \mathbb{P}(Z_n \leq x) = \Phi(x), \quad -\infty < x < \infty.$$

Theorem 11.6 (Lévy–Lindeberg theorem). *Let $X_1, X_2, \dots$ be independent and identically distributed random variables with expectation μ and finite variance σ^2 . Denote*

$$Z_n = \frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{n\sigma^2}}, \quad n = 1, 2, \dots$$

³Paul Pierre Lévy (1886–1971), French mathematician. Jarl Waldemar Lindeberg (1876–1932), Finnish mathematician.

and let $F_n(x)$ denote the distribution function of Z_n . Then

$$\lim_{n \rightarrow \infty} F_n(x) = \Phi(x), \quad -\infty < x < \infty,$$

where $\Phi(x)$ is the distribution function $N(0, 1)$.

Theorem 11.6 is stated without proof (its proof requires the theoretical apparatus of characteristic functions in probability theory); however, we shall show how the de Moivre–Laplace theorem is a special case of the Lévy–Lindeberg theorem.

Proof of Theorem 11.5. We start from Theorem 11.6. Let X_n be random variables with a Bernoulli (alternative) distribution with parameter p . Then Y_n has a binomial distribution with parameters (n, p) , and by Theorem 11.6, statement (11.1) holds. $\square$

Example 11.5. Mr. Smith regularly travels to and from work by tram. It is known that the waiting time for the arrival of a tram ranges between 0 and 3 minutes. What is the probability that the total waiting time of Mr. Smith during 23 working days will be less than 80 minutes?

Solution. The waiting time for the arrival of the tram during the i -th trip, i.e., the random variable X_i , has a uniform distribution with probability density function

$$f(x_i) = \begin{cases} \frac{1}{3} & 0 < x_i < 3, \quad i = 1, \dots, 46, \\ 0 & \text{otherwise.} \end{cases}$$

The total waiting time for commuting to and from work during 23 days is the random variable

$$Y_{46} = \sum_{i=1}^{46} X_i.$$

The expectation and variance of the random variables X_i , for $i = 1, 2, \dots, 46$, are

$$\mathbb{E}X_i = \frac{3-0}{2} = \frac{3}{2}, \quad \text{var } X_i = \frac{3}{4}.$$

Since the conditions of the Lévy–Lindeberg theorem are satisfied, we can determine the required probability as

$$\begin{aligned} \mathbb{P}(Y_{46} < 80) &= \mathbb{P}\left(\frac{\sum_{i=1}^{46} X_i - 46 \cdot \mathbb{E}X_i}{\sqrt{46 \cdot \text{var } X_i}} < \frac{80 - 46 \cdot \mathbb{E}X_i}{\sqrt{46 \cdot \text{var } X_i}}\right) = \\ &= \mathbb{P}\left(Z_{46} < \frac{80 - 69}{\sqrt{34.5}}\right) \approx \mathbb{P}(Z_{46} < 1.87) \approx \Phi(1.87) \approx 0.969. \end{aligned}$$

Table 11.1: Comparison of the distribution function of Z_{12} with the distribution function $N(0, 1)$

x	0	0.5	1.0	1.5	2.0	2.5	3.0
$\Phi(x)$	0.5	0.6915	0.8413	0.9331	0.9772	0.9937	0.9986
$F_{12}(x)$	0.5	0.6894	0.8393	0.9326	0.9777	0.9944	0.9990

```
= NORM.S.DIST(STANDARDIZE(80,69,SQRT(34.5)),TRUE)
```

△

Very generally, the Central Limit Theorem was formulated by Aleksandr Lyapunov⁴. He proved that the distribution of the sum of independent random variables X_i , for $i = 1, 2, \dots, n$, converges to a normal distribution even when the variables X_i are not identically distributed. This theorem is named after him as the Lyapunov theorem. Other variants of the CLT describe the asymptotic behaviour of random vectors or even entire stochastic processes. The assumption of independence of the variables $X_1, X_2, \dots$ can also be weakened. However, the rigorous formulations of these theorems, as well as their mathematical proofs, are rather involved; therefore, we omit them in this textbook and refer the reader, for example, to the literature [8].

How fast does the distribution of sums of independent random variables approach the normal distribution? We will not address this question theoretically, but instead present a numerical comparison for one specific case: let X_k , $1 \leq k \leq 12$, be “random numbers,” i.e., independent random variables with uniform distribution on $[0, 1]$. Denote $Z_{12} = \sum_{k=1}^{12} X_k - 6$ (thus Z_{12} has zero mean and unit variance). In Table 11.1, F_{12} denotes the distribution function of Z_{12} and Φ the distribution function of $N(0, 1)$; due to symmetry, it suffices to consider non-negative values of the argument. The agreement is surprisingly good, considering the small number of summands and the markedly different shapes of the uniform and normal densities. However, such a good approximation cannot always be expected. Roughly speaking, convergence is faster for distributions that are symmetric (around the mean) and slower for asymmetric ones. Let us also mention a practical rule of thumb: the binomial distribution can be approximated by the normal distribution whenever $np(1-p) \geq 9$.

⁴Aleksandr Mikhailovich Lyapunov (1857–1918), Russian mathematician and physicist.

11.4 Exercises

11.1 A shipment contains 3,000 products of a certain type. It is known that the probability of producing a defective product of this type is 0.04.

- Estimate the probability that the absolute deviation between the proportion of defective products in the shipment and the probability of producing a defective product is less than 0.01.
- How does the result change if the probability of producing a defective product is 0.004 and the shipment contains 30,000 products?

11.2 During reliability testing, a product fails with probability $p = 0.05$. What is the probability that, when testing 100 products,

- at least 7 products fail,
- fewer than 5 products fail?

11.3 What is the probability that the relative frequency of occurrence of an event in 100 trials will lie between 0.2 and 0.4, given that the probability of occurrence of the event in a single trial is 0.3?

11.4 Let the time required to detect and repair a failure be a random variable X with expectation $\mathbb{E}X = 40$ minutes and standard deviation $\sigma = 30$ minutes. What total time will be required to handle 100 failures, if we require that this time is not exceeded with probability 0.95?

11.5 We roll a die 100 times. Let S_{100} denote the sum of the obtained outcomes. Using the CLT, determine an approximate value of the probability

$$\mathbb{P}(320 \leq S_{100} \leq 380).$$

11.6 Let v_n denote the relative frequency of heads in n coin tosses. How many trials must be performed so that the probability of the event $|v_n - \frac{1}{2}| \leq 0.05$ is at least 0.95?

- Solve using Chebyshev's inequality.
- Solve using the Central Limit Theorem.



11.7 An insurance company insures 1 000 individuals of the same age. The probability that any one of them dies within a year is 0.01. Each insured person pays a premium of 1 200 CZK. In the event of death, the insurance company pays the family 80 000 CZK. What is the probability that the insurance company incurs a loss? Use the Central Limit Theorem (CLT).

Solutions to the Exercises

11.1 a) 0.872, b) approximately 1.

11.2 a) 0.179, b) 0.324.

11.3 0.971

11.4 4 493.5 minutes

11.5 0.9216

11.6 384 trials

11.7 approximately 0.065

Part II

Mathematical Statistics

Chapter 12

Statistical data processing

12.1 Basic Concepts

Before we begin dealing with basic statistical methods, we define the fundamental concepts from the field of statistical data processing.

Definition 12.1 (Fundamental concepts of statistics)

- a) A statistical population Z is defined as a set of objects classified with respect to a certain common property, called a variable.
- b) Objects belonging to a statistical population are called statistical units, or simply elements of the set.
- c) A variable whose different values or realizations are observed for all elements of a given population, and which is chosen as the basis for their classification, will be referred to as the argument of the population.
- d) The total number of elements in the population under consideration is called the size of the population.
- e) The sum of all values of the argument (variable) of a statistical population is called the total of the argument (total of the variable).



Example 12.1. A group of n individuals classified with respect to their age forms a statistical population of size n . Its argument is the age of the individuals. The same group of n individuals constitutes a different statistical population if we choose height (measured, for example, in centimeters) as the argument.

Different types of measurements on the units of the population produce variables of different character:

- 1) categorical variables arising from nominal measurement, for example the determination of the sex of the statistical units (male, female), occupation, blood groups, etc., for which statistical processing consists primarily in
 - determination of frequencies (absolute and relative frequencies),
 - determination of measures of location (mode),
 - determination of measures of variability (nominal variance),
 - analysis of contingency tables,
 - computation of measures of association (contingency coefficient), etc.
- 2) categorical variables arising from ordinal (rank) measurement, in which numerical values are assigned to the units of the population so as to express their ordering according to a given criterion; for example, in a group of 20 employees, the ranks 1–20 are assigned according to an individual's resistance to stress (1 denotes the most resistant individual, 20 the least resistant), or responses obtained from a questionnaire survey using a Likert scale (thus, the magnitude of differences between units with different variable values cannot be inferred from the variable itself); statistical processing of ordinal data typically consists in
 - determination of frequencies (including cumulative frequencies),
 - determination of measures of location (median, quantiles),
 - determination of measures of variability (e.g. quartile deviation),
 - determination of measures of dependence (e.g. Spearman's rank correlation coefficient),
 - hypothesis testing based on the computation of order statistics (sign test, Wilcoxon tests, Kruskal–Wallis test, etc.).
- 3) continuous variables arising from interval (metric) measurement, in which numerical values are assigned to the units so as to represent the magnitude of differences between them (that is, the differences between values are equal, but there is no true zero in the sense of the absence of the given property), for example the measurement of temperature in degrees Celsius (0°C does not indicate the absence of temperature, but merely one of its possible values); in the statistical processing of interval data we typically use
 - determination of measures of location (arithmetic mean),



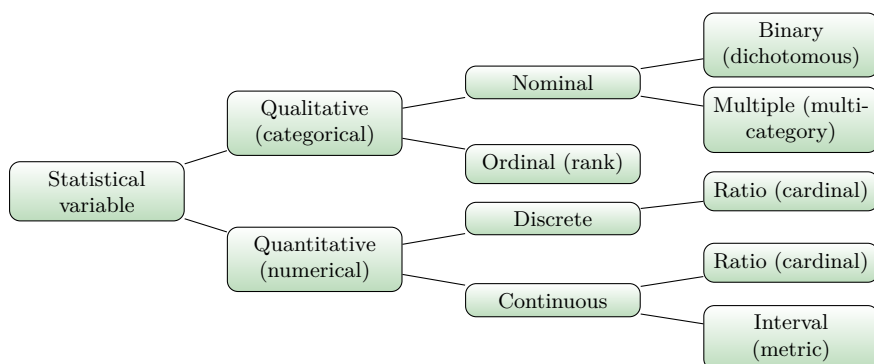


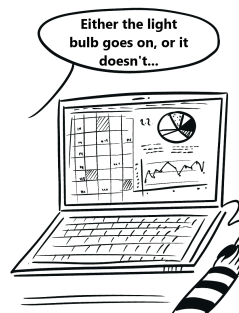
Figure 12.1: Categorization of statistical variables

- determination of measures of variability (standard deviation),
 - parametric tests (Student's t -tests, F -test, analysis of variance),
 - determination of measures of dependence (Pearson's correlation coefficient), etc.
- 4) discrete or continuous variables arising from ratio (cardinal) measurement, in which numerical values assigned to the units express the “amount of a property” being measured (that is, the differences between values are equal and the zero value indicates the absence of the given variable in the object under study), for example the measurement of an individual's weight (0 indicates the absence of weight) or the measurement of the number of insurance events (0 indicates the absence of an event); in the statistical processing of ratio-scale data, methods from all the preceding categories can be employed 1)–3).

The diagram in Figure 12.1 presents the types of variables most commonly encountered by statisticians. The values of a variable obtained by a researcher during data collection in a real-world setting are referred to as data.

12.2 Frequency Distribution and Classification

In statistical populations of larger size, repeated occurrences of identical values of a statistical variable frequently arise. The number of occurrences of a given value of a variable is referred to as the *frequency* of that value. It serves primarily for the complete description of categorical (nominal and



ordinal) data, and also of quantitative (interval and ratio) data when these are pre-classified into so-called class intervals. In the case of discrete statistical variables with a limited number of distinct values, simple (point) classification is employed. Individual values of the statistical variable are referred to as *classes* of the statistical variable.

Definition 12.2 (Types of frequencies)

- a) *The number of elements of the population belonging to the k -th class is called the absolute frequency of the argument in the k -th class (also the absolute class frequency or simply the frequency of the k -th class), and is denoted by f_k .*
- b) *If f_k denotes the absolute class frequency of the k -th class and n the size of the population under consideration, then*
 - (a) *$\frac{f_k}{n}$ is called the relative frequency of the k -th class,*
 - (b) *$100\frac{f_k}{n}$ is called the percentage relative frequency of the k -th class.*
- c) *Cumulative (cumulative-sum) absolute frequency F_k of the k -th class is defined as the sum of all frequencies f_j up to and including the k -th class, i.e.*

$$F_k = f_1 + f_2 + \cdots + f_k = \sum_{j=1}^k f_j.$$

- d) *Cumulative relative frequency R_k of the k -th class is defined as the sum*

$$R_k = \frac{f_1}{n} + \frac{f_2}{n} + \cdots + \frac{f_k}{n} = \sum_{j=1}^k \frac{f_j}{n} = \frac{F_k}{n}.$$

Cumulative frequencies (items c) and d)) are, of course, meaningful for data that can be naturally ordered, that is, in particular for ordinal (but not nominal) categorical variables. Frequencies satisfy certain natural properties summarized in the following theorem.



Theorem 12.1. *Consider a statistical population of size n , which is partitioned into r classes.*

- a) *The sum of all absolute frequencies of all values across all classes is equal to the size of the population, i.e.*

$$f_1 + f_2 + \cdots + f_r = \sum_{k=1}^r f_k = n.$$

- 2) The cumulative absolute frequency of all classes of the population is equal to the size of the population, i.e.

$$F_r = n,$$

- 3) the sum of all relative frequencies of all classes of the population is equal to one, i.e.

$$\frac{f_1}{n} + \frac{f_2}{n} + \cdots + \frac{f_r}{n} = \sum_{k=1}^r \frac{f_k}{n} = 1.$$

Frequencies of a statistical population are typically presented in the form of a table containing the values of the statistical variable along with the corresponding absolute and relative frequencies; this is referred to as a frequency distribution table (or distribution table of frequencies). For the visualization of these frequencies, a variety of graphical tools can be employed, as described in the following list.

- 1) *Bar chart* of the distribution of absolute (or relative) frequencies is constructed by plotting individual classes (values of the categorical statistical variable) on the x -axis and, above each class k , drawing a rectangle whose height equals the corresponding absolute frequency f_k or the relative frequency $\frac{f_k}{n}$.
- 2) *Spike diagram* (or graph) of the distribution of absolute (respectively relative) frequencies is constructed analogously, except that the rectangles are replaced by line segments oriented in the direction of the y -axis.
- 3) *Frequency polygon* (line diagram) is obtained by connecting the end-points of the line segments from the frequency diagram, thereby forming a broken line representing the desired polygon, i.e. the line diagram.
- 4) A bar chart, spike diagram, or polygon of cumulative frequencies is obtained analogously as in items 1, 2, and 3. The polygon of cumulative relative frequencies is called the *ogive* (briefly, ogive).

Example 12.2.  At a telephone exchange, the number of incoming calls per one-minute interval was recorded. Over the course of one hour, during a certain period of the day, the following results were obtained: 3, 2, 2, 3, 1, 1, 0, 4, 2, 1, 1, 4, 0, 1, 2, 3, 1, 2, 5, 2, 3, 0, 2, 4, 1, 2, 3, 0, 1, 2, 1, 3, 1, 2, 0, 7, 3, 2, 1, 1, 4, 0, 0, 1, 4, 2, 3, 2, 1, 2, 2, 3, 1, 4, 0, 2, 1, 1, 5, 3. Construct the table of the distribution of absolute and relative frequencies (including cumulative frequencies) for this statistical population. Furthermore, construct the graph of relative frequencies and the graph of cumulative relative frequencies for this population.

Table 12.1: Frequency distribution table for Example 12.2

Number of calls	0	1	2	3	4	5	6	7	Total
f_k	8	17	16	10	6	2	0	1	60
F_k	8	25	41	51	57	59	59	60	
f_k/n	0.133	0.283	0.266	0.166	0.1	0.033	0.000	0.016	1
R_k	0.133	0.417	0.683	0.850	0.950	0.983	0.983	1	

Solution. In Excel, frequency values are most efficiently computed using the pivot table tool (Insert > Pivot Table): the field “number of calls” is placed simultaneously in the Rows and Values areas, while in the Values area the field settings are adjusted by changing the aggregation criterion to Count – thus obtaining the distribution of absolute frequencies. By further modifying the value display to % of column total (and optionally setting the number format to three decimal places), we obtain the values of relative frequencies. Cumulative frequencies are obtained by selecting the display option Running Total In, or % Running Total In, respectively. A single source field can be placed repeatedly into the Values area of the pivot table, allowing all four types of frequencies to be computed within the same pivot table. The result is shown in Table 12.1.

An alternative approach is to use Excel functions for conditional counting. To do this, it is first necessary to create a block of cells containing the possible values of the categorical variable (in our case, values from 0 to 7, for example in cells C3:C9). The `=COUNTIF` function returns, for each of these values, the number of occurrences in the range of source data (for example in cells A2:A61). For the occurrence of zeros, the formula thus takes the form `=COUNTIF($C$2:$C$61,$E2)`, and for the remaining values it can simply be copied. The result is a table of absolute frequencies. Relative frequencies are obtained by simple division by the total number of observations, and cumulative frequencies by successive summation of absolute frequencies (the construction of these formulas is left to the reader as an exercise).

The graphical representation can be obtained using an appropriate chart type based on the tables prepared above. Either the pivot table or the manually constructed frequency table may be used as the data source. As the chart type, we choose a bar chart or a line chart, respectively. The resulting charts may appear, for example, as shown in Figures 12.2 and 12.3. △

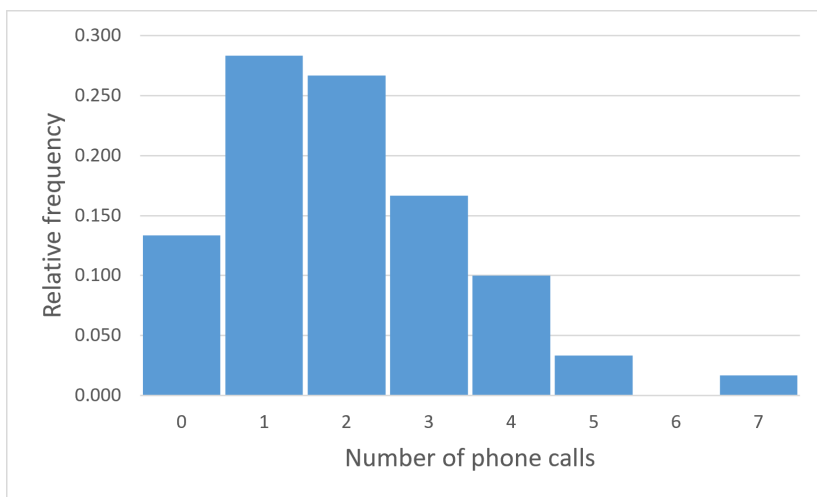


Figure 12.2: Bar chart of relative frequencies for Example 12.2

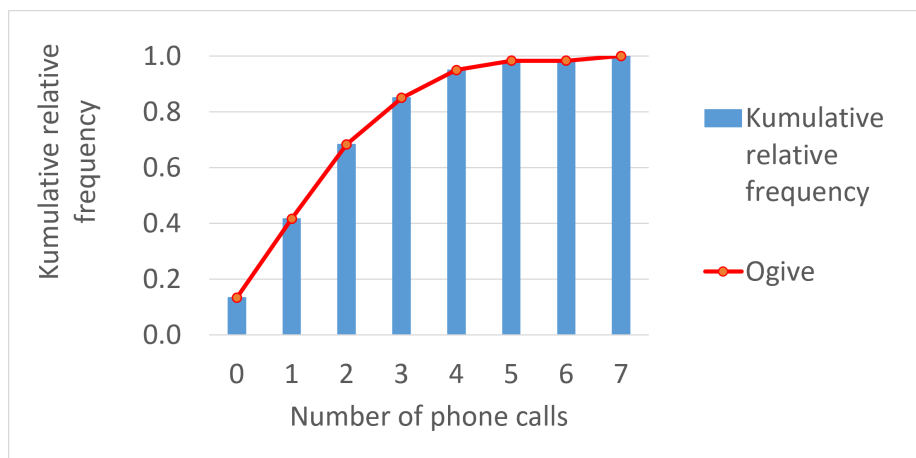


Figure 12.3: Graph of cumulative relative frequencies and the ogive for Example 12.2

Relative frequencies approximate the probability distribution of a random variable, and cumulative relative frequencies approximate its cumulative distribution function. For example, the values of relative frequencies in Table 12.1 represent estimates of the unknown probabilities of a discrete random variable representing the number of telephone calls during a randomly selected one-minute time interval, while the cumulative relative frequencies represent estimates of the values of its distribution function.

Simple classification of quantitative variables does not yield satisfactory results: most values of such variables typically occur only once (or with

low frequencies if the statistical population results from measurements with low precision). Therefore, before computing frequencies, it is necessary to categorize the data into a smaller number of classes (typically class intervals) – this procedure is referred to as interval classification.

Definition 12.3 (Range and interval classification)

Let a denote the minimum value of the argument X and b the maximum value of the argument X of the given statistical population, i.e. $x_{\min} = a$, $x_{\max} = b$.¹

- a) The interval $[a, b]$ is called the range (or also the domain of variability, variability interval) of the argument X of the given statistical population.
- b) The difference $R = b - a$ is called the range (spread) of the argument X of the given statistical population.
- c) The range $[a, b]$ is partitioned into smaller subsets called classes of the argument X . The most common case is the partition of the interval $[a, b]$ into a sequence of consecutive non-overlapping class intervals $(a_k, b_k]$.
- d) The width (length) of the class interval $(a_k, b_k]$ is defined as $h_k = b_k - a_k$. The quantity $\frac{1}{2}(a_k + b_k)$ is called the class midpoint, the number a_k the lower class boundary, and the number b_k the upper class boundary.
- e) A selected value x_k of the argument X , representing all values belonging to a given class, is called the class mark of the k -th class. The class mark is most commonly chosen as the midpoint of the corresponding class.

Two observations are considered equivalent whenever the data (values of the variable) obtained in both measurements fall into the same class. When partitioning the range $[a, b]$ with n observed values into class intervals, the following principles are usually observed:

- 1) The recommended number of class intervals is between 8 and 20. The chosen number depends on the size of the population and the purpose of the statistical table. The number r of class intervals can also be determined, for example, using Sturges' rule²

$$r \approx 1 + \log_2 n \approx 1 + 3.322 \log n,$$

Yule's rule³

$$r \approx 2.5\sqrt[4]{n},$$

or simpler relations $r \approx \sqrt{n}$ or $r \approx 5 \log_{10} n$.

¹It is meaningful to consider maximum and minimum values in the case of metric measurement.

²Herbert Arthur Sturges (1882–1958), American mathematician and statistician.

³George Udny Yule (1871–1951), English mathematician and statistician.

- 2) The number of classes may also be determined by choosing the class width. In such cases, we usually select the same width for all class intervals and typically choose some natural values. As an indicative value, we may take, for example, a width

$$h_k = h \approx \frac{8}{100}(b - a).$$

- 3) Class intervals in Definition 12.3 are formalized as left-open and right-closed. This is not a strict rule; for example, in Excel the function `= FREQUENCY` and histogram charts behave in this way (moreover, the first class interval in these tools is, by definition, also left-closed). Other tools may use alternative conventions for the inclusion of boundary values: classification using the `= VLOOKUP` function, for instance, typically leads to class intervals that are left-closed and right-open. Values lying on class boundaries are usually assigned according to the default behavior of the selected tool.
- 4) If very few values of the argument X occur in the extreme classes (the first or the last), it is advisable to merge these classes with the adjacent class into a single class. (Such a class will, of course, have a different width.)

The number and width of class intervals are not uniquely determined and always depend on the subjective choice of the researcher. The objective, however, is to select such a number of intervals and corresponding widths that the resulting interval frequency distribution reflects the basic characteristics of the variable under investigation. Once the data have been classified into class intervals, frequencies can be computed and frequency graphs constructed, as was already done in the case of simple classification. Some elements can even be constructed directly from the data (the classification into class intervals being performed automatically by the corresponding tool). In particular, a bar chart of absolute or relative frequencies of class intervals is called a *histogram*. It is constructed by plotting the midpoints of the individual classes on the x -axis and, above each segment representing a given class (of width h_k), drawing a rectangle whose height equals the corresponding absolute frequency f_k , or the relative frequency $\frac{f_k}{n}$. In Excel, a histogram of absolute frequencies is created directly from the data as a histogram-type chart. A histogram of relative frequencies, however, must be constructed either as a standard bar chart based on the interval classification table, or by using the Data Analysis tool (specifically, the Histogram analysis tool).

Example 12.3.  The data in Table 12.2 present the numbers of deaths in the Czech Republic per 100 000 inhabitants as of 1 January 2023 for the

Table 12.2: Numbers of deaths per 100 000 inhabitants in individual weeks of the year (from weeks 1 to 52: left to right, top to bottom)

50.08	44.55	42.5	40.68	39.42	38.63	43.45	36.97	39.50	42.26
38.23	38.87	38.55	37.92	38.79	38.00	37.44	39.73	33.41	35.86
32.47	34.05	32.94	32.47	37.13	31.28	33.26	37.52	32.23	34.76
32.78	37.84	34.76	37.29	33.89	35.23	34.05	36.97	30.65	36.34
38.55	35.31	39.50	36.50	39.15	38.79	38.47	43.84	38.47	39.34
43.21	42.18								

Data source: Czech Statistical Office

age group 65 to 74, by week. Construct a table of the interval distribution of absolute and relative frequencies (including cumulative frequencies) for the given statistical population and construct a histogram of the frequencies.



Solution. The size of the data set is $n = 52$ observations. This corresponds to $1 + \log_2(52) \approx 6.7 \approx 7$ class intervals according to Sturges' rule, and the same number of intervals also according to Yule's rule ($2.5\sqrt[4]{52} \approx 6.7$) as well as the square-root rule $\sqrt{52} \approx 7.2$. The value given by the logarithmic rule is higher, $5 \log_{10}(52) \approx 8.6$, which corresponds to eight or nine class intervals. Alternatively, we may proceed from the range $50.08 - 30.65 = 19.43$ and choose a width of approximately $8/100 \cdot 19.43 \approx 1.55$.

On the basis of the above analysis, we decided to classify the data into seven class intervals (which corresponds to a class width of $19.43/7 \approx 2.77 \approx 3$) and additionally to choose class boundaries at “convenient” values starting from 30. With this choice, the minimum and maximum lie within the range defined by the classification (between 30 and 51), but even if this were not the case, any low or high values can be incorporated into the extreme class intervals. The solution is shown in Table 12.3. If we place the data into cells A2:A53, we construct the table of absolute frequencies as follows: first, in the first column of the resulting table (for example in cells C2:C8), we enter the values of the class boundaries, starting at 33 and ending at 48. In the adjacent column, we select cells D2:D9 (one cell more than the number of boundary values), press the key F2, and enter the formula `= FREQUENCY(A2:A53,C2:C8)` and confirm it by pressing **Ctrl+Shift+ENTER** – the formula is thus stored as an array formula. In newer versions of Excel, the formula can be inserted more simply as a dy-

Table 12.3: Table of interval classification of the number of deaths per 100,000 inhabitants (n_k – absolute frequency, F_k – cumulative absolute frequency, f_k – relative frequency, R_k – cumulative relative frequency)

k	interval	n_k	F_k	f_k	R_k
1	(30, 33]	7	7	0.135	0.135
2	(33, 36]	10	17	0.192	0.327
3	(36, 39]	20	37	0.385	0.712
4	(39, 42]	7	44	0.135	0.846
5	(42, 45]	7	51	0.135	0.981
6	(45, 49]	0	51	0.000	0.981
7	(48, 51]	1	52	0.019	1.000

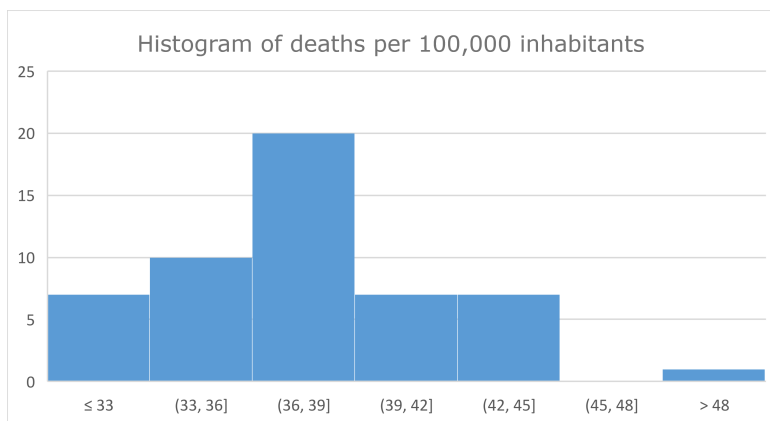


Figure 12.4: Histogram of frequencies for Example 12.3

namic array formula – by entering it in the standard way into cell D2, with Excel automatically expanding it to the remaining cells.

The Histogram tool (Insert > Charts > Histogram) automatically selects the number of intervals and the class width; however, these settings can be adjusted via the formatting of the x -axis options (Format Axis > Axis Options). By setting the Underflow Bin or Overflow Bin options, the positions of the class boundaries can also be modified (these correspond to the boundary between the first and second, and between the penultimate and last class intervals, respectively); this also ensures that low and high values of the variable are included in the extreme class intervals. The resulting output may resemble, for example, Figure 12.4.

The argument of a statistical population is, in probabilistic terms, represented by a random variable X . From the law of large numbers (see, in particular, Theorem 13.2), it follows that the relative frequency $\frac{f_k}{n}$ gives (approximately) the probability that X falls into the k -th class, i.e. $\mathbb{P}(a_k < X \leq b_k) \approx \frac{f_k}{n}$. Thus, the histogram of relative frequencies approximates the probability density function of the random variable X . Similarly, the cumulative relative frequencies represented by the ogive approximate the cumulative distribution function of this random variable.

12.3 Characteristics of Location

The probabilistic behavior of a random variable is uniquely and completely determined by the probability mass function, the probability density function, or the cumulative distribution function. In solving probabilistic problems, it is often advantageous to summarize the essential information about the distribution of a random variable by means of several numerical characteristics, as we have done in Chapter 6 for characteristics of location (mean, mode, median), variability (variance, standard deviation), symmetry (skewness), and concentration (kurtosis). Such characteristics are referred to in statistics as theoretical characteristics. In an analogous manner, data sets can also be characterized: for example, relative frequencies describing categorical data represent the empirical counterpart of the probability mass function of a random variable, while cumulative frequencies are the empirical counterpart of the cumulative distribution function. Likewise, the properties of data variables can be summarized by simple numerical characteristics. Since most data sets originate from samples (empirical data), these characteristics are referred to as sample (empirical) characteristics. In this and the following sections, we therefore focus on sample characteristics of location, variability, symmetry, and concentration of empirical data sets.

Definition 12.4 (Arithmetic, geometric, and harmonic mean)

Let a statistical population be given whose argument x takes values $x_1, x_2, \dots, x_n$, possibly classified into r classes, where f_k denotes the absolute frequency of the k -th class and x_k is a suitably chosen representative of this class.

a) Arithmetic mean $\bar{x}$ is defined by the relations

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \approx \frac{1}{n} \sum_{k=1}^r f_k x_k. \quad (12.1)$$

b) Geometric mean $\bar{x}_g$ is defined for positive values $x_1, x_2, \dots, x_n$ ⁴ by the relation

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdots x_n} \approx \sqrt[n]{\prod_{k=1}^r x_k^{f_k}}. \quad (12.2)$$

c) Harmonic mean $\bar{x}_h$ is defined by the relations

$$\bar{x}_h = \frac{1}{a}, \quad \text{where} \quad a = \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i} \approx \frac{1}{n} \sum_{k=1}^r \frac{f_k}{x_k}. \quad (12.3)$$

In relations (12.1), (12.2) to (12.3), two forms are given. The first form corresponds to an ungrouped data set, and the second to a grouped data set.

The arithmetic mean is evidently the most commonly used measure of location of a data set. However, the value of the arithmetic mean is significantly influenced by extreme values present in the data set (referred to as outliers). Therefore, it is sometimes more appropriate to use another measure of location, such as the median or the trimmed mean.

The geometric mean is applied in economics and financial mathematics, for example in the computation of the average percentage increase (decrease) in the prices of goods (the average rate of growth or decline) over time intervals within the observed period. The geometric mean cannot be used if the argument X assumes zero or negative values.

The harmonic mean can be used when the sum of reciprocal values is meaningful, i.e., when the averaged quantity has the character of a part of a whole, that is, when computing an average of so-called ratio quantities; for example, in the computation of the average speed $\bar{v}$ of a car in km/h, given partial distances s_i and the corresponding speeds v_i , where time is given by $t = \frac{s}{v}$.

In contrast to frequency distribution graphs, which provide relatively extensive information about a statistical population, each measure of location yields only limited partial information about the “average” value of a unit of the statistical population.

The following two theorems are presented without proof.

Theorem 12.2 (On the properties of the arithmetic mean).

⁴If at least one of the values x_i equals zero, then the geometric mean is also equal to zero.

- a) The sum of all deviations of the values of the variable X_i from the arithmetic mean of all values of the data set is equal to zero, i.e.

$$\sum_{i=1}^n (x_i - \bar{x}) = 0.$$

- b) If all values of the variable x_i are equal to the same constant c , then the arithmetic mean of all values of the variable is equal to the same constant c , i.e.

$$\bar{x} = \sum_{i=1}^n \frac{x_i}{n} = \sum_{i=1}^n \frac{c}{n} = c.$$

- c) If the same constant c is added to all values of the variable x_i , then the arithmetic mean of all values of the variable increases by this constant c , i.e. we have

$$\sum_{i=1}^n \frac{x_i + c}{n} = \bar{x} + c.$$

- d) If all values of the variable x_i are multiplied by the same constant c , then the arithmetic mean of these values is multiplied by the same constant c , i.e.

$$\sum_{i=1}^n \frac{c \cdot x_i}{n} = c \cdot \bar{x}.$$

Theorem 12.3 (On the relationship between means, without proof). *For an arbitrary statistical population with variable X , the following holds:*

$$\bar{x}_h \leq \bar{x}_g \leq \bar{x}.$$

Example 12.4.  Consider a stock whose price increased by 10% in the first year, decreased by 20% in the second year, and decreased again by 15% in the third year. What was the average percentage growth rate of the stock price?

Solution. The average growth rate of the stock price is given by the geometric mean of the growth (decline) data. We can write:

$$\bar{x}_g = \sqrt[3]{(1 + 0.1) \cdot (1 - 0.2) \cdot (1 - 0.15)} \approx 0.908,$$

 = GEOMEAN(A1:A3)

thus, the average decline in the stock price was 9.1% after rounding. The value 9.1% represents the average constant rate of decrease in the stock price over the observed years, i.e. $0.908^3 \approx 1.1 \cdot 0.8 \cdot 0.85$. This can also be interpreted as meaning that the total decrease in the price of this stock after three years is the same as if the price had decreased by 9.1% each year. $\triangle$

Example 12.5.  Consider a post office branch with four clerks serving customers at the counter at a given time. The first clerk serves a customer in two minutes, the second in three minutes, the third in five minutes, and the fourth in six minutes. All clerks perform tasks of equal difficulty. What is the average time required to serve one customer?

Solution. The first clerk serves 30 customers per hour, the second 20 customers per hour, the third 12 customers per hour, and the fourth 10 customers per hour. Over a total time of 4 hours, all four clerks together serve 72 customers. The average time required to serve one customer is therefore $\frac{4}{72} = \frac{1}{18}$ hours, which is approximately 3.3 minutes. The fraction $\frac{4}{\frac{72}{60}}$ is the simple harmonic mean of the times required to serve one customer by the individual clerks.

$$\bar{t}_h = \frac{4}{\frac{1}{t_1} + \frac{1}{t_2} + \frac{1}{t_3} + \frac{1}{t_4}} = \frac{4}{\frac{1}{2} + \frac{1}{3} + \frac{1}{5} + \frac{1}{6}} = \frac{4}{\frac{72}{60}} \approx 3.3 \text{ min.}$$

 = HARMEAN(A1:A4)

The average time required to serve one customer is approximately 3.3 min. △

If the researcher suspects that the value of the computed arithmetic mean is substantially influenced by extreme values in the data set (the minimum and maximum values), it is a common practice to exclude these extreme values (or a certain percentage of all values) from the data set; this is referred to as the *trimmed mean*, or to replace them by values closer to the remaining observations in the data set, in which case it is called the *winsorized mean*. Most often, however, the issue of extreme values is addressed by a robust measure of location, namely the median.

Let a statistical population be given whose argument x takes values $x_1, x_2, \dots, x_n$. If we arrange these values in ascending order, we obtain the so-called *ordered statistical population*

$$x_{(1)}, x_{(2)}, \dots, x_{(n)},$$

where $x_{(1)}$ denotes the smallest value, $x_{(2)}$ denotes the second smallest value, etc. In general, $x_{(i)}$ denotes the i -th order statistic.

Definition 12.5 (Median)

The median is defined in two ways depending on the number of elements in the statistical population. In the case of an odd number of values, the median $\tilde{x}$ is taken to be the central value of the ordered data set

$$\tilde{x} = x_{(\lfloor \frac{n}{2} \rfloor + 1)}.$$

If X has an even number of values, the median $\tilde{x}$ is defined as the arithmetic mean of the two central values

$$\tilde{x} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}.$$

The median is a special case of the sample quantile. A sample quantile is a value chosen such that the observations less than this value constitute a prescribed proportion of the sample (e.g., the 10% sample quantile denotes a value that is greater than 10% of the observations in the data set and less than 90% of the observations in the data set).

Definition 12.6 (Quantile)

The sample quantile $\tilde{x}_p$ is a real number such that 100p% of the units of the ordered sample have values less than or equal to $\tilde{x}_p$ and 100(1 - p)% of the units have values greater than or equal to x_p , where $0 < p < 1$.

We distinguish three special cases of the sample quantile: the 25% sample quantile, denoted by $\tilde{x}_{25}$, is called the *lower* (or *first*) *sample quartile* and is the value below which one quarter of all observations lie, i.e. 25% of all values are less than or equal to this quantile. The 50% sample quantile is called the second quartile and is, in fact, the median, denoted in this context by $\tilde{x}_{50}$. Finally, the 75% sample quantile, denoted by $\tilde{x}_{75}$, is called the *upper* (*third*) *sample quartile*, and three quarters of all observations lie below it. Other commonly used quantiles include deciles $\tilde{x}_{10}, \tilde{x}_{20}, \dots, \tilde{x}_{100}$ and percentiles $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_{100}$.

Quantiles (including the median and quartiles) are not uniquely determined by their definition. For example, in the case of an even sample size, the median can in principle be any value between the two “central” observations. The formula given in Definition 12.5 (which averages these two values) is merely the most common choice. For quantiles and quartiles, the situation is less uniform: different statistical software may return different (although similar) values of sample quantiles, and Microsoft Excel even provides two different functions for computing quantiles: `= PERCENTILE.INC` and `= PERCENTILE.EXC`. For larger data sets, however, these differences are negligible in practice.

Definition 12.7 (Mode)

Let the argument of a statistical population take only finitely many values. Then the mode is the value of the argument with the highest absolute frequency.

Table 12.4: Table of absolute frequencies for Example 12.6

Variable x	x_1	x_2	x_3	x_4
Value of the variable	10	16	28	50
Absolute frequency n_i	30	10	3	1

The mode, most commonly denoted by $\hat{x}$, need not be uniquely determined: it may happen that two or more values of the argument have the same absolute frequency. For continuous variables, the mode is usually not defined at all (most values typically occur only once or not at all in the data set). In such cases, however, the statistical population may be classified into class intervals and the mode of the class intervals determined. The mode obtained in this way is referred to as the *modal class*.

In MS Excel, the mode can be computed using the function `= MODE.MULT` (or `= MODE.SNGL` if the mode is unique). Their argument is the range of all processed data, and the output of both functions is the value that occurs most frequently in the data set, i.e. at least twice. If each value in the range occurs only once, the function returns an error. If the mode is multiple, all such values are returned by the former function as a dynamic array.

The mode, the median, and the arithmetic mean typically take different values. Close values of the median and the mean indicate a symmetric frequency distribution. In statistics, data sets with multiple maxima (peaks) are also often encountered. Distributions with a single peak are called unimodal, those with two peaks bimodal, etc.

Example 12.6.  Table 12.4 presents the absolute frequencies of four values of a certain statistical variable. Compute the arithmetic mean, median, and mode. Determine the values of the lower and upper sample quartiles.

Solution. The arithmetic mean is

$$\bar{x} = \frac{30 \cdot 10 + 10 \cdot 16 + 3 \cdot 28 + 1 \cdot 50}{44} = 13.50.$$

`= AVERAGE(A1:A44)`

Since there are 44 values, the median is the mean of the 22nd and 23rd ordered values, i.e. $\tilde{x} = 10$.

`= MEDIAN(A1:A44)`

The mode is the most frequent value of the variable; according to Table 12.4, it is the value 10. It can also be verified that $\hat{x} = \text{MODE.MULT}(10, \dots, 50) = 10$.

The lower sample quartile is the mean of the 11th and 12th ordered values, i.e. 10, and the upper sample quartile is 16.

$$\tilde{x}_{25} = \text{QUARTILE.EXC}(A1:A44, 1) = 10,$$

$$\tilde{x}_{75} = \text{QUARTILE.EXC}(A1:A44, 3) = 16.$$

△

12.4 Measures of Variability (dispersion)

Definition 12.8 (Measures of variability)

- a) The range (variation range) R is defined as the difference between the largest and the smallest value of the variable (in this order), i.e.

$$R = x_{(n)} - x_{(1)}.$$

- b) The variance (dispersion, also population variance) σ_n^2 of a statistical population of size n is defined as the arithmetic mean of the squared deviations $(x_k - \bar{x})^2$ of the values of the variable X from the arithmetic mean $\bar{x}$:

$$\sigma_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{k=1}^r f_k (x_k - \bar{x})^2. \quad (12.4)$$

- c) The sample variance s_n^2 of a statistical population of size n is defined as the mean of the squared deviations $(x_k - \bar{x})^2$ of the values of the variable X from the arithmetic mean $\bar{x}$, adjusted by the so-called degrees of freedom:

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \sum_{k=1}^r f_k (x_k - \bar{x})^2. \quad (12.5)$$

- d) The population and sample standard deviation σ_n , respectively s_n , is defined as

$$\sqrt{\sigma_n^2} = \sigma_n \geq 0. \quad (12.6)$$

for the population version of the variance, respectively

$$\sqrt{s_n^2} = s_n \geq 0 \quad (12.7)$$

for the sample version of the variance.

- e) The mean absolute deviation $\bar{d}$ is defined as the arithmetic mean of the absolute values of the deviations from the arithmetic mean $\bar{x}$, i.e.

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| = \frac{1}{n} \sum_{k=1}^r f_k |x_k - \bar{x}|. \quad (12.8)$$

f) The coefficient of variation⁵ v of a statistical population is defined as

$$v = \frac{\sigma_n}{\bar{x}} \text{ or } v = \frac{s_n}{\bar{x}} \quad (12.9)$$

(depending on the version of the standard deviation used).

g) The variation ratio (for nominal data) is defined as

$$VP = 1 - \frac{n_m}{n}, \quad (12.10)$$

where n_m denotes the frequency of observations in the modal category and n is the total size of the population.

h) Nominal variance is defined by the formula

$$NV = \frac{n^2 - \sum n_k^2}{n^2}, \quad (12.11)$$

where n_k denotes the frequency corresponding to the given category and n is the size of the population.

Characteristics f) and g) are used for nominal data.

The range is the “coarsest” measure of variability, indicating the length of the interval in which all values of the variable lie. The disadvantage of the range is that it depends only on extreme values, which may not be representative of the given variable. Therefore, in the analysis of variability (dispersion) of data, so-called *quantile ranges* are also used, in particular:

- the *interquartile range* defined as the difference between the upper and lower quartile,

$$R_Q = \tilde{x}_{75} - \tilde{x}_{25},$$

- the *decile range*

$$R_D = \tilde{x}_{90} - \tilde{x}_{10},$$

- the *percentile range*

$$R_C = \tilde{x}_{99} - \tilde{x}_1,$$

and *quantile deviations*

- the *interquartile deviation* $Q = \frac{R_Q}{2}$,
- the *decile deviation* $D = \frac{R_D}{8}$,
- the *percentile deviation* $C = \frac{R_C}{98}$.

⁵Also referred to as CV from the English *coefficient of variation*.

For example, the interquartile deviation represents the average length of the two central quartile intervals; similarly, the decile and percentile deviations express the average distance between adjacent deciles and percentiles, respectively. The interquartile range and the interquartile deviation are typically associated with situations in which the median is chosen as the measure of location of the statistical population.

The coefficient of variation serves to compare the relative magnitude of the standard deviation with respect to the arithmetic mean, i.e. what percentage of the mean is represented by the standard deviation. This indicator is dimensionless, usually expressed as a percentage, and is used for comparing the variability of two or more variables measured in different units. The coefficient of variation is not suitable when the mean value is close to zero. On the contrary, it is appropriate when all data are positive. The value $v = 1$ corresponds to 100%.

The variance is defined by formula (12.4), which is suitable for manual computation in MS Excel. For calculations using a calculator, one would use the so-called computational form

$$\sigma_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i^2) - \bar{x}^2 = \frac{1}{n} \sum_{k=1}^r f_k x_i^2 - \bar{x}^2. \quad (12.12)$$

If, however, the underlying data are available, the fastest way to compute the variance is by using the formula = VAR.P.

The following theorem is again stated without proof.

Theorem 12.4 (On the properties of the variance).

- a) *If all values of the statistical variable are equal to the same constant c , then their variance is equal to zero, i.e.*

$$\sigma_n^2 = \sum_{i=1}^n \frac{(c - c)^2}{n} = 0.$$

- b) *If the same constant c is added to all values of the variable x_i , the value of the variance remains unchanged, i.e.*

$$\sum_{i=1}^n \frac{[(x_i + c) - (\bar{x} + c)]^2}{n} = \sigma_n^2.$$

- c) *If all values of the variable X_i are multiplied by a constant c , then the variance is multiplied by the square of this constant, i.e.*

$$\sum_{i=1}^n \frac{(c \cdot x_i - c \cdot \bar{x})^2}{n} = c^2 \cdot \sigma_n^2.$$

Table 12.5: Data set for Example 12.7

23.1	16.7	22.9	23.6	22.9	19.5	46.5	43.4	39.3	31.5	28.0
31.5	31.4	35.1	42.2	36.1	36.1	28.0	31.4	25.6	56	55.9
65.4	35.4	45.4	46.2	38.8	40.9	40.9	38.4	40.9	38.8	38.4
59.2	53.3	38.4	32.2	25.3	32.2	28.0	23.6	40.0	17.0	23.5
23.9	36.3	18.1	17.2	23.4	35.1	35.0	23.6	30.4	28.9	23.6
23.6	40.7	32.2	13.2	28.0	41.1	40.4	39.6	39.3	29.5	49.0
33.7	32.6	31.3	28.0	31.3	33.2	32.9	31.2	35.3	32.3	32.2
38.9	38.2	46.9	27.7	23.4						

The values of a statistical variable are most often realizations of a random variable. For example, the number of telephone calls at an exchange per 1 minute (see Example 12.2) is a random variable that follows a Poisson distribution $X \sim \text{Po}(\lambda)$, with an unknown parameter λ . One of the tasks of statistical analysis (specifically, estimation theory) is to estimate the true value of the parameter λ . If data are available, the arithmetic mean of the statistical population is used to estimate the unknown expected value $\mathbb{E}X = \lambda$ of this random variable. In a similar manner, the variance of the statistical population is used to approximate the variance of the random variable $\text{var } X = \lambda$. However, the variance given in formulas (12.4) and (12.12) actually underestimates the variance of the underlying random variable; therefore, the sample variance (12.5) is used for estimation, or its computational formula

$$s_n^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k^2) - \frac{n}{n-1} \bar{x}^2 = \frac{1}{n-1} \sum_{i=1}^r f_i x_i^2 - \frac{n}{n-1} \bar{x}^2.$$

In Excel, this can also be computed using the functions `=VAR.S`, and `=STDEV.S` for the sample standard deviation. The variance defined in this way no longer underestimates the theoretical value of the variance of the underlying random variable (for details, see Theorem 13.1 on page 275).

Example 12.7.  Compute the measures of variability – the range, upper and lower quartile, interquartile range, interquartile deviation, and coefficient of variation – for the data (see Table 12.5) representing fuel consumption in mpg (miles per gallon) of passenger cars of various international brands.

Solution. The variance σ_n^2 is the arithmetic mean of the squared deviations of the values from the arithmetic mean:

$$\begin{aligned}\sigma_{82}^2 &= \frac{1}{82} \cdot \sum_{k=1}^{82} (x_k - 33.78)^2 = \\ &= \frac{1}{82} \cdot ((23.1 - 33.78)^2 + \cdots + (23.4 - 33.78)^2) \approx 98.87.\end{aligned}$$

$$= \text{VAR.P(D4:D85)}$$

Standard deviation $\sigma_{82} = \sqrt{s_{82}^2} \approx \sqrt{98.87} \approx 9.94$.

$$= \text{STDEV.P(D4:D85)}$$

The sample variance s^2 differs only in the normalizing denominator:

$$s_{82}^2 = \frac{82}{81} \sigma_{82}^2 \approx 100.09.$$

$$= \text{VAR.S(D4:D85)}$$

Sample standard deviation $s_{82} = \sqrt{s_{82}^2} = \sqrt{100.09} \approx 10$.

$$= \text{STDEV.S(D4:D85)}$$

The mean absolute deviation

$$\begin{aligned}\bar{d} &= \frac{1}{82} \sum_{k=1}^{81} |x_k - 33.78| = \\ &= \frac{1}{82} \cdot (|23.1 - 33.78| + \cdots + |23.4 - 33.78|) \approx 7.73.\end{aligned}$$

$$= \text{AVEDEV(D4:D85)}$$

The coefficient of variation

$$v = \frac{\sigma}{\bar{x}} \approx \frac{10}{33.78} \approx 0.296.$$

The quartiles are $x_{25} \approx 27.175$, $x_{75} \approx 39.375$.

$$x_{25} = \text{QUARTILE.EXC(D4:D85, 1)}$$

$$x_{75} = \text{QUARTILE.EXC(D4:D85, 3)}$$

The interquartile range is $R_Q = 12.2$, and the interquartile deviation is $Q = \frac{12.2}{2} = 6.1$. △

12.5 General sample moments*

In statistics, so-called moments are used to describe characteristics of a probability distribution or a data distribution. In mathematics, a moment is understood as a characteristic (property) of a probability distribution or a data

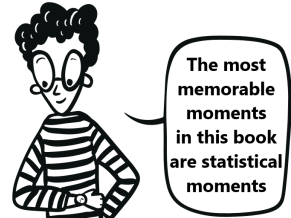
distribution that can provide information about the shape and nature of the distribution. These moments may differ in their ability to describe the distribution of data, and are therefore useful in data analysis and in comparing different data sets.

The so-called *general moment of order k* is defined by the formula

$$m'_k = \frac{1}{n} \sum_{i=1}^n x_i^k, \quad k = 0, 1, \dots \quad (12.13)$$

The so-called *central moment of order k* is defined by the formula

$$m_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k, \quad k = 0, 1, \dots \quad (12.14)$$



Moments include, for example, the mean and the variance:

- the mean is the general moment of the first order ($m'_1 = \bar{x}$),
- the variance is the central moment of the second order ($m_2 = \sigma_n^2$).

Moments of the third and fourth order are also the basis for the characteristics of skewness and kurtosis, which will be introduced in the following section.

12.6 Measures of Symmetry and Concentration

In statistical data analysis, characteristics of location and variability are often not sufficient. Additional information about a data set is provided by measures of symmetry and concentration (“peakedness” or density) of the data. At the level of random variables, these were introduced in Section 6.4. We now introduce the coefficients of skewness and kurtosis as characteristics of a statistical data set.



Definition 12.9 (Coefficient of skewness, asymmetry)

The coefficient of skewness is given by the relation

$$a_3 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\sigma_n^3} = \frac{m_3}{m_2^{3/2}},$$

where m_2 is the second central moment (variance), m_3 is the third central moment, and s_n is the standard deviation of the statistical population.

Skewness can be defined for a data set in several different ways, all of which estimate the skewness of the underlying random variable; however, none is generally preferred over the others. For example, the function `=SKEW` in MS Excel computes skewness according to formula (21.4) given on page 461.

If $a_3 = 0$, then the values below and above the mean are represented equally in the data set, which leads to a symmetric frequency distribution. If $a_3 > 0$, the representation of smaller values is greater than that of larger values; such a frequency distribution is referred to as left-skewed (negatively skewed). If $a_3 < 0$, the frequency distribution is right-skewed (positively skewed).

Definition 12.10 (Coefficient of kurtosis, concentration)

The coefficient of kurtosis is defined by

$$a_4 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\sigma_n^4} - 3 = \frac{m_4}{m_2^2} - 3,$$

where m_2 is the second central moment (variance), m_4 is the fourth central moment, and s_n is the standard deviation of the statistical population.

The coefficient of kurtosis can also be defined in several ways, all of which estimate the kurtosis (more precisely, the “excess” kurtosis) of the underlying random variable. In Excel, the function `=KURT` uses the adjusted formula (21.3).

If $a_4 = 0$, the distribution exhibits normal kurtosis corresponding to the concentration of the normal distribution; such a distribution is termed mesokurtic. If $a_4 > 0$, the values of the statistical population are more concentrated around the mean; a distribution with greater peakedness than the normal distribution is called leptokurtic. Conversely, if $a_4 < 0$, a larger portion of the data lies further from the mean and the distribution becomes flatter. A distribution with less peakedness than the normal distribution is referred to as platykurtic.

Example 12.8.  Determine the coefficients of skewness and kurtosis for the data from Example 12.7.

Solution. $\bar{x} = 33.78$, $\sigma_n \approx 9.94$,

$$a_3 = \frac{\sum_{i=1}^{82} (x_i - 33.78)^3}{82 \cdot 9.94^3} \approx 0.595.$$

= `SKEW.P(A2:A45)`

Since $a_3 > 0$, we expect the frequency distribution to be left-skewed. The modified measure of skewness computed using the function `= SKEW.P` has the value 0.606.

$$a_4 = \frac{\sum_{i=1}^{82} (x_i - 33.78)^4}{82 \cdot 10^4} \approx 0.749.$$

= `KURT(A2:A45)`

The values of the statistical population are more concentrated around the mean than would correspond to the standard normal distribution. $\triangle$

12.7 Exercises

12.1 Construct the graph of the distribution of absolute frequencies, the histogram of relative frequencies, and the ogive for the data from Example 12.3.

12.2 Compute the measures of location and variability, including the quartiles, the interquartile range, and the semi-interquartile range, for the following data set: 2, 7, 1, 9, 2, 4, 4, 5, 2, 8, 2, 6.

12.3 Data on groups of cars classified according to age and their frequencies are shown in Table 12.6. Describe the data set using basic descriptive characteristics, construct a histogram, and compare it with the normal distribution.

Table 12.6: Table of the frequency distribution of cars by age

Age of car	1	2	3	4	5	6	7	8	9	10
Count	10	19	27	10	9	7	8	5	3	2

12.4 During final inspection, ten packages of rice with a declared weight of 480 g and ten packages of lentils, each with a weight of 500 g, were randomly selected; each package was numbered and weighed individually. The weighing results are given in Table 12.7. Compute the variance of the weights of the rice and lentil packages and determine the standard deviations of the package weights. Compare the precision of the two machines used to fill the rice and lentil packages.

Table 12.7: Weights of individual rice packages

Number	1	2	3	4	5	6	7	8	9	10
m [g] Rice	485	482	487	479	476	480	482	488	477	488
m [g] Lentils	485	495	500	510	520	512	489	491	485	488

12.5 Compute $\bar{x}$, R , s_n^2 , s_n , $\bar{d}$, and v for the three data sets in Table 12.8, and compare the variability of the data in all three sets.

Table 12.8: Three data sets

S_1	10	10	10	10	10	10	10	10	10
S_2	6	7	8	9	10	11	12	13	14
S_3	2	4	6	8	10	12	14	16	18

Solutions to the Exercises

12.1

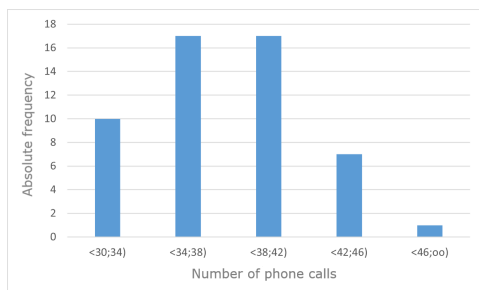


Figure 12.5: Distribution of absolute frequencies

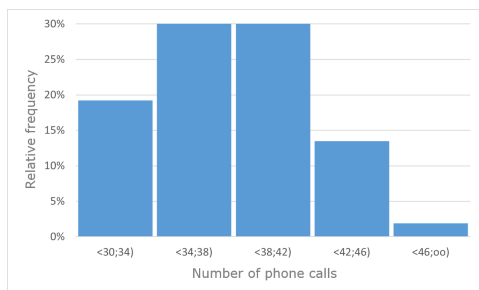


Figure 12.6: Histogram

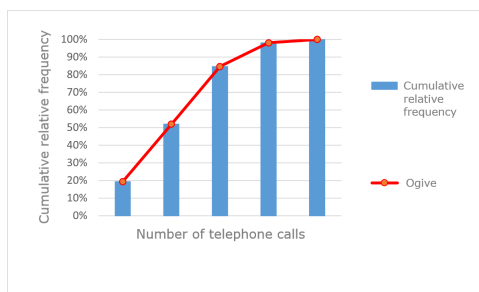


Figure 12.7: Distribution of cumulative relative frequencies and the ogive

12.2 $\bar{x} = 4.\bar{3}$, $\bar{x}_g \approx 3.54$, $\bar{x}_h \approx 2.83$, $\tilde{x} = 4$, $\hat{x} = 2$, $x_{0.25} = 2$, $x_{0.75} = 6.75$, $R_Q = 4.75$, $Q = 2.375$, $R = 8$, $\sigma_{12}^2 = 6.\bar{5}$, $\sigma_{12} \approx 2.56$, $s_{12}^2 = 7.\bar{15}$, $s_{12} \approx 2.67$, $\bar{d} = 2.\bar{2}$.

12.3 $\bar{x} \approx 3.99$, $\tilde{x} = 3$, $\hat{x} = 3$, $\sigma_{10} \approx 2.31$, coefficient of variation $v \approx 57.93\%$, indicating high data dispersion, skewness coefficient $a_3 \approx 0.8135$, kurtosis coefficient $a_4 \approx -0.200235$.

12.4 Rice packages: $\bar{x} = 482.4$ g, $s_{10}^2 = 19.8\bar{2}$, $s_{10} \approx 4.45$, $v \approx 0.0092$, i.e., approximately 0.9%, which indicates a high degree of filling precision of the packaging machine. Lentil packages: $\bar{x} = 497.5$ g, $s_{10}^2 = 155.8\bar{3}$, $s_{10} \approx 12.48$, $v \approx 0.025$, i.e., approximately 2.5%, indicating that this machine also operates with high precision, although its precision is lower than that of the machine used for packaging rice.

12.5

Table 12.9: Descriptive statistics for Exercise 12.5

	$\bar{x}$	R	s_n^2	s_n	$\bar{d}$	v
S_1	10	0	0	0	0	0
S_2	10	8	7.5	2.74	$2.\bar{2}$	0.27
S_3	10	16	30	5.48	$4.\bar{4}$	0.55

Chapter 13

Building Blocks of Statistics

13.1 Population and Sample

In many cases, statistical data analysis cannot be based on the entire population Z (for example, when the population is inaccessible or very large), and we must restrict our attention to some subset of Z . The statistical results obtained from the analysis of such a subset may then be used either

- 1) for generalization to the entire population Z (this generalization is referred to as statistical induction), or
- 2) for testing hypotheses about the entire population (i.e., statistical inference).

This means that we examine only a certain subset of the population under study, which we call a sample. Statistical induction and statistical inference do not provide conclusions about the population with complete certainty, but only with a predetermined probability. The foundation of these methods is the theory of random sampling, which will be discussed in this chapter.

Random samples must be understood from two different perspectives: either as a collection (sequence) of random variables denoted by uppercase letters, or as a collection of realizations of these random variables, denoted by lowercase letters. In the latter case, we are dealing with specific numerical values – we often refer to these as observations, a realization of a random sample, an instance of a random sample, or simply data.

Random samples may be classified, for example, according to the method by which they are obtained:

- 1) *Simple random sampling with replacement* is a sampling procedure in which each selected element of the population is returned to the

population after selection, and the next element is again selected from the entire population.

- 2) *Simple random sampling without replacement* is a sampling procedure in which a selected element is not returned to the population.
- 3) *Stratified sampling* consists in partitioning the population into homogeneous, mutually disjoint subpopulations (strata) and then performing random sampling within each of them. Sufficient information about the population must be available to allow an appropriate choice of the individual strata. Stratified sampling is, for example, employed by agencies conducting pre-election surveys: the sample is required to contain proportional numbers of young people, middle-aged people, and retirees, or people living in urban and rural areas, etc., in proportions corresponding to those in the target population.
- 4) *Systematic (mechanical) random sampling* is a sampling procedure in which the selection rules are fixed in advance, and repeated application of the sampling process to the same population, with the same choice of the initial element, yields an identical sample. For example, the population may first be ordered according to a specified criterion (such as alphabetical order), and then every k -th element is selected. In this case, randomness enters only through the choice of the initial element among the first k elements.

Another classification of random samples may be based on their size:

- 1) a *small sample* – a sample of size $n < 30$,
- 2) a *large sample* – a sample of size $n \geq 30$.

Generally, a random sample is classified as a large sample when the validity of the Central Limit Theorem is already apparent. Thus, the numerical boundary between a small and a large sample is only approximate and depends on the properties of the environment from which the sample is drawn. For example, in economic studies, observational samples are typical (the data are merely observed, and the environments in which they are observed are not identical for all observations), and the usual boundary between a small and a large sample is shifted toward approximately 120 observations.

In the following text, we shall consider only simple random sampling with replacement. From the standpoint of probability theory, a random sample is defined as follows.

Definition 13.1 (Random sample)

Let Z be a population whose variate is represented by the random variable X . A random sample from the distribution of the random variable X is defined as a sequence of n independent realizations of an experiment, represented by

random variables $X_1, X_2, \dots, X_n$, which have the same distribution as the random variable X and are jointly independent.

Thus, a random sample is a sample in which every element of the population has the same and independent probability of being included in the sample.

Definition 13.2 (Theoretical and empirical characteristics)

A characteristic of the population $\mathbf{Z}$ represented by the random variable X will be referred to as a theoretical characteristic. A characteristic derived from a random sample will be referred to as an empirical (sample) characteristic.

Theoretical characteristics typically include the expectation, median, mode, variance, standard deviation, and, in general, any other characteristic of the random variable X that, by itself, describes the selected attribute of the population. By their very nature, the theoretical characteristics of a population are fixed (non-random) quantities, whereas empirical characteristics are random variables, since their values vary from one realization of a random sample to another. Empirical characteristics are also referred to simply as *statistics*. If the type of distribution of the random variable X , representing the variate of the population $\mathbf{Z}$, is known, then, under certain assumptions, empirical characteristics may be used to determine the corresponding theoretical characteristics.

Example 13.1. The statistical population consists of all men in the Czech Republic. The variate is their age. The random variable X represents the age of a randomly selected man from the Czech Republic. To determine the characteristics of the random variable X , we may either examine the entire population of men in the Czech Republic (for example, using population registers), which may be quite costly, or draw a random sample of smaller size n . The age of each selected man is a realization of the random variable X_i , where $i = 1, \dots, n$ indexes the sampled individuals. The empirical characteristics obtained from the sample $X_1, \dots, X_n$ may then be used to estimate the corresponding theoretical characteristics of the random variable X .

Example 13.2.  Let X be a random variable denoting the number of telephone calls received within 1 minute at the telephone exchange mentioned in Example 12.2. The data recorded in Example 12.2 represent a realization of a random sample from the distribution of X . Determine the empirical characteristics of the expectation and variance, and compare them with the theoretical characteristics of the random variable X .

Table 13.1: Comparison of theoretical probabilities with relative frequencies

k	0	1	2	3	4	5	6	7
$\mathbb{P}(X = k)$	0.135	0.271	0.271	0.180	0.090	0.036	0.012	0.003
f_k	0.133	0.283	0.266	0.166	0.100	0.033	0	0.016

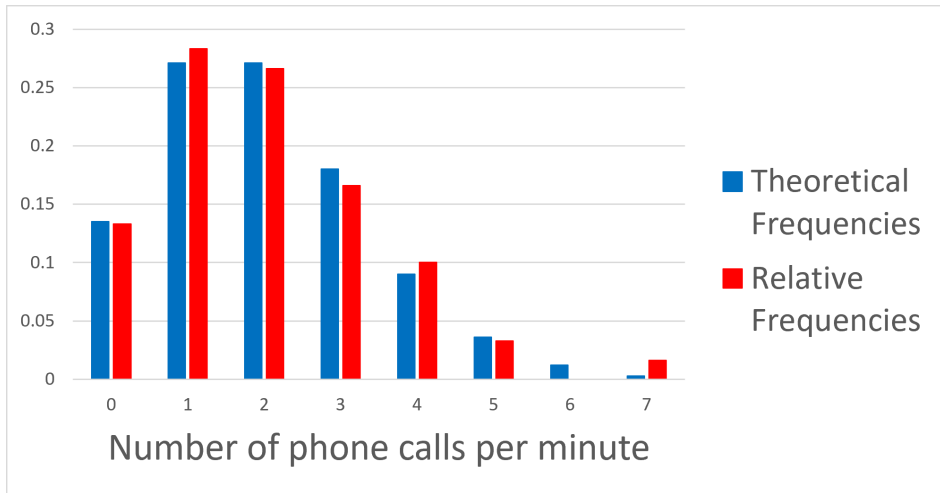


Figure 13.1: Comparison of the relative and theoretical frequencies in Example 12.2

Solution. Mathematical statistics provides answers to the questions raised in Example 13.2, as well as to many others. It will be studied in detail in the following chapters. For the moment, however, we shall only outline some of the standard techniques employed in mathematical statistics.

The empirical mean is typically determined by the arithmetic mean (12.1), and the empirical variance by formula (12.5). The reader may verify by direct calculation that $\bar{x} = 2$ and $s_{60}^2 = 2.136$. In statistics, these values are used as estimates of the unknown theoretical characteristics, namely the expectation and variance of the random variable X .

Assume further that $X \sim \text{Po}(\lambda)$. From probability theory, we know that in this case $\mathbb{E}X = \lambda = \text{var } X$ (see Section 7.8). Let us ask, for example, whether the empirical data are consistent with the hypothesis $X \sim \text{Po}(2)$ concerning the theoretical parameter of the random variable X . A preliminary assessment can be made by comparing the relative frequencies (already given in Table 12.1) with the theoretical probabilities $\mathbb{P}(X = k)$ determined by the distribution $\text{Po}(2)$. The result of this exploratory analysis is presented in Table 13.1 and in the bar chart shown in Figure 13.1. We observe that the relative frequencies and the theoretical probabilities are not sub-

stantially different. However, to verify formally whether the empirical data are consistent with the hypothesis $X \sim \text{Po}(2)$, more rigorous methods of mathematical statistics must be employed, specifically a statistical hypothesis test. $\triangle$

13.2 Distribution of Sample Statistics

Definition 13.3 (Sample mean, variance, and standard deviation) Let $X_1, \dots, X_n$ be a random sample from the distribution of the random variable X . Define the random variables

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, \quad S = \sqrt{S^2}.$$

$\bar{X}$ is called the sample mean, S^2 the sample variance, and S the sample standard deviation.

Theorem 13.1. Let $X_1, \dots, X_n$ be a random sample drawn from a distribution with expectation μ and finite variance σ^2 . Then

$$\mathbb{E}\bar{X} = \mu, \quad \text{var } \bar{X} = \frac{\sigma^2}{n}, \quad \mathbb{E}S^2 = \sigma^2.$$

The random variables $\bar{X}$, S^2 , and S represent empirical characteristics from the perspective of probability theory, serving as estimators of the theoretical characteristics of the random variable X , namely its expectation, variance, and standard deviation. The values of the arithmetic mean $\bar{x}$, the sample variance s_n^2 , and the sample standard deviation s_n computed from a data set are, in this context, realizations of the random variables $\bar{X}$, S^2 , and S . They are also referred to as estimates, or more precisely, estimated values. Theorem 13.1 will subsequently provide, in Chapter 14, the theoretical foundation for assessing the quality of these estimators.

Theorem 13.2 (Strong Law of Large Numbers). Let $X_1, \dots, X_n$ be a random sample drawn from a distribution with expectation μ and finite variance σ^2 . Then

$$\bar{X} \rightarrow \mu \text{ almost surely.}$$

Almost sure convergence means that there exists only a set $A \subset \Omega$ of probability zero ($\mathbb{P}(A) = 0$) on which the given expression fails to converge.

Theorem 13.3 (On a random sample from a normal distribution). Let $X_1, \dots, X_n$ be a random sample drawn from the distribution $\mathcal{N}(\mu, \sigma^2)$, where $\sigma^2 > 0$. Then the following statements hold:

Table 13.2: Notation for quantiles of selected random variables

Distribution	Quantile notation	Quantile value ³	Excel functions
standard normal $N(0, 1)$	$u(\alpha)$	$\Phi^{-1}(\alpha)$	= <code>NORM.S.INV</code>
Pearson χ_k^2	$\chi_k^2(\alpha)$	$F_{Q_k}^{-1}(\alpha)$	= <code>CHISQ.INV</code>
Student t_k	$t_k(\alpha)$	$F_T^{-1}(\alpha)$	= <code>T.INV</code>
Fisher–Snedecor F_{ν_1, ν_2}	$F_{\nu_1, \nu_2}(\alpha)$	$F_F^{-1}(\alpha)$	= <code>F.INV</code>

- a) $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.
b) If $n \geq 2$, then $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$.
c) If $n \geq 2$, then $\bar{X}$ and S^2 are independent.
d) If $n \geq 2$, then $\frac{\bar{X}-\mu}{S}\sqrt{n} \sim t_{n-1}$.

The proofs of the above theorems can be found, for example, in [2].

13.3 Critical Values and Quantiles

The term *critical value of a distribution* was traditionally used in older textbooks to denote a value that is exceeded by a random variable with a prescribed probability α . Thus, the critical value of a given distribution is its $(1 - \alpha)$ -quantile. For example, the 5% critical value of the standard normal distribution is approximately 1.64, which is the 95% quantile of this random variable¹.

Statistical tables were traditionally used to determine critical values of distributions. These tables employed a standard notation that has since become obsolete² and conflicts with the notation for quantiles introduced in Chapter 8. Therefore, we shall henceforth reserve the notion of a critical value exclusively for a property of a statistical test (to be introduced in Chapter 15) and avoid the term *critical value of a distribution*. Instead, Table 13.2 summarizes the notation for quantiles that we have introduced for selected random variables and that will be used in Chapter 15 on statistical hypothesis testing. Let us also recall two important properties of quantiles that will likewise prove useful in hypothesis testing.

¹Thus, $\mathbb{P}(X > 1.64) = 1 - \Phi(1.64) \approx 1 - 0.95 = 0.05$.

²For example, the critical value corresponding to the 97.5% quantile of Student's t -distribution was typically denoted in such tables by $t_k(0.05)$.

³ F^{-1} denotes the quantile function of the corresponding distribution, whereas Φ^{-1} denotes the quantile function of the standard normal distribution.

- 1) For the quantiles of symmetric distributions, namely the standard normal distribution and Student's t -distribution, the following relation holds.

$$u(1 - \alpha) = -u(\alpha), \text{ and similarly } t_k(1 - \alpha) = -t_k(\alpha), \quad (13.1)$$

thus, the quantiles corresponding to the two tails of the distribution differ only in sign and not in magnitude.

- 2) The quantile of the Fisher–Snedecor distribution with interchanged degrees of freedom is the reciprocal of the quantile of the distribution with the original ordering of the degrees of freedom taken from the opposite tail:

$$F_{\nu_2, \nu_1}(\alpha) = \frac{1}{F_{\nu_1, \nu_2}(1 - \alpha)}. \quad (13.2)$$

Example 13.3.  A company manufacturing plastic bottles has determined that bottle weights are normally distributed with mean $\mu = 50$ g and standard deviation $\sigma = 2$ g. An automatic quality-control system classifies bottles as defective if their weight falls within the lightest 5% or the heaviest 5% of all produced bottles. Determine the corresponding lower and upper specification limits for the weight of a bottle.

Solution. The random variable X , representing the weight of a manufactured bottle, follows a normal distribution with parameters $\mu = 50$ and $\sigma^2 = 4$. The threshold values x_1 and x_2 are determined by the inequalities

$$\mathbb{P}(X \leq x_1) \leq 0.05, \quad \text{and similarly} \quad \mathbb{P}(X \geq x_2) \geq 0.05.$$

The values x_1 and x_2 are therefore the 5% and 95% quantiles, respectively, of the prescribed distribution, which can be readily computed using the functions

$$\begin{aligned} x_1 &= \text{= NORM.INV(0.05, 50, 2)} \approx 46.71, \\ x_2 &= \text{= NORM.INV(0.95, 50, 2)} \approx 53.29. \end{aligned}$$

A properly manufactured bottle must have a weight of at least 46.71 g and at most 53.29 g. $\triangle$

13.4 Exercises

13.1 Determine the values of the following quantiles for $\alpha = 0.05$, $\nu = 8$, and $\mu = 11$:

- a) $u(\alpha)$, $u(1 - \alpha)$, $u(\frac{\alpha}{2})$, $u(1 - \frac{\alpha}{2})$,
- b) $t_\nu(\alpha)$, $t_\nu(1 - \alpha)$, $t_\nu(\frac{\alpha}{2})$, $t_\nu(1 - \frac{\alpha}{2})$,
- c) $\chi_\nu^2(\alpha)$, $\chi_\nu^2(1 - \alpha)$, $\chi_\nu^2(\frac{\alpha}{2})$, $\chi_\nu^2(1 - \frac{\alpha}{2})$,
- d) $F_{\nu,\mu}(\alpha)$, $F_{\mu,\nu}(\alpha)$, $F_{\nu,\mu}(1 - \alpha)$, $F_{\nu,\mu}(\frac{\alpha}{2})$, $F_{\nu,\mu}(1 - \frac{\alpha}{2})$.

13.2 For selected continuous distributions, sketch the graph of the probability density function and the graph of the distribution function for chosen parameter values, and indicate the values of the 5% and 95% quantiles.

13.3 Assume that the daily portfolio loss X during a financial crisis follows a normal distribution with mean 0.015 and variance 0.010. Compute the 1% and 5% Value at Risk (VaR) of this portfolio, i.e., the loss level that will be exceeded with probability 1% and 5%, respectively.

13.4 Compute the z -score corresponding to the value $x = 120$ for a normal distribution with mean 110 and variance 100.

Solutions to the Exercises

13.1 Values of the quantiles rounded to three decimal places:

- a) -1.645 , 1.645 , -1.960 , 1.960 ,
- b) -1.860 , 1.860 , -2.306 , 2.306 ,
- c) 2.733 , 15.507 , 2.180 , 17.535 ,
- d) 0.302 , 0.339 , 2.948 , 0.236 , 3.664 .

13.2 Figure 13.2 shows the 5% and 95% quantiles of the standard normal distribution. $X \sim N(0, 1)$.

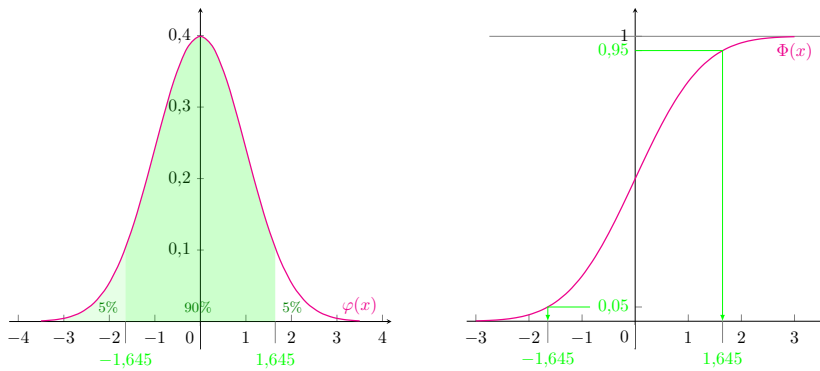


Figure 13.2: 5% and 95% quantiles of the standard normal distribution

13.3 0.249, 0.179.

13.4 $z = 1$

Chapter 14

Parameter Estimation

The characteristics of a population are described by the properties of the probability distribution of the random variable representing the variate of the population. For example, a univariate distribution may be characterized by a measure of location, typically the expectation, and by a measure of variability, such as the standard deviation. In a bivariate or multivariate distribution, the strength of dependence may be characterized, for instance, by the correlation coefficient. The probability distribution of the variate is usually specified by one or more parameters. For example, the normal distribution $N(\mu, \sigma^2)$ is completely determined by its parameters μ and σ^2 , which represent the expectation and the variance. In some cases, the parameters of a distribution coincide with its characteristics, as in the case of the normal distribution mentioned above.

The goal of *statistical inference* is to estimate the characteristics and parameters of a population by means of estimating the parameters of the distribution of the random variable representing the variate of the population. Such estimates are constructed from a sample, and for each characteristic or distribution parameter one can define a corresponding sample counterpart (see Table 14.1).

We distinguish two types of estimators:

- *Point estimators* yield a single numerical estimate of a population characteristic or parameter. For example, the expectation μ of the underlying random variable is estimated by the sample mean $\bar{x}$.
- *Interval estimators*, or *confidence intervals*, estimate a population or parameter by means of an interval. The estimate is therefore an interval, or, more precisely, its endpoints.

Table 14.1: Some population parameters and their sample counterparts

Parameter	Population parameter	Sample counterparts
Expectation	μ	$\bar{x}$
Variance	σ^2	s_n^2
Standard deviation	σ	s_n
Median	$\tilde{\mu}$	$\tilde{x}$
Mode	$\hat{\mu}$	$\hat{x}$
Correlation coefficient	ρ	r

Formally, an estimator T of a parameter θ may be defined as an (abstract) function of a random sample. Estimators should be viewed from two different perspectives, similarly to the treatment of random samples on page 271: either the estimator is a numerical value computed from observed data, or the estimator denotes a random variable derived from a random sample.

Definition 14.1 (Estimator of a parameter)

Let $X_1, \dots, X_n$ be a random sample drawn from a distribution characterized by a parameter θ . An estimator T of the parameter θ is any known function of the random sample

$$T = h(X_1, \dots, X_n).$$

A realized value of the random variable T is called an estimated value.

The term *estimate* is often used ambiguously: it may refer either to the estimator itself or to its realized numerical value. In this text, an estimator is understood as a random variable, whereas an estimated value is a computed numerical quantity.

14.1 Properties of Estimators

According to Definition 14.1, an estimator of a parameter is any function of a random sample. Consequently, even the sample maximum could, in principle, be used as an estimator of the expectation of a random variable. Clearly, however, the sample maximum does not estimate the expectation particularly well – it is not an appropriately chosen estimator of the mean. The quality of an estimator may be assessed from several perspectives, which will be discussed in this section.

Definition 14.2 (Unbiasedness of an estimator)

Let T be an estimator of a parameter θ . If

$$\mathbb{E}T = \theta,$$

then T is called an unbiased estimator of the parameter θ . An estimator that is not unbiased is called a biased estimator, and the difference $\mathbb{E}T - \theta$ is called the bias of the estimator.

Unbiased estimators fluctuate around the theoretical value from one sample to another in such a way that they yield the theoretical value “on average” (in expectation). Thus, unbiasedness is not a property of a particular estimated value but rather a property of the distribution of the random variable T . More specifically, it expresses the fact that the function h defining the estimator T has been chosen appropriately in a certain sense.

Theorem 13.1 shows that the sample mean $\bar{X}$ is an unbiased estimator of the expectation μ , since $\mathbb{E}\bar{X} = \mu$. Similarly, the sample variance S^2 is an unbiased estimator of the population variance σ^2 , because $\mathbb{E}S^2 = \sigma^2$.

Unbiasedness is a very strong property: by no means are all routinely used estimators unbiased. For example, the sample standard deviation S is not an unbiased estimator of the population standard deviation σ , even though the corresponding property holds for the variance. Nevertheless, S is the standard estimator of the standard deviation. The minimum requirement for the usefulness of an estimator is specified by the following definition.

Definition 14.3 (Consistency of an estimator)

Let T_n be an estimator of a parameter θ based on a random sample $X_1, \dots, X_n$ of size n . The estimator T_n is called consistent if, for every $\varepsilon > 0$,

$$\mathbb{P}\{|T_n - \theta| > \varepsilon\} \xrightarrow{n \rightarrow \infty} 0. \quad (14.1)$$

Property (14.1) in Definition 14.3 states that the estimator T_n converges to the true parameter value θ in probability, in the sense of Definition 11.1, part b). From an interpretational point of view, the definition states that, as the number of observations increases, the distribution of the estimator T_n becomes increasingly concentrated around θ : for larger sample sizes, it becomes progressively less likely that the estimated value will be far (by more than ε) from the true parameter value θ .

Every useful estimator should be consistent. Inconsistent estimators do not contribute to estimating the value of θ , even when an unlimited amount of data is available. The sample maximum mentioned above, for example, is

generally not a consistent estimator of the expectation. In many cases, especially for distributions with unbounded support, it does not approach the expected value as the sample size increases. The following useful criterion, stated without proof, facilitates the verification of estimator consistency.

Theorem 14.1 (Consistency of estimators). *If T_n is an unbiased estimator of the parameter θ and*

$$\text{var } T_n \xrightarrow{n \rightarrow \infty} 0,$$

then T_n is a consistent estimator.

Corollary 14.2. *The sample mean $\bar{X}$ is a consistent estimator of the expectation μ .*

Proof. This follows from Theorem 13.1: $\mathbb{E}\bar{X} = \mu$ and $\text{var } \bar{X} = \frac{1}{n}\sigma^2 \rightarrow 0$. $\square$

Definition 12.8 introduced, through formula (12.4), the descriptive statistic known as the (population) variance. It can be shown that the random variable $\Sigma^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ is not an unbiased estimator of the parameter σ^2 , although it is a consistent estimator. Likewise, the sample and population standard deviations S and Σ , respectively, are consistent estimators of the theoretical standard deviation σ .

We have already noted that good estimators need not be unbiased (although they should always be consistent). On the other hand, unbiasedness alone does not necessarily imply a high-quality estimator. For example, the theoretical expectation may be estimated by the first observation in the random sample (i.e., by taking $T_n = X_1$). The estimator T_n is unbiased ($\mathbb{E}T_n = \mu$ by the definition of a random sample), but it is evident that it does not make sufficient use of the information provided by the sample, since all observations except the first are discarded.¹ The estimation error is most naturally characterized by the variance of the estimator.

Definition 14.4 (Efficiency of an estimator)

The efficiency of an estimator T is characterized by the variance $\text{var } T$. Within a specified class of estimators, typically among unbiased estimators, an estimator is said to be efficient if it has the smallest variance.

Continuing with the estimator $T_n = X_1$, its variance is $\text{var } T_n = \text{var } X_1 = \sigma^2$, whereas the estimator defined by the arithmetic mean has variance $\text{var } \bar{X} = \frac{1}{n}\sigma^2$. For example, with ten observations, the variance of the arithmetic mean is ten times smaller than the variance of X_1 , and as the sample size increases, the variance of the arithmetic mean decreases further.

¹It can be shown that X_1 is not a consistent estimator, but that is not the point of the present discussion.

14.2 Point Estimators of the Mean and Variance

Let us recall that Theorem 13.1 states that $\bar{X}$ is an unbiased estimator of the expectation μ , and that S^2 is an unbiased estimator of the theoretical variance σ^2 . Together with their consistency, this gives us reason to believe that $\bar{X}$ and S^2 are good point estimators of the parameters μ and σ^2 , and indeed they are used as such in practice. Similarly, the sample standard deviation S is taken as a point estimator of the theoretical standard deviation σ , although it is not² an unbiased estimator of σ . Nevertheless, with a sufficiently large sample size, the property of consistency prevails here as well, yielding a reasonably accurate estimate of the theoretical standard deviation.

14.3 Interval Estimation

A point estimate is a single numerical value obtained by evaluating a point estimator on the observed sample. If the estimator is consistent, we know that, for a sufficiently large sample size, it lies close to the true theoretical value. If the variance of the estimator is known, we are also able to say, at least approximately, how close it is (using the interpretation of the standard deviation of the estimator). To obtain a more precise assessment of the accuracy (or inaccuracy) of an estimate, we use an *interval estimator*, that is, an interval derived from a random sample³ which covers the true value of the parameter with a prescribed probability.

Definition 14.5 (Confidence interval, interval estimator)

Let Θ_L, Θ_U be statistics derived from a random sample drawn from a distribution with parameter θ such that, for a given $\alpha \in (0, 1)$,

$$\mathbb{P}(\Theta_L \leq \theta \leq \Theta_U) = 1 - \alpha.$$

Then the interval $[\Theta_L, \Theta_U]$ is called a confidence interval for the parameter θ with confidence level $1 - \alpha$.

²In fact, an unbiased estimator of the standard deviation does not exist. Alternative estimators of the theoretical standard deviation designed to reduce estimation bias can be found in the literature, but they involve a number of restrictive assumptions on the population and the properties of the random sample and have never become routinely used.

³Its endpoints are therefore random variables as well.

The random interval $[\Theta_L, \Theta_U]$ is called an *interval estimator* or a *confidence interval procedure*. Its realized value $[\hat{\theta}_L, \hat{\theta}_U]$ is called a $100(1 - \alpha)\%$ *confidence interval estimate* for the parameter θ .

Most commonly, α is chosen to be 0.05 or 0.01; in exceptional cases, smaller values are used when a higher confidence level is required. The price paid for such coverage is, of course, the width of the resulting interval. The terms $100(1 - \alpha)\%$ confidence interval for the parameter θ and confidence interval for the parameter θ with $100(1 - \alpha)\%$ confidence are also commonly used.

Unlike point estimation, the construction of exact confidence intervals usually requires knowledge of the family of distributions of the underlying random variable. In the following sections, we present interval estimators for the expectation and the standard deviation of a normal distribution.

14.3.1 Confidence Interval for the Mean of a Normal Distribution

Suppose that we have obtained a random sample $X_1, \dots, X_n$ from a population having the normal distribution $N(\mu, \sigma^2)$. We wish to estimate the parameter μ and assume that the parameter $\sigma^2 > 0$ is unknown. Then, by Theorem 13.3,

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

and therefore, by the definition of the quantiles of Student's t -distribution,

$$\mathbb{P} \left(-t_{n-1} \left(1 - \frac{\alpha}{2} \right) \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{n-1} \left(1 - \frac{\alpha}{2} \right) \right) = 1 - \alpha.$$

By rearranging the inequalities and using property (13.1), we obtain a two-sided confidence interval for the expectation μ of a normal distribution with confidence level $1 - \alpha$

$$\left[\bar{X} - t_{n-1} \left(1 - \frac{\alpha}{2} \right) \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1} \left(1 - \frac{\alpha}{2} \right) \frac{S}{\sqrt{n}} \right]. \quad (14.2)$$

Similarly, one-sided confidence intervals for μ can be derived:

- lower one-sided confidence interval for μ

$$\left[\bar{X} - t_{n-1}(1 - \alpha) \frac{S}{\sqrt{n}}, \infty \right), \quad (14.3)$$

- upper one-sided confidence interval for μ

$$\left(-\infty, \bar{X} + t_{n-1}(1 - \alpha) \frac{S}{\sqrt{n}}\right]. \quad (14.4)$$

The quantity $\frac{S}{\sqrt{n}}$ is called the *standard error*, while the quantity $t_{n-1}(1 - \alpha) \frac{S}{\sqrt{n}}$ is called the *margin of error* and is often denoted by Δ . The two-sided confidence interval for the expectation then takes the simplified form $[\bar{X} - \Delta, \bar{X} + \Delta]$.

Example 14.1.  Twenty randomly selected employees of a company were surveyed regarding the amount of time (in minutes) they spend commuting to work each day. The observed commuting times are as follows: 45, 38, 50, 42, 47, 40, 43, 49, 41, 44, 45, 38, 50, 42, 47, 40, 43, 49, 41, 44. Assuming that commuting times are normally distributed and independent across employees, compute a two-sided 95% confidence interval for the mean commuting time.

Solution. The sample size is $n = 20$, the sample mean is $\bar{x} = 43.9$, the sample standard deviation is $s \approx 3.796$, the confidence level is $1 - \alpha = 0.95$, the number of degrees of freedom is $n - 1 = 19$, and the quantile of Student's t -distribution is $t_{19}(0.025) \approx -2.093$. The standard error is

$$\frac{s}{\sqrt{n}} = \frac{3.796}{\sqrt{20}} \approx 0.849,$$

the margin of error is $\delta \approx 1.777$, and the required confidence interval is $[42.12, 45.68]$. The average commuting time of the company's employees will lie in the obtained interval with a probability of 95%.

△

14.3.2 Determining the Sample Size Required for a Prescribed Width of a Confidence Interval for the Mean

In some situations, it is necessary to determine the sample size n in such a way that the resulting confidence interval for the mean has a prescribed width.

The length d of the interval is the difference between its upper and lower bounds (in that order). Therefore, from formula (14.2), we obtain the following expression for the width d of a two-sided confidence interval with confidence level $1 - \alpha$:

$$d = 2t_{n-1} \left(1 - \frac{\alpha}{2}\right) \frac{s}{\sqrt{n}},$$

which, after squaring both sides, can be transformed into the form

$$\begin{aligned}d^2 &= 4t_{n-1}^2 \left(1 - \frac{\alpha}{2}\right) \frac{s^2}{n}, \\n &= 4t_{n-1}^2 \left(1 - \frac{\alpha}{2}\right) \frac{s^2}{d^2}.\end{aligned}\tag{14.5}$$

A similar procedure can be applied to the confidence interval for the variance, which will be discussed later.

If the sample size exceeds 30, the quantile $t_{n-1}(1 - \frac{\alpha}{2})$ may be replaced by the corresponding quantile of the standard normal distribution $u(1 - \frac{\alpha}{2})$.

Example 14.2. Return to Example 14.1 on page 287. Determine how many observations must be collected so that, at the same confidence level, the resulting confidence interval has width equal to 1 minute.

Solution. Let the desired width d of the confidence interval with confidence level $1 - \alpha = 0.95$ be 1. In Example 14.1, twenty observations were taken, from which we estimated $s \approx 3.796$ and $\bar{x} = 43.9$. We use the quantile of Student's t -distribution is $t_{19}(0.975) \approx 2.093$.

Substituting into formula 14.5, we obtain

$$n \approx 4t_{19}(0.975)^2 \frac{s^2}{d^2} = 4 \cdot 2.093^2 \cdot \frac{3.796^2}{1} \approx 252.5.$$

It is therefore necessary to perform at least 253 observations in order for the width of the resulting confidence interval to be approximately 1.

In Section 8.8, we noted that Student's t -distribution converges to the normal distribution as the number of degrees of freedom n increases. If $n \geq 30$, the quantile $t_{n-1}(0.975)$ may be replaced by the corresponding quantile of the standard normal distribution, $u(0.975) \approx 1.96$; in that case, the calculation would yield a sample size of 222 observations. In our sample, however, only 20 observations are available, and replacing the quantile in the above manner would therefore lead to an underestimate of the required sample size. $\triangle$

14.3.3 Confidence Interval for the Variance σ^2 of a Normal Distribution

In constructing a confidence interval for the variance σ^2 of a normal distribution, we proceed (similarly to the case of the expectation) from the

known distribution of the random variable

$$(n-1) \frac{S^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Since

$$\mathbb{P} \left(\chi_{n-1}^2 \left(\frac{\alpha}{2} \right) \leq (n-1) \frac{S^2}{\sigma^2} \leq \chi_{n-1}^2 \left(1 - \frac{\alpha}{2} \right) \right) = 1 - \alpha,$$

we obtain by rearranging the inequality the two-sided confidence interval for the variance σ^2 of a normal distribution with confidence level $q = 1 - \alpha$

$$\left[\frac{S^2(n-1)}{\chi_{n-1}^2 \left(1 - \frac{\alpha}{2} \right)}, \frac{S^2(n-1)}{\chi_{n-1}^2 \left(\frac{\alpha}{2} \right)} \right]. \quad (14.6)$$

Again, one-sided confidence intervals for the variance can be derived:

- lower one-sided confidence interval

$$\left[\frac{S^2(n-1)}{\chi_{n-1}^2(1-\alpha)}, \infty \right),$$

- upper one-sided confidence interval

$$\left(0, \frac{S^2(n-1)}{\chi_{n-1}^2(\alpha)} \right].$$

Example 14.3. [1]  During an inspection of a machine designed to fill sugar packages with a nominal weight of 1 kg, the following deviations from the nominal value (in grams) were observed when five packages were reweighed: $-3, 2, -2, 0, -1$. Compute the point estimate and the confidence interval estimate of the standard deviation of the packaging machine.

Solution. A prerequisite for the point and interval estimation of the variance σ^2 is a collection of independent random variables with distribution $N(\mu, \sigma^2)$. The point estimate of the variance σ^2 is computed as

$$S^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)$$

$$s^2 = \frac{1}{4} \{ [(-3)^2 + 2^2 + (-2)^2 + 0^2 + (-1)^2] - 5(-0.8)^2 \} = 3.7.$$

 **VAR.S(B2:B6)** (let the observed data be stored in cells B2:B6).

The point estimate of the standard deviation σ is $s = \sqrt{s^2} \approx 1.9235$.

= **STDEV.S(B2:B6)**

Assuming that the measurement deviations are normally distributed, the interval estimate can be determined. In this case, the quantiles $\chi_{n-1}^2(0.975) \approx 11.14$ and $\chi_{n-1}^2(0.025) \approx 0.48$ are required; these can be obtained using statistical software.

= **CHISQ.INV(0.975,4)**, = **CHISQ.INV(0.025,4)**

The confidence interval for the variance with confidence level $1 - \alpha = 0.95$ is

$$\left[\frac{s^2(n-1)}{\chi_{n-1}^2(0.975)}, \frac{s^2(n-1)}{\chi_{n-1}^2(0.025)} \right] = \left[\frac{3.7 \cdot 4}{11.143}, \frac{3.7 \cdot 4}{0.484} \right] = [1.33, 30.55].$$

The endpoints of the confidence interval for the standard deviation are obtained by taking the square roots of the corresponding endpoints of the confidence interval for the variance. Therefore, the confidence interval for the standard deviation is $[1.15, 5.53]$. $\triangle$

14.4 Confidence Interval for the Mean Based on the CLT

If the underlying random variable and the corresponding random sample do not follow a normal distribution, the confidence intervals defined above cannot be used, because in such a case the random variables

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \quad \text{and} \quad (n-1) \frac{S^2}{\sigma^2} \quad (14.7)$$

do not have the required Student's t -distribution or Pearson's χ^2 -distribution. If, however, the sample is sufficiently large, the Central Limit Theorem can be applied. Roughly speaking, it states that the sum of a large number of random variables behaves approximately as a normally distributed random variable. Consequently, the random variables given in (14.7) have, at least approximately, Student's t -distribution and Pearson's χ^2 -distribution, respectively, and the confidence intervals constructed in the preceding sections may still be used. For this approximation, a sample size of at least $n \geq 30$ is usually recommended.



An alternative approach, also based on the Central Limit Theorem, is to use the following argument. Let $X_1, \dots, X_n$ be a random sample drawn from a distribution with finite expectation μ and finite variance σ^2 . Then, according to the Central Limit Theorem, as $n \rightarrow \infty$, the statistic

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has an asymptotically normal distribution $N(0, 1)$. By the definition of the quantile of the standard normal distribution

$$\mathbb{P} \left(-u \left(1 - \frac{\alpha}{2} \right) \leq \frac{\bar{X} - \mu}{S} \sqrt{n} \leq u \left(1 - \frac{\alpha}{2} \right) \right) = 1 - \alpha.$$

By rearranging the inequalities, we obtain the two-sided confidence interval for the expectation μ with confidence level $1 - \alpha$.

$$\left[\bar{X} - u \left(1 - \frac{\alpha}{2} \right) \frac{S}{\sqrt{n}}, \bar{X} + u \left(1 - \frac{\alpha}{2} \right) \frac{S}{\sqrt{n}} \right]. \quad (14.8)$$

One-sided confidence intervals are given by:

- lower one-sided confidence interval for the expectation μ

$$\left[\bar{X} - u(1 - \alpha) \frac{S}{\sqrt{n}}, \infty \right);$$

- upper one-sided confidence interval for the expectation μ

$$\left(-\infty, \bar{X} + u(1 - \alpha) \frac{S}{\sqrt{n}} \right].$$

Example 14.4.  An experiment was conducted in which a die was rolled 600 times, and a six occurred on 75 occasions. We are interested in estimating the probability that a six is obtained on this die.

Solution. Let $X_1, \dots, X_{600}$ be random variables with Bernoulli distribution $\mathbf{B}$, where a success ($X = 1$) occurs when a 6 is rolled and a failure ($X = 0$) occurs when any of the outcomes 1–5 occurs. We are interested in $p = \mathbb{P}[X = 1]$. Recall that $p = \mathbb{E}X$. Hence, the point estimate of p is $\bar{x} = \frac{75}{600} = 0.125$. The sample variance is

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{599} (75 \cdot (1 - 0.125)^2 + 525 \cdot (0 - 0.125)^2) \approx 0.11.$$

For the computation of the confidence interval with confidence level $1 - \alpha = 95\%$, we additionally require the value $u(0.975) \approx 1.96$.

= `NORM.S.INV(0.975)`

$$\left[\bar{x} - u(0.975) \frac{s}{\sqrt{n}}, \bar{x} + u(0.975) \frac{s}{\sqrt{n}} \right] = [0.098, 0.151].$$

Hence, the true probability of rolling a six with this die lies in the interval $[0.098, 0.151]$ with probability 0.95. Since the value $\frac{1}{6}$ does not belong to the confidence interval, the data provide evidence against the hypothesis that the die is fair. Equivalently, a two-sided test at the significance level 0.05 would reject the hypothesis $p = \frac{1}{6}$. This is the same conclusion as that obtained from the hypothesis test carried out in Example 15.6. $\triangle$

14.5 Exercises

14.1 Determine unbiased point estimators of the mean height of students and of the variance of student height based on the available sample data: 167, 169, 186, 197, 184, 185, 159, 174, 175, 168, 193, 174, 182, 172, 184, 187, 203, 177, 181, 163, 175, 171, 202, 199, 182, 184, 185, 173, 172, 174, 183, 171, 169, 164, 179, 186, 169, 201, 168, 195, 172, 187, 180, 174, 165.

14.2 Based on the data in Exercise 14.1, determine the two-sided, lower one-sided, and upper one-sided confidence intervals with confidence levels of 95%, 90%, and 99% for

- a) the expectation μ ,
- b) the variance σ^2 ,
- c) the standard deviation σ .

14.3 How many observations would be required so that, when estimating the expectation μ of student height from Exercise 14.2, the estimation error exceeds ± 3 cm with only a prescribed probability? Determine the required sample sizes for confidence levels of 90%, 95%, and 99%.

14.4 What is the confidence level of the confidence interval if the data from Exercise 14.1 are used and an estimation error of at most ± 4 cm is considered acceptable?

14.5 During the draining of a fishpond, a random sample of 15 carp was selected and weighed (in kg): 3.0, 3.1, 2.5, 2.5, 4.2, 2.1, 3.25, 2.5, 4.8, 2.3, 4.1, 3.6, 3.0, 3.6, 4.0.

- a) Estimate the mean weight of a carp with a confidence level of 99%.
- b) What minimum guaranteed load capacity must a bag have in order to carry a carp with 95% confidence?
- c) For publication of the fish harvest results, the estimates must be reported with a prescribed precision. How many carp must be weighed in order to estimate the mean carp weight with a confidence level of 99% and a maximum permissible error of 0.25 kg?
- d) With what confidence level can the mean weight of carp in the pond be estimated with a maximum permissible error of 0.25 kg based only on the existing sample of 15 carp?

14.6 During traffic monitoring aimed at enforcing speed limits within a municipality, the speeds of 36 randomly selected vehicles were measured. Compare the difference between the confidence intervals for the mean vehicle speed obtained for $\alpha = 0.05$ and those obtained using the Central Limit Theorem (CLT).

14.7 The weights (in kg) of pigs slaughtered at the age of 182 days were as follows: 124.3, 119.4, 117.8, 131.6, 118.2, 113.9, 135.1, 133.0, 138.4, 125.2, 129.9, 127.9, 131.8, 137.2, 119.7, 122.7, 111.4, 135.2, 129.1, 117.2, 134.3, 133.9, 129.8, 126.4, 123.5, 126.6, 121.0, 127.2, 117.8, 113.5, 126.6, 121.6, 122.1, 124.6, 121.4, 120.8, 122.3, 134.5, 110.1, 127.0, 140.5, 127.7, 137.5, 130.4, 136.5, 136.6, 124.5.

Using a confidence level of 95%, estimate the farmer's revenue from the sale of 60 pigs at the age of 182 days, assuming that the payment is 45 CZK per 1 kg of live weight.

Compare the usual confidence interval with the confidence interval obtained using the Central Limit Theorem (CLT).

14.8 In a forest stand, 80 diameters of spruce tree trunks were selected at random and measured in centimetres:

20, 29, 39, 29, 21, 26, 36, 18, 58, 35, 45, 49, 38, 20, 17, 42, 56, 46, 28, 59, 29, 46, 36, 45, 17, 68, 26, 23, 37, 22, 57, 48, 44, 25, 50, 21, 16, 74, 27, 28, 38, 50, 52, 26, 15, 69, 20, 16, 31, 19, 21, 40, 29, 18, 37, 45, 80, 29, 50, 20, 19, 59, 24, 19, 21, 24, 19, 38, 20, 25, 18, 15, 23, 42, 66, 55, 29, 27, 15, 49.

The sample of spruce trees is assumed to be representative of the entire spruce population, which follows a normal distribution. Estimate, with a confidence level of 95%, the amount of timber that can be obtained from 200 trunks, each 4 m in height. (Neglect the tapering of the trunk.) Compare the usual confidence interval estimate with the confidence interval obtained using the Central Limit Theorem (CLT).

Solutions to the Exercises

14.1 $\bar{x} = 179.\bar{1}$, $s^2 \approx 122.419$.

14.2

- a) confidence intervals for the expectation μ :
 estimate: two-sided, lower one-sided, upper one-sided
 confidence level 90%: $[176.340, 181.882]$, $[176.965, \infty)$, $(-\infty, 181.257]$,

confidence level 95%: $[175.787, 182.435]$, $[176.340, \infty)$, $(-\infty, 181.882]$,
 confidence level 99%: $[174.671, 183.552]$, $[175.129, \infty)$, $(-\infty, 183.093]$,

b) confidence intervals for the variance σ^2 :

estimate: two-sided, lower one-sided, upper one-sided

confidence level 90%: $[89.060, 180.829]$, $[95.060, \infty)$, $[0, 165.802]$,

confidence level 95%: $[83.899, 195.341]$, $[89.060, \infty)$, $[0, 180.829]$,

confidence level 99%: $[74.923, 228.397]$, $[78.394, \infty)$, $[0, 214.190]$,

c) confidence intervals for the standard deviation σ :

estimate: two-sided, lower one-sided, upper one-sided

confidence level 90%: $[9.437, 13.447]$, $[9.77, \infty)$, $[0, 12.876]$,

confidence level 95%: $[9.160, 13.976]$, $[9.437, \infty)$, $[0, 13.447]$,

confidence level 99%: $[8.656, 15.112]$, $[8.854, \infty)$, $[0, 14.635]$.

14.3 Substituting into the formula for the margin of error

$\Delta = t_{m-1} \left(1 - \frac{\alpha}{2}\right) \cdot \frac{S}{\sqrt{n}}$, we determine n . For confidence level $q = 90\%$, the required sample size is 38 observations; for confidence level $q = 95\%$, 55 observations; and for confidence level $q = 99\%$, 120 observations.

14.4 $1 - \alpha \approx 0.979$, i.e., approximately 97.9%.

14.5

- Point estimate of the expectation: 3.237 kg; confidence interval for the expectation: $[2.624, 3.849]$.
- 3.599 kg.
- Approximately 90.057, i.e., about 91 carp.
- 0.756.

14.6 Using the statistic with Student's t -distribution, the confidence interval is $[49.69, 52.66]$; using the CLT, the confidence interval is $[49.74, 52.61]$. Thus, the two estimates differ only slightly.

14.7 The farmer's revenue will, with a confidence level of 95%, lie within the interval $[335\,069.60, 347\,134.20]$ CZK. The confidence interval obtained using the CLT is $[335\,228.20, 346\,975.60]$ CZK. Thus, it can be concluded that the CLT-based estimate is practically identical to the original confidence interval estimate.

14.8 The amount of timber obtained will, with a confidence level of 95%, lie in the interval $[60.3, 91.05] \text{ m}^3$ of timber. The estimate obtained using the CLT differs only slightly, namely $[60.52, 90.8] \text{ m}^3$ of timber.

Chapter 15

Parametric Tests

15.1 Principles of Hypothesis Testing

In scientific research as well as in applications, problems are often formulated in the form of hypotheses. A *statistical hypothesis* is a statement concerning a probability distribution or, alternatively, the relationships among the values of the parameters of a random variable.

Hypotheses can be classified into two types:

- 1) parametric hypotheses – the hypothesis is specified in terms of the parameter of a particular statistical model,
- 2) nonparametric hypotheses – hypotheses concerning the form of a probability distribution, the dependence structure between variables, etc.

Every hypothesis testing problem is formulated in terms of two competing hypotheses: the *null hypothesis* (H_0) and the alternative hypothesis (H_1 or equivalently H_A). In this chapter, we shall consider only parametric tests, i.e., we shall assume that the probability distribution of the random variable under study is known, and the parameter of the given distribution will be the subject of the test.

Suppose that the distribution of a random variable depends on a parameter θ . We suspect that the parameter θ may be equal to a given value θ_0 . In this case, the null hypothesis is written in the form

$$H_0: \theta = \theta_0.$$

The alternative hypothesis H_1 may take one of the following forms

$$H_1: \theta \neq \theta_0, \quad \text{or} \quad H_1: \theta > \theta_0, \quad \text{or} \quad H_1: \theta < \theta_0.$$

Table 15.1: Errors in Hypothesis Testing

Decision True State of Nature	Reject H_0	Fail to Reject H_0
H_0 holds	α Type I error	$1 - \alpha$ correct decision
H_0 does not hold	$1 - \beta$ correct decision	β Type II error

In the first case, we have a so-called two-sided hypothesis; in the second and third cases, we have a one-sided hypothesis (more specifically, a right-sided and a left-sided hypothesis, respectively).

In making a decision regarding the validity of H_1 or H_0 , we may commit one of two types of errors. If we reject H_0 even though it is in fact true, we commit a so-called *Type I error* (false positive). If we fail to reject H_0 even though it is in fact false, we commit a so-called *Type II error* (false negative). In hypothesis testing, it is naturally desirable that the probabilities of both types of errors be as small as possible. The probability of a Type I error is typically denoted by α and is referred to as the *significance level of the test*, whereas the probability of a Type II error is denoted by β . The probability $1 - \beta$ is also referred to as the “power of the test” or the “discriminatory power” of the test and represents the probability of correctly rejecting the null hypothesis H_0 when it is indeed false. The power of a test usually depends on the pre-specified significance level α , in such a way that the power decreases as the significance level decreases. An overview of the possible errors and their probabilities in hypothesis testing is given in Table 15.1.



When deciding whether to accept the null hypothesis or the alternative hypothesis, we rely on a so-called test statistic T . A test statistic is a predefined function of the random sample $X_1, X_2, \dots, X_n$ drawn from the underlying probability distribution. The values of the statistic T may belong to one of two disjoint sets: either the critical region W (the rejection region



of the hypothesis H_0) or the acceptance region V (more precisely, the non-rejection region of the hypothesis H_0). As already mentioned, one of two types of errors may be committed in hypothesis testing. In practice, it is usually required only that the probability of a Type I error be equal to or less than α , where α is a given number from the interval $(0, 1)$. The most common choices in practice are $\alpha = 0.05$ or $\alpha = 0.01$. The rejection region W and the non-rejection region V are separated by the so-called *critical values of the test*. We therefore require that

$$\mathbb{P}(T \in W | H_0) \leq \alpha.$$

Nowadays, most statistical software packages (such as Statistica, R, S+, SAS, and even MS Excel) report the so-called *p-value*, also known as the attained significance level. The *p-value* is the smallest significance level at which the null hypothesis H_0 would still be rejected. Thus, if we choose $\alpha = 0.05$ and the *p-value* is less than or equal to 0.05, we reject the hypothesis H_0 at the significance level $\alpha = 0.05$. If the *p-value* exceeds 0.05, then the hypothesis H_0 cannot be rejected at the significance level $\alpha = 0.05$. However, a “failure to reject” the hypothesis H_0 does not mean that we confirm it or accept it as true, i.e., that the statement it contains is true. Rather, by “failing to reject” the null hypothesis, we merely do not exclude its validity. We also say that the observed data are not inconsistent with H_0 .



Table 15.1 shows that if a test rejects the hypothesis H_0 , then the only possible error is a Type I error, whose probability is controlled by the significance level α and is therefore known to be small. On the other hand, if the test fails to reject H_0 , then the only possible error is a Type II error, which is not controlled by the procedure described above. For this reason, the statement “we reject H_0 ” is stronger than the statement “we fail to reject H_0 ”. This is also reflected in the linguistic asymmetry of these expressions. If we fail to reject H_0 , this may be due to various reasons. Either H_0 is indeed true, or we simply do not have a sufficient amount of data, and hence insufficient information, to reject H_0 .

Hypothesis testing may be compared to the principle of the presumption of innocence: if the sample data do not provide evidence against the null hypothesis, then the null hypothesis cannot be rejected. Similarly, under the presumption of innocence, a defendant is regarded as innocent until

convincing evidence of guilt is presented. In this analogy, the statistical test plays the role of the judge: the statistical test determines whether the sample data are consistent with the null hypothesis, just as a judge determines whether the testimony presented supports the prosecution or the defense. It is required that $\mathbb{P}(T \in W | H_0) = \alpha$, i.e., that the witnesses (the sample data) provide false testimony against the defense when the defendant is in fact innocent (that is, under the condition that H_0 holds) only with probability α . (One difference from the principle of the presumption of innocence is that a judge cannot set the probability level α and does not know its value for a particular case.)

Most commonly, hypothesis testing is used to compare a sample with a given value, or to compare two or more samples with one another. If the test is based on a single random sample, we speak of a one-sample test; if it is based on several samples, we speak of a two-sample test or a multiple-sample test. In the latter case, it is always necessary to distinguish between dependent and independent random samples. For independent samples, it is assumed that the units selected from one population are independent of the units selected from the other population. If the units from one sample are observed on the same objects as those from the other sample, or if they are linked in some other way, then the observations in the two samples are dependent; in the case of two samples, we then speak of a paired test.

15.2 (Student's) One-Sample t -Test for the Mean

Let $X_1, \dots, X_n$ be a random sample from the distribution $N(\mu, \sigma^2)$ of size $n > 1$. The parameter $\sigma^2 > 0$ is assumed to be unknown. In Student's two-sided one-sample t -test, the objective is to test the hypothesis $H_0: \mu = \mu_0$, where μ_0 is a given constant, against the alternative hypothesis $H_1: \mu \neq \mu_0$. The hypothesis H_0 is rejected if $\bar{X}$ is sufficiently far from the value μ_0 . By Theorem 13.3, we know that under the null hypothesis H_0 , the statistic

$$T = \frac{(\bar{X} - \mu_0)}{S/\sqrt{n}} \sim t_{n-1}$$

has a Student t -distribution with $n - 1$ degrees of freedom. By the definition of the critical value of the Student t -distribution, it follows that

$$\mathbb{P}\left(|T| \geq t_{n-1}\left(1 - \frac{\alpha}{2}\right)\right) = \alpha.$$

Thus, the hypothesis H_0 is rejected at the significance level α whenever

$$|t| \geq t_{n-1} \left(1 - \frac{\alpha}{2}\right).$$

For one-sided alternatives $H_1: \mu > \mu_0$ (right-tailed test) and $H_1: \mu < \mu_0$ (left-tailed test), respectively, the hypothesis H_0 is rejected whenever

$$t \geq t_{n-1}(1 - \alpha), \quad \text{respectively} \quad t \leq -t_{n-1}(1 - \alpha).$$

Example 15.1.  Let us return to Example 14.3 on page 289. The objective is to determine whether the filling machine exhibits a systematic bias. Therefore, we need to test the hypothesis $H_0: \mu = 0$ against the alternative $H_1: \mu \neq 0$ at the significance level $\alpha = 0.05$ (i.e., that the deviations fluctuate around zero and are not systematically shifted either towards positive or negative values).

Solution. Given the formulation of the problem and the hypotheses, we use the two-sided version of the test. We have $n = 5$, $\mu_0 = 0$, $\bar{x} = -0.8$, $s \approx 1.9235$,

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \approx -0.93.$$

The critical value of the *t*-distribution with, in general, $n - 1$ degrees of freedom (here $5 - 1 = 4$), namely $t_4(0.975) \approx 2.776$, can be obtained either from statistical tables or by computation in MS Excel.

 = T.INV(0.975,4)

Since $|-0.93| \not\geq t_4(0.975) \approx 2.776$, we fail to reject H_0 . Therefore, the observed data do not contradict the assumption that the filling machine has no systematic bias. Notice that $\mu_0 = 0$ (the mean under the null hypothesis H_0) lies within the 95% confidence interval (see Example 14.3). In other words, 0 is a plausible value of the true population mean, and therefore we cannot reject H_0 . The two approaches to hypothesis testing – the classical hypothesis-testing approach and the confidence-interval approach – are equivalent. $\triangle$

15.2.1 Computation of the Two-Sided and One-Sided *p*-Values of the *t*-Test

We have already mentioned that the *p*-value, also known as the attained level of significance (significance) of a test, is a numerical quantity that can be used as an alternative criterion for decision-making in statistical hypothesis testing. When testing the hypothesis H_0 using a test statistic T , the *p*-value can be interpreted in two equivalent ways:

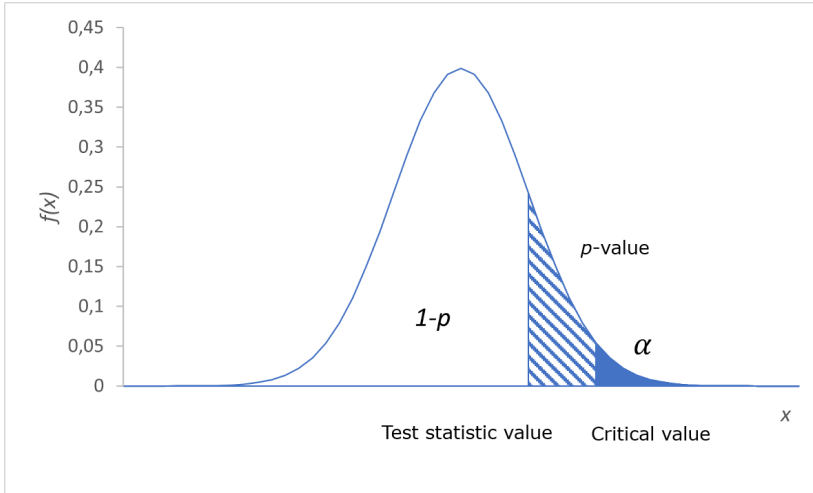


Figure 15.1: Relationship between α and the p -Value for a Right-Tailed Test (Illustrated by the Probability Density Function of the Standard Normal Distribution U)

- 1) The p -value is the smallest significance level at which the null hypothesis H_0 would still be rejected.
- 2) The p -value is the conditional probability that, under the assumption that H_0 is true, the test statistic T attains the observed value t or a value with a greater absolute magnitude (for a two-sided alternative hypothesis)¹.

When solving practical problems, we usually first specify a significance level α , then compute the p -value and compare these two quantities. If the p -value is smaller than the chosen significance level ($p < \alpha$), we reject H_0 ; otherwise ($p > \alpha$), we fail to reject it. Consequently, the smaller the p -value, the less plausible the null hypothesis H_0 appears to be.

According to the type of hypothesis test, we distinguish three cases:

- In the case of a two-sided test with a symmetric distribution of the test statistic, the p -value is equal to twice the area under the probability density function of the test statistic to the right of the absolute value of the observed test statistic. It is computed in general as $2 \cdot [1 - F(|t|)]$. In the case of an asymmetric distribution, the relation $p = 2 \cdot \min\{F(t), 1 - F(t)\}$ applies.
- In the case of a right-tailed test, the p -value is equal to the area under the probability density function of the test statistic to the right of its

¹In the case of a right-tailed alternative, values greater than t are considered; in the case of a left-tailed alternative, values smaller than t are considered.

observed value and is computed as $1 - F(t)$, where $F(t)$ denotes the value of the cumulative distribution function at the observed value t of the test statistic.

- In the case of a left-tailed test, the p -value is equal to the area under the probability density function of the test statistic to the left of its observed value and is computed as $F(t)$.

In all three cases, the probability density function and the cumulative distribution function F of the test statistic refer to the distribution of the test statistic under the assumption that H_0 holds. In particular, the p -values for the one-sample t -test are computed using the formula

$$p = 2 \cdot [1 - F(|t|)]$$

for the two-sided version of the test and by means of

$$p = F(t), \quad \text{respectively} \quad p = 1 - F(t)$$

for the left-tailed and right-tailed versions of the test, respectively.

15.3 Test for the Variance of a Normal Distribution

The one-sample test for variance is used to test a hypothesis concerning the variance of a population, i.e., to determine whether the hypothesized value of the variance is equal to its population counterpart. In other words, it assesses whether, at a given significance level, the observations in the sample exhibit the same degree of variability as the population.

Let $X_1, \dots, X_n$ be a random sample from the distribution $N(\mu, \sigma^2)$, where $n > 1$. The objective is to test the hypothesis

$$H_0: \sigma^2 = \sigma_0^2 \quad \text{against the alternative hypothesis} \quad H_1: \sigma^2 \neq \sigma_0^2,$$

where σ_0^2 is a given positive constant. The hypothesis H_0 is rejected if S^2 is sufficiently far from the value σ_0^2 . From Theorem 13.3, we know that under the null hypothesis H_0 , the statistic Q follows a χ^2 distribution with $n - 1$ degrees of freedom:

$$Q = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi_{n-1}^2.$$

By the definition of the critical values (quantiles) of the χ^2 distribution, it follows that

$$\mathbb{P} \left(\chi_{n-1}^2 \left(\frac{\alpha}{2} \right) \leq (n-1) \frac{S^2}{\sigma_0^2} \leq \chi_{n-1}^2 \left(1 - \frac{\alpha}{2} \right) \right) = 1 - \alpha,$$

Thus, the hypothesis H_0 is rejected at the significance level α if, for the observed value q of the test statistic Q , the following condition holds:

$$q \leq \chi_{n-1}^2\left(\frac{\alpha}{2}\right) \quad \text{or} \quad q \geq \chi_{n-1}^2\left(1 - \frac{\alpha}{2}\right).$$

In the case of the one-sided alternative hypothesis $H_1: \sigma^2 > \sigma_0^2$ (right-tailed test), or $H_1: \sigma^2 < \sigma_0^2$ (left-tailed test), respectively, the hypothesis H_0 is rejected whenever

$$q \geq \chi_{n-1}^2(1 - \alpha), \quad \text{respectively} \quad q \leq \chi_{n-1}^2(\alpha).$$

The p -value of the test for the variance of a sample drawn from a normal distribution is computed, due to the asymmetry of the χ^2 distribution,

$$p = 2 \cdot \min\{F(q); 1 - F(q)\},$$

where $\min$ denotes the minimum of the two computed values and $F(q)$ is the value of the cumulative distribution function evaluated at the observed value of the test statistic q .

The p -values for the one-sided versions of the test (left-tailed and right-tailed, respectively) for the variance of a normal distribution are computed according to the formulas

$$p = F(q), \quad p = 1 - F(q).$$

Example 15.2.  The training of a laboratory technician in operating a particular optical instrument is considered complete once the measurement variance for a specified object does not exceed 0.0196. The following measurements were obtained: 6.82, 6.44, 6.38, 6.21, 6.38, 6.60, 6.32. Determine whether further training is required.

Solution. In this case, a one-sided alternative must be used. We are interested in determining whether further training is necessary, i.e., whether we can conclude with a high degree of confidence that $\sigma^2 > 0.0196$. This statement is therefore chosen as the alternative hypothesis. The null hypothesis H_0 is formulated in the form of an equality. Hence, we test $H_0: \sigma^2 = 0.0196$ against the alternative $H_1: \sigma^2 > 0.0196$. We compute the sample variance $s^2 \approx 0.0406$

$$= \text{VAR.S(B3:B9)}$$

and the test statistic $q \approx \frac{6s^2}{0.0196} \approx 12.44$. The desired degree of confidence is specified through the significance level (here chosen at the conventional level) $\alpha = 0.05$. The critical value of the test is $\chi_6^2(0.05) \approx 12.59$,

`= CHISQ.INV(0.95,6)`

and since $q \not\geq \chi_6^2(0.05)$, we fail to reject H_0 .

Notice that the estimated variance (0.0406) exceeds the required variance (0.0196), yet the observed data do not allow us to reject the hypothesis that the variance is equal to the required value at the significance level $\alpha = 0.05$. In other words, there is a probability greater than 5% that the observed data could have arisen from a normal distribution with variance 0.0196. The same conclusion can also be reached using the p -value approach: $p \approx 0.05286 > 0.05$,

`= 1-CHISQ.DIST(12.44,6,TRUE)`

and therefore we do not reject H_0 at the significance level $\alpha = 0.05$. Note, however, that the p -value exceeds the significance level α only by a very small margin. With only slightly weaker evidence requirements, for example at the significance level $\alpha = 0.06$ (in fact, at any significance level greater than or equal to 0.05286), we would reject the null hypothesis. $\triangle$

15.4 Paired *t*-Test for Equality of Means

Let $(Y_1, Z_1), (Y_2, Z_2), \dots, (Y_n, Z_n)$ be a random sample from a bivariate distribution with mean vector (μ_1, μ_2) . We wish to test the hypothesis

$$H_0: \mu_2 - \mu_1 = \Delta \quad \text{against the alternative hypothesis} \quad H_1: \mu_2 - \mu_1 \neq \Delta$$

where Δ is a given constant, most commonly $\Delta = 0$. Define

$$X_1 = Z_1 - Y_1, \quad X_2 = Z_2 - Y_2, \dots, X_n = Z_n - Y_n.$$

The order of subtraction in the definition of the random variables $Z - Y$ is chosen according to the intended interpretation. For example, if the values Y are measured before the application of a treatment and the values Z after the treatment, then X represents the increase in the measured variable.

The variables $X_1, X_2, \dots, X_n$ are independent. Assume that $X_i \sim N(\mu, \sigma^2)$, $i = 1, 2, \dots, n$. Clearly, $\mu = \mu_2 - \mu_1$. If these assumptions are satisfied, the problem is reduced to a one-sample t -test: from the observed values $x_1, x_2, \dots, x_n$, we compute $\bar{x}$, s^2 , and the value of the test statistic t . The hypothesis H_0 is rejected at the significance level α whenever

$$|t| = \left| \frac{\bar{x} - \Delta}{s/\sqrt{n}} \right| \geq t_{n-1} \left(1 - \frac{\alpha}{2} \right).$$

The paired t -test is used in situations where measurements of two different variables have been obtained on each of n objects. The individual

Table 15.2: Higher of the Measured Systolic Blood Pressure Values

Blood Pressure / Patient	1	2	3	4	5	6
Before Drug Administration	180	160	150	165	170	175
After Drug Administration	150	155	155	150	155	170
Difference	30	5	-5	15	15	5

objects can usually be regarded as independent, whereas the measurements taken on the same object cannot. The paired t -test is employed, for example, when assessing the effectiveness of a treatment on n patients, where Y_i denotes the measurements taken before administration of the treatment and Z_i denotes the measurements taken after administration. Such samples may arise, for instance, in a pretest–posttest design. A paired test is also appropriate when the same variable is measured on two parts of the same object. A typical example is determining whether, after proper wheel alignment, both tires of a vehicle wear at the same rate.

The p -values for the two-sided paired t -test are computed according to the formula

$$p = 2 \cdot (1 - F(|t|)),$$

for the one-sided versions of the test

$$p = 1 - F(|t|), \quad \text{respectively} \quad p = F(|t|).$$

The p -value for both the one-sided and two-sided versions of the test can also be computed using the MS Excel function `= T.TEST(array1,array2,tails,type)`, where the argument “tails” is set to either 1 (one-tailed test) or 2 (two-tailed test), and the argument “type” is set to 1 (paired test).

Example 15.3.  At the significance level $\alpha = 0.05$, it is required to determine whether a drug intended to reduce systolic blood pressure is effective. To this end, 6 patients were selected, and their systolic blood pressure was measured both before administration of the drug and one hour after administration. For each patient, the larger of the two systolic blood pressure measurements is recorded in Table 15.2.



Solution. We shall regard the measurement differences as realizations of independent random variables with distribution $N(\mu, \sigma^2)$, where σ^2 is un-

known. If the drug has no effect on blood pressure, then the null hypothesis $H_0: \mu = 0$ holds; if the drug is effective, then the alternative hypothesis $H_1: \mu > 0$ holds. Thus, we have $n = 6$, $\Delta = 0$, and

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{6} \cdot (30 + 5 - 5 + 15 + 15 + 5) \approx 10.833.$$

= AVERAGE(D4:D9)

$$\begin{aligned} s^2 &= \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) \approx \\ &\approx \frac{1}{5} \cdot (30^2 + 5^2 + 5^2 + 15^2 + 15^2 + 5^2 - 6 \cdot (-10.833)^2) \approx \\ &\approx 144.167. \end{aligned}$$

= VAR.S(D4:D9)

$$s = \sqrt{s^2} \approx \sqrt{144.167} \approx 12.007.$$

= STDEV.S(D4:D9)

We compute the value of the test statistic T :

$$t = \frac{\bar{x} - \Delta}{s/\sqrt{n}} \approx 2.21.$$

We obtain the critical value from statistical tables or compute it in MS Excel: $t_5(0.05) \approx 2.015$.

= T.INV(0.95,5)

Since $t \geq t_5(0.05)$, we reject the null hypothesis H_0 in favor of H_1 on the basis of the observed measurements. We therefore conclude, with a high degree of confidence, that the drug is effective.

The p -value of the one-sided test can be computed directly in MS Excel using the function = T.TEST as follows:

= T.TEST(array1,array2,1,1) $\approx 0.039049 < \alpha$. The probability that, if the drug were ineffective, we would obtain a value of the test statistic equal to 2.21 (or larger) is only 3.9%, which is less than the 5% significance level of the test. Hence, once again, we reject H_0 at this significance level.

The example can also be solved using the Data Analysis tool in MS Excel. On the Data tab, select Data Analysis and then choose the option “t-Test:

t-Test: Paired Two Sample for Means		
	Variable 1	Variable 2
Mean	166.6666667	155.8333333
Variance	116.6666667	54.16666667
Observations	6	6
Pearson Correlation	0.167725574	
Hypothesized Mean Diff	0	
df	5	
t Stat	2.210066303	
P(T<=t) one-tail	0.039048983	
t Critical one-tail	2.015048373	
P(T<=t) two-tail	0.078097966	
t Critical two-tail	2.570581836	

Figure 15.2: Excerpt from the MS Excel Data Analysis Tool for the Paired Two-Sample t -Test

Paired Two Sample for Means”. First, select the corresponding data ranges. The hypothesized mean difference is assumed to be equal to 0 (and therefore need not be entered explicitly), and the significance level α is specified (the default value is 0.05). Next, select the cell where the output table should begin and click OK.

Figure 15.2 shows the output table produced by the Data Analysis tool for the data from Example 15.3. In addition to the value of the test statistic and the critical value corresponding to the selected significance level, the table also contains the previously computed p -values for both the one-sided and the two-sided alternative hypotheses.

△

15.5 Two-Sample t -Test for Equality of Means under the Assumption of Equal Variances

Consider the situation in which we have two samples of unpaired observations of possibly different sizes. Assume that the two populations have equal variances. In such a case, the comparison of the sample means is carried out using the two-sample t -test, sometimes also referred to as the two-sample Student’s t -test.

Let $X_1, X_2, \dots, X_n$ be a random sample from the normal distribution $N(\mu_1, \sigma^2)$ and let $Y_1, Y_2, \dots, Y_m$ be a random sample from the normal distribution $N(\mu_2, \sigma^2)$. Assume that these two samples are independent and note that we assume a common (not necessarily known) variance satisfying $\sigma^2 > 0$. Furthermore, suppose that $n \geq 2$ and $m \geq 2$. We wish to test the

hypothesis

$$H_0: \mu_1 - \mu_2 = \Delta \quad \text{against} \quad H_1: \mu_1 - \mu_2 \neq \Delta,$$

where Δ is a given constant (most commonly $\Delta = 0$). Let $\bar{X}$, S_X^2 and $\bar{Y}$, S_Y^2 denote the sample means and sample variances of the respective samples. It can be shown (see [1]) that, under the null hypothesis H_0 , the test statistic T of the form

$$T = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{(n-1)S_X^2 + (m-1)S_Y^2}} \cdot \sqrt{\frac{nm(n+m-2)}{n+m}} \sim t_{n+m-2},$$

has a Student t -distribution t with $n+m-2$ degrees of freedom. Therefore, the hypothesis H_0 is rejected at the significance level α whenever

$$|t| \geq t_{n+m-2} \left(1 - \frac{\alpha}{2}\right).$$

The two-sample t -test is used in situations where, for example, the effect of drug A is studied on n patients and the effect of drug B on another m patients. The purpose of the study is to determine whether the effects of the two drugs are the same.

The paired t -test and the two-sample t -test are often confused, which constitutes a serious error. The two-sample t -test may be used only when the independence of all random variables $X_1, X_2, \dots, X_n$, $Y_1, Y_2, \dots, Y_m$ is ensured. Confusing these two tests will generally lead to meaningless results.

The p -value for the two-sided version of the test is computed according to the formula

$$p = 2 \cdot (1 - F(|t|)),$$

for the one-sided versions of the test

$$p = 1 - F(|t|), \quad \text{respectively} \quad p = F(|t|).$$

For the direct computation of the p -value for both the one-sided and two-sided versions of the test, the MS Excel function `= T.TEST(array1,array2,tails,type)` can also be used, where the argument “tails” is set to either 1 (one-tailed test) or 2 (two-tailed test), and the argument “type” is set to 2 (two samples assuming equal variances).

When solving problems involving the testing of equality of means using the two-sample t -test, the Data Analysis tool in MS Excel can be employed. On the Data tab, select Data Analysis and then choose the option “t-Test: Two-Sample Assuming Equal Variances”. First, select the corresponding

data ranges. The hypothesized mean difference may be left at its default value of 0 (it is not necessary to enter it explicitly), and the significance level α may likewise be left at its default value of 0.05 (unless a different significance level is desired). Finally, select the cell where the output table is to begin and click OK.

For the tests described in Sections 15.2 through 15.5, certain assumptions must be satisfied. The most important of these is the independence of the individual random variables and, in the case of the two-sample test, the independence of the two samples as well. Violation of this assumption has serious consequences and renders conclusions based on the tests invalid. In practice, this assumption is violated when the sample is not randomly selected and therefore does not represent the population of interest (for example, when certain subgroups of the population are systematically omitted).

Another assumption underlying these tests is the normality of the distribution. Owing to the validity of the Central Limit Theorem (CLT) and the laws of large numbers, violations of this assumption are generally not serious when the sample size is sufficiently large, and all parametric tests are routinely applied in such situations. Section 15.8 also presents, for large samples, an alternative test based on the CLT that does not require the assumption of normality. When normality is seriously violated and the sample size is small, it is preferable to use an appropriate nonparametric test. Tests for assessing the normality of a random sample are presented in Sections 20.2 and 20.4.

For the two-sample t -test, there is an additional requirement that the variances of the two populations be equal. In such cases, we always seek to compare data sets for which the difference between the variances is not excessively large. Equality of variances can be tested using the procedure described in the following section.

15.6 Test for Equality of Two Variances (F -Test)

Let $X_1, X_2, \dots, X_m$ be a random sample from $N(\mu_1, \sigma_1^2)$ and let $Y_1, Y_2, \dots, Y_n$ be a random sample from $N(\mu_2, \sigma_2^2)$. Assume that these two samples are independent. Furthermore, suppose that $n \geq 2$, $m \geq 2$, $\sigma_1^2 > 0$, and $\sigma_2^2 > 0$. We wish to test the hypothesis

$$H_0: \sigma_1^2 = \sigma_2^2 \quad \text{against} \quad H_1: \sigma_1^2 \neq \sigma_2^2.$$

Since S_X^2 is an unbiased estimator of the parameter σ_1^2 and S_Y^2 is an unbiased estimator of the parameter σ_2^2 , it is reasonable to expect that, under the

Table 15.3: Average Field Yields [t/ha] for Example 15.4

Fertilizer A	62	54	55	60	53	58
Fertilizer B	52	56	49	50	51	-

null hypothesis H_0 , the ratio

$$F = \frac{S_X^2}{S_Y^2}$$

will be close to 1. Consequently, evidence against H_0 is provided either by values close to zero or by sufficiently large values.

According to Part b) of Theorem 13.3 and the defining formula (8.19), the statistic F , being the ratio of two independent random variables having χ^2 distributions, follows the Fisher–Snedecor distribution F with $n - 1$ and $m - 1$ degrees of freedom.

The hypothesis H_0 is rejected if

$$f = \frac{s_X^2}{s_Y^2} \leq F_{n-1, m-1}\left(\frac{\alpha}{2}\right) \quad \text{or} \quad f = \frac{s_X^2}{s_Y^2} \geq F_{n-1, m-1}\left(1 - \frac{\alpha}{2}\right).$$

where $F_{n-1, m-1}(\cdot)$ denotes the quantiles of the Fisher–Snedecor distribution F with $n - 1$ and $m - 1$ degrees of freedom.

The p -values for the two-sided F -test are computed according to the formula

$$p = 2 \cdot \min\{F(f), 1 - F(f)\},$$

and for the one-sided F -test according to the formula

$$p = 1 - F(f) \quad \text{for } f > 1, \quad p = F(f) \quad \text{for } f < 1.$$

For the direct computation of the p -value of the two-sided F -test, the MS Excel function `= F.TEST(array1,array2)` can be used.

Example 15.4.  A farmer cultivated eleven fields, applying fertilizer A to six of them and fertilizer B to the remaining five fields. After harvesting, the average yields (in tonnes per hectare) were determined for each field (see Table 15.3). It is required to determine whether the two fertilizers are equally effective.

Solution. We regard the average yields in the first group as a sample drawn from the distribution $N(\mu_1, \sigma_1^2)$ and the average yields in the second group

as a sample drawn from $N(\mu_2, \sigma_2^2)$. We first test the hypothesis of equality of variances of the two samples

$$H'_0: \sigma_1^2 = \sigma_2^2 \quad \text{against} \quad H'_1: \sigma_1^2 \neq \sigma_2^2.$$

We compute $s_X^2 = 12.8$, $s_Y^2 = 7.3$, and the value of the test statistic

$$f = \frac{s_X^2}{s_Y^2} = \frac{12.8}{7.3} \approx 1.753.$$

The hypothesis of equal variances would be rejected if either

$$f \leq F_{5,4}(0.025) \approx 0.135,$$

`= F.INV(0.025,5,4)`, which is not the case, or

$$f \geq F_{(5,4)}(0.975) \approx 9.364,$$

`= F.INV(0.975,5,4)`, which is also not the case². The p -value of the two-sided F -test is computed as follows:

`= F.TEST(A2:A7,B2:B6)` ≈ 0.606345 (with the data stored in cells A2:A7 and B2:B6), or equivalently

$$\text{= 2*MIN(1-F.DIST(1.753,5,4);F.DIST(1.753,5,4))} \approx 0.606345.$$

Because the p -value of the test exceeds α , we fail to reject H_0 at the chosen significance level.

The example can also be solved using the Data Analysis tool in MS Excel. On the Data tab, select Data Analysis and then choose the option “F-Test Two-Sample for Variances”. First, select the appropriate data ranges. The hypothesized difference between the means is assumed to be equal to 0 (and therefore need not be entered), and the significance level alpha is selected (its default setting is 0.05). Next, choose the cell where the output table is to begin and click OK. The table shown in Figure 15.3 presents the results obtained for Example 15.4. However, the table does not report either the critical value or the p -value for the required two-sided test; nevertheless, the two-sided p -value can be obtained simply as twice the reported one-sided p -value.

We therefore fail to reject the hypothesis of equal variances and may proceed to the two-sample t -test of the hypothesis

$$H_0: \mu_1 = \mu_2 \quad \text{against} \quad H_1: \mu_1 \neq \mu_2$$

²Owing to the symmetry property (13.2) of the quantiles of the Fisher–Snedecor distribution, this argument can be simplified: if the samples are labeled so that the first sample is the one with the larger variance, it is sufficient, without loss of generality, to verify only the second of the two inequalities mentioned above.

F-Test Two-Sample for Variances		
	Variable 1	Variable 2
Mean	57	51.6
Variance	12.8	7.3
Observations	6	5
df	5	4
F	1.753424658	
P(F<=f) one-tail	0.303172533	
F Critical one-tail	6.256056502	

Figure 15.3: MS Excel *F*-Test Output for Example 15.4

under the assumption of equal variances. We have $n = 6$, $m = 5$, $\Delta = 0$, $\bar{x} = 57$, $\bar{y} = 51.6$, and the test statistic T has the value

$$t = \frac{\bar{x} - \bar{y} - \Delta}{\sqrt{(n-1)s_X^2 + (m-1)s_Y^2}} \cdot \sqrt{\frac{nm(n+m-2)}{n+m}} \approx 2.7712.$$

We compare the value of the test statistic T with the critical value

$$= \text{T.INV}(0.975, 9) = t_9(0.975) \approx 2.26,$$

and since $|t| \geq t_9(0.975)$, we reject the hypothesis H_0 of equal effectiveness of fertilizers A and B. The p -value for the two-sided version of the t -test is

$$= 2 * (1 - \text{T.DIST}(\text{ABS}(2.7712), 9, \text{TRUE})) \approx 0.02171.$$

For the direct computation of the p -value of either the one-sided or the two-sided version of the t -test, the function

$$= \text{T.TEST}(\text{array1}, \text{array2}, 2, 2) \approx 0.02171$$

can also be used. If one-sided p -values are required³, they can be computed as follows:

$$= \text{T.DIST}(2.7712, 9, \text{TRUE}) \approx 0.989145 \text{ (left-tailed test), respectively}$$

$$= 1 - \text{T.DIST}(2.7712, 9, \text{TRUE}) \approx 0.010855 \text{ (right-tailed test).}$$

The example can also be solved using the Data Analysis tool in MS Excel. On the Data tab, select Data Analysis and then choose the option “t-Test: Two-Sample Assuming Equal Variances”. First, select the appropriate data ranges. The hypothesized difference between the means is assumed to be equal to 0 (and therefore need not be entered), and the significance level alpha is selected (its default value is 0.05). Next, choose the cell where the output table is to begin and click OK. The table shown in Figure 15.4

³For example, when testing the hypothesis that fertilizer A is more effective than fertilizer B.

t-Test: Two-Sample Assuming Equal Variances		
	Variable 1	Variable 2
Mean	57	51.6
Variance	12.8	7.3
Observations	6	5
Pooled Variance	10.35555556	
Hypothesized Mean	0	
df	9	
t Stat	2.77122216	
P(T<=t) one-tail	0.010855041	
t Critical one-tail	1.833112933	
P(T<=t) two-tail	0.021710083	
t Critical two-tail	2.262157163	

Figure 15.4: MS Excel t -Test Output for Example 15.4

presents the results of testing the equality of means for the data from Example 15.4.

△

15.7 Testing the Equality of Means under Unequal Variances

If the variances of the two samples drawn from normal distributions are found to be unequal on the basis of the F -test ($\sigma_1^2 \neq \sigma_2^2$), we do not assume equal variances. Likewise, if the sample sizes differ ($m \neq n$), alternative tests for comparing the means of two independent samples should be used, such as the Cochran–Cox test or Welch’s⁴ (also known as the Aspin–Welch) test.

Assume that $X_1, X_2, \dots, X_n$ is a random sample from $N(\mu_1, \sigma_1^2)$ and that $Y_1, Y_2, \dots, Y_m$ is a random sample from $N(\mu_2, \sigma_2^2)$ independent of the first sample. If, when testing the equality of variances of the two samples, we find that $\sigma_1^2 \neq \sigma_2^2$, then the means can be compared using an approximate

⁴Bernard Lewis Welch (1911–1989), British statistician.

test. First, we compute

$$S_X^2 = \frac{1}{m-1} \left(\sum_{i=1}^m X_i^2 - m\bar{X}^2 \right), \quad S_Y^2 = \frac{1}{n-1} \left(\sum_{j=1}^n Y_j^2 - n\bar{Y}^2 \right),$$

$$S = \sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}, \quad v_x = \frac{S_X^2}{m}, \quad v_y = \frac{S_Y^2}{n}.$$

and subsequently compute the test statistic

$$T = \frac{\bar{X} - \bar{Y} - \Delta_0}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}}.$$

We test $H_0: \mu_1 - \mu_2 = 0$ against $H_1: \mu_1 - \mu_2 \neq 0$. Under the null hypothesis H_0 , the statistic T is approximately Student t -distributed. According to the Cochran–Cox criterion, we reject H_0 whenever

$$|t| \geq \frac{v_x t_{m-1}(1 - \frac{\alpha}{2}) + v_y t_{n-1}(1 - \frac{\alpha}{2})}{v_x + v_y}. \quad (15.1)$$

This test has an approximate significance level equal to α . In MS Excel, the test is carried out using an approximation of the degrees of freedom given by

$$\tilde{\mu} = \frac{\left(\frac{S_X^2}{m} + \frac{S_Y^2}{n} \right)^2}{\frac{(S_X^2/m)^2}{m-1} + \frac{(S_Y^2/n)^2}{n-1}}.$$

The p -values for Welch's test are then computed in the standard manner using the formulas

$$p = 2 \cdot (1 - F(|t|))$$

for the two-sided test and

$$p = F(t), \quad p = 1 - F(t)$$

for the one-sided versions of the test, where F denotes the cumulative distribution function of the Student t -distribution with $\tilde{\mu}$ degrees of freedom.

The function `=T.TEST` in MS Excel computes the p -value of Welch's test using the approximated number of degrees of freedom $\tilde{\mu}$. The Data Analysis tool as well as a manual computation using the function `=T.DIST` return a slightly less accurate p -value because any non-integer degrees of freedom are truncated to an integer value. However, the difference is very small and usually has no effect on the decision regarding the validity of the null hypothesis H_0 .

Table 15.4: Average Student Scores in Example 15.5

SIM	90	85	88	89	94	91	79	83	87	88	91	90
Without SIM	67	90	71	95	88	83	72	66	75	86	93	84

Example 15.5.  Table 15.4 presents the test scores obtained by two groups of different students in a mathematics examination. The first group used supplementary instructional material (SIM) provided by the teacher during their preparation for the test, whereas the second group did not. At the significance level $\alpha = 0.05$, test whether the teacher's supplementary instructional material contributed to higher average test scores.

Solution. The numbers in the two samples come from different students; hence, the samples are independent. We assume that the samples are normally distributed. First, we test the equality of the variances of the two samples. We formulate the auxiliary hypotheses $H'_0: \sigma_1 = \sigma_2$ (i.e., the variances of the score results are the same in both groups) against $H'_1: \sigma_1 \neq \sigma_2$ (the variances of the score results in the two groups are different). In MS Excel, we obtain the value of the test statistic $f \approx 0.156$ and critical values of approximately 0.288 and 3.474; thus, the value f of the test statistic lies in the critical region and we reject H'_0 . The p -value also

$$= \text{F.TEST}(A2:A13, B2:B13) \approx 0.004625 \leq 0.05$$

confirms, with a high degree of confidence, that the variances of the two samples are not equal.

For testing the means, we formulate the null hypothesis $H_0: \mu_1 = \mu_2$ (the score results are the same in both groups) against the right-tailed alternative $H_1: \mu_1 > \mu_2$ (the score results of the group that used the supplementary material are, on average, better than those of the group without aids). After rounding, the test statistic has the value $t \approx 2.236$. The critical value for the right-tailed Cochran–Cox test⁵ is, after rounding, 1.796; the Data Analysis tool returns the critical value 1.761 (the more accurate value with non-integer degrees of freedom is 1.758, but this is not easy to compute in Excel). In all cases, we reject H_0 in favor of the alternative hypothesis H_1 . The p -value of Welch's test also

$$= \text{T.TEST}(A2:A13, B2:B13, 1, 3) \approx 0.02085 \leq \alpha = 0.05$$

confirms that the teacher's supplementary material helped the students achieve better results. $\triangle$

⁵In formula (15.1), we use the $(1 - \alpha)$ -quantiles of the Student t -distribution.

15.8 Test for the Mean Based on the CLT

In cases where the random variables substantially deviate from normality, the preceding tests cannot be applied, since the test statistics T , Q , and F no longer follow the required distributions. However, if the number of random variables is sufficiently large, we may invoke the Central Limit Theorem (hereafter abbreviated as CLT), according to which the sum of a large number of random variables behaves approximately as a normal distribution. For the CLT approximation to be applicable, a sample size of at least $n \geq 20$ is commonly recommended.

Let $X_1, \dots, X_n$ be a random sample from a distribution with finite mean μ and finite variance σ^2 . The objective is to test the hypothesis $H_0: \mu = \mu_0$, where μ_0 is a given constant, against the alternative $H_1: \mu \neq \mu_0$. The hypothesis H_0 is rejected whenever $\bar{X}$ is sufficiently far from the value μ_0 . According to the CLT, the statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}$$

under the null hypothesis H_0 , has an asymptotic standard normal distribution. By the definition of the quantile of the standard normal distribution, it follows asymptotically that

$$\mathbb{P}\left(|Z| \leq u\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha.$$

Thus, the hypothesis H_0 is rejected at the significance level α whenever

$$|z| \geq u\left(1 - \frac{\alpha}{2}\right).$$

For the one-sided alternatives $H_1: \mu > \mu_0$ and $H_1: \mu < \mu_0$, respectively, the hypothesis H_0 is rejected whenever

$$z \geq u(1 - \alpha), \quad \text{respectively} \quad z \leq u(\alpha).$$

If the variance σ_0^2 is unknown, we replace it in the computation of the statistic Z by its unbiased estimator S^2 . The resulting modified statistic also has an asymptotic standard normal distribution.

The p -value for the two-sided version of the test is computed according to the formula

$$p = 2 \cdot (1 - \Phi(|z|)),$$

The p -values for the one-sided versions of the test are computed according to the formulas

$$p = \Phi(z), \quad \text{respectively} \quad p = 1 - \Phi(z).$$

Example 15.6.  Let us return to Example 14.4. The objective is to determine whether the die is fair with respect to the probability of rolling a six. Therefore, we test the hypothesis $H_0: \mu = \frac{1}{6}$ against the alternative hypothesis $H_1: \mu \neq \frac{1}{6}$ at the significance level $\alpha = 0.05$.

Solution. We have $n = 600$, $\mu_0 = \frac{1}{6}$, $\bar{x} = 0.125$, $s^2 = 0.109$. The variance of the random variable $X_i \sim B(\mu_0)$ under the null hypothesis H_0 is $\sigma_0^2 = \mu_0(1 - \mu_0) = \frac{5}{36}$. We compute

$$z = \frac{\bar{x} - \mu_0}{\sigma_0 / \sqrt{n}} \approx -2.74. \quad (15.2)$$

We compute the critical value $u(0.975) \approx 1.96$.

 `NORM.S.INV(0.975)`

Since $|z| \geq u(0.975)$, we reject H_0 . Therefore, the observed data are inconsistent with the assumption that the die is fair. Note that in this test we used σ_0^2 , whereas in Example 14.4 we used s^2 . If, in formula (15.2), we were also to replace the parameter σ_0^2 by the statistic s^2 , the test would be equivalent to the confidence-interval approach. From a practical point of view, however, both approaches are equivalent.

For computing the p -value of the test, the  `Z.TEST` function in MS Excel can be used. This function has three arguments: the data range, the hypothesized mean (x), and the known population standard deviation (*sigma*). In our case, we use the range A4:A603, which contains 75 ones and 525 zeros. The p -value for the two-sided test is equal to twice the minimum of the values

 `Z.TEST(A4:A603,1/6,SQRT(1/6*5/6))` and

 `1-Z.TEST(A4:A603,1/6,SQRT(1/6*5/6))`.

$p \approx 0.00617$, and therefore we reject H_0 . △

Nowadays, the above tests based directly on the Central Limit Theorem are rarely used in practice. Instead, the Central Limit Theorem serves as the theoretical foundation and justification for the routine use of (exact) parametric tests (t -tests, χ^2 -tests, and F -tests) in the case of large samples, even when the underlying distribution is not normal. All of these tests are asymptotically equivalent.

15.9 Exercises

15.1 For a random sample of eight cigarettes of the same brand, the mean nicotine content was found to be 4.2 mg with a sample standard deviation of 1.3 mg. Determine whether the manufacturer violates its advertising claim that the average nicotine content of this type of cigarette does not exceed 3.5 mg. Use a significance level of 0.01 and assume that the nicotine content is normally distributed.

15.2 When inspecting a machine designed to fill 1 kg sugar packages, the deviations from the target weight observed for five packages were $-3, 2, -2, 0, -1$ grams. At the significance level 0.01, determine whether the machine exhibits a systematic deviation from the target value. (The individual deviations are assumed to be realizations of independent random variables with distribution $N(\mu, \sigma^2)$, where σ^2 is an unknown parameter.)⁶

15.3 The automatic dosing system is set to dispense a vitamin premix at a level of 100 g per 50 kg of feed mixture. During a technical inspection, fifty samples were selected at random and the exact amount of premix was measured. The results are given in Table 15.5. Determine whether the amount of premix dispensed differs significantly from the target setting at the significance levels $\alpha = 0.05$ and $\alpha = 0.01$.

Table 15.5: Data for Example 15.3

Dose [g]	96	98	100	102	104	Total
Number of Samples	7	29	9	3	2	50

15.4 For a group of individuals, body temperature was measured before and after treatment. The measurement results are shown in Table 15.6. At the significance level 0.01 (chosen to increase the power of the test), compare the temperatures before and after treatment.



⁶The problem, including the data, is taken from [4], p. 98.

Table 15.6: Results of Body Temperature Measurements t Before and After Treatment

$t [^{\circ}\text{C}]$	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
Before	36.2	36.4	35.9	37.1	38.2	39.1	37	38.4	36.8	37.1
After	37.3	36.5	37.2	38.2	39.2	39.5	38.4	38.5	36.8	36.9

15.5 Two voltmeters were used simultaneously to measure a series of voltages, yielding different readings with a standard deviation of $\sigma = 10$ mV (see Table 15.7). At the 2% significance level, assess the hypothesis that the mean readings of the two voltmeters are equal.

Table 15.7: Results of Measurements (M) btained Usir [mV]

M 1	3.02	3.51	4	4.57	4.25	3.6	3	2.9	2.7
M 2	3.005	3.65	4.002	4.512	3.91	3.04	3.2	3.3	2.62



15.6 The objective is to determine whether an increase in temperature reduces the productivity of workers in a workshop. For this purpose, fifteen workers were tested, and their productivity during a shift (measured as the number of products manufactured per shift; see Table 15.8) was recorded. At the significance level $\alpha = 0.05$, test whether the productivity of the workers is the same under different temperature conditions.

Table 15.8: Numbers of Products Manufactured by Individual Workers as a Function of Temperature

Worker No.	19°C	23°C	Worker No.	19°C	23°C
1	52	50	9	67	61
2	58	51	10	68	57
3	59	60	11	70	64
4	65	51	12	63	64
5	71	58	13	51	50
6	55	52	14	56	51
7	59	60	15	62	53
8	63	59			

15.7 In a statistics examination, it was investigated whether students achieved the same results in the written and oral parts of the examination. Ten students were tested, and their scores are given in Table 15.9.

Table 15.9: Scores Obtained by Individual Students in the Written and Oral Parts of the Examination

Score / Student No.	1	2	3	4	5	6	7	8	9	10
Written Examination	28	16	11	24	26	30	14	16	26	35
Oral Examination	25	10	0	30	28	25	8	12	27	30

15.8 Eight tobacco plants had their second leaf removed. For each leaf, one randomly selected half was treated with preparation A and the other half with preparation B. The leaves were then inoculated with a suspension of the pathogen, and the number of lesions on each half was recorded. At the significance level $\alpha = 0.05$, test whether the numbers of lesions shown in Table 15.10 differ significantly between preparations A and B.

Table 15.10: Number of Lesions on the Leaves

Number of Lesions / Plant No.	1	2	3	4	5	6	7	8
Preparation A	9	17	31	18	7	8	20	10
Preparation B	10	11	18	14	6	7	17	5

15.9 Determine whether the weights of heads of two different cabbage varieties differ significantly (in kg, see Table 15.11). Test the hypothesis at the significance level 0.05.

Table 15.11: Weights of Cabbage Heads of the Two Varieties z_1 and z_2 (kg)

z_1	5.50	6.20	4.00	4.80	5.30	4.40	3.90	6.40	5.30
z_2	5.80	6.50	5.30	5.60	4.80	7.30	-	-	-

15.10 The effectiveness of two anticorrosive agents was compared. The first agent was applied in 10 cases and the second in 11 cases. After a prescribed period, the degree of corrosion damage was assessed, yielding the following results: $\bar{x}_1 = 82.4 \mu\text{m}$, $s_1^2 = 12.1 \mu\text{m}^2$, $\bar{x}_2 = 80.0 \mu\text{m}$, and $s_2^2 = 10.5 \mu\text{m}^2$. Compare the average effectiveness of the two agents.

15.11 It was necessary to determine whether the addition of a particular alloying additive increases the strength of steel. The results of strength measurements (in MPa) for steel without the additive and with the additive are shown in Table 15.12. At the significance level $\alpha = 0.05$, test the hypothesis that the addition of the special additive has no effect on the strength of the steel.

Table 15.12: Strength Measurements of Steel with and without the Additive (MPa)

Without Add.	5.6	5.9	5.9	5.7	5.8	5.7	6	5.5	5.7	5.5
With Add.	6.2	5.7	5.6	6	6.3	5.8	5.7	6	6	5.8

15.12 At a concrete production plant, eleven specimens manufactured from two different concrete mixtures were tested. At the significance level $\alpha = 0.05$, test whether the compressive strengths (measured in MPa) of the two mixtures can be regarded as equal.

Table 15.13: Strength Measurements of Two Concrete Mixtures (MPa)

Mixture 1	18	19	19	21	22	20	19	21	22	18	19
Mixture 2	19	20	21	20	18	22	21	19	21	20	21

15.13 Using the data collected on orange nectars from manufacturers listed in Table 15.14, test at the significance level $\alpha = 0.05$ whether the orange nectars produced by Company A

- differ significantly in their orange juice content from the products of Company B,
- contain, on average, less orange juice than the products of Company B.

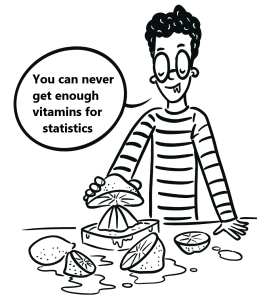


Table 15.14: Orange Juice Content of Company A and Company B Nectars

A	58.8	50.2	53.5	49.4	51.7	47.9	50.3	53.8
B	51.1	49.7	53.8	52.3	50.5	49.1	52.8	57.1

15.14 An important indicator of flax variety quality is the fiber weight measured in grams. For two flax varieties, measurements of fiber weight were obtained (Table 15.15). At the significance level $\alpha = 0.05$, test whether there are significant differences in fiber weight between the two flax varieties.

Table 15.15: Fiber Weight Measurements of Flax (g)

Variety I	47.5	57.7	47.1	38.6	45.2	49.8	43.4
	50.8	41.5	38.8	47.1	51.2	52.6	46.1
Variety II	49.2	44.1	38.1	40.9	46.4	57.1	36.8
	42.3	50.5	52.5	39.5	40.9	41.2	50.1

15.15 Twenty-six piglets of the same age were divided into two groups. The first group consisted of fourteen piglets and was fed according to feeding plan A, while the second group consisted of twelve piglets and received feeding plan B. After six months, their weight gains were recorded (see Table 15.16).

Test the equality of variances of the two samples. Then, using an appropriate statistical test, determine whether the mean weight gains of the piglets in the two groups differ significantly.

Table 15.16: Weight Gains of Piglets (kg)

Plan A	62	54	55	60	53	58	60	57	63	70	58	64	61	55
Plan B	52	56	49	50	51	47	60	53	57	60	62	70		

Solutions to the Exercises

15.1 From the problem statement, we have $n = 8$, $\bar{x} = 4.2$, and $s_x = 1.3$. We test the hypothesis $H_0: \mu = 3.5$ against $H_1: \mu > 3.5$, since the alternative is that the nicotine content exceeds the specified limit. Applying the one-sample t -test for the mean, we obtain $t \approx 1.4246$, $p = 0.01$, and critical values of approximately ± 2.9979 . Consequently, we fail to reject H_0 , namely the hypothesis that the mean nicotine content of these cigarettes does not exceed 3.5 mg. Therefore, the manufacturer does not violate the claim made in the advertisement.

15.2 From the problem statement, we have $n = 5$ and $\mu = 0$. We compute $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = -0.8$ and $s^2 = \frac{1}{n-1} (\sum_{i=1}^n x_i^2 - n\bar{x}^2) \approx 3.7$, hence $s \approx 1.9235$. We test the hypothesis $H_0: \mu = 0$ against $H_1: \mu \neq 0$.

Applying the one-sample t -test for the mean, we obtain the value of the test statistic $|t| \approx 0.93$, $p = 0.01$, and critical values of approximately ± 3.747 . The decision is to fail to reject H_0 , namely the hypothesis that the mean deviation in the filling process is 0 mg. Therefore, based on the observed data, the machine does not exhibit a systematic bias.

15.3 $\bar{x} \approx 98.56$, $s_X^2 \approx 3.435$, and the value of the test statistic is $t \approx -5.49$. The critical values are 2.009 and 2.678, respectively. Decision regarding H_0 : we reject H_0 at both significance levels $\alpha = 0.05$ and $\alpha = 0.01$. Therefore, the amount of premix dispensed in 50 kg of feed mixture differs significantly from the required target value.

15.4 We test the hypothesis H_0 : the treatment has no effect on the change in temperature against H_1 : the treatment affects the change in temperature,

i.e., the temperature either increases or decreases. Applying the paired two-sided t -test for the mean, we obtain a test statistic value of approximately -3.28 and a critical value of approximately -3.25 . Therefore, H_0 is rejected at the significance level 0.01 (the p -value of the two-sided version of the test is approximately $0.0095 < 0.01$ as well). The conclusion is that, at the chosen significance level of 0.01 , the treatment has a statistically significant effect on human body temperature.

15.5 Applying the paired two-sided t -test for the mean, we obtain a value of the test statistic of approximately 0.364 and a critical value of approximately 2.896 . The hypothesis H_0 is therefore not rejected at the significance level 0.02 (the p -value of the two-sided version of the test is approximately $0.725 > 0.02$ as well). The conclusion is that H_0 is not rejected at the 0.02 significance level. There is insufficient evidence to conclude that the mean measurements differ; in other words, the two voltmeters provide equivalent measurement results.

15.6 Paired t -test: the value of the test statistic is $t \approx 4.089$. We reject H_0 ; the differences in worker productivity are statistically significant.

15.7 Paired t -test: $t \approx 2.017$. We fail to reject H_0 that the students' performances in the two parts of the examination are equivalent.

15.8 Paired t -test: H_0 : there is no difference in the effects of preparations A and B, against H_1 : there is a difference in the effects of preparations A and B. We obtain $t \approx 2.625$, with a critical value of 2.365 . Therefore, we reject H_0 , and the sample provides evidence of a difference in the effects of preparations A and B.

15.9 Two-sample (independent-samples) t -test. We test the hypothesis H_0 : the difference in the weights of the cabbage heads is not statistically significant (i.e., there is no significant difference) against H_1 : the difference in the weights of the cabbage heads is statistically significant (i.e., there is a significant difference).

15.10 F -test for equality of variances: $H_0 : \sigma_1^2 = \sigma_2^2$, $H_1 : \sigma_1^2 \neq \sigma_2^2$, $f \approx 0.152 \not> F_{9,10}(0.975) \approx 3.799$, $p \approx 0.411$. At the significance level $\alpha = 0.05$, H_0 cannot be rejected; therefore, the variances cannot be considered significantly different.

Two-sample t -test: $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$, $|t| = |1.637| \not\geq t_{19}(0.975) \approx 2.093$. Thus, H_0 cannot be rejected at the significance level $\alpha = 0.05$. Consequently, there is no evidence that the effectiveness of the two agents differs significantly. Furthermore, $p = \mathbb{P}(|T| > 1.637) \approx 0.118$, which means that H_0 would be rejected only at a significance level of approximately 11.8% or higher.

15.11 F -test for equality of variances: $H_0 : \sigma_1^2 = \sigma_2^2$, $H_1 : \sigma_1^2 \neq \sigma_2^2$, $\sigma_1^2 \approx 0.029$, $\sigma_2^2 \approx 0.052$, $f \approx 0.556 \not\geq F_{9,9}(0.975) \approx 4.025$, $p \approx 0.478$. At the significance level $\alpha = 0.05$, H_0 cannot be rejected; thus, the variances may be regarded as equal, or equivalently, there is no evidence that the variances differ significantly.

Two-sample t -test: $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$, $\bar{x}_1 \approx 5.73$, $\bar{x}_2 \approx 5.91$, $|t| = |-1.998| \not\geq t_{18}(0.975) \approx 2.1$. The null hypothesis H_0 cannot be rejected at the significance level $\alpha = 0.05$. Therefore, it has not been demonstrated that the additive produces a statistically significant increase in the strength of the steel. Furthermore, $p = \mathbb{P}(|T| > 1.998) \approx 0.061$, which means that H_0 would not be rejected until the significance level reaches approximately 6.1% or higher.

15.12 F -test for equality of variances: $H_0 : \sigma_1^2 = \sigma_2^2$, $H_1 : \sigma_1^2 \neq \sigma_2^2$, $\sigma_1^2 \approx 2.164$, $\sigma_2^2 \approx 1.364$, $f \approx 1.587 \not\geq F_{10,10}(0.975) \approx 3.717$, $p \approx 0.478$. At the significance level $\alpha = 0.05$, H_0 cannot be rejected; therefore, the variances are not significantly different.

Two-sample t -test: $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$, $\bar{x}_1 \approx 19.818$, $\bar{x}_2 \approx 20.182$, $|t| = |-0.642| \not\geq t_{20}(0.975) \approx 2.086$. The null hypothesis H_0 cannot be rejected at the significance level $\alpha = 0.05$. Therefore, it has not been demonstrated that the strength, and hence the quality, of the two concrete mixtures differs significantly.

Furthermore, $p = \mathbb{P}(|T| > 0.642) \approx 0.528$, which means that H_0 would not be rejected at any practically meaningful significance level (up to approximately 52.8% and higher).

15.13 F -test for equality of variances: $H_0 : \sigma_A^2 = \sigma_B^2$, $H_1 : \sigma_A^2 \neq \sigma_B^2$, $n_A = n_B = 8$, $\sigma_A^2 \approx 11.642$, $\sigma_B^2 \approx 6.674$, $f \approx 1.744 \not\geq F_{7,7}(0.975) \approx 4.995$, $p \approx 0.48$. At the significance level $\alpha = 0.05$, H_0 cannot be rejected; therefore, the variances are not significantly different.

Two-sample t -test: $H_0 : \mu_A = \mu_B$, $H_1 : \mu_A \neq \mu_B$, $\bar{x}_A \approx 51.95$, $\bar{x}_B \approx 52.05$, $|t| = |-0.066| \not\geq t_{14}(0.975) \approx 2.145$. The hypothesis H_0 cannot be rejected at the significance level $\alpha = 0.05$. Thus, it has not been demon-

strated that the two nectars differ significantly in their orange juice content, respectively that the quality of the two mixtures differs significantly. Furthermore, $p = \mathbb{P}(|T| > 0.066) \approx 0.948$, which means that H_0 cannot be rejected at any practically meaningful significance level (up to a level of 94.8% or higher).

15.14 Two-sample t -test without assuming equal variances of the two samples. H_0 : The flax fiber weights are equal for the two flax varieties. Against H_1 : The flax fiber weights differ between the two flax varieties. We obtain $t \approx 0.92$, with a critical value of 2.05. Therefore, we fail to reject H_0 concerning the equality of flax fiber weights for the two flax varieties.

15.15 $s_X^2 = 21.14$, $s_Y^2 = 43.53$, $\bar{x} \approx 59.285$, $\bar{y} \approx 55.583$, $t \approx 1.679$. We fail to reject H_0 concerning the equality of the mean weight gains of piglets under feeding plans A and B.

Chapter 16

Nonparametric Tests

Parametric tests are based on several assumptions. One of these is the assumption that the sample originates from a distribution whose form is known up to certain parameters. If the underlying distribution is normal, we use the parametric tests introduced in Chapter 15. Moreover, violations of normality do not affect the conclusions of these tests when the sample size is sufficiently large, because in such cases we can rely on the Central Limit Theorem and the laws of large numbers. In practice, however, we often encounter small samples drawn from markedly “non-normal” populations. In such situations, the parametric tests presented in Chapter 15 cannot be applied, because the corresponding test statistics do not possess the required distributions (Student’s, Pearson’s, or Fisher–Snedecor’s distributions). In these cases, we employ so-called nonparametric tests. These tests require only very general assumptions and are mathematically less demanding. The price paid for this generality, however, is their lower statistical power in comparison with parametric tests.

Before introducing several nonparametric tests, we first define the concept of a rank. Let $x_1, x_2, \dots, x_n$ be distinct real numbers. The rank R_i of the number x_i is defined as the number of values among $x_1, x_2, \dots, x_n$ that are less than or equal to x_i . For example, consider the numbers 5, 8, 9, 3, 2, 1. The number 5 has rank 4, because the numbers 5, 3, 2, and 1 are less than or equal to 5. The corresponding ranks are summarized in Table 16.1.

Table 16.1: Table of Selected Numbers with Their Assigned Ranks

Numbers x_i	5	8	9	3	2	1
Rank R_i	4	5	6	3	2	1

It may happen that the numbers $x_1, x_2, \dots, x_n$ are not distinct, that is, some of them are equal and form so-called ties. In such a case, all observations involved in a tie are assigned the average of the ranks corresponding to that tied group. For example, the numbers 5, 5, 5, 9, 9, 2, 1 are assigned the ranks $R_1, R_2, \dots, R_7$ shown in Table 16.2.

Table 16.2: Table of Values with Tied Ranks

Numbering of Values x_i	1	2	3	4	5	6	7
Ascending ordered values x_i	1	2	5	5	5	9	9
Rank R_i	1	2	4	4	4	6.5	6.5

For determining the ranks of observations in a data set of arbitrary size, the MS Excel function `= RANK.AVG` can be used. This function returns the rank of a specified value within a list of numbers. The first argument is the value whose rank is to be determined, the second argument is the range of data within which the ranking is performed, and the third argument allows the choice between 1 and 0, corresponding to ascending or descending order, respectively. If multiple values share the same rank, the function returns their average rank.

16.1 Sign Test

Let $X_1, X_2, \dots, X_n$ be a random sample from a distribution with a continuous cumulative distribution function, and let $\tilde{\mu}$ denote the median of this distribution. Then,

$$\mathbb{P}(X < \tilde{\mu}) = \mathbb{P}(X > \tilde{\mu}) = \frac{1}{2}.$$

We wish to test the hypothesis

$$H_0: \tilde{\mu} = \tilde{\mu}_0 \quad \text{against the alternative hypothesis} \quad H_1: \tilde{\mu} \neq \tilde{\mu}_0,$$

where $\tilde{\mu}_0$ is a given constant (most commonly equal to zero). We first form the differences

$$X_1 - \tilde{\mu}_0, X_2 - \tilde{\mu}_0, \dots, X_n - \tilde{\mu}_0.$$

The random variable Y then denotes the number of differences having a positive sign. Under the null hypothesis H_0 , the random variable Y follows a binomial distribution with parameters n and $\frac{1}{2}$. For this reason, the sign test is also referred to as the binomial test. In the two-sided version of the

Table 16.3: Amount of Milk Produced [l] and the Calculation of Difference Ranks for Example 16.1

Cow No.	1	2	3	4	5	6	7	8	9	10
Lactation I	14	12	15	12	12	14	11	12	15	10
Lactation II	15	16	19	15	14	11	11	15	14	13
Difference	-1	-4	-4	-3	-2	3	0	-3	1	-3
Sign of the Difference	-	-	-	-	-	+		-	+	-

test, the critical region consists of both very small values of Y (i.e., values close to zero) and very large values of Y (i.e., values close to n). For small sample sizes (that is, for small values of n), the critical values of the test, denoted by k_1 and k_2 , are determined so that

$$\mathbb{P}(Y \leq k_1) \leq \frac{\alpha}{2}, \quad \mathbb{P}(Y \geq k_2) \leq \frac{\alpha}{2}.$$

The critical values k_1 and k_2 can be found in statistical tables; for $\alpha = 0.05$ and $\alpha = 0.01$, they are listed in Table 21.29 on page 463. The null hypothesis H_0 is therefore rejected whenever it is observed that $Y \leq k_1$ or $Y \geq k_2$. For large random samples (it is sufficient that $n \geq 20$), we compute the statistic

$$Z = \frac{2Y - n}{\sqrt{n}}.$$

Under the null hypothesis H_0 , the statistic Z has, by the CLT, an asymptotic standard normal distribution $N(0, 1)$. Therefore, the hypothesis H_0 is rejected whenever

$$|z| \geq u\left(1 - \frac{\alpha}{2}\right).$$

The sign test can also be applied as a paired test. Unlike the paired t -test, it is not necessary to know the exact values of X_i and Y_i , $i = 1, 2, \dots, n$, in order to perform the sign test; it is sufficient to know whether the difference $X_i - Y_i$ is positive or negative. For this reason, the sign test can also be used when only qualitative comparisons are available, for example when treatment A is judged to be more effective than treatment B. In the sign test, some differences may be equal to zero. For instance, in a qualitative comparison, a subject may be unable to decide in favor of either alternative. In such cases, it is recommended to omit the zero differences and take n to be the number of non-zero differences only. The sign test is most commonly used for the preliminary evaluation of pilot studies.

Example 16.1.  For ten cows of the same breed, the amount of colostrum produced on the third day after calving was measured during the first and

second lactation periods (see the first three rows of Table 16.3). Determine whether there is a statistically significant difference between the observed data.

Solution. The last row of Table 16.3 contains the signs of the differences¹; the zero difference for Cow 7 is omitted, hence $n = 9$. The number of positive signs (the value of the test statistic Y) is $y = 2$. The lower critical value for $\alpha = 0.05$ and $n = 9$ is $k_1 = 1$ (see Table 21.29 on page 463). Owing to symmetry, the upper critical value is $k_2 = n - k_1 = 8$; for simplicity, it is therefore not listed in the table of critical values. Since $2 \not\leq 1$ and $2 \not\geq 8$, we fail to reject the null hypothesis H_0 . Thus, according to the sign test, the difference in the amount of milk produced between the two lactation periods is not statistically significant. $\triangle$

16.2 One-Sample Wilcoxon Test

The one-sample Wilcoxon² test is the nonparametric counterpart of the one-sample Student's t -test and the paired Student's t -test. Its implementation is somewhat more involved than that of the sign test; however, it is also more sensitive.

When applying the Wilcoxon test as a nonparametric version of the one-sample t -test, assume that $X_1, X_2, \dots, X_n$ is a random sample from a continuous distribution with cumulative distribution function $F(x)$ and a density symmetric about zero, i.e.

$$F(x) = 1 - F(-x), \quad -\infty < x < \infty.$$

In this case, zero is the median of the underlying distribution. We test the hypothesis

$$H_0: \tilde{\mu} = 0 \quad \text{against the alternative} \quad H_1: \tilde{\mu} \neq 0,$$

or, alternatively, against the one-sided alternatives

$$H_1: \tilde{\mu} < 0, \quad \text{respectively} \quad H_1: \tilde{\mu} > 0.$$

Arrange $X_1, X_2, \dots, X_n$ in ascending order according to the magnitudes of their absolute values, i.e.

$$|X|^{(1)} < |X|^{(2)} < \dots < |X|^{(n)}.$$

¹In Excel, these can also be determined using the `=SIGN` function.

²Frank Wilcoxon (1892–1965), Irish-born American chemist and statistician.

With this ordering, let R_i^+ denote the rank of $|X_i|$, and define the quantities

$$S^+ = \sum_{X_i \geq 0} R_i^+, \quad S^- = \sum_{X_i < 0} R_i^+$$

which represent the sum of the ranks corresponding to the nonnegative values X_i and the negative values, respectively. If the quantities S^+ and S^- have been determined correctly, then

$$S^+ + S^- = \frac{n(n+1)}{2},$$

since the ranks from 1 to n are being summed.

To assess the validity of the null hypothesis against a two-sided alternative, we use the statistic $W = \min(S^+, S^-)$. If this statistic is less than or equal to the tabulated critical value, the null hypothesis is rejected. For a left-sided alternative, the test statistic is $W = S^+$, whereas for a right-sided alternative, the test statistic is $W = S^-$. The critical values are given in statistical tables. For $n \geq 30$, we use the test statistic Z , which has an asymptotic standard normal distribution $N(0, 1)$ and is given by

$$Z = \frac{S^+ - \frac{1}{4}n(n+1)}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}.$$

In the case where

$$|z| \geq u\left(1 - \frac{\alpha}{2}\right),$$

we reject the hypothesis at a significance level that is asymptotically equal to α . Finally, it should be noted that the Wilcoxon test is based on the assumption of symmetry; if the underlying distribution is substantially asymmetric, the test may reject the null hypothesis even when the median is equal to zero.

Example 16.2. In a fatigue test of metals conducted at a stress level of 560 MPa, the following numbers of cycles to failure (in thousands) were observed in eight trials: 3 322, 14 713, 763, 46 296, 2 845, 9 411, 1 532, 24 023. At the significance level $\alpha = 0.05$, test the hypothesis $H_0: \tilde{\mu} = 8 \cdot 10^6$ against the alternative $H_1: \tilde{\mu} > 8 \cdot 10^6$.

Solution. The one-sample Wilcoxon test was introduced above for testing a zero median. Under the null hypothesis, this condition is satisfied by the variable $Z_i = X_i - \tilde{\mu}$, whose values and corresponding ranks R_i are given in Table 16.4. The values a_i assign the value 1 or 0 to the difference $X_i - \tilde{\mu}$

Table 16.4: Results of Fatigue Testing of Metals (Thousands of Cycles to Failure)

X_i	$Z_i = X_i - 8\,000$	R_i	a_i
3 322	-4 678	2	0
14 713	6 713	5	1
763	-7 237	6	0
46 296	38 296	8	1
2 845	-5 155	3	0
9 411	1 411	1	1
1 532	-6 468	4	0
24 023	16 023	7	1

according to whether it is positive or negative, respectively. We test against the right-sided alternative $H_1: \tilde{\mu} - 8 \cdot 10^6 > 0$, for which small values of S^- provide evidence in favor of the alternative. Therefore, in the right-sided version of the test, we use S^- as the test statistic and compare it with the critical value of the one-sided test. From Table 16.4, we obtain $s^- = 2 + 6 + 3 + 4 = 15$, so the test statistic for the right-sided version of the test is $w = s^- = 15$. The critical value is $w_8(0.05) = 5$. Since $15 \not\leq 5$, we fail to reject H_0 concerning the number of cycles to failure. $\triangle$

When applying the one-sample Wilcoxon test as the nonparametric counterpart of the paired t -test, assume that $(X_1, Y_1), \dots, (X_n, Y_n)$ is a random sample from a continuous bivariate distribution. We test the hypothesis that the medians of the two random variables are equal against a two-sided alternative or, respectively, against one-sided alternatives: $H_0: \tilde{\mu}_X - \tilde{\mu}_Y = 0$ against $H_1: \tilde{\mu}_X - \tilde{\mu}_Y \neq 0$, or, respectively, $H_1: \tilde{\mu}_X - \tilde{\mu}_Y < 0$, or $H_1: \tilde{\mu}_X - \tilde{\mu}_Y > 0$. We compute the differences $Z_i = X_i - Y_i$, $i = 1, 2, \dots, n$, and then proceed in the same way as for the nonparametric version of the one-sample t -test.

Example 16.3.  A special mental arithmetic training program was tested on eleven pupils. Table 16.5 shows the times (in seconds) required by the pupils to solve control tasks before and after the training. Can we conclude that the training improves the pupils' problem-solving ability at the significance level $\alpha = 0.05$?

Solution. We test the hypothesis $H_0: \tilde{\mu}_X - \tilde{\mu}_Y = 0$ (the training has no effect on problem-solving ability) against the alternative $H_1: \tilde{\mu}_X - \tilde{\mu}_Y > 0$ (the

Table 16.5: Results of the Time Required to Solve the Control Tasks in Example 16.3

Before (x_i)	87	61	98	90	93	74	83	72	81	75	83
After (y_i)	50	45	79	90	88	65	52	79	84	61	52
Differences	37	16	19	0	5	9	31	-7	-3	14	31
Absolute Values	37	16	19	0	5	9	31	7	3	14	31
Rank	11	7	8	1	3	5	9.5	4	2	6	9.5

training improves the pupils' problem-solving ability). The differences computed in Table 16.5 are not normally distributed (the reader may construct a histogram or verify normality using the test described in Section 20.2); therefore, we apply the nonparametric Wilcoxon test. We compute the statistics

$$s^+ = 1 + 3 + 5 + 6 + 7 + 8 + 9 + 10 + 11 = 60, \quad s^- = 2 + 4 = 6,$$

$$w = \min(s^+, s^-) = 6$$

The critical value of the one-sided right-tailed one-sample Wilcoxon test is $w_{11}(0.05) = 13$ (see the table of critical values on page 465), from which it follows that we reject the null hypothesis H_0 at the significance level 0.05.

This problem can also be solved using the sign test; however, in doing so, not all of the available information is utilized, which may lead to misleading conclusions. The number of positive differences is $y = 9$. The critical values of the sign test are $k_1 = 1$ and $k_2 = 10$. Since $k_1 < y < k_2$, we fail to reject the null hypothesis H_0 . In this case, the two tests lead to different conclusions. The sign test does not provide sufficient evidence to reject H_0 , because it uses only the number of negative values, whereas the Wilcoxon test additionally exploits the information that the negative values are relatively small. We therefore say that the Wilcoxon test is more powerful than the sign test. $\triangle$

16.3 Two-Sample Wilcoxon Test

This test is most commonly used as a substitute for the two-sample t -test. Once again, the assumptions imposed on the distribution functions of the random samples are generalized. Let $X_1, X_2, \dots, X_m$ and $Y_1, Y_2, \dots, Y_n$ be two independent samples drawn from continuous distributions. We wish to test the hypothesis that the distribution functions of the two populations are identical. The observations $X_1, X_2, \dots, X_m$ and $Y_1, Y_2, \dots, Y_n$ are pooled

Table 16.6: Lengths of Larvae in Rice and Millet and the Corresponding Ranks for Example 16.4

Larval Length in Rice	7	10	14	15	12	16	12
Larval Length in Millet	11	12	16	13	18	15	-
Ranks of Larval Lengths in Rice	1	2	8	9.5	5	11.5	5
Ranks of Larval Lengths in Millet	3	5	11.5	7	13	9.5	-

and arranged together as a single sample in ascending order. We then compute the sum of the ranks of the observations $X_1, X_2, \dots, X_m$ within the combined sample and denote this sum by T_1 . Similarly, we compute the sum of the ranks of the observations $Y_1, Y_2, \dots, Y_n$ and denote it by T_2 . We then calculate

$$U_1 = mn + \frac{m(m+1)}{2} - T_1, \quad U_2 = mn + \frac{n(n+1)}{2} - T_2.$$

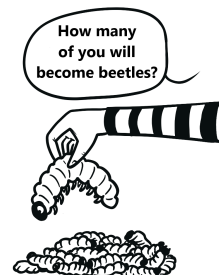
Since $T_1 + T_2 = \frac{(m+n+1)(m+n)}{2}$, it follows that $U_1 + U_2 = mn$. If $\min(u_1, u_2)$ is less than or equal to the critical value given in the table, we reject the hypothesis that the distribution functions of the two populations are identical. For large sample sizes, we again switch to a test statistic that, under the null hypothesis, has an asymptotic standard normal distribution $N(0, 1)$ and is given by

$$Z = \frac{U_1 - \frac{mn}{2}}{\sqrt{\frac{mn}{12}(m+n+1)}}.$$

If $|z| \geq u(1 - \frac{\alpha}{2})$, we reject the hypothesis at a significance level asymptotically equal to α .

Example 16.4.  Table 16.6 contains the lengths of beetle larvae living in winter rice and millet seeds.³ Compare the distributions of these lengths at the significance level $\alpha = 0.05$.

Solution. The sum of the ranks for the rice sample is $t_1 = 42$, and for the millet sample it is $t_2 = 49$. Furthermore, $n_1 = 7$ and $n_2 = 6$. We compute $u_1 = 42 + \frac{56}{2} - 42 = 28$, $u_2 = 42 + \frac{42}{2} - 49 = 14$. The critical value of the two-sample Wilcoxon test at the significance level $\alpha = 0.05$ for sample sizes 7



³The problem, including the data, was taken from [17], p. 185.

and 6 is $w_{7,6}(0.05) = 6$ (see the table on page 466). We fail to reject the hypothesis that the distributions of the two samples are identical, because $\min(u_1, u_2) \not\leq w_{7,6}(0.05)$. $\triangle$

16.4 Exercises

16.1 Assess the effectiveness of a new training method based on the differences between the maximum performances of members of a ski-jumping team before the introduction of the new method and after three months of its use. The performance differences for the individual ski jumpers are given in Table 16.7.

Table 16.7: Differences in Performance (cm)

Jumper No.	1	2	3	4	5	6	7	8
Performance difference	+5	-10	+15	-1	+8	+21	+7	-6
Jumper No.	9	10	11	12	13	14	15	16
Performance difference	+4	+12	-3	+7	+9	-2	+3	+7

16.2 For twenty randomly selected shipments, the number of days between delivery of the ordered shipment to the customer's residence and its collection by the customer was recorded: 1, 2, 3, 0, 5, 7, 10, 4, 5, 0, 7, 8, 5, 4, 1, 12, 3, 5, 10, 1. At the 5% significance level, test the null hypothesis that the median delivery collection time is 5 days.

16.3 Urine samples were collected from sixteen patients. Furantoin (F) was applied to the first half of each sample and Penicillin (P) to the second half. After 24 hours, the number of bacteria was determined in the first and second half of each sample. It was found that in ten cases the bacterial count was lower for treatment F than for treatment P (i.e., $F < P$), in four cases the bacterial count was higher for treatment F than for treatment P (i.e., $F > P$), and in two cases no difference could be determined (i.e., $F = P$). Determine whether Furantoin and Penicillin differ in their effectiveness. Using the sign test, test the hypothesis that treatments F and P have the same effect.

16.4 For eighteen restaurant customers, the waiting time to place an order (d_o) and the waiting time for the waiter to arrive so that payment could be made (d_z) were measured. The results, expressed in minutes, are presented in Table 16.8. At the significance level 0.05, test the hypothesis that the two waiting times are equal.

Table 16.8: Waiting Times in Minutes for Ordering and Paying in a Restaurant

d_o	7, 0, 5, 2, 4, 5, 3, 7, 9, 3, 1, 4, 3, 5, 5, 4, 6, 3
d_z	5, 4, 6, 3, 8, 5, 2, 6, 3, 7, 2, 8, 2, 7, 1, 8, 2, 2

16.5 Urine samples were collected from eight patients and each sample was divided into two equal parts. The first half was treated with substance A and the second half with substance B. Subsequently, differences in the numbers of bacterial colonies in the A and B series were determined. The results are presented in Table 16.9. Evaluate the results of testing the samples treated with substances A and B.

Table 16.9: Numbers of Bacterial Colonies in Samples Treated with Substances A and B

A	2, 0, 4, 5, 2, 1, 2, 3
B	1, 2, 6, 3, 5, 3, 0, 4

16.6 At the 5% significance level, compare two methods for measuring compression pressure (in 10^5 Pa) in the combustion chamber of a tractor engine. For six engines, compression pressure was measured using both methods, and the differences between the obtained results were determined (see Table 16.10).

Table 16.10: Results of Compression Pressure Measurements (10^5 Pa)

Engine No.	1	2	3	4	5	6
Diference d_i	-0.1	-0.2	0.2	0.1	-0.2	-0.1

16.7 Test the equality of the distributions of two independent samples V_1 and V_2 (see Table 16.11), representing the mathematics test scores of pupils.

Table 16.11: Mathematics Test Scores of the Pupils

V_1	17, 15, 13, 2, 5, 11, 30, 24, 19, 23, 21, 24, 29, 25, 15, 22, 6, 8
V_2	5, 14, 26, 30, 18, 15, 2, 26, 29, 17, 12, 8, 12, 7, 21, 4, 12, 22

16.8 Using a nonparametric test, determine whether two different post-harvest drying methods affect the germination rate of malting barley. For each method, ten samples containing one hundred seeds were taken, yielding the results shown in Table 16.12.⁴

⁴The problem, including the data, was taken from [19], p. 59.

Table 16.12: Numbers of Germinated Seeds under Drying Method A and Drying Method B

Sample No.	1	2	3	4	5	6	7	8	9	10
Method A	89	88	92	94	90	81	93	90	87	83
Method B	88	77	83	76	79	82	80	72	78	72

16.9 On four combine harvesters, a new sowing method was tested, while the remaining six combine harvesters used the standard sowing method. After harvest, the field yields in metric centners per hectare were recorded, as shown in Table 16.13. Determine whether the new sowing method affects the average field yield per hectare compared with the original sowing method.

Table 16.13: Field Yields in Metric Centners per Hectare

New Sowing Method	51	52	49	55		
Original Sowing Method	45	54	48	44	53	50

Solutions to the Exercises

16.1 Sign test: H_0 : The new training method is equally effective as the original method. Against H_1 : The new training method is more effective than the original method.

$m_+ = 11$, $m_- = 5$, $n = 16$, and the test statistic is $m = \min(11, 5) = 5$.

The critical value for $\alpha = 0.05$ and $n = 16$ paired observations is 4. Since $5 > 4$, we fail to reject H_0 .

16.2 $m_+ = 6$, $m_- = 10$, $n = 16$, test statistic: $m = \min(6, 10) = 6$. The critical value for $\alpha = 0.05$ and $n = 16$ observations is 3. Since $m > 3$, i.e. $6 > 3$, $\Rightarrow$ we fail to reject the hypothesis that the median is equal to 5. In other words, no statistically significant difference ($p < 0.05$) between the observed values and the hypothesized median was demonstrated.

16.3 $m_+ = 10$, $m_- = 4$, $n = 14$, test statistic: $m = \min(10, 4) = 4$. The critical value for $\alpha = 0.05$ and $n = 14$ paired observations is 2. Since $m > 2$, that is, $4 > 2 \Rightarrow$ we fail to reject the hypothesis that the two treatments have the same effect. In other words, no statistically significant difference ($p < 0.05$) in the effectiveness of Furantoin and Penicillin on bacterial growth in the patients' urine samples has been demonstrated.

16.4 $m_+ = 8$, $m_- = 9$, $n = 17$, test statistic: $m = \min(8, 9) = 8$.

The critical value for $\alpha = 0.05$ and $n = 17$ paired observations is 4. Since $m > 4$, that is, $8 > 4$, we fail to reject the hypothesis that the two waiting times are equal. In other words, no statistically significant difference ($p < 0.05$) between the two waiting times has been demonstrated.

16.5 $n = 8$, there are no zero differences, the sum of ranks corresponding to the positive differences is $w^+ = 11.5$, the sum of ranks corresponding to the negative differences is $w^- = 24.5$. The test statistic is $w = \min(11.5, 24.5) = 11.5$. The critical value is $w_8(0.05) = 3$. Since $11.5 > 3$, we fail to reject the null hypothesis of equal effects of substances A and B. In other words, no significant difference in the effects of substances A and B on the collected samples has been demonstrated.

16.6 The measurement results are assigned ranks according to Table 16.14, and the one-sample Wilcoxon test is applied.

Table 16.14: Results of Compression Pressure Measurements (10^5 Pa)

Engine No.	1	2	3	4	5	6
Difference d_i	-0.1	-0.2	0.2	0.1	-0.2	-0.1
Rank d_i^+			5	2		
Rank d_i^-	2	5			5	2

$w^+ = 5 + 2 = 7$, $w^- = 2 + 5 + 5 + 2 = 14$, the critical value of the Wilcoxon test for $\alpha = 0.05$ and $n = 7$ is 2, while the test statistic is $w = \min(7, 14) = 7$. Since $7 > 2$, we fail to reject the null hypothesis that there is no significant difference between the two methods of measuring compression pressure.

16.7 Wilcoxon test for two independent samples: $n_1 = n_2 = 18$, $T_1 = 352$, $T_2 = 314$, $U_1 = 143$, $U_2 = 181$, and the test statistic is $W = \min(143, 181) = 143$.

The critical value is approximately 200. Since $143 < 200$, we reject the null hypothesis H_0 that the distributions (distribution functions) of the two samples are identical.

Thus, the students' results follow different distributions and therefore cannot be regarded as equivalent.

16.8 For the two-sample Wilcoxon test, ranks are assigned to all observations (see Table 16.15). We then compute the sums of ranks in the first and second groups and subsequently calculate the value of the test statistic.

Table 16.15: Numbers of Germinated Seeds under Drying Method A and Drying Method B

Sample No.	1	2	3	4	5	6	7	8	9	10
Method A	89	88	92	94	90	81	93	90	87	83
Rank x	15	13.5	18	20	16.5	8	19	16.5	12	10.5
Method B	88	77	83	76	79	82	80	72	78	72
Rank y	13.5	4	10.5	3	6	9	7	2	5	1

The sums of ranks are $T_x = 149$ and $T_y = 61$, yielding $U_x = 6$ and $U_y = 94$. The critical value of the test statistic for $\alpha = 0.05$ and $n_1 = n_2 = 10$ is 23.

Since $6 < 23$, we reject the null hypothesis. Therefore, there is a significant difference in seed germination rates depending on the drying method used.

16.9 After assigning ranks to all ordered observations (see Table 16.16), we obtain $T_1 = 27$, $T_2 = 28$, $U_1 = 7$, $U_2 = 17$, with $m = 4$, $n = 6$, and $\alpha = 0.05$. Since

$$\min(7, 17) = 7,$$

and the critical value is 2, we have

$$7 > 2.$$

Therefore, we fail to reject the null hypothesis. We conclude that the new sowing method does not have a greater effect on the average yield per hectare than the original sowing method.

Table 16.16: Field Yields in Metric Centners per Hectare

Field Yield	44	45	48	49	50	51	52	53	54	55
Rank	1	2	3	4	5	6	7	8	9	10

Chapter 17

Comparison of Multiple Samples

17.1 One-Way Analysis of Variance

This test is a generalization of the two-sample t -test to the case of ($I \geq 3$) samples. Consider therefore I independent samples,

$$\begin{aligned} Y_{11}, \dots, Y_{1n_1} & \text{ is a sample from } N(\mu_1, \sigma^2), \\ Y_{21}, \dots, Y_{2n_2} & \text{ is a sample from } N(\mu_2, \sigma^2), \\ & \text{and so on up to} \\ Y_{I1}, \dots, Y_{In_I} & \text{ is a sample from } N(\mu_I, \sigma^2). \end{aligned}$$

We wish to test the hypothesis

$$H_0: \mu_1 = \dots = \mu_I$$

against the alternative

$$H_1: \text{At least two of the means are not equal.}$$

Sometimes the above situation is expressed by the model

$$Y_{ij} = \mu + \alpha_i + e_{ij},$$

where $\mu + \alpha_i = \mu_i$ and $e_{ij} \sim N(0, \sigma^2)$ denotes the error arising either from measurement inaccuracy or from random variation about the mean. The

hypothesis H_0 can then be rewritten in terms of a simpler model, which is satisfied whenever the hypothesis H_0 :

$$Y_{ij} = \mu + e_{ij}.$$

The test is carried out as follows. First, let us denote the sample means of the individual samples by

$$\bar{Y}_i = \frac{Y_{i1} + \cdots + Y_{in_i}}{n_i},$$

for $i = 1, \dots, I$, and let the overall mean of all observations be denoted by

$$\bar{Y} = \frac{\sum_i \sum_j Y_{ij}}{n},$$

where $n = n_1 + \cdots + n_I$. Next, we compute the total sum of squares S_T (i.e., the total quadratic error of the model under the null hypothesis H_0 , that is, in the case where $\mu_1 = \mu_2 = \cdots = \mu_I = \mu$). The parameter μ is estimated by the overall mean $\bar{Y}$.

$$S_T = \sum_i \sum_j (Y_{ij} - \bar{Y})^2 = \sum_i \sum_j Y_{ij}^2 - n\bar{Y}^2.$$

The residual sum of squares S_e is the total quadratic error of the model under the assumption that the null hypothesis H_0 does not hold, that is, in the case where $\mu_1 \neq \cdots \neq \mu_I$. The parameter μ_i is estimated by $\bar{Y}_i$.

$$S_e = \sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2 = \sum_i \sum_j Y_{ij}^2 - \sum_i n_i \bar{Y}_i^2.$$

The quantity $S_A = S_T - S_e$ is interpreted as the sum of squares attributable to treatment differences. This quantity is always nonnegative, because the error of the more general model, S_e , is always smaller than the error of the simpler model, S_T . If S_A is small, the two models are similar, and therefore we will not reject the hypothesis H_0 . If S_A is large, the more general model explains a substantial portion of the total variation S_T , and consequently we reject H_0 .

Under the null hypothesis H_0 , the statistic

$$F_A = \frac{S_A/(I-1)}{S_e/(n-I)} \sim F_{I-1, n-I}$$

follows the Fisher–Snedecor distribution F with $I-1$ and $n-I$ degrees of freedom. Therefore, the hypothesis H_0 is rejected at the significance level α whenever

$$f_A \geq F_{I-1, n-I}(1 - \alpha).$$

Table 17.1: Table of Analysis of Variance Results

Variation	Sum of Squares S	Degrees of Freedom df	Mean Square S/df	F
Treatments	S_A	$df_A = I - 1$	S_A/df_A	F_A
Residual	S_e	$df_e = n - I$	S_e/df_e	-
Total	S_T	$df_T = n - 1$	-	-

The results of the entire test are customarily summarized in tabular form (see Table 17.1).

The assumptions of this test are similar to those of the two-sample t -test (see page 310). The most important assumption is, again, the independence of the individual samples. The normality assumption may be violated provided that the sample sizes are sufficiently large to justify the use of the Central Limit Theorem (CLT). If this is not the case, it is preferable to apply the nonparametric counterpart of this test, known as the Kruskal–Wallis test. A final assumption is the equality of variances across all samples. If the hypothesis of equal variances (homoscedasticity) is rejected, one may use either the Kruskal–Wallis test or a more general weighted version of ANOVA that allows unequal variances to be modeled. Equality of variances can be assessed, for example, using Bartlett’s test or Levene’s test (see, e.g., [2]).

The quantity $S^2 = \frac{S_e}{(n-I)}$ is called the residual variance and is an unbiased estimator of the variance σ^2 .

At first sight, it might appear that the above test could be replaced by a collection of two-sample t -tests performed for every pair of samples. However, this would require $\frac{I(I-1)}{2}$ separate tests. If each were carried out at the significance level α , the overall significance level would become substantially larger than α . On the other hand, if the significance level of each individual test were reduced to $\frac{2\alpha}{I(I-1)}$, the overall significance level would become considerably smaller than α . It turns out that this approach does not lead to satisfactory results.

When the hypothesis H_0 is rejected, it is often necessary to determine for which pairs of indices the relation $\mu_i \neq \mu_j$ holds. This problem is addressed by Tukey’s¹ multiple comparison procedure.

Since $\bar{Y}_i$ is an estimator of μ_i , a table of pairwise differences $\bar{Y}_i - \bar{Y}_j$ is first constructed (see Table 17.2).

¹John Wilder Tukey (1915–2000), American mathematician and statistician.

Table 17.2: Differences between Means

i	j			
	2	3	...	I
1	$\bar{Y}_1 - \bar{Y}_2$	$\bar{Y}_1 - \bar{Y}_3$	...	$\bar{Y}_1 - \bar{Y}_I$
2		$\bar{Y}_2 - \bar{Y}_3$	...	$\bar{Y}_2 - \bar{Y}_I$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$I - 1$				$\bar{Y}_{I-1} - \bar{Y}_I$

The statistic

$$\frac{|\bar{Y}_i - \bar{Y}_j|}{S\sqrt{\frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim q_{I,n-I}$$

follows a distribution known as the studentized range distribution. The critical value $q_{I,f}(\alpha)$ of the studentized range distribution is defined as the number satisfying

$$\mathbb{P}[Q \geq q_{I,f}(\alpha)] = \alpha.$$

These critical values are tabulated. Therefore, if

$$|\bar{y}_i - \bar{y}_j| \geq s q_{I,n-I}(\alpha) \sqrt{\frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)},$$

we reject the hypothesis that $\mu_i = \mu_j$. If this procedure is applied to all pairs of groups, the overall significance level of the test is less than or equal to α . Equality holds when all samples have the same size.

To facilitate the interpretation of the results of Tukey's method, an asterisk is added in Table 17.2 to each difference that is significant at the 0.05 level. Two asterisks indicate significance at the 0.01 level, and three asterisks indicate significance at the 0.001 level.

17.1.1 Levene's Test

The so-called Levene's² test is used to assess the equality of variances for a variable measured in two or more samples. Assuming that the individual samples are independent, we test the hypothesis

$$H_0: \sigma_1^2 = \sigma_2^2 = \dots$$

²Howard Levene (1914–2003), American statistician and geneticist.

Table 17.3: Table of Wheat Yields from Five Fields

Year	Field No. / Yield (tonnes)				
	1	2	3	4	5
2024	105	98	95	101	105
2023	89	121	100	102	104
2022	120	110	99	98	99
2021	111	98	98	105	104
2020	100	130	100	95	92
2019	98	110	120	110	100
2018	97	120	104	94	92
2017	125		90	112	98
2016	103		99		100
2015	90				102

against the alternative

H_1 : The variances of at least one pair of random samples are unequal.

For the original data set Y_{ij} , we compute

$$Z_{ij} = |Y_{ij} - \bar{Y}_i|,$$

that is, the absolute differences between the observed values and the mean of the sample to which they belong, where $\bar{Y}_i$ denotes the arithmetic mean of the observations in the i th sample. A one-way ANOVA is then applied to the resulting table of transformed data.

Bartlett's test may also be used to assess the equality of variances. The forms of the test statistics for both tests can be found, for example, in [2].

Example 17.1.  Table 17.3 contains data on wheat yields (in tonnes) from five fields over the last 10 years. However, not all farmers provided all of the required data. At the significance level 0.05, determine whether there are significant differences among the yields.

Solution. Using Levene's test, we first test the equality of variances across all samples. We compute the mean yields for each field over all years. We then construct a new table containing the absolute values of the differences between the original observations and the mean of the corresponding sample

Table 17.4: Table for Levene's Test (the data are absolute differences rounded to three decimal places)

Year	Field No. / Yield (tonnes)				
	1	2	3	4	5
2024	1.200	14.429	5.556	1.125	5.400
2023	14.800	8.571	0.556	0.125	4.400
2022	16.200	2.429	1.556	4.125	0.600
2021	7.200	14.429	2.556	2.875	4.400
2020	3.800	17.571	0.556	7.125	7.600
2019	5.800	2.429	19.444	7.875	0.400
2018	6.800	7.571	3.444	8.125	7.600
2017	21.2		10.556	9.875	1.600
2016	0.800		1.556		0.400
2015	13.800				2.400

<i>F</i>	<i>P-value</i>	<i>F crit</i>
2.459707	0.061311426	2.612305612

Figure 17.1: Excerpt from the ANOVA Section of the MS Excel Data Analysis Tool (ANOVA: Single Factor) for Levene's Test

<i>F</i>	<i>Hodnota P</i>	<i>F krit</i>
2.459707	0.061311426	2.612305612

Figure 17.2: Excerpt from the ANOVA Section of the MS Excel Data Analysis Tool (ANOVA: Single Factor) for Testing the Equality of Mean Yields

(see Table 17.4). The data from Table 17.4 are analyzed using the “ANOVA: Single Factor” procedure from the Data Analysis tool in MS Excel. From the ANOVA section, we obtain the value of the test statistic W , the critical value for the chosen significance level α , and the corresponding p -value (see Figure 17.1). Since the test statistic $F \approx 2.282$ is smaller than the critical value of approximately 2.61, we fail to reject H_0 concerning the equality of variances of wheat yields across all fields. Likewise, the value $p \approx 0.078$ supports the decision not to reject H_0 .

Having established equality of variances among the samples, we may proceed to test the equality of the sample means, again using the “ANOVA: Single Factor” procedure from the MS Excel Data Analysis tool. In this case, however, we use the original data. The ANOVA results indicate that

we fail to reject H_0 concerning the equality of the mean yields across the selected fields (see Figure 17.2). $\triangle$

Example 17.2. The effectiveness of three anticorrosive agents is being investigated. The first agent was applied in 20 cases, the second in 25 cases, and the third in 22 cases. After a specified period, the degree of corrosion damage was measured, yielding the following results: $\bar{y}_1 = 82.4$, $s_1^2 = 12$, $\bar{y}_2 = 80$, $s_2^2 = 10$, $\bar{y}_3 = 85.8$, $s_3^2 = 12$. Unfortunately, the individual measurements have not been preserved. Determine whether there is a statistically significant difference in effectiveness at the significance level 0.01.

Solution. We first compute $n = 67$,

$$\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2 + n_3\bar{y}_3}{n} \approx 82.64.$$

Using the definition of the sample variance,

$$s_i^2 = \frac{1}{n_i - 1} \sum_j y_{ij}^2 - \frac{n_i}{n_i - 1} \bar{y}_i^2,$$

we obtain

$$\sum_j y_{1j}^2 = 136\,023, \quad \sum_j y_{2j}^2 = 160\,240, \quad \sum_j y_{3j}^2 = 162\,208.$$

From these quantities, we can construct the ANOVA table:

Table 17.5: Analysis of Variance Table

Variation	S	df	S/df	F
Between Treatments	183.5	2	91.75	8.15**
Residual	720	64	11.25	-
Total	903.5	66	-	-

The critical value is $F_{2,64}(0.01) \approx 4.98$. Therefore, we reject the hypothesis that all anticorrosive agents are equally effective.

It is now necessary to determine which equalities are violated. To this end, we construct a table of pairwise mean differences (see Table 17.6).

Table 17.6: Table of Mean Differences

Differences	Agent 2	Agent 3
Agent 1	2.4	-3.4**
Agent 2		-5.8**

From the tables, we obtain the critical value of the studentized range: $q_{3,64}(0.01) \approx 4.28$, and we then compute the corresponding rejection values for each pair of treatments separately (see Table 17.7).

Table 17.7: Critical Values for Individual Pairwise Comparisons

Critical Rejection Values	Agent 2	Agent 3
Agent 1	2.93	3.01
Agent 2		2.85

It can be seen that the third anticorrosive agent differs significantly from the other two at the significance level 0.01. In contrast, no significant difference between the first and second agents was demonstrated at the 0.01 significance level. $\triangle$

17.2 Kruskal–Wallis Test

The Kruskal³–Wallis⁴ test is the nonparametric counterpart of one-way analysis of variance and is a generalization of the two-sample Wilcoxon test to the case of I samples ($I \geq 3$). Consider therefore I independent samples of sizes $n_1, n_2, \dots, n_I$. Let $n = n_1 + n_2 + \dots + n_I$. Assume that each sample originates from a distribution with a continuous cumulative distribution function, and therefore we cannot rely on the assumption that all samples come from normal distributions. We wish to test the hypothesis that all samples originate from the same distribution. This test is particularly sensitive to location shifts between the distributions.

As in the two-sample Wilcoxon test, we arrange all n observations from the I samples into a single ascending sequence and assign a rank to each observation. Let T_i denote the sum of the ranks corresponding to observations belonging to the i th sample ($i = 1, 2, \dots, I$). Since the total number

³William Henry Kruskal (1919–2005), American mathematician and statistician.

⁴Wilson Allen Wallis (1912–1998), American economist and statistician.

Table 17.8: Operating Times of Devices Without Failure (Hours)

A	48	16	75	29	96	67	89	22
B	46	94	20	87	66	25	14	75
C	58	17	65	34	26	63	74	106

of observations across all samples is n , the following identity must hold:

$$T_1 + T_2 + \cdots + T_I = \frac{n(n+1)}{2}.$$

This relationship can be used to verify the correctness of the calculations of the values T_i . Under the null hypothesis, the statistic

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^I \frac{T_i^2}{n_i} - 3 \cdot (n+1)$$

has, asymptotically as $n \rightarrow \infty$, a χ^2 distribution with $I - 1$ degrees of freedom. If $q \geq \chi_{I-1}^2(1 - \alpha)$, we reject the hypothesis at a significance level asymptotically equal to α . If the null hypothesis is rejected, we conclude that not all samples originate from the same distribution. In that case, we are interested in identifying which samples differ from one another.

If all samples have the same size, Neményi's multiple comparison procedure may be used for this purpose. If the quantity $|T_i - T_j|$ is greater than or equal to the tabulated critical value (the critical values used by this procedure are available in statistical tables), the hypothesis that the i th and j th samples come from the same distribution is rejected. This procedure is applied to all $\frac{I(I-1)}{2}$ quantities $|T_i - T_j|$.

If the sample sizes are not equal, we define

$$t_i = \frac{T_i}{n_i}, \quad i = 1, \dots, I,$$

and conclude that the distribution functions of the i th and j th samples differ significantly whenever

$$|t_i - t_j| > \sqrt{\frac{1}{12} \left(\frac{1}{n_i} + \frac{1}{n_j} \right) n(n+1) \chi_{I-1}^2(1 - \alpha)}.$$

Example 17.3.  The operating time without failure (in hours) of eight devices from three different manufacturers, A, B, and C, was observed. The results are summarized in Table 17.8. Compare the quality of the three brands in terms of their failure-free operating times.

Table 17.9: Ordered Data from All Samples

	Ranks in the Combined Data Set								Total
A	11	2	18.5	8	23	16	21	5	104.5
B	10	22	4	20	15	6	1	18.5	96.5
C	12	3	14	9	7	13	17	24	99

Solution. It may reasonably be assumed that the individual samples follow exponential distributions. Therefore, analysis of variance cannot be used to compare the reliability of the devices produced by the different manufacturers. Instead, we apply the Kruskal–Wallis test. The data are combined into a single sample and ranked; the results are shown in Table 17.9. We obtain $q = 0.08375$. The critical value is $\chi_2^2(0.95) \approx 5.99$.

$$= \text{CHISQ.INV}(0.95, 2)$$

Since $q < \chi_2^2(0.95)$, we fail to reject the null hypothesis H_0 . The data do not provide evidence of a significant difference in reliability among the three manufacturers. $\triangle$

17.3 Two-Way Analysis of Variance

Suppose that we are investigating the influence of two factors, A and B , on a dependent variable Y . For example, we may wish to determine whether the level of financial literacy measured by scores on an achievement test (the random variable Y) is affected by the type of instructional method (factor A) and by parental education (factor B). We assume that factor A has J levels (i.e., the number of instructional methods used), while factor B has I levels (i.e., the highest educational attainment of at least one parent: primary education, vocational training, secondary education with a school-leaving examination, bachelor's degree, master's degree, ...). Suppose further that we have n_{ij} observations such that, under the i th level of factor A and the j th level of factor B , a student obtains a certain test score. The results (i.e., the test scores) of these n_{ij} observations are denoted by $Y_{ij1}, Y_{ij2}, \dots, Y_{ijn_{ij}}$. If we restrict ourselves to the case where the number of observations in each cell is $n_{ij} = c \geq 1$, (the so-called balanced design), then we obtain Table 17.10.

In this book, we shall focus only on two-way classification without replication, i.e., for each subclass there is a single observation (each cell in Table 17.10 contains only one value). Furthermore, we assume that the column factor A and the row factor B do not interact (for example, this means that

Table 17.10: Results of the Financial Literacy Achievement Test

B	Factor A				
	1	2	3	...	J
1	$Y_{111}, \dots, Y_{11c}$	$Y_{121}, \dots, Y_{12c}$	$Y_{131}, \dots, Y_{13c}$	...	$Y_{1J1}, \dots, Y_{1Jc}$
2	$Y_{211}, \dots, Y_{21c}$	$Y_{221}, \dots, Y_{22c}$	$Y_{231}, \dots, Y_{23c}$	...	$Y_{2J1}, \dots, Y_{2Jc}$
...	...	...	...	...	...
I	$Y_{I11}, \dots, Y_{I1c}$	$Y_{I21}, \dots, Y_{I2c}$	$Y_{I31}, \dots, Y_{I3c}$	...	$Y_{IJ1}, \dots, Y_{IJc}$

the instructional method affects a student's performance independently of the parents' educational attainment; the effects of the factors are additive). This model is known as the two-way classification model without interaction and without replication.

Consider the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \text{ where } i = 1, \dots, I, j = 1, \dots, J, \quad (17.1)$$

where μ , α_i for $i = 1, \dots, I$, and β_j for $j = 1, \dots, J$ are unknown parameters, and $e_{ij} \sim N(0, \sigma^2)$ denotes an error term arising from measurement inaccuracy or random deviation from the mean. Thus, the observed values Y_{ij} depend both on the column and on the row in which they occur. Moreover, each row contains the same number of observations. As an example, suppose that the blood pressure of J patients is measured at I different times, such as morning, noon, and evening. Each patient has a different average blood pressure level $\mu + \beta_j$, while the daily fluctuations are determined by the parameters α_i .

The parameters of model (17.1) are not uniquely determined. In particular, all values α_i may be shifted by an arbitrary constant $\tilde{\alpha} \in \mathbb{R}$ and, similarly, all values β_j may be shifted by a constant $\tilde{\beta} \in \mathbb{R}$, while the sum $\tilde{\alpha} + \tilde{\beta}$ is subtracted from μ . The resulting model remains unchanged. To avoid this non-identifiability, we impose the additional constraints

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0.$$

We now wish to test the hypothesis $H_0 : \alpha_1 = \dots = \alpha_I = 0$ (that is, the row classification has no effect), which can be rewritten as the simpler model corresponding to one-way classification:

$$Y_{ij} = \mu + \beta_j + e_{ij}.$$

The test is carried out as follows. First, let us denote the sample means of the individual groups by

$$\begin{aligned}\bar{Y}_{i.} &= \frac{Y_{i1} + \cdots + Y_{iJ}}{J} \text{ for } i = 1, \dots, I, \\ \bar{Y}_{.j} &= \frac{Y_{1j} + \cdots + Y_{Ij}}{I} \text{ for } j = 1, \dots, J\end{aligned}$$

and let the overall mean of all observations be denoted by

$$\bar{Y} = \frac{\sum_i \sum_j Y_{ij}}{n},$$

where $n = IJ$. Next, we compute the total sum of squares S_T , that is, the total quadratic error in the case where $\alpha_1 = \cdots = \alpha_I = \beta_1 = \cdots = \beta_J = 0$. The parameter μ is estimated by the overall mean $\bar{Y}$.

$$S_T = \sum_i \sum_j (Y_{ij} - \bar{Y})^2 = \sum_i \sum_j Y_{ij}^2 - n\bar{Y}^2.$$

The sum of squares attributable to row effects is denoted by

$$S_A = J \sum_i \bar{Y}_{i.}^2 - n\bar{Y}^2.$$

The sum of squares attributable to column effects is denoted by

$$S_B = I \sum_j \bar{Y}_{.j}^2 - n\bar{Y}^2.$$

The residual sum of squares $S_e = S_T - S_A - S_B$ is the total quadratic error of model 17.1.

The numbers of degrees of freedom associated with the individual sums of squares are:

$$df_T = n - 1, \quad df_A = I - 1, \quad df_B = J - 1, \quad df_e = n - I - J + 1.$$

We reject the hypothesis H_0 at the significance level α whenever

$$F_A = \frac{S_A/df_A}{S_e/df_e} \geq F_{df_A, df_e}(1 - \alpha).$$

Similarly, we proceed when testing the hypothesis $H'_0: \beta_1 = \cdots = \beta_J = 0$ (i.e., that the column classification has no effect), which can be rewritten as the following simpler model corresponding to one-way classification:

$$Y_{ij} = \mu + \alpha_i + e_{ij}.$$

We reject the hypothesis H'_0 at the significance level α whenever

$$F_B = \frac{S_B/df_B}{S_e/df_e} \geq F_{df_B, df_e}(1 - \alpha).$$

The results of the entire analysis are conveniently summarized in tabular form (see Table 17.11).

Table 17.11: Table of Results for Two-Way Analysis of Variance

Variation	Sum of Squares S	Degrees of Freedom df	Mean Square S/df	F
Row Factor	S_A	$df_A = I - 1$	S_A/df_A	F_A
Column Factor	S_B	$df_B = J - 1$	S_B/df_B	F_B
Residual	S_e	$df_e = n - I - J + 1$	S_e/df_e	-
Total	S_T	$df_T = n - 1$	-	-

The residual variance $S^2 = \frac{S_e}{df_e}$ is an unbiased estimator of the variance σ^2 .

If the hypothesis H_0 is rejected, it is often necessary to determine for which pairs of indices the equality does not hold. As in one-way classification, this problem is addressed by Tukey's multiple comparison procedure.

The equality $\alpha_i = \alpha_l$ is rejected whenever

$$|\bar{Y}_{i.} - \bar{Y}_{l.}| \geq sq_{I, n-I-J+1}(\alpha) \sqrt{\frac{1}{J}}.$$

The equality $\beta_j = \beta_l$ is rejected whenever

$$|\bar{y}_{.j} - \bar{y}_{.l}| \geq sq_{J, n-I-J+1}(\alpha) \sqrt{\frac{1}{I}}.$$

As mentioned at the beginning of this chapter, model (17.1) can be generalized further. For example, for each pair of indices i, j , we may have C observations.

$$Y_{ijc} = \mu + \alpha_i + \beta_j + e_{ijc},$$

where $i = 1, \dots, I$, $j = 1, \dots, J$, $c = 1, \dots, C$.

If the factors influencing the response variable interact with one another, we use model (17.2), which explicitly incorporates interactions between rows

and columns. This is known as the two-way classification model with interactions:

$$Y_{ijc} = \mu + \alpha_i + \beta_j + \lambda_{ij} + e_{ijc}. \quad (17.2)$$

Furthermore, it is also possible to study the dependence of a response variable on three factors, leading to the so-called three-way classification. These models are discussed, for example, in [1] and [2]. When performing two-way analysis with one observation in each subclass, the corresponding nonparametric analogue is the Friedman test, which serves as the nonparametric counterpart of the two-way classification model described above.

Example 17.4.  The effect of three pharmaceutical preparations on blood coagulation was investigated. Among other indicators, the so-called thrombin time was measured. For each subject, a baseline value K was first obtained, representing the thrombin time before the beginning of the experiment. Subsequently, preparations A , B , and C were administered, each after a sufficiently long interval to ensure that the effects of the previous preparation had disappeared. The measurements for 10 subjects are presented in Table 17.12.⁵

Table 17.12: Results of Thrombin Time Measurements

Person	Preparation				$\bar{Y}_{i\cdot}$
	K	A	B	C	
1	11.3	11.2	11.4	11.0	11.225
2	11.9	12.1	11.8	9.5	11.325
3	11.8	13.2	12.0	11.1	12.025
4	12.1	12.8	12.0	12.5	12.35
5	11.2	13.5	11.5	8.4	11.15
6	11.3	12.5	11.5	9.0	11.075
7	10.8	10.7	10.9	9.7	10.525
8	12.0	13.8	11.6	12.2	12.4
9	11.5	12.9	11.3	10.3	11.5
10	11.7	11.9	11.3	8.2	10.775
$\bar{Y}_{\cdot j}$	11.56	12.46	11.53	10.19	11.435

Determine whether, at the significance levels 0.05 and 0.01, all tested subjects respond equally to preparations A , B , and C with respect to thrombin time.

⁵The problem and the data in Table 17.12 are taken from [1], p. 233.

Solution. From these quantities, we can now construct the analysis of variance table (see Table 17.13).

Table 17.13: Analysis of Variance Table

Variation	S	df	S/df	F
Row Factor (Subjects)	14.606	9	1.62	2.58*
Column Factor (Preparations)	26.253	3	8.75	13.9**
Residual	16.992	27	0.63	—
Total	57.851	39	—	—

We compute the critical value $F_{9,27}(0.05) \approx 2.25$.

= F.INV(0.95,9,27)

Therefore, we reject the hypothesis that all subjects have the same thrombin time at the significance level 0.05.

Next, we compute the critical value $F_{3,27}(0.01) \approx 4.6$. Therefore, we reject the hypothesis that the preparations have no effect on thrombin time at the significance level 0.01.

In MS Excel, comparisons of means from multiple samples can be carried out either by performing the numerical calculations directly from the formulas (see the accompanying MS Excel file) or by using the “ANOVA: Two-Factor Without Replication” procedure from the Data Analysis tool. The ANOVA section of the output displays the same information as Table 17.13 (see Figure 17.3).

ANOVA				
Source of Variation	SS	df	MS	F
Rows	14.606	9	1.622888889	2.578742938
Columns	26.253	3	8.751	13.90519068
Error	16.992	27	0.629333333	
Total	57.851	39		

Figure 17.3: Excerpt from the MS Excel Two-Factor ANOVA Output from the Data Analysis Tool

In the right-hand part of the same table, information is provided on the values of the F test statistics, the corresponding critical values, and the p -values separately for rows and columns (see Figure 17.4).

<i>F</i>	<i>Hodnota P</i>	<i>F krit</i>
2.578742938	0.027464187	2.250131477
13.90519068	1.13131E-05	2.960351318

Figure 17.4: Excerpt from the MS Excel Two-Factor ANOVA Output from the Data Analysis Tool

The hypothesis concerning equality of thrombin times across subjects is not of primary interest here. Its invalidity was already anticipated at the outset, which is why two-way classification was chosen. We now need to determine which equalities among the preparations are violated. To this end, let us construct a table of mean differences (see Table 17.14).

Table 17.14: Differences in Mean Thrombin Times

Differences	<i>A</i>	<i>B</i>	<i>C</i>
K	−0.9	0.03	1.37**
A		0.93	2.27**
B			1.34**

From the tables, we obtain the critical values of the studentized range:

$$q_{4,27}(0.01) \approx 4.85,$$

$$q_{4,27}(0.05) \approx 3.875,$$

and compute the corresponding rejection thresholds:

$$s \cdot q_{4,27}(0.01) \sqrt{\frac{1}{10}} \approx 1.22,$$

$$s \cdot q_{4,27}(0.05) \sqrt{\frac{1}{10}} \approx 0.97.$$

We observe that preparation *C* significantly reduces thrombin time both in comparison with the baseline condition *K* and in comparison with preparations *A* and *B*. $\triangle$

17.4 Friedman's Test

Friedman's⁶ test is the nonparametric counterpart of two-way analysis of variance. Suppose we have IJ independent observations, which are arranged in a table (see Table 17.15).

Table 17.15: Observed Values

Columns	1	2	3	...	J
Rows					
1	Y_{11}	Y_{12}	Y_{13}	...	Y_{1J}
2	Y_{21}	Y_{22}	Y_{23}	...	Y_{2J}
.	...	...	...	...	
.	...	...	...	...	
.	...	...	...	...	
I	Y_{I1}	Y_{I2}	Y_{I3}	...	Y_{IJ}

The random variables Y_{ij} have continuous distribution functions F_{ij} , $i = 1, 2, \dots, I$, $j = 1, 2, \dots, J$, and are mutually independent.

We shall test the hypothesis

$$H_0: F_{i1} = F_{i2} = \dots = F_{iJ}.$$

In other words, we test whether the distribution functions F_{ij} depend on the column index j , while allowing them to depend on the row index i .

For each row, we arrange the observations in ascending order and assign the corresponding ranks $R_{i1}, R_{i2}, \dots, R_{iJ}$, $i = 1, 2, \dots, I$ (see Table 17.16).

⁶Milton Friedman (1912–2006), American economist and statistician.

Table 17.16: Ranks and Sums of the Observed Values

Rows	Columns					Row Totals
	1	2	3	...	J	
1	R_{11}	R_{12}	R_{13}	...	R_{1J}	$\frac{J(J+1)}{2}$
2	R_{21}	R_{22}	R_{23}	...	R_{2J}	$\frac{J(J+1)}{2}$
.	...	...	...	...		
.	...	...	...	...		
.	...	...	...	...		
I	R_{I1}	R_{I2}	...		R_{IJ}	$\frac{J(J+1)}{2}$
Column Totals	$R_{.1}$	$R_{.2}$	...		$R_{.J}$	$R_{..} = \frac{IJ(J+1)}{2}$

Here $R_{.j} = \sum_i R_{ij}$, $R_{..} = \sum_i \sum_j R_{ij}$.

The test statistic of Friedman's test is given by

$$Q = \frac{12}{IJ(J+1)} \sum_{j=1}^J R_{.j}^2 - 3I(J+1). \quad (17.3)$$

We reject the hypothesis H_0 whenever Q exceeds the critical value given in the tables. For larger values of I , the critical value is taken to be $\chi_{J-1}^2(1-\alpha)$.

If the hypothesis is rejected, we are interested in determining which columns differ from one another. For this purpose, we construct a table of the values $|R_{.j} - R_{.m}|$ for all pairs satisfying $j < m$.

If any of the values $|R_{.j} - R_{.m}|$ is greater than or equal to the tabulated critical value, the hypothesis that $F_{ij} = F_{im}$ is rejected at the corresponding significance level.

Example 17.5. Solve Example 17.4 using Friedman's test.

Solution. We assign the ranks R_{ij} (see Table 17.17).

Table 17.17: Rank

Person	Preparation			
	<i>K</i>	<i>A</i>	<i>B</i>	<i>C</i>
1	3	2	4	1
2	3	4	2	1
3	2	4	3	1
4	2	4	1	3
5	2	4	3	1
6	2	4	3	1
7	3	2	4	1
8	2	4	1	3
9	3	4	2	1
10	3	4	2	1
Total	25	36	25	14

Using formula 17.3, we obtain

$$q = \frac{12}{10 \cdot 4 \cdot 5} (25^2 + 36^2 + 25^2 + 14^2) - 3 \cdot 10 \cdot 5 = 14.52^{**}.$$

From the tables, we find that the critical value for Friedman's test at the 0.01 significance level is 10.53. Since $q > 10.53$, we reject the hypothesis that thrombin time is independent of preparations *K*, *A*, *B*, and *C* at the significance level 0.01.

To determine which of the preparations *K*, *A*, *B*, and *C* differ from one another, we compute the values $|R_{.j} - R_{.m}|$ (see Table 17.18).

Table 17.18: Absolute values of the differences

	<i>A</i>	<i>B</i>	<i>C</i>
<i>K</i>	11	0	11
<i>A</i>		11	22 ^{**}
<i>B</i>			11

The critical value for multiple comparisons in the Friedman test at the significance level of 0.01 is 18.0; therefore, only the difference between treatments *A* and *C* appears to be statistically significant.

Comparing these results with Example 17.4, we observe that the Friedman test is less powerful than the two-way analysis of variance. On the other hand, the Friedman test can also be applied in the case of non-normal distributions or when only the ranking of observations is available. $\triangle$

17.5 Exercises

17.1 Compare the performance of five randomly selected graduates of grammar schools (G), secondary technical schools (SPŠ), and vocational training schools (OU) in the entrance examination (see Table 17.19). Conduct the test at the significance level $\alpha = 0.05$. If differences among the population means exist, determine between which groups these differences occur. Table 17.19: Scores achieved by students from different types of schools in the entrance examination.

Student	G	SPŠ	OU
s_1	55	54	47
s_2	54	50	53
s_3	58	51	49
s_4	61	51	50
s_5	52	49	46

17.2 Determine whether there is a statistically significant difference in fertility among any of the four sows A to D . If such a difference exists, identify between which sows the significant difference actually occurs. Table 17.20 presents the numbers of piglets born to the individual sows.

Table 17.20: Number of piglets born to sows A – D during the years 2000–2009

Year / Sow	A	B	C	D
2000	10	8	18	10
2001	12	3	13	12
2002	8	6	10	10
2003	13	7	12	8
2004	7	8	14	10
2005	10	7	12	5
2006	11	7	12	4
2007	8	6	10	-
2008	8	-	11	-
2009	-	-	14	-

17.3 On randomly selected plots belonging to six farmers, cereal yields (in t/ha) were measured over a period of several years (see Table 17.21). Determine between which farmers there are statistically significant differences in cereal yield.

Table 17.21: Cereal yields of individual farmers in the years 2000–2009

Year / Farmer	A	B	C	D	E	F
2000	4	8	6	7	5	4
2001	6	3	2	6	7	5
2002	8	6	3	8	6	4
2003	5	7	5	8	5	5
2004	7	8	6	4	5	5
2005	3	7	7	5	7	2
2006	2	7	1	4	2	3
2007	5	6	2		6	1
2008	6				5	2
2009					4	

17.4 It was investigated whether four different potato varieties (A, B, C, D) have equal yields (in tonnes). Each variety was grown on a total of five plots. The observed yields are given in Table 17.22.

Table 17.22: Yields of potato varieties on five different fields (in tonnes)

Yield of the variety (in tonnes) / Field No.	1	2	3	4	5
A	22.1	24.0	23.6	20.8	21.5
B	25.9	28.0	26.1	26.7	25.4
C	26.6	27.2	26.1	26.9	27.1
D	28.3	27.9	29.1	27.2	28.6

Using one-way analysis of variance, determine whether there are statistically significant differences in cereal yields.

17.5 A group of neurotic children was divided into groups A, B, C, and D according to the respective syndromes under investigation. To provide an objective supplement to the psychiatric diagnosis, a blood analysis was performed, examining certain physiological parameters of blood composition. Suppose that the analysis yielded the results presented in Table 17.23.

Table 17.23: Group Data Table

Patient No.	Group A	Group B	Group C	Group D
1	13	13	12	10
2	12	14	10	11
3	11	15	9	12
4	15	14	9	-
5	14	-	10	-
$\bar{x}_i$	13	14	10	11

Determine whether the observed values differ significantly among the respective groups to such an extent that the given analysis could be used for the diagnosis of the syndromes under investigation.

17.6 In a workshop, the length of work experience and the number of products manufactured by individual workers during one week were recorded. The results are summarized in Table 17.24. At the 5% and 1% significance levels, test whether the length of work experience has a statistically significant effect on labour productivity. The productivity measurements corresponding to different levels of work experience constitute three samples drawn from three populations with distributions $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$ and $N(\mu_3, \sigma^2)$.

Table 17.24: Productivity Data Table

Work experience	Productivity	n_i	Sum x_i	Mean $\bar{x}_i$	$\sum_j x_{ij}^2$
< 10 years	76, 67, 80, 70	4	293	73.25	21 565
10–25 years	82, 102, 98, 91	4	373	93.25	35 013
> 25 years	81, 77, 75, 85, 69	5	387	77.4	30 101

17.7 At five selected locations with a speed limit of 50 km/h, the speeds of passing vehicles were measured (see Table 17.25). At the significance level $\alpha = 0.05$, test the equality of the means of all measurements, assuming that the samples do not originate from normal distributions.

Table 17.25: Speeds of vehicles in km/h

M_1	M_2	M_3	M_4	M_5
55.4	68.4	52.1	62.3	67.5
61.2	57.9	49.5	63.5	54.3
65.8	58	47.5	63.4	56.3
59.3	47.8	58	56.2	55.5
62.5	63.8	37.8	51	57.2
58.4	52.3	57.2	41.8	65.5
48.1	62.5	50.5	44.5	60.2

17.8 An entomologist is studying the vertical distribution of a particular fly species in a deciduous forest and obtained five samples of individuals from each of three vegetation strata (herb layer, shrub layer, and canopy layer). At the significance level $\alpha = 0.05$, test the hypotheses:

H_0 : The vertical distribution of the species is the same across all three vegetation strata.

H_1 : The vertical distribution of the species is not the same across all three vegetation strata.

Table 17.26: Table of numbers of individuals per m^3 of vegetation stratum (rank in parentheses)

Individual No.	Vertical distribution per m^3		
	Herb layer	Shrub layer	Canopy layer
1	14.0 (15)	8.4 (11)	6.9 (8)
2	12.1 (14)	5.1 (2)	7.3 (9)
3	9.6 (12)	5.5 (4)	5.8 (5)
4	8.2 (10)	6.6 (7)	4.1 (1)
5	10.2 (13)	6.3 (6)	5.4 (3)

17.9 Table 17.27 presents measured values of the specific draft resistance of a plough body on medium-heavy soil at different operating speeds: $v_1 = 4$ km/h, $v_2 = 6$ km/h, and $v_3 = 8$ km/h.

Table 17.27: Table of Specific Draft Resistance Values of the Plough Body in 10^5 Pa

Speed	Specific Draft Resistance (Measurement No.)			
	M 1	M 2	M 3	M 4
v_1	0.71	0.66	0.63	0.65
v_2	0.59	0.64	0.67	0.69
v_3	0.70	0.72	0.75	0.73

Assess whether the effect of operating speed on the specific draft resistance is statistically significant or not.

17.10 Table 17.28 contains the values of three different samples. Determine whether all three samples could have been drawn from the same population with an identical probability distribution.

Table 17.28: Table of Values from Three Different Samples

Sample No.	Values
1	2, 3, 10, 12, 18, 20.5, 26, 28
2	1, 4.5, 10, 10, 15, 16, 19, 23, 26, 30
3	4.5, 6, 7, 8, 13, 14, 17, 20.5, 22, 24, 26, 29

17.11 In 2006, a test of mayonnaise samples from different manufacturers was conducted. For the ranking test, the assessors were given five samples in beakers simultaneously and ranked them from the most palatable to the least palatable. The results of the assessment are shown in Table 17.29. Are the individual samples statistically significantly different ($\alpha = 0.05$)?

Table 17.29: Results of the Sensory Evaluation of Mayonnaise Samples from Different Manufacturers (Mean Scores on a 100-Point Scale)

Evaluated Attribute / Sample	1	2	3	4	5
Appearance of colour	74	43	76	46	58
Colour intensity	62	37	77	37	44
Odour acceptability	71	56	54	48	43
Odour intensity	48	34	55	29	33

17.12 A company operates four bicycle retail stores. At the 5% significance level, test whether the differences in bicycle sales among the individual stores

are statistically significant. Data on the numbers of bicycles sold in each store during the last five months are available (see Table 17.30).

Table 17.30: Table of Numbers of Bicycles Sold in Individual Stores

Month / Store	A	B	C	D
I	23	25	22	18
II	18	15	20	16
III	16	16	17	20
IV	20	19	15	24
V	21	22	18	25

Solutions to the Exercises

17.1 Test for equality of variances: One-way ANOVA: the test statistic is $1.391 <$ the critical value of approximately 3.885, therefore, the hypothesis of equal variances is not rejected (likewise, the p -value of the F -test is approximately $0.286 > 0.05$).

Assuming homogeneity of variances, we test the equality of means: H_0 : There is no statistically significant difference among the test results, against the alternative hypothesis H_1 : There is a statistically significant difference among the test results.

One-way ANOVA: the test statistic is approximately $F \approx 8.298 >$ the critical value of approximately 3.885 (likewise, the p -value is $0.00546 < 0.05$); therefore, a statistically significant difference exists among the means of the three groups.

Tukey's test: the critical value of the test statistic is approximately 3.773. A statistically significant difference is detected only between the mean scores of grammar school graduates and vocational training school graduates.

17.2 Test for equality of variances: one-way ANOVA: the F -statistic is approximately $1.175 <$ the critical F -value of approximately 2.922 (likewise, the p -value is approximately $0.336 > 0.05$), therefore, the variances can be considered equal at the chosen significance level.

We test the hypothesis H_0 : There is no statistically significant difference in the number of piglets among the sows under investigation.

Against the alternative hypothesis H_1 : There is a statistically significant difference in the number of piglets among the sows under investigation.

Test of means: one-way ANOVA. The test statistic is approximately $F \approx 11.443 >$ the critical F -value of approximately 2.922 (likewise, the p -value is approximately $0.000\,036 < 0.05$).

Tukey's test: the critical value of the test statistic is approximately $q = 3.093$. Statistically significant differences in fertility are found between sows AB , AC , BC , and CD .

17.3 The critical value of Tukey's test for 6 groups at the significance level 0.05 and with 45 degrees of freedom is approximately 4.208. A statistically significant difference in mean cereal yields is found only between farmers B and F (test statistic approximately -5.119).

Table 17.31: Table of intermediate results of the analysis

x_i	x_i^2	$\sum x_{ij}^2$	$\bar{x}_i$
112	12 544	2 516.3	22.4
132.1	17 450.41	3 494.1	26.4
133.9	17 929.21	3 586.6	26.8
141.1	19 909.21	3 983.9	28.2
$\sum x_{i0} = 519.1$	$\sum x_i^2 = 67\,832.83$	$\sum \sum x_{ij}^2 = 13\,581$	

Number of groups (rows): $m = 4$, number of columns: $n = 5$.
Sums of squares: $S_l = \frac{1}{n} \sum x_i^2 - c \approx 93.325\,5$, $S_r = S - S_l \approx 14.304$,
 $S = \sum \sum x_{ij}^2 - c \approx 107.629\,5$, $c = \frac{\sum x_{i0}^2}{m \cdot n} \approx 13\,473.240\,5$.

Table 17.32: ANOVA summary table

Variation	Sum of Squares	Degrees of Freedom	Variance
Between groups	93.325, 5	$m - 1$ (3)	31.107
Within groups (residual)	14.304	$m \cdot (n - 1)$ (16)	0.894
Total	107.629 5	$mn - 1$ (19)	

F -test: $F = 34.795$, $F_{0.05}(3, 16) = 3.24$, $F > F_{0.05}(3, 16)$; therefore, we reject the null hypothesis of no significant differences among the yields of the individual varieties.

17.5 $H_0 : \mu_A = \mu_B = \mu_C = \mu_D$. $S_A = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2 = 44$, $S_R = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{x}_{ij} - \bar{x}_i)^2 = 20$. The sample means are drawn from populations with identical means; that is, the blood composition of all populations corresponding to the different syndromes has the same mean physiological values.

17.6 We test $H_0: \mu_1 = \mu_2 = \mu_3$ against H_1 : Not all μ_i are equal. From Table 17.24, we have $n = 13$, $\sum x_i = 1053$, $\sum_i \sum_j x_{ij}^2 = 86679$. The test statistic $F = \frac{n-k}{k-1} \cdot \frac{S_A}{S_e}$ has distribution $F(k-1, n-k) = F(2, 10)$. The critical value is $F_{0.05}(2, 10) \approx 4.103$, the critical regions are $(4.103, \infty)$, $F_{0.01}(2, 10) \approx 7.559$, and the critical region is $(7.559; \infty)$. The value of the test statistic: $S_T = 1386$, $S_A = 905.3$, $S_e \approx 480.7$. The value of the test statistic is computed from the information in the analysis of variance table (see Table 17.33).

Table 17.33: Analysis of Variance Table

Source of variation	Sum of squares	Degrees of Freedom	Mean square	F
Length of work experience	905.3	2	452.65	$F = \frac{452.65}{48.07} \approx 9.416$
Residual	480.7	10	48.07	
Total	1386	12	-	

At the significance level 0.05, $F = 9.416 > 4.103$; therefore, we reject H_0 . At the significance level 0.01, $F = 9.416 > 7.559$; therefore, H_0 is also rejected.

The study demonstrated a statistically significant effect of the length of work experience on labour productivity. In particular, the productivity of workers with 10 to 25 years of work experience differs significantly from that of the other groups.

17.7 Kruskal–Wallis test: $T \approx 6.995 < \chi^2(4) \approx 9.49$; therefore, we do not reject H_0 of equality of population means. Hence, at the 0.05 significance level, vehicles passed through locations M_1 – M_5 with comparable average speeds.

17.8 $n_1 = n_2 = n_3 = 5$, $N = 15$, Kruskal–Wallis test: $R_1 = 64$, $R_2 = 30$, $R_3 = 26$, $T \approx 8.720 > \chi^2(2) \approx 5.78$; therefore, we reject H_0 . Hence, the vertical distribution of the species is not the same across all three vegetation strata. The herb layer differs from the other strata.

17.9 The measured values are combined into a single sample, and the ranks of the individual observations are determined (see Table 17.34).

Table 17.34: Table of Plough Body Draft Resistance Values

v_1	0.71	0.66	0.63	0.65	
Rank	9	5	2	4	$T_1 = 20$
v_2	0.59	0.64	0.67	0.69	
Rank	1	3	6	7	$T_2 = 17$
v_3	0.70	0.72	0.75	0.73	
Rank	8	10	12	11	$T_3 = 41$

$k = 3$, $n_1 = n_2 = n_3 = n_4 = 3$, $n = 12$, Kruskal–Wallis test: $T \approx 6.577 > \chi^2(2) \approx 5.991$; therefore, we reject H_0 of equal specific draft resistance at the different operating speeds.

17.10 $R_1 = 119.5$, $\bar{R}_1 = 14.94$, $R_2 = 154.5$, $\bar{R}_2 = 15.45$, $R_3 = 191$, $\bar{R}_3 = 15.92$, $T = 0.07$, $p = 4$ groups of tied ranks; the first contains ranks 4 and 5, which we replace by the average rank 4.5, thus $T_1 = 2^3 - 2 = 6$. For the second and fourth groups, $T_2 = T_4 = 3^3 - 3 = 24$. For the third group, $T_3 = T_1 = 6$. The correction factor is equal to $1 - \frac{60}{26970} = 0.998$. Since $T = 0.07/0.998 \approx 0.07 < \chi^2(2) = 5.99$, we do not reject H_0 that all three samples come from the same distribution.

17.11 For the Friedman test, the values in Table 17.29 are replaced by their ranks within each row, and the rank sums for the individual samples are then computed (see Table 17.35).

Table 17.35: Table of Processed Results of the Sensory Evaluation of Mayonnaise Samples from Different Manufacturers

Evaluated Attribute / Sample	1	2	3	4	5	Total
Appearance of colour	4	3	3	3	4	17
Colour intensity	2	2	4	2	3	13
Odour acceptability	3	4	1	4	2	14
Odour intensity	1	1	2	1	1	6

Because $T = 10.785 \geq 7.81$, we reject H_0 of equal quality levels among the tested mayonnaise samples.

17.12 For the Friedman test, the values in Table 17.30 are replaced by their ranks within each row, and the rank sums for the individual stores are then calculated (see Table 17.36).

Table 17.36: Table of Ranks of Bicycles Sold in Individual Stores

Month / Store	A	B	C	D
I	3	4	2	1
II	3	1	4	2
III	1.5	1.5	3	4
IV	3	2	1	4
V	2	3	1	4
R_1	12.5	11.5	11	15

$T = 1.14$, the number of samples (corresponding to the months) is 5. We approximate the distribution of the test statistic T by a χ^2 distribution with 3 degrees of freedom ($\nu = k - 1 = 3$). The critical value is the quantile $\chi^2_{0.05}(3) \approx 7.81$. Since $T < 7.81$, we do not reject H_0 of equality of the population means. Therefore, no statistically significant difference in the numbers of bicycles sold among the individual stores was demonstrated.

Chapter 18

Correlation Analysis

18.1 Pearson Sample Correlation Coefficient

If a linear relationship exists between the variables under study, the sample correlation coefficient is used to assess the strength of the association. Recall that in Section 9.2 we introduced the concept of the correlation coefficient for a pair of random variables X and Y with nonzero variances by the formula

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var } X} \sqrt{\text{var } Y}} = \frac{\mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)}{\sqrt{\text{var } X} \sqrt{\text{var } Y}} \quad (18.1)$$

If the unknown variances and the unknown covariance are replaced by their (unbiased) estimators

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, \quad S_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2, \\ S_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

we obtain the sample version of the correlation coefficient, which is formally introduced in the following definition.

Definition 18.1 (Pearson Sample Correlation Coefficient)

Let $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ be a random sample from a bivariate distribution. The Pearson sample correlation coefficient is defined as

$$R_P = \frac{S_{XY}}{\sqrt{S_X^2 \cdot S_Y^2}}. \quad (18.2)$$

After a straightforward algebraic simplification, we obtain a formula that is more convenient for computation:

$$R_P = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{(\sum X_i^2 - n\bar{X}^2)(\sum Y_i^2 - n\bar{Y}^2)}}.$$

From the Cauchy–Schwarz–Bunyakovsky inequality, it follows directly that $-1 \leq R_P \leq 1$. Pearson's correlation coefficient is used when we wish to assess whether a linear relationship exists between the variables X and Y . Its sample value r_P is a dimensionless number belonging to the interval $[-1; 1]$. The sample correlation coefficient is not an unbiased estimator of ρ , although the sample variance S^2 and the sample covariance S_{XY} are unbiased estimators.

Now suppose that $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ is a random sample from a bivariate normal distribution, and that $\text{var } X > 0$, $\text{var } Y > 0$, and $|\rho| < 1$. Let us now test the hypothesis $H_0: \rho = 0$ against the alternative $H_1: \rho \neq 0$. Under the null hypothesis H_0 and the assumptions given above, the statistic

$$T = \frac{R_P}{\sqrt{1 - R_P^2}} \sqrt{n - 2} \sim t_{n-2}$$

has a Student's t distribution with $n - 2$ degrees of freedom. Therefore, we reject the null hypothesis H_0 at significance level α if

$$|t| \geq t_{n-2} \left(1 - \frac{\alpha}{2}\right).$$

The normality assumption is essential for this test. If there is uncertainty regarding this assumption, it is preferable to use Spearman's rank correlation coefficient for testing independence.

When assessing the strength of a linear relationship between two numerical variables, the following generally accepted guidelines may be used:

- If $r_P = 0$, the variables under study are uncorrelated.
- If $|r_P| = 1$, the variables under study are linearly dependent: for $\rho = 1$, the relationship is a positive linear dependence, whereas for $\rho = -1$, it is a negative linear dependence.
- If $|r_P| < 0.3$, the degree of correlation is considered weak.
- If $0.3 \leq |r_P| < 0.5$, the degree of correlation is considered moderate.
- If $0.5 \leq |r_P| < 0.7$, the degree of correlation is considered substantial.
- If $0.7 \leq |r_P| < 0.9$, the degree of correlation is considered strong.

Table 18.1: Weight of Twins (g)

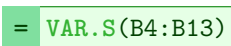
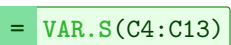
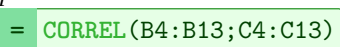
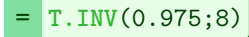
Older Twin	Younger Twin	Older Twin	Younger Twin
2 440	2 700	2 690	2 350
3 500	3 080	2 750	3 500
2 820	2 200	2 750	2 500
2 540	2 700	2 650	2 420
2 650	2 550	2 200	2 520

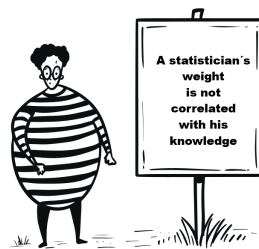
	Sloupec 1	Sloupec 2
Sloupec 1	1	
Sloupec 2	0.338520934	1

Figure 18.1: Excerpt from MS Excel Showing the Correlation Output Generated by the Data Analysis Tool

- If $0.9 \leq |r_P| < 1$, the degree of correlation is considered very strong.

Example 18.1.  The weights (in grams) of ten pairs of twins were recorded (see Table 18.1). Test whether the weights of the twins are correlated.

Solution. We successively compute $s_X^2 \approx 111\,965.6$ , and similarly $s_Y^2 = 145\,240$ , $n = 10$, $s_{XY} = 43\,169$, and $r_P^2 \approx 0.191$. The value of the correlation coefficient is 0.3385 . We compute the value of the test statistic $t \approx 0.550$. We choose $\alpha = 0.05$ and compute the critical value $t_8(0.975) \approx 2.306$ . Since $|t| \not\geq t_8(0.975)$, we do not reject the hypothesis that the weights of the twins are uncorrelated at the significance level $\alpha = 0.05$. The value of the correlation coefficient is not sufficiently high to allow us to reject the hypothesis of no correlation between the birth weights of twins.



The value of the correlation coefficient can also be computed in MS Excel using the *Correlation* option available in the *Data Analysis* tool. Figure 18.1 shows the resulting output table. 

18.2 Spearman's Rank Correlation Coefficient

Spearman's¹ correlation coefficient r_s is a nonparametric estimator of the correlation coefficient and is used when the relationship between the variables under study (continuous random variables) is not purely linear but rather some monotonic function. Because the dependence between the variables is expressed through their ranks, this coefficient is also referred to as Spearman's rank correlation coefficient.

Consider $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ to be a random sample from some bivariate distribution. To construct Spearman's correlation coefficient, it is sufficient to know only the ranks of $X_1, X_2, \dots, X_n$ and the ranks of $Y_1, Y_2, \dots, Y_n$. If the rankings are very similar, this indicates a dependence between X_i and Y_i . Unlike Pearson's correlation coefficient, Spearman's correlation coefficient is robust to outliers.

Definition 18.2 (Spearman's Rank Correlation Coefficient)

Let $R_1, R_2, \dots, R_n$ denote the ranks of $X_1, X_2, \dots, X_n$, and let $Q_1, Q_2, \dots, Q_n$ denote the ranks of $Y_1, Y_2, \dots, Y_n$. Spearman's rank correlation coefficient is defined by

$$R_S = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^n (R_i - Q_i)^2.$$

As with Pearson's correlation coefficient, the value of the coefficient r_s lies within the interval $[-1; 1]$. A value close to 1 indicates a strong positive association, whereas a value close to -1 indicates a strong negative association.

When testing the hypothesis $H_0: \rho = 0$ against the alternative $H_1: \rho \neq 0$, critical values of r_s are tabulated for $n \leq 30$. For $n > 30$, the null hypothesis H_0 is rejected whenever

$$|r_s| \geq \frac{u(1 - \frac{\alpha}{2})}{\sqrt{n - 1}},$$

where $u(1 - \frac{\alpha}{2})$ is the critical value of the $N(0; 1)$ distribution.

Example 18.2.  (see [1]) Ten students were observed. Based on a psychological assessment, the students were ranked according to their degree of emotional instability (the more emotionally unstable a student was, the higher the assigned rank R_i). In addition, the students were assigned ranks Q_i according to their performance in mathematics (the best student in mathematics received rank 1). The results are presented in Table 18.2. Test

¹Charles Edward Spearman (1863–1945), British psychologist and statistician.

Table 18.2: Ranking of Students According to Neuroticism and Mathematics Achievement

R_t (Neuroticism)	1	2	3	4	5	6	7	8	9	10
Q_t (Mathematics)	9	3	8	5	4	2	10	1	7	6
$R_t - Q_t$	-8	-1	-5	-1	1	4	-3	7	2	4

whether there is a dependence between emotional instability and performance in mathematics.

18.3 Correlation versus Causation

Correlation and causation (or causality) are two key concepts frequently encountered in statistics and scientific research. Correlation refers to the relationship between two variables, as discussed in the preceding chapters. Changes in the values of one variable are associated with changes in the values of the other variable. However, a correlation between variables does not automatically imply that a change in one variable is the cause of a change in the other. When such a causal relationship exists, we speak of causation. This means that one event is a consequence of the occurrence of another event; that is, a cause-and-effect relationship exists between them. Although correlation does not imply causality, the absence of correlation excludes it. In other words, if two variables are uncorrelated, we cannot speak of causality.

When a strong correlation (association) exists between two variables, but changes in neither variable cause changes in the other, the observed relationship may be purely coincidental or may be influenced by a third variable affecting both variables simultaneously. We shall present several examples.

Example 18.3 (Positive Correlation without Causation). During the summer months, warmer weather leads to an increase in ice cream sales and, at the same time, an increase in the number of people swimming, which may result in a higher number of drownings. These two variables (ice cream sales and the number of drownings) are positively correlated because both increase during the same period. However, ice cream sales do not cause drownings. The common factor affecting both variables is warm weather.

Example 18.4 (Erroneous Causal Interpretation). It is often claimed that exposure to violence in video games or television (phenomenon A) causes children to become more aggressive (phenomenon B). Such a statement

implies a causal relationship. However, one could also argue that children with a natural tendency toward aggressive behaviour are more likely to seek out games and television programs containing violent content. Therefore, we cannot determine with certainty whether A causes B , or whether B causes A . In such a situation, we can speak only of correlation, not of causation.

Example 18.5 (Correlation Induced by a Third Variable – Ice Cream Sales and Revenue). Data may show a positive correlation between higher sales at an ice cream stand and an increased incidence of heatstroke among people. At first glance, one might conclude that higher sales cause heatstroke (or vice versa). In reality, however, there is a third variable that influences both factors and provides a better explanation of the observed phenomenon – the weather. During hot weather, people seek ways to cool down and are more likely to visit ice cream stands, leading to increased sales. At the same time, hot weather increases the risk of heatstroke. Thus, weather is the key variable that affects both of these phenomena.

Example 18.6 (Spurious Correlation Due to Chance – Number of Nicolas Cage Films and Swimming Pool Drownings). Data may suggest that in years when Nicolas Cage appeared in a larger number of films, the number of drownings in swimming pools also increased. This correlation is purely coincidental, and there is no logical reason why the number of films featuring Nicolas Cage should influence the number of swimming pool drownings. This example illustrates how some correlations may arise entirely by chance and without any causal relationship. Such examples demonstrate the importance of careful data analysis and the danger of drawing conclusions solely on the basis of correlation.

Example 18.7 (Causation without a Third Variable – Exercise and Physical Fitness). When a person exercises regularly, their physical fitness improves. In this case, exercise is the cause and improved physical fitness is the effect. There is no third variable that explains this relationship. We can directly observe that regular exercise leads to improved physical fitness.

18.4 Exercises

18.1 Table 18.3 presents the numbers of research and development (R&D) personnel, measured in thousands, and the total wage expenditure, measured in billions of CZK, for the years 2010–2020. Examine the dependence between the number of employees and wage expenditure. Construct a scatter plot and compute both Pearson’s and Spearman’s correlation coefficients.

Table 18.3: Numbers of R& D Personnel in the Czech Republic (in Thousands) and Wage Expenditure (in Billions of CZK), 2010–2020

Year	2010	2011	2012	2013	2014	2015
Numbers of Personnel	52	57	60	62	64	67
Wages	26	25	30	29	45	46
Year	2016	2017	2018	2019	2020	
Numbers of Personnel	65	70	75	79	81	
Wages	35	35	71	70	69	

18.2 Table 18.4 presents the average number of defective products per shift for individual employees and their length of service with the company. Determine Pearson’s and Spearman’s correlation coefficients. Does the length of employment affect the number of defective products produced by individual employees? At the significance level 0.05, test the hypothesis that the number of defective products is independent of the length of employment with the company.

Table 18.4: Table of Numbers of Defective Items (z) and Length of Employment (d) of Individual Employees in the Company

d	1	2	3	10	5	25	15	3	10	0	4	40	4	11	8	30	2	5
z	4	6	1	2	1	4	0	7	2	5	2	2	3	1	5	2	0	1

18.3 Table 18.5 presents the average number of defective products per shift produced on machines of different ages. By calculating Pearson’s and Spearman’s correlation coefficients, assess the strength of association between these variables. Construct a scatter plot of the two variables, display the regression line and its equation, and test the hypothesis of no correlation at the significance level 0.05.

Table 18.5: Table of Numbers of Defective Products (z) and the Age of Manufacturing Machines (s)

s	10	2	7	12	20	25	11	3	0	5	16	22
z	4	6	1	2	1	4	0	7	2	5	2	2

18.4 Based on a random sample of size 24, we wish to assess the strength of the relationship between age and blood pressure in a group of women. The data are presented in Table 18.6. At the significance level 0.05, test whether the variables V and T are uncorrelated.

Table 18.6: Age of Women (v) in Years and Their Blood Pressure (t) in mm Hg

v	46	52	30	34	60	45	63	39	49	41	31	55
t	130	140	113	119	145	133	151	117	145	135	115	136
v	46	50	30	63	51	42	35	70	45	50	25	37
t	133	141	110	137	132	135	111	165	125	135	112	115

18.5 Table 18.7 presents data on store turnover O (in millions of CZK) and retail floor area S . The stores are located in comparable locations within the same city and have similar customer traffic.

- Calculate the correlation coefficient.
- At the significance level 0.01, test the hypothesis that there is a positive relationship between retail floor area and turnover.

Table 18.7: Turnover o (in Millions of CZK) and Retail Floor Area s (in m^2)

s	55	52	66	58	52	45	68	50	49	60
o	8, 53	8, 30	11, 52	9, 50	8, 99	7, 90	11, 49	8, 20	8, 55	10, 72

Solutions to the Exercises

18.1 Figure 18.2 shows a scatter plot of wage expenditure as a function of the number of employees. Pearson's correlation coefficient is $r_P \approx 0.987$, while Spearman's rank correlation coefficient is $r_S \approx 0.991$. Both coefficients, as well as their similarity, indicate a very strong positive correlation between the number of employees engaged in research and development and the total amount of wages paid.

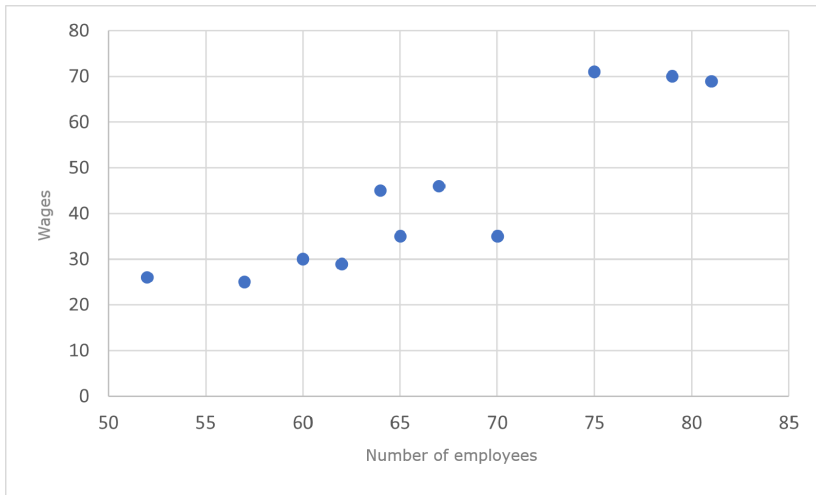


Figure 18.2: Scatter Plot of Wage Expenditure (in Billions of CZK) as a Function of the Number of Employees (in Thousands)

18.2 From the values $r_P \approx -0.190$ and $r_S \approx -0.426$, we infer a weak negative correlation between the variables. Visually, we also observe a high degree of dispersion in the data (see Figure 18.3). Since $|r_S|$ does not exceed the critical value 0.4716, the hypothesis of no correlation between the variables cannot be rejected: it cannot be concluded unambiguously that the number of defective products decreases as the length of employment with the company increases.

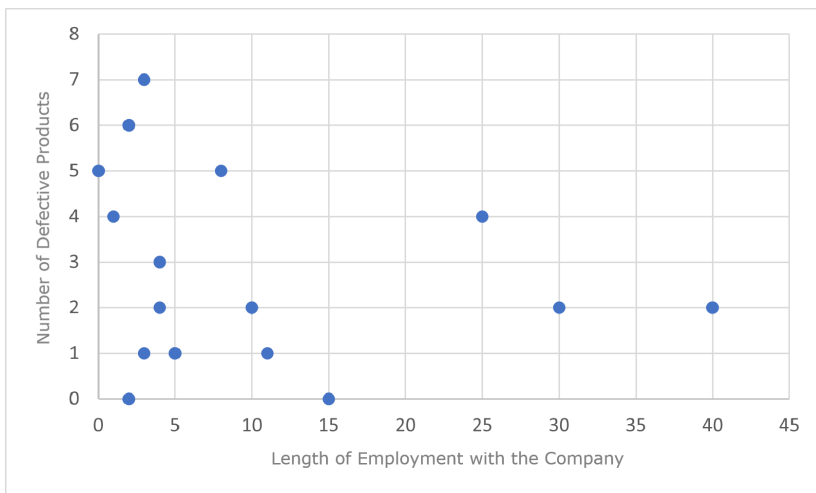


Figure 18.3: Scatter Plot of the Dependence of the Number of Defective Products on the Length of Employment with the Company (in Years)

18.3 From the values $r_P \approx 0.993$ and $r_S \approx 0.994$, we infer a very strong positive correlation between the variables. Thus, it may be concluded that the older the machine, the greater the number of defective products it produces. This conclusion is also supported by the low dispersion of the data (see Figure 18.4). Since $|r_S|$ exceeds the critical value 0.5804, we reject the hypothesis that the two variables are uncorrelated.

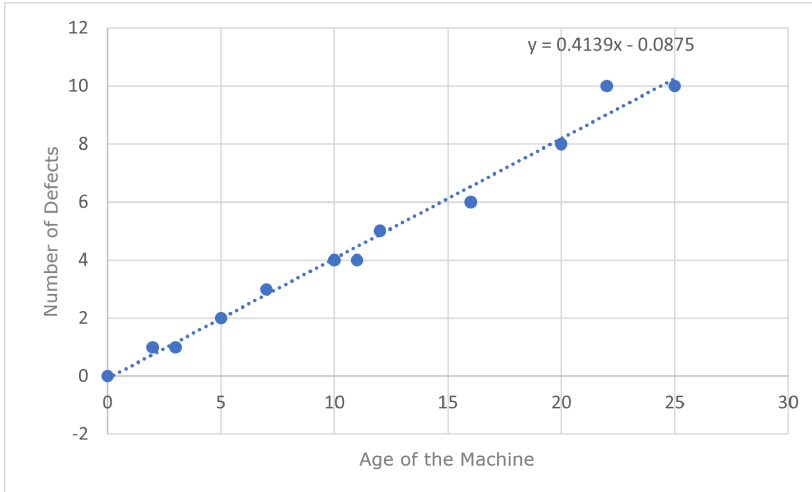


Figure 18.4: Scatter Plot of the Dependence of the Number of Defective Products on Machine Age (in Years)

18.4 Assuming that both variables are normally distributed in the population, we obtain $r_P \approx 0.912$.

Test of independence: $H_0: \rho_{V,T} = 0$, $H_1: \rho_{V,T} \neq 0$.

Since $|t| \approx 7.991 > t_{22}(0.975) \approx 2.16$, we reject the null hypothesis H_0 of no correlation between the variables V and T . Therefore, the variables V and T are correlated, indicating the existence of a linear relationship between age and blood pressure in women.

18.5

- a) $r_P \approx 0.952$;
- b) Testing for a positive association: $H_0: \rho_{O,S} = 0$, $H_1: \rho_{O,S} > 0$. Assuming that S and O are normally distributed, the test statistic is $t \approx 8.76$, and the critical value is $t_8(0.95) \approx 1.86$. Since $t > t_k$, we reject H_0 of no correlation between the variables S and O . The variables S and O are positively correlated; thus, it can be concluded that an increase in retail floor area is associated with an increase in store turnover.

Chapter 19

Regression Analysis

In practice, some random variables are readily available, for example through simple measurement, whereas other variables are difficult to determine or become available only after a considerable time delay. If a relationship exists between these two types of variables, the values of one can be used to estimate or predict the values of the other. When specifying a functional relationship, the primary objectives are to demonstrate the existence of the relationship, estimate its parameters, provide an empirical representation of large data sets, interpolate missing values, generate predictions, and so forth. Linear regression methods serve these purposes, and their fundamental principles are described in this chapter.

19.1 Simple Linear Regression with Single Explanatory Variable

The regression model

$$Y = f(x)$$

describes the dependence of the variable Y on the values of the variable x through the regression function f . The objective of regression analysis is to determine the regression function f when n observed pairs

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n),$$

are known, where x_i are the values of the independent (explanatory) variable x , and y_i are the values of the dependent variable (i.e., the observed values of the response variable Y).

Assume that the values y_i are measured with some error e_i . In deriving all hypothesis tests and confidence intervals throughout this chapter, we

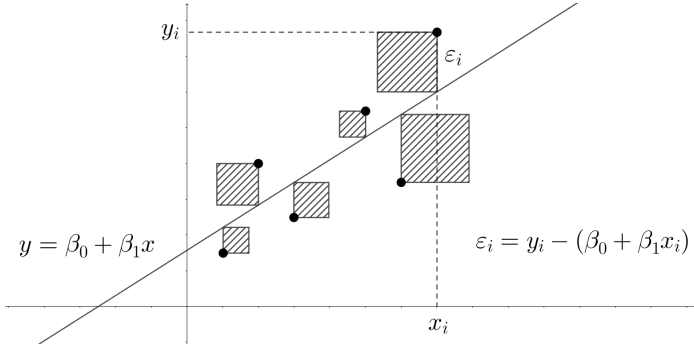


Figure 19.1: Least Squares Method

assume that the errors e_i follow a normal distribution with zero mean and constant variance, i.e. $N(0, \sigma^2)$. This assumption is not required for the derivation of point estimates. In other words, we have n observations of the response variable Y at n known values of the explanatory variable x , and therefore we have n equations:

$$Y_i = f(x_i) + e_i, \quad i = 1, 2, \dots, n.$$

The simplest form of regression is simple linear regression, which assumes a linear relationship between two variables expressed by the equation

$$f(x) = \beta_0 + \beta_1 x.$$

The objective of linear regression is to estimate the parameters β_0 and β_1 . The regression coefficient β_1 represents the change in $f(x)$ associated with a one-unit change in x . Consequently, its value depends on the units in which x and y are measured.

The problem of determining the equation of the regression line is solved using the least squares method (see Figure 19.1). This method consists in finding the parameters β_0 and β_1 for which the sum of squared model errors is minimized. In other words, we seek the minimum of the function L (from the English term *loss function*)

$$L(\beta_0, \beta_1) = \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 x_i))^2.$$

Therefore, we solve the following system of equations:

$$\frac{\partial L(\beta_0, \beta_1)}{\partial \beta_0} = 0, \quad \frac{\partial L(\beta_0, \beta_1)}{\partial \beta_1} = 0.$$

After simplification, we obtain the following estimates:

$$b_1 = \frac{\sum (x_i - \bar{x})Y_i}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i Y_i - n\bar{x}\bar{Y}}{\sum x_i^2 - n\bar{x}^2}, \quad b_0 = \bar{Y} - b_1\bar{x}, \quad (19.1)$$

where $\bar{x} = \frac{1}{n} \sum x_i$ and $\bar{Y} = \frac{1}{n} \sum Y_i$. The estimators b_0 and b_1 are the best linear unbiased estimators, that is, the estimators b_0 and b_1 are unbiased ($\mathbb{E}b_0 = \beta_0$, $\mathbb{E}b_1 = \beta_1$) and have the smallest variance among all unbiased estimators.

Minimum of the loss function L , given by

$$S_e = L(b_0, b_1) = \sum (Y_i - (b_0 + b_1 x_i))^2 = \sum Y_i^2 - b_0 \sum Y_i - b_1 \sum x_i Y_i,$$

is called the *residual sum of squares*. The estimator of the error variance σ^2 is

$$S^2 = \frac{S_e}{n-2},$$

its square root S is called the *standard error of the regression*. Finally, the *total sum of squares*

$$S_T = \sum (Y_i - \bar{Y})^2$$

expresses the total variability of the response variable.

The adequacy of the model is assessed by the *coefficient of determination*

$$R^2 = 1 - \frac{S_e}{S_T} = \frac{S_T - S_e}{S_T},$$

which expresses the proportion of the total variability S_T explained by the regression model. The error term S_e thus contains the part of the variability that cannot be explained by the regression model. The coefficient of determination can also be calculated using the formula

$$R^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2},$$

where $\hat{Y}_i = \hat{f}(x_i) = b_0 + b_1 x_i$ is the regression estimate of the value of the regression function at the point x_i . It is evident that the closer R^2 is to one, the better the regression model fits the observed data. It is sometimes stated that if the coefficient of determination exceeds 0.85, the model may be regarded as well fitted. However, in disciplines that typically deal only with observational (non-experimental) data (e.g. econometrics or the social sciences), such high values of the coefficient of determination are achieved

rather rarely. Nevertheless, even models with substantially lower coefficients of determination may still possess considerable explanatory value.

Most often, we are interested in determining whether the model can be simplified so that the values Y_i do not depend on x_i at all. Consequently, we test the hypothesis

$$H_0: \beta_1 = 0 \text{ against } H_1: \beta_1 \neq 0.$$

Under the null hypothesis H_0 , the test statistic

$$T = \frac{b_1}{S} \sqrt{\sum x_i^2 - n\bar{x}^2} \sim t_{n-2}$$

follows Student's t distribution with $n - 2$ degrees of freedom. Therefore, if $|T| \geq t_{n-2}(\frac{\alpha}{2})$, we reject the null hypothesis H_0 at the significance level α . By rejecting H_0 in this test, we have in fact confirmed the existence of a linear dependence of Y_i on x_i , obscured by the random errors e_i .

We may also test the hypothesis $H_0: \beta_0 = 0$ against $H_1: \beta_0 \neq 0$. If H_0 is not rejected, we have demonstrated that the regression line passes through the origin. Tests of the significance of the regression coefficients are collectively referred to as *partial t-tests*.

A standard $(1 - \alpha)$ confidence interval for the parameter β_1 is given by

$$\left[b_1 - \frac{t_{n-2} \left(1 - \frac{\alpha}{2}\right) S}{\sqrt{\sum x_i^2 - n\bar{x}^2}}, b_1 + \frac{t_{n-2} \left(1 - \frac{\alpha}{2}\right) S}{\sqrt{\sum x_i^2 - n\bar{x}^2}} \right],$$

A $(1 - \alpha)$ confidence interval for the parameter β_0 is given by

$$[b_0 - t_{n-2}s_{b_0}, b_0 + t_{n-2}s_{b_0}], \quad (19.2)$$

where s_{b_0} is the standard error of the estimator b_0 .

The values of the above sums of squares (S_e , S_T) are usually reported by statistical software in the form of an ANOVA table, which is associated with testing the hypothesis of whether the specified relationship between the variables exists at all. In contrast, the partial t -tests are used to assess the statistical significance of the individual regression coefficients.

In some cases, we seek a confidence interval for $\beta_0 + \beta_1 x$ for a given value of x . This interval can be derived as

$$\left[b_0 + b_1 x - t_{n-2} \left(1 - \frac{\alpha}{2}\right) S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2}}, \right. \\ \left. b_0 + b_1 x + t_{n-2} \left(1 - \frac{\alpha}{2}\right) S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2}} \right].$$

Table 19.1: Data on the number of operating hours x_i of the production line and the maintenance costs Y_i , in thousands of CZK

x_i	275	350	250	325	375	400	300
Y_i	149	170	140	164	192	200	165

This interval contains the value $\beta_0 + \beta_1 x$ with probability $1 - \alpha$. If such intervals are constructed for all $x \in [\min x_i, \max x_i]$, they form the so-called confidence band around the regression line. The boundaries of the band are given by two branches of a hyperbola.

The relationship between regression and the correlation of two variables X and Y is expressed by the relationship between Pearson's correlation coefficient and the regression coefficient b_1 :

$$b_1 = R_P \frac{S_Y}{S_X}.$$

Thus, the regression coefficient b_1 of the regression line can be computed from the correlation coefficient R_P , and vice versa. This implies that the interpretative power of simple regression is equivalent to that of correlation analysis.

Example 19.1.  Table 19.1 contains records of the number of operating hours of a production line (x_i) and the corresponding maintenance costs (Y_i), in thousands of CZK, during the first seven months of the year.

- Using MS Excel, create a scatter plot illustrating the dependence of maintenance costs (Y_i) on the number of operating hours of the production line x_i (this plot is referred to as a scatter diagram) and select an appropriate linear regression model.
- Verify the adequacy of the selected model (coefficient of determination and partial t -tests).
- If the operating time of the production line increases by one hour, by how much will the maintenance costs increase?
- Interpret the fitted regression model.

Solution. We first determine the regression line $Y = b_0 + b_1 x$. Substituting into formulas 19.1, we obtain $b_0 = 42.75$ and $b_1 = 0.387$. Therefore, the equation of the regression line is $y = 42.75 + 0.387x$. Next, we calculate the residual sum of squares, $S_e = 148.821$; hence the estimate of the error variance is $s^2 = 29.76$. The total sum of squares can be easily computed as $S_T = (n - 1)S_Y^2$, where S_Y^2 is the sample variance of Y . We obtain $S_T = 2771.71$. Therefore, the coefficient of determination is $R^2 \approx 0.9463$.

SUMMARY OUTPUT	
<i>Regression Statistics</i>	
Multiple R	0.9727832
R Square	0.9463071
Adjusted R Square	0.9355685
Standard Error	5.4556655
Observations	7

Figure 19.2: Main Output of the Regression Analysis Tool in MS Excel for Example 19.1

The regression line can be added to the scatter plot in MS Excel by selecting *Trendline* from the options available under the + icon. The values of b_0 and b_1 can be read directly from the trendline equation, which can be displayed by selecting the *Display Equation on Chart* option on the *Format Trendline* pane. The same pane also contains the option *Display R-squared value on chart*, which displays the coefficient of determination.

If we wished to determine the intercept b_0 of the regression line with the y -axis without explicitly estimating the regression equation, the Excel function `= INTERCEPT` could be used. In this example, $b_0 = 42.75$.

`= INTERCEPT(C4:C10,B4:B10)`

The most comprehensive tool for regression analysis in MS Excel is the *Regression* procedure, which is included in the *Analysis ToolPak*. After activating the add-in, select *Regression* from the *Data Analysis* menu. After specifying the required inputs and output options, the mandatory information includes the range containing the dependent variable y and the values of one or more independent variables x . The output generated by the *Regression* tool for Example 19.1 is shown in Figure 19.2.

The first statistic (*Multiple R*) represents Pearson's correlation coefficient, whose square equals the coefficient of determination. The second statistic (*R Square*) is the coefficient of determination itself. The third statistic (*Adjusted R Square*) is the adjusted coefficient of determination, which may be used when comparing models with different numbers of parameters. The fourth statistic (*Standard Error*) represents the standard error of the regression (an estimate of the standard deviation σ), and the final item (*Observations*) denotes the sample size (number of observations).¹

¹Precise definitions of these concepts can be found in the statistical literature, for example [3],[4].

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	2622.89	2622.893	88.1222	0.000231306
Residual	5	148.821	29.76429		
Total	6	2771.71			

Figure 19.3: ANOVA Table for Example 19.1

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95%</i>
Intercept	42.75	13.56100163	3.1524	0.02531	7.89033553	77.6096645	7.8903
X Variable 1	0.3871429	0.041240954	9.3873	0.00023	0.28112961	0.49315611	0.2811

Figure 19.4: Table of Estimates of the Regression Line Coefficients and Their Confidence Intervals for Example 19.1

If it is necessary to calculate the standard error of the regression separately, the MS Excel function `= STEYX` can be used. For Example 19.1, the following expression may be entered in MS Excel:

`= STEYX(C4:C10,B4:B10)` ≈ 5.456 .

Another possible output is an ANOVA table for the observed values of the dependent variable Y : the total sum of squares S_T is equal to the sum of the regression sum of squares² and the residual sum of squares S_e (see Figure 19.3). The total sum of squares can also be obtained in MS Excel using the function `= DEVSQ`. For Example 19.1, the corresponding calculation is `= DEVSQ(C4:C10)` ≈ 2771.714 , as also reported in the table shown in Figure 19.3.

The principal output of the *Regression* analysis tool is a table containing the estimates of the regression coefficients b_0 and b_1 (*Coefficients*), estimates of the standard errors of these estimators (listed in the table as *Standard Error*), the t -test statistic (*t Stat*), the smallest significance level at which H_0 can still be rejected, i.e. the p -value (*P-value*), and the 95% confidence intervals (*Lower 95%*, *Upper 95%*). (See Figure 19.4.)

The *Regression* tool also produces a table of residuals (see Figure 19.5) and their standardized counterparts. These data can be used for a graphical assessment of the assumptions underlying linear regression, particularly the independence of residuals, homoscedasticity, and the linearity of the relationship between Y and X . If the residual plot exhibits a trend or some recognizable *pattern*, then at least one of the assumptions is likely to be

²This quantity is not introduced in the theoretical section; however, it is used in regression analysis to assess how well the regression equation fits the observed data. For further details, see the MS Excel Help documentation.

RESIDUAL OUTPUT			
<i>Observation</i>	<i>Predicted Y</i>	<i>Residuals</i>	<i>Standard Residuals</i>
1	149.21429	-0.214285714	-0.043026509
2	178.25	-8.25	-1.656520593
3	139.53571	0.464285714	0.093224103
4	168.57143	-4.571428571	-0.917898857
5	187.92857	4.071428571	0.817503669
6	197.60714	2.392857143	0.480462683
7	158.89286	6.107142857	1.226255504

Figure 19.5: Table of Residuals for Example 19.1

violated. For testing graphical time trends and for computing the Durbin³–Watson⁴ test, which is not discussed here, residual plots can be used.

Let us now consider the hypothesis $H_0 : \beta_1 = 0$. We compute (or read from the table in Figure 19.4) the test statistic $t \approx 9.38$ and compare it with the corresponding quantile value $t_5(0.975) \approx 2.57$ = T.INV(0.975,5). Therefore, we reject the null hypothesis H_0 at the 5% significance level. Both the coefficient of determination and this test confirm the adequacy of the linear model.

Let us also examine the confidence intervals. The 95% confidence interval for the parameter β_1 is [0.2811, 0.4931]. This interval likewise shows that the hypothesis H_0 should be rejected. The confidence band around the regression line is not shown in Figure 19.6.

The remaining task is to interpret the model. The intercept term b_0 of the model represents the fixed monthly costs (costs incurred each month regardless of the operating time of the production line), whereas the linear term b_1x provides an estimate of the variable costs directly proportional to the operating time of the production line. $\triangle$

19.2 Multiple Linear Regression with Multiple Explanatory Variables

Regression is one of the fundamental statistical methods. Its objective is to find a regression function, which attempts to explain a larger set of observed

³James Durbin (1923–2012), British statistician and econometrician.

⁴Geoffrey S. Watson (1921–1998), Australian statistician.

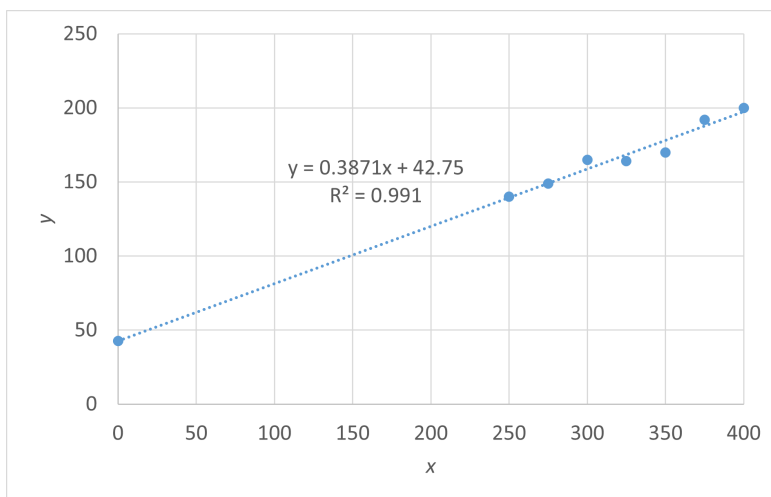


Figure 19.6: Dependence of Operating Costs on Operating Time (points represent the observed values, and the solid line denotes the estimated regression line) $Y = b_0 + b_1x$

random variables $Y_1, Y_2, \dots, Y_n$ by means of known effects x_{ij} and a relatively small number of parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_k$. A regression model is considered linear if the dependence on the parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ is linear.

For a single response variable Y , we have available k explanatory variables. Therefore, in this section we will work with the model

$$Y = \beta_0 + \beta_1x_1 + \dots + \beta_kx_k. \quad (19.3)$$

If we have n observations, we obtain n equations in $k + 1$ unknown parameters of the form

$$Y_i = \beta_0 + \beta_1X_{i1} + \beta_2X_{i2} + \dots + \beta_kX_{ik} + e_i, \quad (19.4)$$

where $i = 1, 2, \dots, n$ and e_i are random errors. For the derivation of all hypothesis tests and confidence intervals throughout this chapter, we also assume that the errors e_i follow a normal distribution $N(0, \sigma^2)$. Thus, linear regression relies on four assumptions: linearity of the relationship between $\mathbf{X}$ and Y , independence of observations, homoscedasticity (constant variance σ^2 across all observations), and normality of the errors. For large sample sizes, the normality assumption may be relaxed owing to the CLT. The remaining assumptions, however, are important for model fitting and will be assessed using residual plots. The normality assumption is not required for the derivation of point estimates.

The matrix representation of this model has the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & \dots & X_{1k} \\ 1 & X_{21} & \dots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \dots & X_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

The objective of linear regression is to estimate the parameters of the model $\beta_0, \beta_1, \beta_2, \dots, \beta_k$. The most commonly used method for estimating these parameters is the *least squares method*, which consists in minimizing the function L

$$L(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik}))^2. \quad (19.5)$$

A necessary condition for the existence of an extremum is that all partial derivatives vanish. Since the given function is convex over its domain, this is also a sufficient condition. Differentiating the function with respect to all variables and setting the resulting partial derivatives equal to zero yields the following system of equations

$$\begin{aligned} \frac{\partial L(\beta_0, \beta_1, \dots, \beta_k)}{\partial \beta_0} &= -2 \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik})) = 0 \\ \frac{\partial L(\beta_0, \beta_1, \dots, \beta_k)}{\partial \beta_j} &= -2 \sum_{i=1}^n X_{ij} (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik})) = 0, \end{aligned}$$

where $j = 1, 2, \dots, k$.

After some algebraic manipulation, we obtain the system

$$\begin{aligned} n\beta_0 + \beta_1 \sum_{i=1}^n X_{i1} + \beta_2 \sum_{i=1}^n X_{i2} + \dots + \beta_k \sum_{i=1}^n X_{ik} &= \sum_{i=1}^n Y_i \\ \beta_0 \sum_{i=1}^n X_{i1} + \beta_1 \sum_{i=1}^n X_{i1}^2 + \beta_2 \sum_{i=1}^n X_{i2} X_{i1} + \dots + \beta_k \sum_{i=1}^n X_{ik} X_{i1} &= \sum_{i=1}^n Y_i X_{i1} \\ &\vdots \\ \beta_0 \sum_{i=1}^n X_{ik} + \beta_1 \sum_{i=1}^n X_{ik} X_{i1} + \beta_2 \sum_{i=1}^n X_{ik} X_{i2} + \dots + \beta_k \sum_{i=1}^n X_{ik}^2 &= \sum_{i=1}^n Y_i X_{ik}. \end{aligned}$$

Solving this system yields the estimators $b_0, b_1, \dots, b_k$ of the parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_k$.

The above system is called the *system of normal equations*. The matrix form of the system of normal equations is

$$(\mathbf{X}^T \mathbf{X}) \cdot \boldsymbol{\beta} = \mathbf{X}^T \mathbf{Y}. \quad (19.6)$$

If the matrix $(\mathbf{X}^T \mathbf{X})$ is nonsingular, then its inverse matrix exists (denoted by $(\mathbf{X}^T \mathbf{X})^{-1}$). The estimator of the parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_k)^T$ is then given by

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}. \quad (19.7)$$

The minimum of the function L is called the *residual sum of squares* and is calculated as

$$S_e = L(\mathbf{b}) = \sum (Y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + b_k x_{ik}))^2 = \sum (Y_i - \hat{Y}_i)^2,$$

where $\hat{Y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + b_k x_{ik}$ is the regression estimate of the value Y_i . The estimator of the error variance σ^2 is $S^2 = \frac{S_e}{n-k-1}$ and is called the *residual variance*; its square root is referred to as the *standard error of the regression*. The *total sum of squares* $S_T = \sum (Y_i - \bar{Y})^2$ expresses the total variability of the response variable.

The adequacy of the model is assessed by the coefficient of determination

$$R^2 = 1 - \frac{S_e}{S_T} = \frac{S_T - S_e}{S_T},$$

which expresses the proportion of the total variation S_T explained by the regression model. (The residual component S_e represents the part of the variability that cannot be explained by the regression model.) The coefficient of determination can also be computed using the formula

$$R^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2},$$

where $\hat{Y}_i = \hat{f}(x_i) = b_0 + b_1 x_{i1} + \dots + b_k x_{ik}$ is the regression estimate of Y_i . It is evident that the closer R^2 is to one, the better the regression model fits the observed data.

Using the least squares method, we obtain point estimates of the regression parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_k$. In some situations, however, we are also interested in constructing $(1 - \alpha)$ confidence intervals for the parameters

$\beta_0, \beta_1, \beta_2, \dots, \beta_k$. A $(1 - \alpha)$ confidence interval for the parameter β_i is the interval

$$\left[b_i - t_{n-k-1}(\alpha) \cdot s \sqrt{(\mathbf{X}^T \mathbf{X})_{ii}^{-1}}, b_i + t_{n-k-1}(\alpha) \cdot S \sqrt{(\mathbf{X}^T \mathbf{X})_{ii}^{-1}} \right], \quad (19.8)$$

where $(\mathbf{X}^T \mathbf{X})_{ii}^{-1}$ is the element of the matrix $(\mathbf{X}^T \mathbf{X})^{-1}$ located in the i th row and the i th column.

It is evident that the fewer explanatory variables are included, the simpler the model becomes. Consequently, we are often interested in determining whether the model can be simplified to the extent that the values Y_i do not depend on the explanatory variables x_{ij} at all. Therefore, we test the hypothesis

$$H_0: \beta_j = 0 \quad \text{against} \quad H_1: \beta_j \neq 0$$

Under the null hypothesis H_0 , the test statistic has the distribution

$$T = \frac{b_j}{S \cdot \sqrt{(\mathbf{X}^T \mathbf{X})_{jj}^{-1}}} \sim t_{n-k-1}. \quad (19.9)$$

Student's t distribution with $n - k - 1$ degrees of freedom. Therefore, if

$$|t| \geq t_{n-k-1} \left(1 - \frac{\alpha}{2} \right),$$

we reject the null hypothesis H_0 at the significance level α . If the null hypothesis H_0 is rejected, we have in fact confirmed the existence of a linear dependence of Y_i on the j th explanatory variable, obscured by the random errors e_i .

Sometimes we wish to determine whether the model can be simplified by removing more than one parameter. In such a case, the two previous tests are not appropriate, since their joint significance level would no longer be α . Instead, we use the following test and consider the hypothesis

$$H_0: \beta_{j_1} = \beta_{j_2} = \dots = \beta_{j_l} = 0, \quad 1 \leq j_1, \dots, j_l \leq k,$$

against the alternative that the reduced model is not valid (i.e. at least one $\beta_{j_i} \neq 0$). Here, the number l denotes the number of parameters that we attempt to remove from the model. The matrix representation of the reduced model has the form

$$\mathbf{Y} = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \tilde{\mathbf{e}},$$

where the matrix $\tilde{\mathbf{X}}$ is obtained from the matrix $\mathbf{X}$ by removing the columns corresponding to the parameters $\beta_{j_1}, \beta_{j_2}, \dots, \beta_{j_l}$, the vector $\tilde{\boldsymbol{\beta}}$ is obtained

from the vector β by removing the parameters $\beta_{j_1}, \beta_{j_2}, \dots, \beta_{j_l}$, and the vector $\tilde{e}$ is defined analogously.

The parameters of the reduced model $\tilde{\beta}$ are estimated using

$$\tilde{b} = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{Y}. \quad (19.10)$$

The residual sum of squares for the reduced model is then computed as

$$\tilde{S}_e = \sum (Y_i - \hat{\tilde{Y}}_i)^2,$$

where $\hat{\tilde{Y}}_i$ is the regression estimate of Y_i in the reduced model. It is evident that $\tilde{S}_e \geq S_e$, since S_e is the minimum of the function $L(\beta)$ without any restrictions on the vector β , whereas $\tilde{S}_e$ is the minimum of the function $L(\beta)$ subject to the constraint $\beta_{j_1} = \beta_{j_2} = \dots = \beta_{j_l} = 0$.

Under the null hypothesis H_0 , the test statistic follows an F-distribution with l and $n - k - 1$ degrees of freedom:

$$F = \frac{(\tilde{S}_e - S_e)/l}{S_e/(n - k - 1)} \sim F_{l, n-k-1}.$$

If $f \geq F_{l, n-k-1}(1 - \alpha)$, we reject the null hypothesis H_0 at the significance level α ; consequently, the model cannot be simplified.

Example 19.2. In the 1960s, the following study was conducted in Great Britain. In 30 counties, the following variables were measured: A , representing the change in population over the previous ten years; B , the number of employees in agriculture; C , the amount of property taxes; D , the percentage of households with telephone service; E , the percentage of residents living in rural areas; and F , the average age of the population. These variables were used to explain the variable Y , which denotes the percentage of residents living below the poverty line.

Solution. Based on the data available in the online help documentation of the Statistica⁵ software package, we obtain the following linear regression model

$$Y_i = \beta_0 + \beta_A A_i + \beta_B B_i + \beta_C C_i + \beta_D D_i + \beta_E E_i + \beta_F F_i + e_i.$$

The matrix $\mathbf{X}$ will contain seven columns: the first column will consist entirely of ones, the second will contain the values of variable A , the third

⁵<https://docs.tibco.com/pub/stat/14.0.0/doc/html/UsersGuide/GUID-4F43645C-C174-48E2-99C5-45BD091A60B81.html>

the values of B , and so on. We then compute the estimates of the individual parameters using formula (19.7)

$$\begin{aligned}\mathbf{b} &= (b_0, b_A, b_B, b_C, b_D, b_E, b_F)^T = \\ &= (31.26, -0.39, 0.0007, 1.23, -0.083, 0.16, -0.42)^T.\end{aligned}$$

Next, we compute the residual sum of squares $S_e = 265.66$ and the total sum of squares $S_T = 1197.72$. Hence, we obtain $R^2 = 0.78$.

We are now interested in determining whether some of the explanatory variables can be removed from the model, i.e. we test the hypothesis $H_0: \beta_j = 0$. For this purpose, we compute the value of the test statistic T for each variable using formula (19.9).

$$\begin{aligned}\mathbf{t} &= (t_0, t_A, t_B, t_C, t_D, t_E, t_F)^T = \\ &= (2.35, -4.87, 1.69, 0.38, -0.63, 2.67, -1.64)^T.\end{aligned}$$

These values are compared with the quantile $t_{23}(0.975) = 2.068$

$$= \text{T.INV}(0.975, 23)$$

and we observe that the null hypothesis of a zero coefficient is rejected for β_0 , β_A , and β_E . These tests indicate that variables B , C , D , or F may potentially be omitted from the model. However, they do not tell us whether all of these variables can be removed simultaneously. This question is addressed by the following F -test.

Let us therefore consider the following reduced model (submodel)

$$Y_i = \beta_0 + \beta_A A_i + \beta_E E_i + e_i.$$

For the reduced model, we compute the estimates of the individual parameters

$$\tilde{\mathbf{b}} = (\tilde{b}_0, \tilde{b}_A, \tilde{b}_E)^T = (16.67, -0.40, 0.13)^T.$$

Next, we compute the residual sum of squares for the reduced model, $\tilde{S}_e \approx 393.03$. Consequently, we obtain $R^2 = 0.67$. We now construct the F -statistic and calculate its value.

$$f = \frac{(30 - 6 - 1)(\tilde{S}_e - S_e)}{3S_e} = 3.67,$$

we compute the critical value $F_{3,23}(0.05) \approx 3.03$

$$= \text{F.INV}(0.95, 3, 23)$$

and compare the two values. Since $3.67 > 3.03$, we reject H_0 ; that is, we cannot remove all four variables simultaneously.

Therefore, we must add one of the variables to the reduced model. Let us add variable B , because $t_B > t_C, t_D, t_F$ (i.e. variable B is more significant in the model than C , D , and F). Let us therefore consider the reduced model

$$Y_i = \beta_0 + \beta_A A_i + \beta_B B_i + \beta_E E_i + e_i.$$

For the reduced model, we compute the estimates of the individual parameters

$$\tilde{\mathbf{b}} = (\tilde{b}_0, \tilde{b}_A, \tilde{b}_B, \tilde{b}_E)^T = (10.99, -0.40, 0.001, 0.19)^T.$$

Next, we compute the residual sum of squares for the reduced model, $\tilde{S}_e \approx 318.83$. Consequently, we obtain $R^2 \approx 0.73$. We then construct the F -statistic and calculate its value.

$$f = \frac{(30 - 6 - 1)(\tilde{S}_e - S_e)}{4S_e} = 1.15.$$

We compute the critical value $F_{4,23}(0.05) \approx 2.80$.

$$= \text{F.INV}(0.95, 4, 23)$$

From the inequality $1.15 \not\geq 2.80$, it follows that H_0 is not rejected; that is, the variables C , D , and F can be omitted from the original model. We therefore describe the variable Y by means of the variables A , B , and E . $\triangle$

19.3 Polynomial Regression, Quadratic Regression

The simplest case of polynomial regression is the so-called quadratic regression. By quadratic regression we mean the model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i, \quad i = 1, 2, \dots, n,$$

where $e_i \sim \mathbf{N}(0, \sigma^2)$, $n \geq 4$. Here, the variables Y_i depend quadratically on the variables X_i .

If we define $Z_i = X_i^2$, $i = 1, 2, \dots, n$, we obtain the model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + e_i, \quad i = 1, 2, \dots, n.$$

In this model, the random variable Y_i depends linearly on the variables X_i and Z_i . In other words, the problem of quadratic regression has been transformed into a problem of multiple linear regression with two explanatory variables.

Table 19.2: Data on Advertising Expenditure (in Millions of CZK) and the Resulting Profit (in Millions of CZK)

Expenditure	0	2	4	5	8	9	12	14	16
Profit	-10	4	13	29	37	32	15	21	11

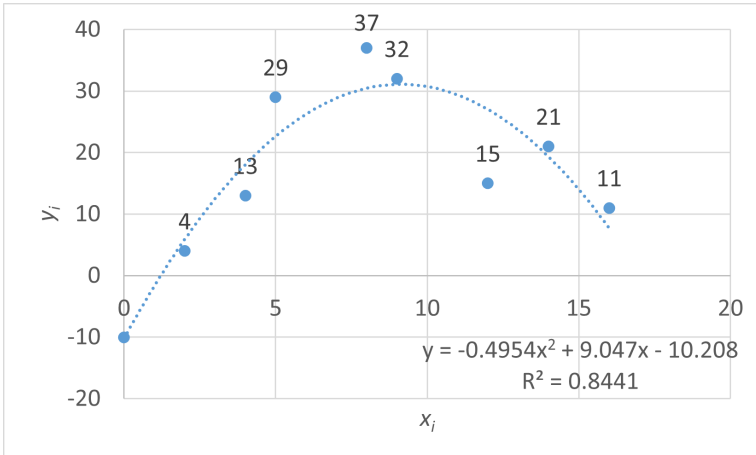


Figure 19.7: Scatter Plot with the Quadratic Regression Curve, Displayed Trendline Equation, and Coefficient of Determination

b_1	b_2	b_0
9.046994	-0.495438	-10.20839192

Figure 19.8: Excerpt from MS Excel Demonstrating the Use of the `=LINEST` Function

Example 19.3.  Table 19.2 contains data on the firm's profit as a function of its advertising expenditure.

- Find and, using MS Excel, graphically display a model describing the dependence of the firm's profit (in millions of CZK) on advertising expenditure (in millions of CZK). Verify the adequacy of the selected regression model using the coefficient of determination and the ANOVA test.
- Estimate the profit if advertising expenditure amounts to 6 million CZK.
- For what level of advertising expenditure will the firm achieve maximum profit?

ANOVA	
	<i>df</i>
Regression	2
Residual	6
Total	8
	<i>Coefficients</i>
Intercept	-10.2083919
X Variable 1	9.046993641
X Variable 2	-0.49543769

Figure 19.9: Excerpt from MS Excel Showing the ANOVA Table Generated by the Regression Tool within the Data Analysis Add-in

Solution. The scatter plot in Figure 19.7, created in MS Excel, graphically illustrates the dependence of the firm's profit on advertising expenditure. The regression model describing this relationship is a segment of a parabola with the approximate analytical form $y = -0.5x^2 + 9x - 10$ (after rounding the coefficients to whole numbers).

The coefficients in the quadratic regression equation can be obtained in MS Excel using the `=LINEST` function. The first argument is the range containing the y_i values, while the second argument is the range containing the columns of x_i^2 and x_i , in this order. The third and fourth arguments are both set to `TRUE`: the third specifies a model with an intercept term, and the fourth requests additional regression statistics in the output. The function is an array formula and must be confirmed by pressing `Ctrl+Shift+Enter` simultaneously.

`= LINEST(array1,array2,TRUE,TRUE)`.

The output produced by this array function is shown in Figure 19.8.

The coefficients of the quadratic regression equation can also be obtained using the *Regression* procedure within the *Data Analysis* tool. For the present example, the estimated coefficients are highlighted in Figure 19.9. The coefficient of determination is $R^2 \approx 0.844$. This value can be obtained either by displaying it directly in the MS Excel chart or from the Regression Statistics table generated by the *Data Analysis* tool (see Figure 19.10). The coefficient of determination R^2 appears as the second item in this table.

SUMMARY OUTPUT	
<i>Regression Statistics</i>	
Multiple R	0.918742955
R Square	0.844088617
Adjusted R Square	0.792118156
Standard Error	6.722010763
Observations	9

Figure 19.10: Excerpt from MS Excel Showing the Regression Statistics Table Generated by the Regression Tool within the Data Analysis Add-in

<i>MS</i>	<i>F</i>	<i>Významnost F</i>
733.888158	16.24169958	0.00378995
45.1854287		

Figure 19.11: Excerpt from MS Excel Showing the ANOVA Table Generated by the Regression Tool within the Data Analysis Add-in

<i>P-value</i>
0.110407665
0.001364245
0.002131367

Figure 19.12: Excerpt from MS Excel Showing the ANOVA Section of the Regression Output Generated by the Data Analysis Tool

The adequacy of the selected polynomial regression model can be assessed using the ANOVA p -value reported in the output of the *Regression* procedure within the *Data Analysis* tool (see Figure 19.11). We test the hypotheses

H_0 : The regression model is not statistically significant.

against the alternative

H_1 : The regression model is statistically significant.

Since $p \approx 0.0038 < 0.05$, we reject H_0 of the overall statistical insignificance of the selected regression model.

The statistical significance of the individual coefficients of the quadratic regression function can also be tested. For this purpose, the p -values of the corresponding t -tests, reported in the *P-value* column (see Figure 19.12)

of the *Regression* output generated by the *Data Analysis* tool, are used. The first value corresponds to the statistical significance of the coefficient b_0 . In this case, $p \approx 0.11 > 0.05$; therefore, we do not reject the null hypothesis that this coefficient is statistically insignificant. The second and third values correspond to the statistical significance of the coefficients b_1 and b_2 , respectively. Since both p -values are smaller than the significance level $\alpha = 0.05$, both b_1 and b_2 are statistically significant components of the fitted quadratic regression model.

The estimated profit for an advertising expenditure of 6 million CZK is obtained by substituting this value into the regression equation, yielding $y(6) \approx 26.2$ million CZK.

To determine the advertising expenditure that maximizes profit, one may, for example, apply differential calculus. This yields an optimal expenditure of approximately 9.13 million CZK. $\triangle$

Similarly to quadratic regression, higher-order polynomial regression models are handled in an analogous manner. One remaining issue is the choice of the degree of the regression polynomial. Consider the model

$$Y_i = \beta_0 + \beta_1 X_i + \cdots + \beta_p X_i^p + e_i, \quad i = 1, 2, \dots, n,$$

where the degree p of the regression polynomial is unknown. Let the number of parameters in this model be denoted by $k = p + 1$.

Assume that the true degree of the polynomial is p_0 , and therefore the true number of model parameters is $k_0 = p_0 + 1$. Let S_k^2 denote the residual variance of a model with k parameters (the residual variance is defined on p. 391). It can be shown that

$$\mathbb{E}S_k^2 > \sigma^2 \text{ pro } k < k_0,$$

and

$$\mathbb{E}S_k^2 = \sigma^2 \text{ pro } k \geq k_0.$$

It is therefore necessary to identify the point at which the sequence s_k^2 changes from a decreasing sequence to an oscillating one. Since this is a difficult task, we transform it into the problem of finding the minimum of a sequence. We construct the sequence

$$A_k = S_k^2 \left(1 + \frac{k}{\sqrt[4]{n}} \right),$$

which assigns a greater penalty to larger values of k . The value of k for which A_k attains its minimum is then taken as an estimate of the true number of parameters k_0 .

19.4 Nonlinear Regression

Consider the nonlinear regression model

$$Y_i = f(X_i, \boldsymbol{\beta}) + e_i, \quad i = 1, 2, \dots, n,$$

where f is the regression function and $\boldsymbol{\beta}$ is the vector of unknown parameters. The least-squares estimate of the parameter vector $\boldsymbol{\beta}$ is obtained by minimizing the expression

$$g(\boldsymbol{\beta}) = \sum_{i=1}^n (Y_i - f(X_i, \boldsymbol{\beta}))^2.$$

This optimization problem is solved iteratively by various statistical and mathematical software packages. Such programs, however, require an initial approximation of the vector $\boldsymbol{b}$ (an estimate of the parameter vector $\boldsymbol{\beta}$).

An initial approximation can be obtained quite easily for so-called linearizable models, i.e. models that can be transformed into a linear model. As an example, consider a model whose regression function is exponential:

$$Y_i = \beta_0 e^{\beta_1 X_i} + e_i, \quad i = 1, 2, \dots, n.$$

For the purpose of obtaining an initial approximation, we may neglect the errors e_i and take logarithms of the model

$$\ln Y_i = \ln \beta_0 + \beta_1 X_i, \quad i = 1, 2, \dots, n.$$

If we introduce the new parameters $\alpha_0 = \ln \beta_0$ and $\alpha_1 = \beta_1$, we obtain the linear regression model

$$\ln Y_i = \alpha_0 + \alpha_1 X_i, \quad i = 1, 2, \dots, n,$$

which can be solved using the methods described in Section 19.2, yielding the estimates a_0 and a_1 . As an initial approximation for the parameter estimates of the original model, we then take the estimates

$$b_0 = e^{a_0}, \quad b_1 = a_1.$$

Finally, let us present several examples of linearizable models:

- 1) $Y = e^{\beta_0 + \beta_1 X}$,
- 2) $Y = \beta_0 X^{\beta_1}$,
- 3) $Y = \ln(\beta_0 + \beta_1 X)$,
- 4) $Y = \frac{1}{\beta_0 + \beta_1 X}$.

Table 19.3: Data on Machine Age (in Years) and Maintenance Costs (in Hundreds of CZK)

Machine Age (x_i)	2	1.2	14.8	8.3	8.4	3
Costs (y_i)	12	8	76.4	17	21.3	10
Machine Age (x_i)	4.8	15.6	16.1	11.5	14.2	14
Costs (y_i)	12.5	97.3	88	25	38.6	47.3

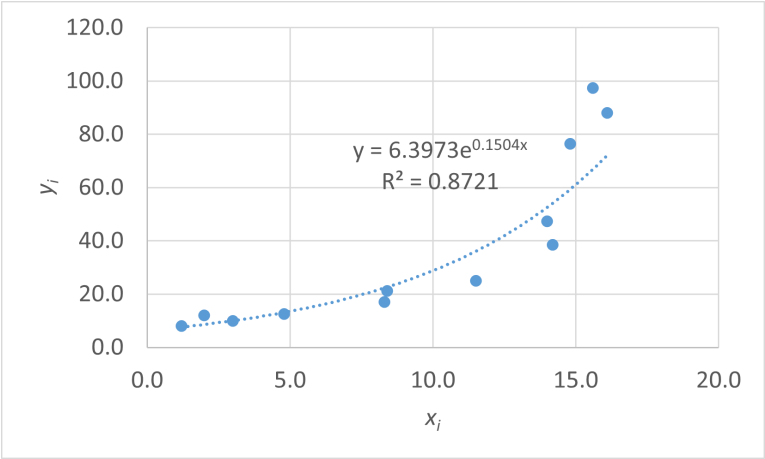


Figure 19.13: Scatter Plot of Maintenance Costs as a Function of Machine Age

LINEST		
b_2	ln(b_1)	b_1
0.15042814	1.855881	6.3973327

Figure 19.14: Excerpt from MS Excel Demonstrating the Use of the `= LINEST` Function

Example 19.4.  Table 19.3 contains data on maintenance costs y (in hundreds of CZK) and machine age (in years). Using an appropriate regression model, describe the dependence of maintenance costs on machine age.

Solution. In MS Excel, we create a scatter plot (see Figure 19.13). From the arrangement of the points in the plot, an exponential regression model can be suggested. We add the exponential regression curve to the plot, together with its equation and the coefficient of determination. The equation of the regression curve is $y = 6.4 \cdot e^{0.15 \cdot x}$ after rounding to two decimal places.

	Koeficienty
Hranice	1.855881132
Soubor X 1	0.150428145

Figure 19.15: Excerpt from MS Excel Showing the Regression Output Table Generated by the Data Analysis Tool

MS	F	Významnost F
7.859582041	113.0815083	9.02369E-07
0.06950369		

Figure 19.16: Excerpt from MS Excel Showing the ANOVA Section of the Regression Output Generated by the Data Analysis Tool

The coefficients b_1 and b_2 of the regression function $y = b_1 \cdot e^{b_2 x}$ can also be obtained in MS Excel using the `= LINEST` function. For this purpose, a column containing the values $\ln y$ is needed, and the equation $y = b_1 \cdot e^{b_2 x}$ is transformed into the form $\ln y = \ln b_1 + b_2 x$. The use of the array function `= LINEST` is shown in Figure 19.14, where the first two cells are the outputs of the array function, and the last cell b_1 is calculated as e raised to the generated value $\ln b_1$.

Both coefficients b_2 and $\ln(b_1)$ can be generated in the Regression table produced by the Data Analysis tool (see Figure 19.15). According to the same table, the coefficient of determination is $R^2 \approx 0.872$.

When testing the statistical significance of the selected exponential regression model, we again use the information contained in the ANOVA section of the Regression table produced by the Data Analysis tool (see Figure 19.16). We find that the p -value for the significance of the selected model is $p \approx 9.02 \cdot 10^{-7} < 0.05$; therefore, we reject the null hypothesis according to which the selected model is statistically insignificant. $\triangle$

Remark 19.1. *Since an additive model is assumed, i.e. the residuals are added to the regression function, the linearization procedure provides only an approximation to the exact solution. In many applications, however, a multiplicative model is assumed, i.e. the residuals are multiplied by the regression function. In this case, linearization transforms the multiplicative exponential model directly into a linear regression model, and the resulting calculations obtained through linearization are exact. Multiplicative models are frequently used in financial analysis, where the magnitude of fluctuations is proportional to the value of the underlying asset.*

19.5 Exercises

19.1 Table 19.4 presents paired observations of the heights and weights of 20 individuals. Determine the equation of the regression line.

Table 19.4: Data on the height v (in cm) and weight m (in kg) of 20 individuals

v	175	175	184	180	173	173	184	179	168	183	178
m	75	73	74	82	77	70	88	68	60	82	79
v	185	180	175	173	180	185	170	178	166	168	160
m	86	64	70	64	91	84	81	72	67	66	45

19.2 The concentration of lactic acid (in mg per 100 ml of blood) was measured in first-time mothers (values x_i) and in their newborns (values y_i) immediately after delivery. Table 19.5 presents the observed data.

Table 19.5: Data on the Lactic Acid Concentration in the Blood of Mothers

x_i	40	64	34	15	57	45	27	35	51
y_i	33	46	23	12	56	40	18	32	47

- Describe the dependence of y_i on x_i using an appropriate regression model. Assess the adequacy of the model using the coefficient of determination.
- Determine the lactic acid concentration in 100 ml of a newborn's blood if the mother's concentration was 60 mg.
- By how many mg will the lactic acid concentration in 100 ml of a newborn's blood increase if the mother's concentration increases by 5 mg?

19.3 Table 19.6 presents data on the value of output, measured in 100 000 CZK (variable y), and the amount of investment, measured in 10 000 CZK (variable x), for the year 1998 in a sample of 12 selected privately owned engineering firms employing more than 24 workers.⁶

⁶The exercise, including the data and solution, is adapted from [9], p. 188.

Table 19.6: Data on Investment and Output

Firm	1	2	3	4	5	6
y_i	52.8	48.4	54.2	50	54.9	53.9
x_i	16.3	16.8	18.5	16.3	17.9	17.4
y_i	53.1	52.4	53	52.9	53.1	60.1
x_i	16.1	16.2	17	16.7	17.5	19.1

- Determine the equation of the regression line describing the dependence of output value on the amount of investment.
- Determine by how much, on average, the value of output increases when the amount of investment increases by 10 000 CZK.

19.4 The effect of plant height on tomato yield was investigated. The observed results are presented in Table 19.7.

Table 19.7: Data on Plant Height (mm) and Tomato Yield (g)

Height	210	330	180	260	280	220	270	260	290	310
Yield	270	420	260	350	390	310	380	330	360	410

- Determine the regression function that models the dependence of yield on plant height. Assess the adequacy of the model using the coefficient of determination and the ANOVA test.
- What yield can be expected for a plant with a height of 320 mm?
- What average change in yield can be expected if plant height increases by 5 cm?

19.5 The relationship between production costs per hectare planted with wheat (c , measured in thousands of CZK) and wheat yield per hectare (v , measured in tonnes) was investigated. The study involved 11 private farmers. The results are presented in Table 19.8.

Table 19.8: Data on Production Costs and Yield

c	29.5	40.5	23.8	34	46.1	41.5	31.9	38.1	31	27	33
v	3.5	4	2.9	3.7	4.5	4.3	3.5	3.9	3.9	3.2	3.9

- Find a model describing the dependence of wheat yield per hectare on production costs per hectare. Assess the adequacy of the model using the coefficient of determination and the ANOVA test.
- Determine by how much the yield per hectare increases if production costs increase by 5 thousand CZK.

19.6 A new type of insulin has been developed, and the relationship between the reduction in a patient's blood glucose level and the amount of insulin administered some time before measurement is being investigated. Eight randomly selected patients received different doses of insulin (n in ml), and after a certain period the reduction in blood glucose level (s in %) was measured. The measurement results are presented in Table 19.9.

Table 19.9: Data on the Amount of Insulin Administered and the Reduction in Blood Glucose Level

n	150	200	250	300	350	400	450	500
s	8	12	30	20	55	58	44	65

- Construct a scatter plot and select an appropriate linear regression model to describe the dependence of the reduction in blood glucose level on the amount of insulin administered.
- Verify the adequacy of the selected model (using the coefficient of determination and ANOVA).
- Estimate by what percentage the blood glucose level will decrease in a patient who is administered 325 ml of insulin.

19.7 The Trabant, an iconic symbol of East German automotive history, has long been associated with the simplicity and practicality of socialist engineering. In the 1980s, the fuel consumption y_i of this vehicle (measured in litres per 100 km) was investigated as a function of its speed x_i (km/h). The tests were conducted on a level road under identical conditions. The vehicle was driven continuously in fourth gear. The observed results are presented in Table 19.10.

Table 19.10: Data on Vehicle Speed and Fuel Consumption

x_i	40	50	60	70	80	90	100
y_i	6.1	5.8	6	6.5	6.8	8.1	10

- Construct a regression model describing the dependence of fuel consumption on speed and assess its adequacy (goodness of fit).
- Estimate the fuel consumption when the vehicle is travelling at 85 km/h.
- At what speed will the vehicle achieve minimum fuel consumption?

19.8 A health-food supplement store conducted observations on 22 randomly selected days. On each day, the number of customers visiting the store during business hours (x) and the daily sales revenue (y , measured in thousands of CZK and rounded) were recorded. At the same time, the store management monitored whether the current number of sales assistants

was sufficient as the number of customers increased. The observed data are presented in Table 19.11.

Table 19.11: Data on Sales Revenue and Number of Customers

x_i	18	19	24	25	26	27	28	29	30	32	33
y_i	11	12	13	13	14	15	16	17	18	19	19
x_i	35	36	37	40	42	46	47	49	52	55	58
y_i	20	20	21	22	21	21	20	19	19	18	17

- Find a regression function describing the dependence of sales revenue on the number of customers. Assess the adequacy of the model using the coefficient of determination.
- Using the fitted regression function, estimate the daily sales revenue when the store is visited by 23 and, respectively, 38 customers.
- Determine the optimal number of customers that the staff is able to serve so that an increasing number of customers does not lead to a decline in sales revenue.

Solutions to the Exercises

19.1 $y = 0.4743x + 141.12$

19.2

- $y = 0.8974x - 2.5814$, $R^2 = 0.8811$,
- 51.26 mg,
- by 4.49 mg.

19.3

- $y = 16.1509 + 2.1622x$,
- by 216 220 CZK.

19.4

- $y = 1.1467x + 48.709$, $R^2 = 0.9204$, ANOVA test: $F(1, 8) = 92.415$, $p = 0.000011$,
- 415.65 g,
- 57.335 g.

19.5

- $y = 0.0646x + 1.544$, $R^2 = 0.8724$, ANOVA test: $F(1, 9) = 61.55$, $p = 0.000026$,

- b) by 323 kg.

19.6

- a) $y = -15.96 + 0.16x$,
b) $R^2 = 0.805$, ANOVA test: $F(1, 6) = 24.73$, $p = 0.0025$,
c) by 36.5%.

19.7

- a) $y = 0.0019x^2 - 0.2073x + 11.393$, $R^2 = 0.9849$,
b) 7.5 l,
c) 54.5 km/h.

19.8

- a) $y = -0.0173x^2 + 1.498x - 11.736$, $R^2 = 0.9122$,
b) approximately 13 566 CZK, approximately 20 207 CZK,
c) approximately 43 customers.

Chapter 20

Goodness-of-Fit Tests

In the preceding chapters, we introduced several statistical tests, and we observed that these tests are generally based on the assumption that the population from which the sample was drawn follows a particular probability distribution. In this chapter, we shall therefore study tests that allow us to determine whether a random sample follows a specified distribution or not.

20.1 Pearson's χ^2 Test

Pearson's χ^2 test is a universal tool for determining whether a random sample originates from a particular discrete or continuous probability distribution. This procedure is referred to as the χ^2 *goodness-of-fit test*.

Let $X_1, \dots, X_n$ be a random sample, where the variables X_j , $j = 1, \dots, n$, may take the values $1, \dots, k$. The quantities X_i representing the numbers of occurrences of outcome i are called empirical frequencies. The random vector $(X_1, \dots, X_k)^T$ has a multinomial distribution. We wish to test the hypothesis H_0 that the true probabilities of the multinomial distribution are equal to the values $p_1, \dots, p_k$. The quantities np_i will be referred to as theoretical frequencies. Under the null hypothesis H_0 , the statistic

$$\chi^2 = \sum_{i=1}^k \frac{(X_i - np_i)^2}{np_i} \sim \chi_{k-1}^2$$

has asymptotically a χ^2 distribution with $k - 1$ degrees of freedom. Whenever

$$\chi^2 \geq \chi_{k-1}^2(1 - \alpha),$$

Table 20.1: Results of Random Number Generator Testing [1]

Digit	0	1	2	3	4	
x_i	59 889	59 796	59 969	60 056	60 303	
p_i	0.1	0.1	0.1	0.1	0.1	
np_i	60 000	60 000	60 000	60 000	60 000	
$\frac{(x_i - np_i)^2}{np_i}$	0.20535	0.6936	0.0160	0.0523	1.53015	
Digit	5	6	7	8	9	Total
x_i	60 048	60 234	59 750	60 224	59 731	600 000
p_i	0.1	0.1	0.1	0.1	0.1	1
np_i	60 000	60 000	60 000	60 000	60 000	600 000
$\frac{(x_i - np_i)^2}{np_i}$	0.0384	0.9126	1.04167	0.8363	1.2061	6.53233

we reject the null hypothesis H_0 at significance level α .

It should be kept in mind that the χ^2 test is an asymptotic procedure and therefore can be recommended only for sufficiently large sample sizes n . The literature usually states that the following condition should hold:

$$np_i \geq 5q \quad \text{for all } i = 1, \dots, k, \quad k \geq 3,$$

where q denotes the proportion of classes for which $np_i < 5$.

This test can be used, for example, to assess the fairness of a die, to evaluate random number generators (each digit $0, 1, \dots, 9$ should occur with probability $\frac{1}{10}$ – see Example 20.1 below), and in many other applications.

Example 20.1.  When testing the fairness of a random number generator, a sequence of six-digit random numbers of length 100 000 was examined. Consequently, the total number of digits is $n = 600\,000$. The observed frequencies of the individual digits are given in Table 20.1. Using a significance level of 0.05, test whether the observed empirical probabilities of occurrence of the individual digits correspond to the theoretical probabilities of occurrence of the digits in a fair random-number generator.

Solution. We test the null hypothesis H_0 that the empirical probabilities are equal to the theoretical probabilities. From Table 20.1, it follows that the value of the test statistic is $\chi^2 \approx 6.53233$, which is compared with the critical value $\chi_9^2(0.05) \approx 16.92$.

= CHISQ.INV(0.95,9)

Based on the observed data, the null hypothesis that the generator is random cannot be rejected; therefore, with a confidence level of 95%, the random-number generator may be regarded as fair. $\triangle$

For solving Example 20.1 and similar examples, it is convenient to use the = CHISQ.TEST function in MS Excel.

= CHISQ.TEST(observed,expected) ≈ 0.68568 . The resulting p -value is greater than the significance level 0.05; therefore, we do not reject H_0 , namely that, with 95% confidence, the observed and theoretical probabilities of occurrence of the individual digits are equal.

20.2 Tests of Normality

In statistics, tests of normality are used to determine whether a data set is adequately modelled by a normal distribution and to assess how likely it is that the random variable underlying the data is normally distributed.

Let $Z_1, \dots, Z_n$ be a random sample. We wish to test the hypothesis H_0 that the sample originates from a distribution $N(\mu, \sigma^2)$, where the parameters μ and σ^2 are unknown. We first construct the classes

$$(-\infty, b_1), [b_1, b_2), [b_2, b_3), \dots, [b_{k-2}, b_{k-1}), [b_{k-1}, \infty),$$

where $k \geq 4$. For brevity, let the i th class be denoted by J_i . The empirical frequencies of the individual classes will again be denoted by X_i .

The probability p_i that the variable Z_j , $j = 1, \dots, n$, falls into the class J_i is equal to

$$p_i = p_i(\mu, \sigma) = \int_{J_i} f(x) dx = \int_{J_i} \frac{1}{\sqrt{2\pi}\sigma} e^{\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]} dx.$$

If the parameters μ and σ^2 were known, the problem would be reduced to the multinomial case with known probabilities $p_1, \dots, p_k$. This approach could be used, for example, when testing whether a random sample follows the distribution $N(0, 1)$.

In the general case, however, the parameters are unknown, and therefore this approach is not suitable. It is then necessary to obtain appropriate estimates of μ and σ^2 . It is not advisable to use the usual estimators, namely the sample mean and sample variance, because this would substantially alter the distribution of the χ^2 statistic. Consequently, the estimates of μ and

σ^2 must be obtained iteratively so that they satisfy the following system of equations

$$\mu = \frac{1}{n} \sum_{i=1}^k \frac{X_i}{p_i} \int_{J_i} x f(x) dx, \quad \sigma^2 = \frac{1}{n} \sum_{i=1}^k \frac{X_i}{p_i} \int_{J_i} (x - \mu)^2 f(x) dx,$$

where the sample mean and sample variance are used as the initial approximations for μ and σ^2 . Recall that both p_i and $f(x)$, appearing on the right-hand sides of these equations, depend on μ and σ . Let the solution of this system be denoted by $\hat{\mu}$ and $\hat{\sigma}$. These estimates are then used to compute the probabilities p_i . The statistic

$$\chi^2 = \sum_{i=1}^k \frac{[X_i - np_i(\hat{\mu}, \hat{\sigma})]^2}{np_i(\hat{\mu}, \hat{\sigma})} \sim \chi_{k-3}^2,$$

has a χ^2 distribution with $k - 3$ degrees of freedom. If $\chi^2 > \chi_{k-3}^2(1 - \alpha)$, we reject the null hypothesis H_0 at significance level α .

Example 20.2.  Fifty observations of a certain variable were obtained, and their frequencies are presented in Table 20.2. Using the χ^2 goodness-of-fit test at the significance level 0.05, determine whether this random sample can be regarded as originating from a population with a normal distribution.

Table 20.2: Values of the Variable and Their Frequencies X_i

Value of the Variable	1	2	3	4	5	6	7
Frequency x_i	2	6	8	15	7	7	5

Solution. The measured variable is a continuous random variable X . The estimate of μ is the sample mean $\bar{x} \approx 4.2$, and the estimate of σ^2 is the sample variance $s^2 \approx 2.61$, i.e. $s \approx 1.62$. We test

H_0 : The random variable X has a normal distribution $X \sim \mathbf{N}(4.2, 1.62)$

against the alternative

H_1 : The random variable X does not have a normal distribution $X \sim \mathbf{N}(4.2, 1.62)$.

The sample is divided into seven intervals. The calculation of the test statistic is shown in Table 20.3.

Table 20.3: Values of the Variable and their Frequencies

b_{j-1}	b_j	n_j	p_j	np_j	np_j	n_j	$\frac{(n_j - np_j)^2}{np_j}$
$-\infty$	1	2	0.0238	1.1929	-	-	-
1	1.857	0	0.0497	2.4860	-	-	-
1.857	2.714	6	0.1054	5.2681	8.9470	8	0.1002
2.714	3.571	8	0.1696	8.4817	8.4817	8	0.0274
3.571	4.428	15	0.2075	10.3759	10.37595	15	2.0608
4.428	5.285	7	0.1929	9.6448	9.6448	7	0.7253
5.285	6.142	7	0.1362	6.8121	12.5506	12	0.0242
6.142	6.999	0	0.0731	3.6557	-	-	-
6.999	∞	5	0.0417	2.0823	-	-	-
Total		50	1	50	50	50	2.9378

Because four expected frequencies were less than 5, the number of intervals was reduced to 5. The value of the test statistic after rounding is 2.9378. Since two parameters (the mean and the variance) were estimated, the resulting number of degrees of freedom for the χ^2 test is $\nu = 5 - 2 - 1 = 2$. For the chosen significance level 0.05, the critical value is $\chi_{0.05}^2(2) \approx 5.99$.

= CHISQ.INV(0.95,2)

Since $\chi^2 < \chi_{0.95}^2(5)$, we do not reject H_0 , and the normal distribution $N(4.2, 1.62)$ provides an appropriate model for describing the random variable X .

In MS Excel, the function = CHISQ.TEST(observed,expected) can be used to determine the p -value of the test. In this example, we obtain $p \approx 0.09 > 0.05$, and therefore we again do not reject H_0 . $\triangle$

20.3 Test of the Poisson Distribution

Let $X_1, \dots, X_n$ be a random sample from a distribution defined on the set of nonnegative integers. We wish to test the null hypothesis H_0 that the sample originates from a Poisson distribution $Po(\lambda)$, i.e. that the observations $x_1, x_2, \dots$ come from a population having a Poisson distribution, where the parameter λ is unknown. The alternative hypothesis H_1 states that the sample observations $x_1, x_2, \dots$ do not originate from a Poisson distribution. The test is performed in a manner analogous to the test of normality. We first construct classes. One possible choice is the following: the first class

Table 20.4: Call Frequency

k	0	1	2	3	4	5	6	7	8	9	10	>10
n_k	1	6	8	10	6	6	7	1	1	2	2	0

contains all observations less than or equal to some number r . The remaining classes are formed by the individual values $r+1, r+2, \dots, r+k-2$. The last class contains all values greater than or equal to $r+k-1$. In this way, k classes are created, where $k \geq 3$, and their frequencies are denoted by $Y_r, Y_{r+1}, \dots, Y_{r+k-1}$. Let

$$q_i = \mathbb{P}[X_j = i] = \frac{\lambda^i e^{-\lambda}}{i!}, \quad i = 0, 1, 2, \dots$$

The probabilities corresponding to the individual classes are then

$$\begin{aligned} p_r &= \sum_{i=0}^r q_i, \\ p_i &= q_i \quad \text{for } i = r+1, \dots, r+k-2, \\ p_{r+k-1} &= \sum_{i=r+k-1}^{\infty} q_i. \end{aligned}$$

The probabilities p_i again depend on the parameter λ , which is unknown and must therefore be estimated. As in the test of normality, we solve the equation iteratively

$$\lambda = \frac{1}{n} \left[Y_r \frac{\sum_{i=0}^r i q_i}{\sum_{i=0}^r q_i} + \sum_{i=r+1}^{r+k-2} i Y_i + Y_{r+k-1} \frac{\sum_{i=r+k-1}^{\infty} i q_i}{\sum_{i=r+k-1}^{\infty} q_i} \right],$$

where the sample mean of the observations x_j , $j = 1, \dots, n$, is used as the initial approximation for λ . Let the solution of this equation be denoted by $\hat{\lambda}$. The statistic

$$\chi^2 = \sum_{i=r}^{r+k-1} \frac{[Y_i - np_i(\hat{\lambda})]^2}{np_i(\hat{\lambda})} \sim \chi_{k-2}^2 \quad (20.1)$$

has a χ^2 distribution with $k-2$ degrees of freedom. If $\chi^2 > \chi_{k-2}^2(1-\alpha)$, we reject the null hypothesis H_0 at significance level α .

Example 20.3.  Using the χ^2 goodness-of-fit test at the significance level 0.05, test whether the number of telephone calls X received by a hotel switchboard during a 15-minute interval follows a Poisson distribution

Table 20.5: Observed and Expected Frequencies of Telephone Calls

i	1 $(-\infty, 1]$	2 $(1, 2]$	3 $(2, 3]$	4 $(3, 4]$	5 $(4, \infty)$
n_i	7	8	10	6	19
$\mathbb{E}_i$	9.957	11.202	11.202	8.4016	9.2222

Po(3). Table 20.4 presents the numbers of calls (k) and the numbers of 15-minute intervals n_k in which exactly k calls were received.

Solution. We first construct Table 20.5 containing the grouped frequency distribution (we choose 5 classes) together with the observed frequencies (n_i) and expected frequencies ($\mathbb{E}_i$) for $i = 1, \dots, 5$. The expected frequencies $\mathbb{E}_i$ are obtained from the Poisson distribution Po(3); for example,

$$\begin{aligned}\mathbb{E}_1 &= n \cdot p_1 = n \cdot \mathbb{P}(X \in (-\infty, 1]) = n \cdot (\mathbb{P}(X = 0) + \mathbb{P}(X = 1)) = \\ &= 50 \cdot 0.199 = 9.95,\end{aligned}$$

where $\mathbb{P}(X = 0) = \frac{e^{-\lambda}\lambda^0}{0!} = \frac{e^{-3}}{1} \approx 0.050$ and similarly for $i = 2, \dots, 5$.

The value of the test statistic is calculated using the formula (20.1):

$$\chi^2 = \sum_{i=1}^5 \frac{(n_i - \mathbb{E}_i)^2}{\mathbb{E}_i} = \frac{(7 - 9.957)^2}{9.957} + \dots + \frac{(19 - 9.222)^2}{9.222} \approx 12.976.$$

We compute the critical value of the test as the quantile of the χ^2 distribution with $n - 1$ degrees of freedom:

$$\chi_4^2(0.05) \approx 9.488.$$

$$= \text{CHISQ.INV}(0.95, 4)$$

We compare this critical value with the test statistic:

$$12.976 > \chi_4^2(0.05).$$

At the chosen significance level 0.05, we therefore reject the null hypothesis H_0 , i.e. the number of calls does not follow Po(3).

The function CHISQ.TEST can also be used to determine the p -value of the test: $\text{CHISQ.TEST}(7, \dots, 19, 9.957, \dots, 9.222)$. We obtain $p \approx 0.0114 < 0.05$, and therefore we again reject H_0 at the chosen significance level. $\triangle$

20.4 Kolmogorov–Smirnov¹ One-Sample Goodness-of-Fit Test

The Kolmogorov–Smirnov test is used to determine whether the “shape” of the empirical distribution agrees with that of a specified population distribution of a continuous random variable. The test is suitable even for small sample sizes. The null hypothesis H_0 states that the random sample originates from a distribution with a continuous cumulative distribution function $F(x)$, whereas the alternative hypothesis H_1 expresses that the sample does not originate from the specified distribution.

We first introduce the concept of the empirical distribution function.

Definition 20.1 (Empirical Distribution Function)

Let $X_1, \dots, X_m$ be a random sample from a distribution with cumulative distribution function F . For $i = 1, \dots, m$, define the random variables

$$\xi_i(x) = \begin{cases} 1, & \text{if } X_i < x, \\ 0, & \text{if } X_i \geq x. \end{cases}$$

Then, the empirical distribution function is defined as

$$F_m(x) = \frac{1}{m} \sum_{i=1}^m \xi_i(x).$$

The empirical distribution function (often also called the empirical cumulative distribution function) is a “step” function with jump height $\frac{1}{m}$ at each of the m observed values and serves as an estimator of the cumulative distribution function that generated the sample observations.

Example 20.4. A random number generator for the standard normal distribution $N(0, 1)$ produced the following twenty values arranged in ascending order: $-2.63, -1.28, -1.23, -0.92, -0.91, -0.78, -0.77, -0.50, -0.41, -0.35, -0.11, -0.01, 0.02, 0.23, 0.56, 0.75, 0.84, 0.87, 1.46, 1.62$. Construct the empirical distribution function and compare it with the graph of the cumulative distribution function of the standard normal distribution.

Solution. There are 20 observations in total; therefore, the height of each jump is $\frac{1}{20} = 0.05$. The widths of the steps are determined by the boundaries defined by the observed values. Figure 20.1 shows the cumulative distribution function of the standard normal distribution $N(0, 1)$ (blue curve) together with the empirical distribution function constructed from the above observations (red line segments). $\triangle$

¹Vladimir Ivanovich Smirnov (1887–1974), Russian mathematician.

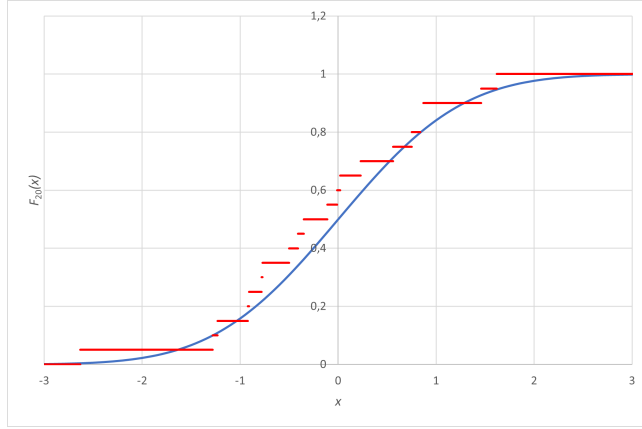


Figure 20.1: Distribution Function $\Phi(x)$ (Blue Curve) and the Empirical Distribution Function of the Standard Normal Distribution $F_{20}(x)$ (Red Step Segments)

The following theorem states that the empirical distribution function provides a good approximation of the cumulative distribution function.

Theorem 20.1. *For every x ,*

$$F_m(x) \rightarrow F(x)$$

almost surely as $m \rightarrow \infty$.

If we define

$$D_m = \sup_x |F_m(x) - F(x)|,$$

then

$$\mathbb{P} \left(\lim_{m \rightarrow \infty} D_m = 0 \right) = 1.$$

Let $X_1, \dots, X_n$ be a random sample from a distribution with a continuous cumulative distribution function. We wish to test the hypothesis H_0 that this cumulative distribution function is F . Let F_m denote the empirical distribution function corresponding to the sample $X_1, \dots, X_m$. Theorem 20.1 implies that large values of the quantity D_m provide evidence against the null hypothesis H_0 . For small values of m , the critical values $D_m(\alpha)$ are obtained from tables. For larger values of m , the critical values are approximated by

$$D_m(\alpha) \approx \sqrt{\frac{1}{2m} \ln \frac{2}{\alpha}}. \quad (20.2)$$

Thus, we reject H_0 whenever $D_m \geq D_m(\alpha)$.

Because of the monotonicity of $F(x)$, when computing the quantity D_m it is sufficient to consider only the endpoints of the intervals on which the empirical distribution function F_m is constant. Therefore, if x is a jump point of F_m , we evaluate the difference from the left, $F_m(x) - F(x)$, and from the right, $\lim_{y \rightarrow x+} F_m(y) - F(y)$.

Similarly to the χ^2 test, the cumulative distribution function $F(x)$ must be known exactly. Therefore, this test can be applied without modification when testing whether a random sample comes from distributions such as $N(0, 1)$, $R(0, 1)$, and so forth. However, when testing, for example, the normality of a random sample, we cannot simply estimate the parameters and substitute these estimates into the distribution function $F(x)$. Doing so would alter the distribution of the test statistic D_m , and consequently the critical values at which the hypothesis is rejected. These modified critical values have been determined through simulation studies and are available in specialized tables. Tests of normality based on the Kolmogorov–Smirnov test are also implemented in statistical software packages.

Example 20.5.  For the sample in Example 20.4, determine the value of the statistic D_m . Compute the empirical and theoretical values of the distribution function. At which jump does the maximum deviation occur? Use MS Excel.

Solution. As already noted, we examine all points at which the empirical distribution function has a jump, considering both the left-hand and right-hand limits. The maximum value, approximately 0.142, occurs at the thirteenth jump, corresponding to the value 0.02, when the right-hand limit is considered:

$$|F_{20}(0.02) - F(0.02)| = |0.65 - 0.50798| \approx 0.142.$$

The critical value is $D_{20}(0.05) = 0.294$; therefore, we do not reject the hypothesis that the sample originates from the standard normal distribution. The approximation of the critical value computed using formula (20.2) is $D_{20}(0.05) = 0.304$. $\triangle$

20.5 Exercises

20.1 Table 20.6 presents 150 observations of a certain variable. Using the χ^2 goodness-of-fit test at the significance level 0.05, determine whether this random sample can be regarded as originating from a population with a normal distribution.

Table 20.6: Values of the Variable and Their Frequencies

Value of the Variable	20	21	22	23	24	25	26	27
Frequency	5	15	20	35	29	32	10	4

20.2 Table 20.7 presents the results of 51 rolls of a die. Using the χ^2 goodness-of-fit test at the significance level 0.05, determine whether the die can be considered fair.

Table 20.7: Results of Die Rolls and Their Frequencies

Die Roll Outcome	1	2	3	4	5	6
Frequency	10	10	2	25	0	4

20.3 The data 8, 9, 10, 15, 17, and 22 are given. Using the one-sample Kolmogorov–Smirnov test at the significance level 0.05, determine whether these data originate from a population with a normal distribution.

20.4 A group of 90 randomly selected primary-school pupils answered the questionnaire item: “Which school subject do you like most?” The survey yielded the following preferences: mathematics – 35 pupils, physics – 28 pupils, and chemistry – 27 pupils. Determine whether there are statistically significant differences in the popularity of mathematics, physics, and chemistry.

20.5 In a given year, 114 children died in the Czech Republic as a result of road traffic accidents. These fatalities were distributed across the individual months of the year as shown in Table 20.8. Determine whether the number of fatalities is the same in all months of the year or whether the numbers differ in a statistically significant manner.

Table 20.8: Number of Child Fatalities in Road Traffic Accidents by Month in a Given Year

Month	1	2	3	4	5	6	7	8	9	10	11	12
Number of Fatalities	7	7	8	10	10	9	14	12	10	12	7	8

20.6 A chain of ice cream shops offering four flavours of ice cream opened a new outlet in a different location. In the chain's existing outlets, the sales distribution by flavour had been as follows: vanilla 62%, chocolate 18%, strawberry 12%, and pistachio 8%. After opening the new outlet, the following sales were recorded: vanilla 120, chocolate 40, strawberry 18, and pistachio 22. Using an appropriate statistical test, determine whether the sales distribution in the new location is consistent with, or differs from, the sales structure observed in the chain's existing outlets.

20.7 In order to assess staff workload, the hospital administration is interested in whether patients are discharged uniformly throughout the week. The numbers of patients discharged on each day of a particular week are shown in Table 20.9. Using an appropriate statistical test, determine whether patients are discharged uniformly throughout the week.

Table 20.9: Number of Patients Discharged During a Single Week

Day of the Week	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Disch. Patients	78	90	94	89	110	84	44

20.8 In order to assess staff workload, the hospital administration is interested in whether patients are discharged uniformly during the working days of the week. The numbers of patients discharged on the individual working days of a given week are shown in Table 20.10. Using an appropriate statistical test, determine whether patients are discharged uniformly across the working days.

Table 20.10: Number of Patients Discharged During the Working Days of a Single Week

Day of the Week	Mon	Tue	Wed	Thu	Fri
Number of Discharged Patients	78	90	94	89	110

20.9 In MS Excel, one hundred random numbers were generated using the random number generator. Test whether the digits are used with equal probability. Table 20.11 contains the frequency distribution of the individual digits in the generated numbers.

Table 20.11: Number of Digits Occurring in the Generated Numbers

Digit	0	1	2	3	4	5	6	7	8	9
Count	75	87	90	102	97	86	113	89	95	99

20.10 Table 20.12 presents the distribution of the numbers of point defects observed in a sample of twenty products. Determine whether the number of defects per product can be regarded as a random variable with a Poisson distribution.

Table 20.12: Number of Products with the Corresponding Number of Point Defects

Number of Defects	0	1	2	3	4	5	6	7
Number of Products	2	3	2	4	2	2	4	1

Solutions to the Exercises

20.1 The measured variable is a continuous random variable X . The estimate of μ is the sample mean $\bar{x} \approx 23.49$, and the estimate of σ^2 is the sample variance $\text{var } X \approx 2.68$, giving $s \approx 1.64$. H_0 : The random variable X follows a normal distribution $X \sim N(23.49, 2.68)$. H_1 : The random variable X does not follow a normal distribution $X \sim N(23.49, 2.68)$. The sample data are divided into eight intervals; the interval frequencies are equal to the observed absolute frequencies. The calculation of the test statistic is shown in Table 20.13.

Table 20.13: Values of the Variable and Their Frequencies

x_{j-1}	x_j	n_j	p_j	np_j	np_j	n_j	$\frac{(n_j - np_j)^2}{np_j}$
$-\infty$	20	0	0.0164	2.4667	-	-	-
20	20.875	5	0.0385	5.7689	8.2356	5	0.0718
20.875	21.75	15	0.0886	13.29095	13.29095	15	0.2198
21.75	22.625	20	0.1544	23.1660	23.1660	20	0.4327
22.625	23.5	35	0.2037	30.55105	30.55105	35	0.6479
23.5	24.75	29	0.2032	30.4862	30.4862	29	0.0724
24.375	25.25	32	0.1535	23.0188	23.0188	32	3.5042
25.25	26.125	10	0.0877	13.1505	13.1505	10	0.7548
26.125	27	4	0.0379	5.6838	8.1008	4	0.34996
27	∞	0	0.0161	2.4171	-	-	-
Total							6.0534

Because two expected frequencies were smaller than 5, the number of intervals was reduced to 8. The value of the test statistic is 6.0534. Since two parameters (the mean and the variance) were estimated, the resulting number of degrees of freedom for the χ^2 test is $\nu = 8 - 2 - 1 = 5$. For the chosen significance level 0.05, the critical value is $\chi_5^2(0.05) \approx 11.07$.

= CHISQ.INV(0.95,5)

Since $\chi^2 < \chi_5^2(0.05)$, we do not reject H_0 , and the normal distribution $N(23.49, 2.68)$ provides an appropriate model for describing the random variable X .

20.2 The value of the test statistic is $\chi^2 \approx 48.4$, which exceeds the critical value $\chi^2(5) \approx 11.07$. Therefore, the die cannot be considered fair. The corresponding p -value is approximately $3.45 \cdot 10^{-12} < 0.05$.

20.3 $n = 6$, test statistic $D_6 \approx 0.466$, $D_6(0.05) \approx 0.519$. We do not reject the null hypothesis that the population from which the sample was drawn has a normal distribution.

20.4 $\chi^2 \approx 1.266$, $\chi_2^2(0.05) \approx 5.991$.

20.5 $\chi^2 \approx 6$, $p \approx 0.873362$.

20.6 $\chi^2 \approx 4.32$, critical value 7.815, $p \approx 0.228589$.

20.7 $\chi^2 \approx 29.388$, $p \approx 0.000051$.

20.8 $\chi^2 \approx 5.822$, $p \approx 0.212834$

20.9 $\chi^2 \approx 10.3976$, $\chi_9^2(0.05) \approx 16.918$.

20.10 $\chi^2 \approx 1.239086$, $\chi_1^2(0.05) \approx 3.8415$, $p \approx 0.265496$.

Chapter 21

Contingency Tables

This chapter is devoted to basic tests for contingency tables, of which there are many. Before introducing these tests, we first define the concept of a contingency table.

Consider the random vector $Z = (X, Y)$ having a discrete distribution. The random variable X takes the values $1, \dots, r$, and the random variable Y takes the values $1, \dots, c$. The random variables X and Y represent characteristics of a statistical population (e.g. sex, educational attainment, etc.). Although the values of a characteristic are formally assumed to be positive integers, the characteristic itself need not be numerical, as is evident from the examples given in parentheses. In many cases, the numbers $1, 2, \dots$ are assigned merely as labels. For example, educational attainment may be coded as: 1 – primary education, 2 – secondary education, 3 – university education. Thus, the characteristics may be

- qualitative,
- discrete quantitative,
- continuous quantitative, with values grouped into classes.

For the random vector $Z = (X, Y)$, let

$$\begin{aligned}p_{ij} &= \mathbb{P}(X = i, Y = j), \\p_{i.} &= \mathbb{P}(Y = i) = \sum_{j=1}^c p_{ij}, \\p_{.j} &= \mathbb{P}(Z = j) = \sum_{i=1}^r p_{ij}.\end{aligned}$$

Assume that a sample of size n has been drawn from this distribution. Let n_{ij} denote the number of observations in the sample for which the pair

Table 21.1: Probability Matrix (Left) and Contingency Table (Right)

Y	$\mathbf{Z}$ $1 \dots c$	Σ
1	$p_{11} \dots p_{1c}$	$p_{1.}$
$\dots$	$\dots \dots \dots$	$\dots$
r	$p_{r1} \dots p_{rc}$	$p_{r.}$
Σ	$p_{.1} \dots p_{.c}$	1

Y	$\mathbf{Z}$ $1 \dots c$	Σ
1	$n_{11} \dots n_{1c}$	$n_{1.}$
$\dots$	$\dots \dots \dots$	$\dots$
r	$n_{r1} \dots n_{rc}$	$n_{r.}$
Σ	$n_{.1} \dots n_{.c}$	n

(i, j) occurred (i.e. the absolute frequency). The random variables n_{ij} have a joint multinomial distribution with parameter n and probabilities p_{ij} . A contingency table is then defined as the matrix (n_{ij}) . A contingency table is shown in Table 21.1, together with the matrix of probabilities (p_{ij}) , where

$$n_{i.} = \sum_{j=1}^c n_{ij}, \quad n_{.j} = \sum_{i=1}^r n_{ij}, \quad n = \sum_{i=1}^r \sum_{j=1}^c n_{ij}$$

and the following identities hold

$$n = \sum_{j=1}^c n_{.j} = \sum_{i=1}^r n_{i.} = \sum_{i=1}^r \sum_{j=1}^c n_{ij}.$$

If the data are arranged in a contingency table, where the categories of one variable define the rows and the categories of the other variable define the columns, as illustrated in Table 21.1, the following hypotheses may be tested:

- the hypothesis of independence of two random variables X and Y ,
- the hypothesis of homogeneity of multinomial distributions,
- the hypothesis of symmetry.

21.1 Test of Independence

For each element of a population, two characteristics are observed. Our goal is to test the null hypothesis

$$H_0: \text{The random variables } X \text{ and } Y \text{ are independent.}^1$$

against the alternative hypothesis

$$H_1: \text{The random variables } X \text{ and } Y \text{ are not independent.}$$

Theorem 21.1. *The random variables X and Y are independent if and only if*

$$p_{ij} = p_{i.}p_{.j}, \quad i = 1, \dots, r, \quad j = 1, \dots, c.$$

By virtue of Theorem (21.1), the hypothesis of independence can be rewritten in the form

$$H_0: p_{ij} = p_{i.}p_{.j} \text{ for } i = 1, \dots, r, \quad j = 1, \dots, c.$$

Under the null hypothesis H_0 , the test statistic

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} \quad (21.1)$$

has asymptotically a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom. Formula (21.1) can be rewritten in the form

$$\chi^2 = n \sum_{i=1}^r \sum_{j=1}^c \frac{n_{ij}^2}{n_{i.}n_{.j}} - n. \quad (21.2)$$

We reject the null hypothesis H_0 of independence of the variables X and Y whenever

$$\chi^2 \geq \chi_{(r-1)(c-1)}^2(1-\alpha).$$

Convergence to the limiting distribution requires that all expected frequencies $\frac{n_{i.}n_{.j}}{n}$ exceed 5. In practice, it is usually required that at least 80% of the expected frequencies be greater than 5 and that all expected frequencies be greater than 1. If this condition is not satisfied, some rows or columns are usually combined. However, this approach is not possible for so-called 2×2 contingency tables. In such cases, Fisher's exact test is used.

To solve the following example, a specially prepared MS Excel workbook is used, in which the computation of the χ^2 test is presented in detail.

Example 21.1.  We are testing for independence between mathematics test results and the study programme to which an applicant has applied. Applicants may apply for the Bachelor's programme in Financial Mathematics, the five-year Master's degree programme in Mathematics Teacher Education for Primary Schools, or the five-year Master's degree programme in Mathematics Teacher Education for Secondary Schools. The programmes are ordered according to their academic difficulty, from the least demanding

Table 21.2: Contingency Table of Observed and Expected Frequencies for Field of Study and Test Results (FM – Financial Mathematics, PST – Primary School Teacher, SST – Secondary School Teacher)

Assessment	Teaching Specialization			Total
	FM	PST	SST	
1 (60–80 points)	9 (9.7)	7 (17.9)	40 (28.4)	56
2 (40–59 points)	10 (17.2)	31 (31.6)	58 (50.2)	99
3 (20–39 points)	17 (13)	29 (24)	29 (38)	75
4 (0–19 points)	14 (10)	25 (18.5)	19 (29.4)	58
Total	50	92	146	288

(the Bachelor’s programme in Financial Mathematics) to the most demanding (the five-year Master’s programme for Secondary School Teachers). This raises the question of whether applicants take this fact into account when choosing a programme, given their previous academic performance. Put simply, do “better” students apply for the more demanding programmes, while “weaker” students apply for the less demanding ones? Let us test whether there is an association between the mathematics test results and the programme to which an applicant applies.

An applicant can obtain at most 80 points on the test. The variable X (test result) takes four values: 1 – score 60–80 points, 2 – score 40–59 points, 3 – score 20–39 points, 4 – score 0–19 points. The variable Y (study programme) takes three values: 1 – Financial Mathematics, 2 – Teacher Education for Primary Schools, 3 – Teacher Education for Secondary Schools. All data are given in Table 21.2, including the expected frequencies (shown in parentheses).

Solution. The expected frequencies are computed as the product of the corresponding row total and column total divided by the overall sample size.

The value of the test statistic, $\chi^2 \approx 27.56$, exceeds the critical value $\chi^2_{6}(0.05) \approx 12.59$; therefore, we reject the null hypothesis H_0 of independence.

`= CHISQ.INV(0.95,6)`

The p -value of the test, $p \approx 0.000114$, also indicates that H_0 should be rejected.

`= CHISQ.TEST(observed,expected)`

Thus, a statistically significant association (H_1) between the mathematics test result and the programme to which the applicant applies has been demonstrated. $\triangle$

21.2 Test of Homogeneity of Multinomial Distributions

This test is also referred to as the *test of equality of structure* or the *test of homogeneity*. It is used to test whether one of the observed characteristics has the same distribution under different conditions, which are represented by the categories of the second characteristic. For example, we may be interested in determining whether the age distribution of hospitalized patients is the same in two different hospitals. In general, the null hypothesis has the form:

H_0 : Probability $q_{i1}, \dots, q_{ic}$ does not depend on the row index i ,

or equivalently, all rows of the matrix (q_{ij}) are identical, where the theoretical probabilities $q_{i1}, \dots, q_{ic}$ correspond to the relative marginal frequencies in the i th row of the contingency table $\frac{n_{i1}}{n_{i.}}, \dots, \frac{n_{ic}}{n_{i.}}$, with $q_{i1} + \dots + q_{ic} = 1$, and where the marginal row frequencies $n_{i.}$ are assumed to be fixed in advance.

When testing homogeneity, we again use the χ^2 statistic computed according to formula 21.1 or 21.2. Under the null hypothesis H_0 , the statistic χ^2 has asymptotically a χ^2 distribution with $(r - 1)(c - 1)$ degrees of freedom. The null hypothesis H_0 of homogeneity of multinomial distributions is rejected whenever $\chi^2 \geq \chi^2_{(r-1)(c-1)}(1 - \alpha)$.

Example 21.2.  The contingency table 21.3 presents the results of a medical experiment conducted in the 1940s that investigated the effect of streptomycin in the treatment of pulmonary tuberculosis. The radiological assessment after six months was compared according to whether the patient belonged to the treatment group or the control group. Is there an association between the treatment and the outcome?



Table 21.3: Results of Radiological Examination (RA – Radiological Assessment)

RA / Treatment	Streptomycin	Control	Total
Significant Improvement	28 (16.45)	4 (15.55)	32
Moderate/Minor Improvement	10 (11.82)	13 (11.18)	23
No Change	2 (2.57)	3 (2.43)	5
Moderate/Minor Deterioration	5 (8.74)	12 (8.26)	17
Significant Deterioration	6 (6.17)	6 (5.83)	12
Death	4 (9.25)	14 (8.75)	18
Total	55	52	107

Solution. Since the test statistic is based on the expected frequencies $\frac{n_{i.}n_{.j}}{n}$, these frequencies must first be computed. They are shown in parentheses next to the observed frequencies in Table 21.3.

The value of the test statistic is $\chi^2 \approx 26.96$. Since $\chi_5^2(0.95) \approx 11.07$,

`= CHISQ.INV(0.95,5)`

we have $\chi^2 \geq \chi_5^2(1 - \alpha)$, and therefore reject the null hypothesis of homogeneity. Thus, at an asymptotic significance level of 0.05, we have demonstrated the existence of an association between treatment and outcome.

Using `= CHISQ.TEST(observed,expected)`, we obtain $p \approx 0.000058$, which is smaller than $\alpha = 0.05$. Therefore, we also reject H_0 of independence between the outcomes of the two groups under study. $\triangle$

21.3 χ^2 Test for 2×2 Contingency Tables

As noted above, when $r \times c = 2 \times 2$, we obtain a so-called 2×2 contingency table. Such a table has the form shown in Table 21.4.

For a 2×2 contingency table, we may again test either independence or homogeneity. The test statistic remains the same as for a general $r \times c$ contingency table. Since the summation indices take only two values, the χ^2 test statistic can be simplified to the following form:

$$\chi^2 = n \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}.$$

If $\chi^2 \geq \chi_1^2(1 - \alpha)$, we reject the hypothesis of independence. The same procedure is used to test the homogeneity of two binomial distributions (a

Table 21.4: 2×2 Contingency Table

n_{11}	n_{12}	$n_{1.}$
n_{21}	n_{22}	$n_{2.}$
$n_{.1}$	$n_{.2}$	n

Table 21.5: 2×2 Contingency Table of Parental Obesity among Obese Children

	Obese Father	Non-Obese Father	Total
Obese Mother	15	9	24
Non-Obese Mother	7	19	26
Total	22	28	50

binomial distribution being a special case of a multinomial distribution), provided that the row (or column) marginal frequencies are fixed.

Example 21.3.  In a random sample of fifty obese children aged 6–14 years, the obesity status of both the mother and the father was determined for each child. The data are recorded in Table 21.5. We are interested in whether parental obesity is associated.

Solution. Substituting into the test statistic, we obtain

$$\chi^2 = 50 \cdot \frac{(15 \cdot 19 - 9 \cdot 7)^2}{24 \cdot 22 \cdot 26 \cdot 28} \approx 6.41.$$

The critical value is $\chi_1^2(0.05) \approx 3.84$.

$$= \text{CHISQ.INV}(0.95, 1)$$

Since $\chi^2 \geq \chi_1^2(1 - \alpha)$, we reject the hypothesis of independence; that is, parental obesity is significantly associated.

Rejection of H_0 also follows from the $= \text{CHISQ.TEST}$ function, which yields $p \approx 0.0113 < 0.05$. $\triangle$

21.4 Fisher's Exact Test

As noted above, convergence to the limiting χ_1^2 distribution requires that all expected frequencies $\frac{n_{i.}n_{.j}}{n}$ be greater than 5. If this condition is not satisfied, rows or columns are usually combined. However, this is not possible in a 2×2 contingency table, and therefore Fisher's exact test is employed.

Table 21.6: Mathematics Achievement (MA) and Whether Students Learn to Play a Musical Instrument

MA / Music Instruction	Learns to Play	Does Not Learn	Total
Good	6	4	10
Poor	1	13	14
Total	7	17	24

This test makes it possible to test the hypothesis of independence even when the frequencies are small. The test is carried out in the following steps.

- 1) Construct the set of all contingency tables having the same marginal frequencies as the original contingency table.
- 2) For each table in this set, compute the probabilities

$$P = \frac{n_{1.}! \cdot n_{2.}! \cdot n_{.1}! \cdot n_{.2}!}{n! \cdot n_{11}! \cdot n_{12}! \cdot n_{21}! \cdot n_{22}!}$$

and the quantity $d = \ln b$, where $b = \frac{n_{11}n_{22}}{n_{12}n_{21}}$. The quantity d is called the log-odds ratio of the given table.

- 3) Sum the probabilities $\mathbb{P}$ of all tables having the same marginal frequencies as the original table and whose log-odds ratios are greater than or equal to $|d|$ in absolute value (where d is the log-odds ratio of the observed table).
- 4) If the sum of these probabilities is less than or equal to α (the significance level of the test), reject the hypothesis of independence.

The entire procedure is illustrated in the following example.

Example 21.4.  For a random sample of 24 students, information was collected on whether they had good or poor achievement in mathematics and whether or not they were learning to play a musical instrument. The observed results are presented in Table 21.6. The objective is to test the hypothesis that achievement in mathematics and the fact that a child is learning to play a musical instrument are independent.

Solution. We construct all tables having the same marginal frequencies as the observed table. For each table, we compute the log-odds ratio d and the probability $\mathbb{P}$ (see Table 21.7).

The observed contingency table has a log-odds ratio whose absolute value is equal to 2.97. Therefore, we sum the probabilities of all tables whose log-odds ratios have absolute values greater than or equal to 2.97. The sum of

Table 21.7: Individual Probabilities

$\begin{array}{c} 0 \quad 10 \\ 7 \quad 7 \\ \hline d = -\infty \\ \mathbb{P} \approx 0.009916 \end{array}$	$\begin{array}{c} 1 \quad 9 \\ 6 \quad 8 \\ \hline d = -1.91 \\ \mathbb{P} \approx 0.086766 \end{array}$	$\begin{array}{c} 2 \quad 8 \\ 5 \quad 9 \\ \hline d = -0.80 \\ \mathbb{P} \approx 0.260297 \end{array}$	$\begin{array}{c} 3 \quad 7 \\ 4 \quad 10 \\ \hline d = 0.07 \\ \mathbb{P} \approx 0.347063 \end{array}$
$\begin{array}{c} 4 \quad 6 \\ 3 \quad 11 \\ \hline d = 0.89 \\ \mathbb{P} \approx 0.220858 \end{array}$	$\begin{array}{c} 5 \quad 5 \\ 2 \quad 12 \\ \hline d = 1.79 \\ \mathbb{P} \approx 0.066258 \end{array}$	$\begin{array}{c} 6 \quad 4 \\ 1 \quad 13 \\ \hline d = 2.97 \\ \mathbb{P} \approx 0.008495 \end{array}$	$\begin{array}{c} 7 \quad 3 \\ 0 \quad 14 \\ \hline d = \infty \\ \mathbb{P} \approx 0.000347 \end{array}$

these probabilities is 0.018758. Since this sum is not greater than $\alpha = 0.05$, we reject the hypothesis of independence. $\triangle$

21.5 McNemar's Test

In the statistical analysis of contingency tables, the objective is not always to perform a classical test of independence or homogeneity. Another test that can be applied to a 2×2 contingency table is McNemar's² test. This test is used when the presence or absence of a particular characteristic is observed in a sample of n randomly selected objects. Subsequently, some intervention is applied to the same sample, and the presence or absence of the characteristic is again determined for each object. The goal is to determine whether the intervention has altered the probability of occurrence of the characteristic.

After statistical therapy,
a student is usually
unrecognizable



Let the symbol $+$ denote the presence of the characteristic under study and the symbol $-$ denote its absence. We obtain tables 21.8 and 21.9 of the following forms, where X = before intervention, Y = after intervention, and $X = +, -, Y = +, -$. Furthermore, for example, $p_{11} = \mathbb{P}(X = +, Y = +)$.

We test the hypothesis

$$H_0: p_{1\cdot} = p_{\cdot 1}.$$

This hypothesis is equivalent to the hypothesis

$$H_0: p_{12} = p_{21}.$$

²Quinn Michael McNemar (1900–1986), American psychologist and statistician.

Table 21.8: Table of Absolute Frequencies

Before Intervention	After Intervention		Total
	+	−	
+	n_{11}	n_{12}	$n_{1.}$
−	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

Table 21.9: Table of Probabilities

Before Intervention	After Intervention		Total
	+	−	
+	p_{11}	p_{12}	$p_{1.}$
−	p_{21}	p_{22}	$p_{2.}$
Total	$p_{.1}$	$p_{.2}$	1

Table 21.10: Comparison of the Effectiveness of Drugs A and B

Drug B	Drug A		Total
	Success	Failure	
Success	1	3	4
Failure	9	5	14
Total	10	8	18

That is, the proportion of positive outcomes before the intervention is the same as the proportion of positive outcomes after the intervention.

The test statistic is

$$\chi^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

and has asymptotically a χ_1^2 distribution.

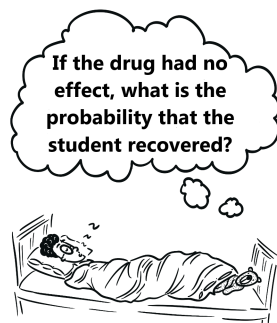
The null hypothesis H_0 is rejected whenever $\chi^2 \geq \chi_1^2(1 - \alpha)$. The asymptotic χ^2 approximation with 1 degree of freedom can be used provided that $(n_{12} + n_{21}) \geq 8$. If this condition is not satisfied, the above test statistic should not be used. Tests applicable for small values of $(n_{12} + n_{21})$ can be found, for example, in [1].

Example 21.5.  We observe a random sample of 18 patients who were treated with two different antihypertensive drugs, A and B. Each patient received drug A for one month and, after any residual effects had disappeared,

Table 21.11: Table of Type $c \times c$

$n_{11} \dots n_{1c}$	$n_{1.}$
$n_{21} \dots n_{2c}$	$n_{2.}$
$\dots$	$\dots$
$n_{c1} \dots n_{cc}$	$n_{c.}$
$n_{.1} \dots n_{.c}$	n

received drug B for one month. The outcome was classified as either a success or a failure. Our goal is to test whether the success rates of the two drugs are the same. The results of the study are presented in Table 21.10.



Solution. Substituting into the corresponding test statistic, we obtain

$$\chi^2 = \frac{(3 - 9)^2}{3 + 9} = 3.$$

Since the corresponding critical value is $\chi_1^2(0.05) \approx 3.84$,

$$= \text{CHISQ.INV}(0.95, 1)$$

and $\chi^2 < \chi_1^2(0.05)$, we do not reject the null hypothesis H_0 . Thus, based on the data under study, no difference in the effectiveness of the two drugs can be demonstrated. $\triangle$

21.6 Test of Symmetry

Consider a square contingency table 21.11 of size $c \times c$. The corresponding table of probabilities has the form shown in Table 21.12. We wish to test the hypothesis

$$H_0: p_{ij} = p_{ji} \quad \text{for all pairs } (i, j), \quad i, j = 1, 2, \dots, c.$$

This test is a generalization of the 2×2 case. McNemar's test is therefore a special case of the test of symmetry. In this setting, the test statistic has the form

$$\chi^2 = \sum_{i < j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}.$$

Table 21.12: Table of Probabilities

$p_{11} \cdots p_{1c}$	$p_{1.}$
$p_{21} \cdots p_{2c}$	$p_{2.}$
$\cdots$	$\cdots$
$p_{c1} \cdots p_{cc}$	$p_{c.}$
$p_{.1} \cdots p_{.c}$	p

Table 21.13: Frequencies of Marriages by the Previous Marital Statuses of the Partners [1]

Groom/Bride	Single	Widowed	Divorced	Total
Single	75 564	824	3 463	79 851
Widowed	1 370	904	798	3 072
Divorced	4 603	590	2 943	8 136
Total	81 537	2 318	7 204	91 059

Under the null hypothesis H_0 (the hypothesis of symmetry), the statistic χ^2 has asymptotically a χ^2 distribution with $\frac{c(c-1)}{2}$ degrees of freedom. We reject the hypothesis of symmetry whenever $\chi^2 \geq \chi^2_{\frac{c(c-1)}{2}}(1 - \alpha)$.



Example 21.6.  Table 21.13 contains data on the marital status of marrying couples. The objective is to determine whether the probability of a marriage between a single groom and a widowed bride is the same as the probability of a marriage between a single bride and a widowed groom, and whether an analogous equality holds for the probabilities of all other combinations of the partners' previous marital statuses.

Solution. Substituting into the test statistic for the test of symmetry, we obtain

$$\chi^2 = \frac{(824 - 1\,370)^2}{824 + 1\,370} + \frac{(3\,463 - 4\,603)^2}{3\,463 + 4\,603} + \frac{(798 - 590)^2}{798 + 590} \approx 328.17.$$

Since $\chi^2 \geq \chi^2_3(0.95) \approx 7.81$,

$$= \text{CHISQ.INV}(0.95, 3)$$

we reject the hypothesis of symmetry. △

21.7 Exercises

21.1 Using a test of independence at the 5% significance level, determine whether vaccination against influenza is effective. A total of 500 persons were examined. Of these, 120 had been vaccinated against influenza, and among them 24 became ill. Among the unvaccinated individuals, 97 became ill. Construct a 2×2 contingency table of absolute frequencies. Determine whether the difference between the proportion of individuals who became ill despite being vaccinated and the proportion of individuals who became ill without being vaccinated is statistically significant.



21.2 For a sample of four hundred university students, the year of study attended and the level of alcohol consumption were recorded (see Table 21.14). Test the null hypothesis that there is no association between the year of study and the level of alcohol consumption. Use a significance level of 5%.

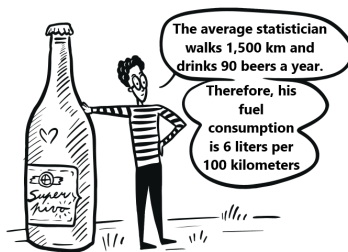


Table 21.14: Empirical and Expected Frequencies of Alcohol Consumption by Year of Study (values in parentheses are the expected frequencies rounded to one decimal place)

Group	Year of Study				Total
	1	2	3	4	
Drinker	29 (38)	41 (34.1)	33 (31.4)	28 (27.5)	131
Consumer	32 (39.4)	29 (35.4)	36 (32.6)	39 (28.6)	136
Abstainer	55 (38.6)	34 (34.6)	27 (31.9)	17 (27.9)	133
h_j	116	104	96	84	400

21.3 For students who discontinued their studies, the reason for withdrawal and the type of secondary school previously attended were recorded. The results are presented in the contingency table 21.15. Using a significance level of 5%, test the hypothesis of independence for this contingency table.

Table 21.15: Empirical Frequencies of Reasons for Discontinuing Studies by Type of Secondary School (G – Grammar School, SEBS – Secondary School of Economics and Business, O – Other, 1 – Academic Difficulty, 2 – Health Reasons, 3 – Social Reasons, 4 – Other)

Type of Secondary School	Reason for Discontinuing Study				h_i
	1	2	3	4	
G	5	35	10	0	50
SEBS	30	5	0	5	40
O	10	0	0	0	10
h_j	45	40	10	5	100



21.4 Using a test of homogeneity of multinomial distributions at the 5% significance level, determine whether the distribution of the population across the individual categories was the same in 2011 and 2021. The required data are given in Table 21.16.

Table 21.16: Empirical and Expected Frequencies of Population Age Groups in 2011 and 2021

Age Group / Year	2011	2021	Total
< 14	1.49 (1.58)	1.69 (1.60)	3.18
16–64	7.27 (6.93)	6.68 (7.02)	13.95
> 65	1.64 (1.88)	2.15 (1.91)	3.79
Total	10.4	10.52	20.92

21.5 Table 21.17 presents the numbers of motorcycle accidents resulting in different levels of injury severity according to the response distance of the emergency medical service. At the 95% confidence level, test whether the severity of injuries is independent of the response distance of the emergency medical service.

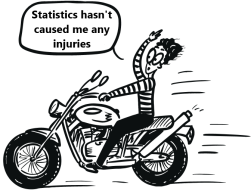


Table 21.17: Contingency Table of Numbers of Individual Types of Injuries and Emergency Medical Service Response Times (in Minutes) (RT – Response Time,

RT	Type of Injury				
	No Injury	Minor In.	Severe In.	Fatal In.	Total
≤ 10	1 768	807	189	47	2 811
10–20	946	1 387	746	53	3 132
> 20	115	438	288	16	857
Total	2 829	2 632	1 223	116	6 800

21.6 A total of 460 adults participated in an influenza vaccination study, of whom 240 received the influenza vaccine (the vaccinated group), while 220 received a placebo (the control group). At the end of the experiment, 100 individuals contracted influenza. Twenty of them belonged to the vaccinated group and eighty to the control group. The observed frequencies are given in Table 21.18.

Is there sufficient evidence that the vaccine was effective, or could the differences between the vaccinated and control groups have arisen by chance? Use the significance level $\alpha = 0.001$.

Table 21.18: Table of Absolute and Relative Frequencies (in %) of Individuals Who Contracted Influenza

Influenza	Vaccinated	Placebo	Total
Yes	20 (8.3)	80 (36.4)	100 (21.7)
No	220 (91.7)	140 (63.6)	360 (78.3)
Total	240 (100.0)	220 (100.0)	460 (100.0)

21.7 The contingency table 21.19 presents the results of a famous medical experiment conducted in the 1940s that investigated the effectiveness of streptomycin in the treatment of pulmonary tuberculosis. The radiological assessment after six months was compared according to whether the patient belonged to the treatment group or the control group. Is there a statistically significant association between treatment and outcome ($\alpha = 0.01$)?

Table 21.19: Results of Tuberculosis Treatment

Radiological Assessment	Streptomycin	Control	Total
Significant Improvement	28	4	32
Minor Improvement	10	13	23
No Change	2	3	5
Minor Deterioration	5	12	17
Marked Deterioration	6	6	12
Death	4	14	18
Total	55	52	107

21.8 For a random sample of 20 students, information was collected on whether they considered the school climate to be good or poor and whether they wished or did not wish to participate in a team-building event. The observed results are presented in Table 21.20. The objective is to test the hypothesis that students' evaluation of the school climate and participation in the team-building event are independent.

Table 21.20: Participation in the Team-Building Event According to Students' Evaluation of the School Climate (SC)

SC / Event	Will Participate	Will Not Participate	Total
Good	5	3	8
Poor	6	6	12
Total	11	9	20

21.9 A total of 20 applicants are competing for a job position. They undergo a psychological test and are also evaluated on the basis of an interview with the manager. In both assessments, applicants are classified either as "passed" (score 1) or "failed" (score 0). The results are shown in Table 21.21. Using McNemar's test, test at the 5% significance level the symmetry (difficulty) of the two parts of the selection procedure. (Dependence of the interview outcome on success in the psychological test is to be expected.)

Table 21.21: Assessment of Applicants in the Selection Process

Interview / Psychological Test	0	1	Total
0	3	6	9
1	3	8	11
Total	6	14	20

21.10 Table 21.22 presents the results of a student evaluation of the quality of dining services in two cafeterias, A and B, using the categories “poor”, “average”, and “good”. At the significance level 0.05, test the hypothesis that the ratings of the two dining facilities are symmetric and that the two evaluations may therefore be considered comparable.



Table 21.22: Evaluation of the Quality of Dining Services in Two Cafeterias, A and B

Mensa A/B	Poor	Average	Good	Total
Poor	23	27	15	65
Average	45	86	92	223
Good	9	110	115	234
Total	77	223	222	522

21.11 Francis Galton, a renowned English natural scientist, recorded the eye colours of fathers and their sons in 1,000 cases. Eye colour was coded as follows: 1 – light blue, 2 – blue-green or grey, 3 – dark grey or light brown, 4 – dark brown. The results of the study are presented in Table 21.23. Test the hypothesis of symmetry of eye colour between fathers and their sons.³

³The problem, including the data, is taken from [3], p. 298.

Table 21.23: Table of Observed Eye-Colour Frequencies

Son's Eye Colour	Father's Eye Colour				
	1	2	3	4	Total
1	194	70	41	30	335
2	83	124	41	36	284
3	25	34	55	23	137
4	56	36	43	109	224
Total	358	264	180	198	1000

21.12 In a small survey, the sexual activity of married couples was observed before and after the birth of a child. The results are presented in Table 21.24. At the 5% significance level, test whether the birth of a child has an effect on the sexual life of married couples.

Table 21.24: Table of Observed Results on the Sexual Activity of Married Couples Before and After the Birth of a Child

Before / After	Unsatisfactory	Satisfactory	Very Good	Total
Unsatisfactory	24	9	6	39
Satisfactory	6	16	11	33
Very Good	1	8	20	29
Total	31	33	37	101

Solutions to the Exercises

21.1 Table 21.25 is a 2×2 contingency table. The difference between the observed proportions in the two groups is 4.2%. H_0 : The two variables are independent. The test statistic $\chi^2 = 500 \cdot \frac{(24 \cdot 283 - 97 \cdot 96)^2}{120 \cdot 380 \cdot 121 \cdot 379} \approx 1.52$, while $\chi^2(0.95, 1) \approx 3.84$. Since $1.52 < 3.84$, we do not reject H_0 , i.e. the percentage difference is not statistically significant at the significance level 0.05. Equivalently, the effectiveness of the influenza vaccine has not been demonstrated statistically.

Table 21.25: Frequencies of Healthy/Ill and Vaccinated/Unvaccinated Individuals During an Influenza Outbreak

Vaccination Status / Health Status	Healthy	Ill	Total
Vaccinated	96	24	120
Unvaccinated	283	97	380
Total	379	121	500

21.2 Test of independence in a contingency table: H_0 : There is no association between the year of study attended and the level of alcohol consumption. H_1 : There is an association between the year of study attended and the level of alcohol consumption. $\alpha = 0.05$, critical value $\chi^2_{\alpha}((m-1)(n-1)) = \chi^2_{0.05}(2 \cdot 3) \approx 12.59$.

Since $\chi^2 \approx 22.291 > \chi^2_{0.05}(6)$, we reject H_0 . Therefore, at the 5% significance level, there is a statistically significant association between students' year of study and their level of alcohol consumption.

21.3 $\chi^2 \approx 68.65$, critical value $\chi^2_{0.05}(6) \approx 12.5916 < \chi^2$, we reject H_0 of independence in the contingency table; that is, there is a statistically significant association between the type of secondary school and the reason for discontinuing studies.

21.4 H_0 : The frequency distributions are identical. The test statistic is $\chi^2 \approx 1.69$, $\chi^2(0.95, 1) \approx 3.84$, and since $1.69 < 3.84$, we do not reject H_0 . Therefore, the differences between the years 2011 and 2021 in the population age groups are not statistically significant at the 0.05 significance level.

21.5 H_0 : The severity of injury and the emergency medical service response time are independent. H_1 : The severity of injury and the emergency medical service response time are not independent.

The value of the test statistic is $\chi^2 \approx 1073.5$, which is far greater than the critical value of the distribution $\chi^2_{(3-1) \cdot (4-1)} = \chi^2_6 \approx 12.59$ at the chosen significance level 0.05. Therefore, we reject H_0 of independence of the two variables, which is also confirmed by the p -value $p \approx 1.124 \cdot 10^{-228} < 0.05$.

21.6 H_0 : The percentage of influenza cases is the same in the vaccinated group and in the control group. The expected frequencies are given in parentheses in Table 21.26.

Table 21.26: Table of Absolute and Relative Frequencies (in %) of Individuals Who Contracted Influenza

Influenza	Vaccinated	Placebo	Total
Yes	52.17 (21.7)	47.83 (21.7)	100 (21.7)
No	187.83 (78.3)	172.17 (78.3)	360 (78.3)
Total	240 (100.0)	220 (100.0)	460 (100.0)

If there is no relationship between vaccination and influenza, then the observed and expected frequencies should be close to one another, and any deviation would be attributable solely to chance. The best way to assess the differences between the observed and expected frequencies is to compute the χ^2 statistic, where in this case we sum over all cells of the table. $\chi^2 = 53$, the number of degrees of freedom is $df = (m - 1) \cdot (n - 1) = 1$, and $\chi^2_{0.001}(1) \approx 10.83 < \chi^2$. This means that the probability that such large observed percentage differences in influenza incidence could have arisen purely by chance, if there were in fact no real differences between the vaccine and the placebo, is less than 0.1%. Therefore, we may conclude that the vaccine is more effective than the placebo.

21.7 If there is no difference between the groups, we obtain the expected frequencies shown in Table 21.27.

Table 21.27: Expected Frequencies for the Tuberculosis Treatment Outcomes (%)

Radiological Assessment	Streptomycin	Control	Total
Significant Improvement	28 (16.45)	4 (15.55)	32
Minor Improvement	10 (11.82)	13 (11.18)	23
No Change	2 (2.57)	3 (2.43)	5
Minor Deterioration	5 (8.74)	12 (8.26)	17
Significant Deterioration	6 (6.17)	6 (5.83)	12
Death	4 (9.25)	14 (8.75)	18
Total	55	52	107

$\chi^2 = 26.96$, $df = (6 - 1)(2 - 1) = 5$. For 5 degrees of freedom, the 1% critical value is given by the quantile $\chi^2_{1-0.01}(5) \approx 15.09$. The computed value of the test statistic $\chi^2 \approx 26.96$ substantially exceeds the critical value 15.09. Therefore, we conclude that at the significance level $\alpha = 0.01$ there is sufficient evidence of an association between the treatment and the radiological outcome.

21.8 We construct all tables having the same marginal frequencies as the observed table. For each table, we compute the log-odds ratio d and the probability $\mathbb{P}$ (see Table 21.28).

Table 21.28: Individual Probabilities

0 8 11 1	1 7 10 2	2 6 9 3	3 5 8 4
$d = -\infty$ $\mathbb{P} \approx 0.000071$	$d = -3.55$ $\mathbb{P} \approx 0.003144$	$d = -2.19$ $\mathbb{P} \approx 0.036675$	$d = -1.20$ $\mathbb{P} \approx 0.165039$
4 4 7 5	5 3 6 6	6 2 5 7	7 1 4 8
$d = -0.33$ $\mathbb{P} \approx 0.330078$	$d = 0.51$ $\mathbb{P} \approx 0.308073$	$d = 1.43$ $\mathbb{P} \approx 0.132031$	$d = 2.64$ $\mathbb{P} \approx 0.023577$

The observed contingency table has a log-odds ratio whose absolute value is equal to 0.51. Therefore, we sum the probabilities of all tables whose log-odds ratios have absolute values greater than or equal to 0.51. The sum of these probabilities is 0.66861. Since this sum is greater than $\alpha = 0.05$, we do not reject the hypothesis of independence. Thus, students' evaluation of the school climate is not associated with participation in the team-building event.

21.9 Both parts of the selection procedure will be equally difficult if $p_{11} + p_{12} = p_{11} + p_{21}$, that is, $p_{12} = p_{21}$. Since $n_{12} + n_{21} = 9 > 8$, McNemar's test can be applied. $H_0 : p_{12} = p_{21}$, $H_1 : p_{12} \neq p_{21}$. The test statistic is $\chi^2 = \frac{(6-3)^2}{9} = 1$, while $\chi^2(0.95, 1) \approx 3.84$. Since $1 < 3.84$, we do not reject H_0 . Therefore, at the significance level 0.05, there is no evidence that the two parts of the selection procedure differ in difficulty.

21.10 $H_0 : p_{ij} = p_{ji}$ for all pairs $(1, j), i, j = 1, 2, 3$,

$$\chi^2 = \sum_{i>j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}} = \frac{(45 - 27)^2}{45 + 27} + \frac{(9 - 15)^2}{9 + 15} + \frac{(110 - 92)^2}{110 + 92} = 7.6,$$

The number of degrees of freedom for this test is $c = \frac{3(3-1)}{2} = 3$, and $\chi^2(0.95, 3) \approx 7.81$. Since $7.6 < 7.81$, the null hypothesis H_0 cannot be rejected at the significance level 0.05. Therefore, the evaluations of the two dining facilities may be regarded as comparable.

21.11 $\chi^2 = 19.56$, $\chi_6^2(0.05) = 12.59$. Since $19.56 > 12.59$, we reject the hypothesis of symmetry; that is, we reject the claim that the probability that “the father has eye colour i and the son has eye colour j ” is equal to the probability that “the father has eye colour j and the son has eye colour i ”.

21.12 $\chi^2 = 4.64$, $df = \frac{k(k-1)}{2} = 3$, $\chi_3^2(0.05) = 7.81$. Since $4.64 < 7.81$, we fail to reject H_0 . Therefore, at the chosen significance level, there is no evidence that the birth of a child affects the sexual life of married couples.

Conclusion

The publication you have just finished reading, intended for readers and students interested in probability and statistics, is the collective work of members of the Department of Data Science and Information Systems of the Faculty of Agriculture and Technology and the Department of Mathematics of the Faculty of Education at the University of South Bohemia in České Budějovice. It is hoped that their vision of a textbook serving as a valuable aid to students in their studies has been fulfilled.



The value of the publication lies in the inclusion of a larger number of worked examples with detailed solutions, as well as numerous exercises for practice accompanied by answers or guidance on the solution process. A further valuable addition is the inclusion of references to MS Excel functions and tools that can be used when solving problems in probability and statistics, since MS Excel is among the most widely available software packages used by both students and the general public.

The technical exposition is complemented by numerous explanatory remarks in everyday language, intended to make the meaning of the material accessible to all readers and to illustrate practical applications of the concepts presented. The textbook is by no means exhaustive in its coverage of probability and statistics; therefore, readers are encouraged to continue their studies using the publications listed in the bibliography.

The authors would like to thank the reviewers for their valuable comments and suggestions.

Index

- σ -algebra
 - Borel, 32
- σ -algebra, 21
 - Borel, 22, 30
- analysis
 - correlation, 371
 - regression, 381
- analysis of variance
 - one-way, 341
 - two-way, 350
- argument of the population, 243
- bar chart, 247
- Bernoulli process, 50
- Bernoulli scheme, 50
- Borel sets, 90
- causation, 375
- class, 250
- class midpoint, 250
- class boundary
 - lower, 250
 - upper, 250
- class length, 250
- class mark, 250
- class width, 250
- classification
 - simple, 243
 - three-way, 354
 - two-way, 350
- coefficient
 - kurtosis, 266
 - of determination, 383, 391
 - of variation, 261
 - rank correlation, 374
 - sample correlation, 371
 - skewness (asymmetry), 265
 - Spearman rank correlation, 372, 374
- comparison of multiple samples, 341
- confidence band of the regression line, 385
- confidence level, 285
- contingency table, 423
- convergence
 - almost surely, 230
 - in mean square, order 2, 230
 - in probability, 230
 - of random variables, 230
- convolution, 218
 - of binomial distributions, 221
 - of densities, 220
 - of exponential distributions, 221
 - of normal distributions, 221
 - of Poisson distributions, 221
 - of uniform distributions, 220
- correlation, 375
- covariance, 196

definition

- σ -algebra, 21
- absolutely continuous
 - distribution of a random vector, 193
- absolutely continuous
 - random variables, 85
- arithmetic mean, 254
- axiomatic definition of
 - probability, 19
- classical probability space, 25
- coefficient of kurtosis, 266
- coefficient of skewness, 265
- conditional probability, 42
- confidence interval, 285
- continuous probability space, 32
- convergence of random
 - variables, 230
- covariance of random
 - variables, 196
- cumulative distribution
 - function, 77
- discrete distribution of
 - a random vector, 191
- discrete probability space, 32
- discrete random variable, 81
- distribution function of
 - random vector, 190
- empirical distribution
 - function, 416
- expected value of a random
 - variable, 101
- geometric mean, 254
- harmonic mean, 254
- independent events, 47
- measures of variability, 260
- median, 257
- mode, 258
- mutually independent events, 48

- mutually independent
 - random variables, 202
- of the n -th moment, 113
- pairwise independent events, 47
- Pearson sample correlation
 - coefficient, 371
- probability of a random
 - event, 24
- quantile, 258
- quantile function, 106
- random sample, 272
- random variable, 73
- random vector, 189
- range and classification, 250
- sample mean, 275
- Spearman's rank correlation
 - coefficient, 374
- standard deviation, 275
- theoretical and empirical
 - characteristics, 273
- types of frequencies, 246
- variance, 275
- variance of a random
 - variable, 110
- variance-covariance and
 - correlation matrix of a random vector, 198
- density of a random variable
 - joint, 195
 - marginal, 195
- density of a random vector, 194
- density of a vector
 - marginal, 195
- deviation
 - decile, 261
 - interquartile, 261
 - mean absolute, 260
 - percentile, 261
 - quantile, 261
 - standard, 111, 185

- diagram
 - line, 247
 - spike, 247
- dispersion, 260
- distribution
 - alternative, 121
 - Bernoulli, 121
 - binomial, 123
 - Cauchy, 98, 180
 - empirical, 229
 - probability, 76
 - relative frequencies, 229
 - Simpson, 221
- distribution of a random variable,
 - 90
 - F, 180
 - absolutely continuous, 106
 - asymptotically normal, 233
 - Bernoulli, 121
 - Bernoulli (0–1), 121
 - bimodal, 259
 - binomial, 123
 - Cauchy, 103
 - continuous, 84
 - continuous uniform, 151
 - discrete, 106
 - Erlang, 159
 - exponential distribution, 153
 - Fisher–Snedecor, 180, 181
 - gamma, 162
 - general bivariate normal, 201
 - general normal, 164
 - geometric, 129
 - hypergeometric, 137
 - leptokurtic, 114, 266
 - lognormal, 173
 - mesokurtic, 114, 266
 - multinomial, 200
 - multivariate, 200
 - multivariate normal, 201
 - negative binomial, 133
 - Pascal, 133
 - Pearson χ^2 , 176
 - Pearson χ^2 , 176
 - platykurtic, 114, 266
 - Poisson, 139
 - Pólya, 133
 - standard (normalized)
 - normal distribution, 168
 - Student's t , t distribution, 178
 - unimodal, 259
- distribution of a random vector
 - absolutely continuous, 193
 - discrete, 191
- error
 - margin of, 287
 - standard, 287
 - standard regression, 383, 386, 391
 - Type I, 298
 - Type II, 298
- estimate
 - confidence interval for the mean, 290
- estimator
 - consistent, 283
 - efficient, 284
 - interval, 285
 - one-sided confidence interval, 286
 - point, 281
 - two-sided confidence interval, 286
 - unbiased, 283
 - variance, 288
- event space, 21
- events
 - independent, 47, 50
 - mutually (totally)
 - independent, 48
 - pairwise independent, 47

- expectation, 102
- expected value, 101
- experiment
 - deterministic, 12
 - random, 13
- false negative, 298
- false positive, 298
- Fatou's lemma, 104
- formula of total probability, 62
- frequency, 245
 - absolute class, 246
 - cumulative
 - (cumulative-sum), 246
 - cumulative relative, 246
 - empirical, 409, 411
 - expected, 425, 429
 - percentage relative, 246
 - relative, 22, 246
 - theoretical, 409
- frequency polygon, 247
- function
 - beta, 178
 - cumulative distribution, 77
 - discrete distribution, 220
 - distribution of a random vector, 191
 - empirical distribution, 416
 - Euler, 178
 - joint distribution, 190, 195
 - marginal distribution, 195
 - measurable, 91
 - of random variables, 213
 - probability, 76
 - probability mass function of
 - a discrete random variable, 81
 - quantile, 106
 - regression, 381, 388
- Gaussian curve, 170
- histogram, 251
- hypothesis, 65, 297
 - alternative, 297
 - null, 297
 - of homogeneity, 424
 - of independence of random variables, 424
 - of symmetry, 424
 - one-sided, 298
 - right-sided, 298
 - two-sided, 298
- independence
 - components of a random vector, 203
 - of random variables, mutual, 202
 - of random variables, pairwise, 202
- independence of random events, 46
- inequality
 - Cauchy–Schwarz–Bunyakovsky, 372
 - Cauchy–Schwarz–Bunyakovsky, 196
 - Chebyshev, 231
 - Chebyshev's, 112
- integral
 - Euler of the first kind, 178
 - Lebesgue, 105
 - Lebesgue–Stieltjes, 103
 - Riemann, 105
- interaction, log-linear, 430
- interval
 - class, 250
 - confidence, 285
- kurtosis, 113
- law of large numbers, 229
 - strong, 232, 275

- weak, 231
- log-odds ratio of a 2×2
 - contingency table, 430
- matrix
 - correlation, 198
 - variance–covariance, 198
- mean
 - arithmetic, 254
 - geometric, 255
 - harmonic, 255
 - sample mean, 275
 - trimmed, 257
 - winsorized, 257
- measure, 89
 - Dirac, 121
 - Lebesgue, 30, 91
 - Lebesgue–Stieltjes, 90
 - probability, 90
- median, 107, 257, 258
- method
 - least squares, 382, 390
 - linearization, 400
 - Neményi's, 349
 - Tukey, 343
 - Tukey's multiple comparison
 - procedure, 353
- mode, 108, 258
- model
 - linear, 389
 - linearizable, 400
 - nonlinear, 400
 - regression, 381
- moment
 - n -th, 113
- moment of a random variable,
 - 113
 - absolute, 113
 - central, 113
- ogive, 247
- population size, 243
- power of the test, 298
- probability, 22
 - conditional, 41
 - geometric, 30
 - of an event, 24
 - posterior, 65
 - prior, 65
 - probability measure, 24
- probability density function, 85
- probability density function of
 - a random variable, 87
- probability space, 24
- Probability Theory, 11
- problem
 - Banach's, 57
 - De Montmort's matching, 57
- process
 - event
 - in continuous time, 139
 - in discrete time, 139
 - Poisson, 139
- quantile, 106, 276
 - sample, 258
- quartile
 - lower, 107
 - lower sample, 258
 - upper, 107
 - upper sample, 258
- random event, 21
 - certain, 21
 - complementary, 28
 - elementary, 19
 - sure, 21
- random sampling
 - simple, with replacement, 271
 - simple, without replacement,
 - 272
 - systematic (mechanical), 272
- random variable, 73

- absolutely continuous, 85
- discrete, 76, 81
- random variables
 - uncorrelated, 218
- random vector, 189
- range
 - decile, 261
 - interquartile, 261
 - percentile, 261
 - quantile, 261
 - studentized range, 344
 - variation, 250
- rank, 327
- region
 - critical, 298
 - non-rejection, 299
 - rejection, 298
- regression
 - linear, 381
 - linear, with multiple explanatory variables, 388
 - linearizable, 400
 - nonlinear, 400
 - polynomial, 395, 399
 - quadratic, 395
- sampling
 - stratified, 272
- significance, 301
- significance level, 298
- skewness, 113
 - left, 266
 - right, 266
- standard deviation, 260
- statistical induction, 271
- statistical inference, 271
- statistical population, 243
 - ordered, 257
 - population, 271
 - sample, 271
- statistical unit, 243
- Stochastics, 11
- subevent, 20
- submodel, 394
- sum of squares
 - columns, 352
 - residual, 342, 383, 391
 - rows, 352
 - total, 342, 383, 391
- system of normal equations, 391
- table
 - 2×2 contingency, 428
 - frequency distribution, 247
- test
 - t , one-sample, 300
 - t , two-sample, 308
 - Durbin–Watson, 388
 - equality of two variances, 310
 - Fisher’s exact, 429
 - for the mean based on the CLT, 317
 - Friedman’s, 354, 357
 - Kolmogorov–Smirnov
 - one-sample
 - goodness-of-fit test, 416
 - Kruskal–Wallis, 343, 348
 - Levene’s, 344
 - McNemar’s, 431
 - nonparametric, 327
 - normality test, 411
 - of equality of distributions, 427
 - of homogeneity, 427
 - of independence, 424
 - of symmetry, 433
 - paired t -test, 305
 - Pearson’s, 409
 - Poisson distribution, 413
 - sign, 328
 - Wilcoxon, one-sample, 330
 - Wilcoxon, two-sample, 333
- test statistic, 298

- theorem
 - Bayes', 64
 - Bayes' theorem, 65
 - central limit (CLT), 233
 - de Moivre–Laplace, 233
 - law of total probability, 61
 - Lévy–Lindeberg, 236
 - Lyapunov, 238
 - multiplication rule for
 - probabilities, 44
 - on a random sample from
 - a normal distribution, 275
 - on properties of the
 - distribution function, 190
 - on the independence of
 - events, 48
 - on the relationship between
 - means, 256
 - on total probability, 62
 - properties of the expected
 - value, 104
 - tied ranks, 328
 - total of the argument, 243
 - trial
 - Bernoulli, 50, 121
 - value
 - p -value, 299
 - critical, 276
 - variable
 - of the population, 243
 - variance, 110
 - population, 260
 - residual, 343, 391
 - sample, 260
 - variance–covariance matrix of
 - a random vector, 198
 - z-score, 169

Microsoft Excel Functions and Tools for Probability and Statistics

= ABS(number) returns the absolute value of the selected number. The function can be used, for example, when calculating the mean absolute deviation.

Data Analysis, a tool on the Data tab used for hypothesis testing. This tool is not displayed on the MS Excel ribbon by default. If it needs to be enabled, go to File, Options, Add-ins, Manage Excel Add-ins, and click Go. A separate dialogue box will then appear, where the Analysis ToolPak option should be selected and confirmed by clicking OK. The Data Analysis command will subsequently be available in the Analysis group on the Data tab. Data Analysis also returns the value of Pearson's correlation coefficient and Spearman's rank correlation coefficient r_s for the obtained ranks.

- **ANOVA: Single Factor:** this tool performs a one-way analysis of variance for two or more samples. The analysis tests the hypothesis that each sample is drawn from the same probability distribution. If there are only two samples, the **= T.TEST** function can be used instead.
- **ANOVA: Two-Factor With Replication:** this tool is used for testing data collected twice from the same set of subjects and for verifying whether each pair belongs to the same population. Otherwise, the alternative hypothesis that there are additional effects beyond those being studied cannot be rejected.



- **ANOVA: Two-Factor Without Replication:** similar to ANOVA: Two-Factor With Replication, except that only one observation is assumed for each pair.
- **Correlation:** used to calculate the correlation coefficient between two measured variables.
- **Covariance:** used to determine the dependence between two measured variables. Dependence means that high values of one variable correspond to high values of the other variable (positive covariance), or that low values of one variable correspond to high values of the other variable (negative covariance). If the values of the two variables are independent, the covariance will be close to zero.
- **Descriptive Statistics:** used to determine the number of values in a dataset, the sum, minimum and maximum values, arithmetic mean, median, mode, sample standard deviation, sample variance, skewness, kurtosis, and the difference between the maximum and minimum values (range).
- **Regression:** performs linear regression by fitting a straight line to a set of observations using the least squares method. Regression makes it possible to analyze how the values of one or more independent variables affect the value of a dependent variable.
- **F-Test Two-Sample for Variances:** this tool performs an F -test for two samples and compares their variances at a selected significance level. An F value close to 1 supports the equality of the population variances.
- **t-Test: Paired Two Sample for Means:** used to test the means of paired observations in each sample.
- **t-Test: Two-Sample Assuming Equal Variances:** performs Student's t -test for two datasets under the assumption that the data come from populations with equal variances.
- **t-Test: Two-Sample Assuming Unequal Variances:** performs Student's t -test for two datasets under the assumption that the data come from populations with different variances.
- **z-Test: Two Sample for Means:** this test is used to test the null hypothesis that there is no difference between two population means.

= BINOM.DIST(number, trials, probability, cumulative) returns the value of the cumulative distribution function F_X or the probability mass function p_X of the binomial distribution of the discrete random variable X .

= CEILING.MATH(number, significance, mode) rounds a number up to the nearest integer or to the nearest multiple of significance.

= CONFIDENCE.NORM(alpha, standard deviation, size) returns the confidence interval for the population mean assuming a normal distribution. Here,

“size” denotes the sample size.

= **CORREL**(array1,array2) returns the value of Pearson’s correlation coefficient r_s .

= **COVARIANCE**(array1,array2) returns the covariance, the average of the products of deviations for each pair of data points in two datasets.

= **COVARIANCE.P**(array1,array2) returns the population covariance, the average of the products of deviations for each pair of data points in two datasets.

= **COVARIANCE.S**(array1,array2) returns the sample covariance, the average of the products of deviations for each pair of data points in two datasets.

= **CONFIDENCE.NORM**(alpha,standard_dev,size) returns the confidence interval for the population mean using the normal distribution.

= **CONFIDENCE.T**(alpha,standard_dev,size) returns the confidence interval for the population mean using Student’s t distribution. Here, “size” denotes the sample size.

= **CRITBINOM**(trials,probability,alpha) returns the smallest value for which the cumulative binomial distribution is greater than or equal to a criterion value. Here, “trials” denotes the number of Bernoulli trials and “probability” the probability of success on each trial.

= **FREQUENCY**(data,values) calculates the frequency of values occurring within a range of values and returns a vertical array of numbers having one more row than the bins array. The function is used to create a frequency distribution table and can also be employed to determine interval frequencies by specifying a column of upper interval limits.

When processing and analyzing data, the first step is often to sort the data by magnitude, either from the smallest value to the largest or vice versa. This can be done by right-clicking any cell in a column of data and selecting Sort and Smallest to Largest or Largest to Smallest, respectively. Another method leading to the same result is to use the Sort command on the Data tab.

= **DEVSQ**(number1,number2, . . .) returns the sum of squares of deviations of the specified values from their mean.

= **EXP**(number) returns the value of Euler’s⁴ number e raised to the specified power ($e \approx 2.718$).

= **EXPON.DIST**(x,lambda,cumulative) returns the value of the cumulative distribution function F_X or the probability density function f_X of the exponential distribution of the continuous random variable X .

= **F.DIST**(x,degree of freedom1,degree of freedom2,cumulative) returns the left-tailed probability value of the F distribution of the random variable X .

⁴Leonhard Paul Euler (1707–1783), Swiss mathematician and physicist.

= **F.DIST.RT**(x ,degree of freedom1,degree of freedom2) returns the right-tailed probability value of the F distribution of the random variable X .

= **F.INV**(probability,degree of freedom1,degree of freedom2) returns the quantile value of the cumulative distribution function F_X of the F distribution of the random variable X .

= **F.TEST**(array1,array2) returns the result of an F -test as the one-tailed probability associated with testing the equality of variances for the data in array1 and array2 (the test statistic itself is not returned, so the degree of agreement cannot be assessed directly).

= **FACT**(number) returns the factorial of the specified number.

= **FORECAST.LINEAR**(x ,known_y's,known_x's) is used to calculate or predict a future value based on existing values. The future value is the value of y corresponding to a given value of x . Existing values are known as the x - and y -values, and the future value is predicted using linear regression. Here, x denotes the data point for which the value of y is to be predicted, "known_y's" is the dependent array or data range, and "known_x's" is the independent array or data range.

= **GAMMA.DIST**(x ,alpha,beta,cumulative) returns the value of the Gamma distribution.

= **GAMMA.INV**(probability,alpha,beta) returns the value of the inverse of the cumulative Gamma distribution function.

= **GEOMEAN**(number1,number2, ...) returns the geometric mean of a set of positive numbers.

Charts and graphs in MS Excel can be created using the wide range of tools available on the *Insert* tab in the *Charts* group.

= **HARMEAN**(number1,number2, ...) returns the harmonic mean of a set of specified numbers.

= **HYPGEOM.DIST**(sample_s,number_sample,population_s,number_pop,cumulative) returns the value of the cumulative distribution function F_X or the probability mass function p_X of the hypergeometric distribution of the discrete random variable X .

= **CHISQ.DIST**(x ,deg_freedom,cumulative) returns the left-tailed probability value of the χ^2 distribution of the random variable X .

= **CHISQ.DIST.RT**(x ,deg_freedom) returns the right-tailed probability value of the χ^2 distribution of the random variable X .

= **CHISQ.TEST**(actual_range,expected_range) performs a test of independence and returns the probability associated with the χ^2 distribution for the observed test statistic and the corresponding degrees of freedom.

= **INTERCEPT**(known_y's,known_x's) returns the y -intercept of the linear regression line fitted through the specified data points.

= **COMBIN**(number,number_chosen) returns the number of combinations

of k elements chosen from n elements.

PivotTable, a tool that can be used to create frequency distribution tables. A PivotTable is created from the **Insert** tab by selecting the *PivotTable* command.

= **KURT**(number1,number2,...) returns the kurtosis excess of a set of specified numbers. The KURT function computes the kurtosis coefficient (denoted by a_4^*) using a formula different from that given in Definition 12.10:

$$a_4^* = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_n} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}. \quad (21.3)$$

= **LARGE**(array,k) returns the k -th largest value in a set of specified numbers.

= **LINEST**(known_y's,known_x's,const,stats) calculates, using the least squares method, the statistics for the straight line $y = mx + b$ that best fits the data and returns an array containing the parameters m and b .

= **TREND**(known_y's,known_x's,new_x's,const) returns values along a linear trend that fits the known data points using the least squares method.

= **LN**(number) returns the natural logarithm of the specified number.

= **LOG10**(number) returns the base-10 logarithm of the specified number.

= **LOG**(number,base) returns the logarithm of the specified number to the given base.

= **LOGNORM.DIST**(x,mean,standard_dev,cumulative) returns the value of the cumulative distribution function F_X or the probability density function f_X of the lognormal distribution of the continuous random variable X .

= **LOGNORM.INV**(probability,mean,standard_dev) returns the inverse of the cumulative distribution function of the lognormal distribution of the random variable X , where $\ln X$ has a normal distribution with parameters μ and σ^2 .

= **MAX**(number1,number2, ...) returns the largest value in a set of specified numbers.

= **MEDIAN**(number1,number2, ...) returns the median (middle value) of a set of specified numbers.

= **MIN**(number1,number2, ...) returns the smallest value in a set of specified numbers.

= **MODE.MULT**(number1,number2, ...) returns the mode (most frequently occurring value) of a set of specified numbers.

= **MODE.SNGL**(number1,number2, ...) returns the mode (most frequently occurring value) of a set of specified numbers.

= **MULTINOMIAL**(number1,number2, ...) returns the multinomial coefficient, that is, the number of permutations with repetitions. The function

can be used to calculate arrangements with repetition, where the individual numbers represent the multiplicities of identical elements.

= **NEGBINOM.DIST**(number_f,number_s,probability_s,cumulative) returns the value of the cumulative distribution function F_X or the probability mass function p_X of the negative binomial distribution of the discrete random variable X .

= **NORM.DIST**(x,mean,standard_dev,cumulative) returns the value of the cumulative distribution function F_X or the probability density function f_X of the normal distribution of the continuous random variable X .

= **NORM.S.DIST**(z,cumulative) returns the value of the cumulative distribution function F_X or the probability density function f_X of the standard normal distribution of the continuous random variable X .

= **NORM.INV**(probability;mean;standard_dev) returns the value of the quantile function F_X^{-1} corresponding to the specified probability for a random variable X having a normal distribution.

= **NORM.S.INV**(probability) returns the value of the quantile function Φ^{-1} corresponding to the specified probability for a random variable X having a standard normal distribution.

= **SQRT**(number) returns the square root of the specified value.

= **PEARSON**(array1,array2) returns the value of Pearson's correlation coefficient r .

= **PERCENTILE.EXC**(array,k) returns the k -th percentile of a dataset, where $k \in (0; 1)$.

= **PERCENTILE.INC**(array,k) returns the k -th percentile of a dataset, where $k \in (0; 1)$.

= **PERMUT**(number;number_chosen) returns the number of variations of class k from n elements or, equivalently, the number of permutations of n elements.

= **PHI**(x) returns the value of the probability density function of the standard normal distribution.

= **COUNT**(number1,number2, ...) returns the number of elements in a set of specified numbers.

= **POISSON.DIST**(x,mean,cumulative) returns the value of the cumulative distribution function F_X or the probability mass function p_X of the Poisson distribution of the discrete random variable X .

= **POWER**(number,power) returns a number raised to a specified power. The function can be used, for example, when calculating variations with repetition.

= **PROB**(x_range,prob_range,lower_limit,upper_limit) returns the probability that the values in a range fall between two limits. If the upper limit is omitted, the function returns the probability that the values in the spec-

ified range are equal to the lower limit.

= AVERAGE(number1,number2, ...) returns the arithmetic mean.

= AVEDEV(number1,number2, ...) returns the average of the absolute deviations of a set of specified numbers.

= QUARTILE.INC(array,quart) **= QUARTILE.EXC**(array,quart)

returns the quartile value of a dataset expressed in terms of percentiles, $k \in \{0, 1, 2, 3, 4\}$, where $k = 0$ is the minimum value, $k = 1$ is the first (lower) quartile (the 25% percentile), $k = 2$ is the median (the 50% percentile), $k = 3$ is the third (upper) quartile (the 75% percentile), and $k = 4$ is the maximum value.

= RANDBETWEEN(bottom,top) returns a random integer from the specified interval.

= RANK.AVG(number,ref,order) returns the rank of a number in a list of numbers, that is, its relative position among the other values in the list. If several values share the same rank, the average rank is returned. The third argument equal to 1 indicates ascending order, while 2 indicates descending order.

= SKEW(number1,number2, ...) returns the skewness of the distribution of a set of specified values. The SKEW function computes the skewness coefficient using a formula different from that given in Definition 12.9:

$$a_3^* = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_n} \right)^3. \quad (21.4)$$

= SKEW.P(number1,number2, ...) returns the skewness of a population.

= SLOPE(known_y's,known_x's) returns the slope of the linear regression line fitted to the specified data points.

= SMALL(array;k) returns the k -th smallest value in a set of specified values.

= STDEV.P(number1,number2, ...) returns the standard deviation of a population.

= STDEV.S(number1,number2, ...) returns the sample standard deviation of a set of specified numbers. The standard deviation can also be calculated as the square root of the sample variance using **= SQRT**.

= SUMPRODUCT(array1, ...) returns the sum of the products of corresponding values. It can be used, for example, to calculate the mean value $\mathbb{E}X$ of a discrete random variable.

= STANDARDIZE(x,mean,standard_dev) returns a normalized value from a distribution characterized by a specified mean and standard deviation.

= STEYX(known_y's,known_x's) returns the standard error of the predicted y -value for each x in a linear regression. The standard error is deter-

mined by the amount of error in estimating y from individual values of x . The `=STEYX` function uses the following formula to compute the standard error of the estimate of y :

$$\sqrt{\frac{1}{n-2} \left[\sum (y - \bar{y})^2 - \frac{[\sum (x - \bar{x})(y - \bar{y})]^2}{\sum (x - \bar{x})^2} \right]},$$

where x and y are the independent and dependent values from the input arrays, respectively, and n is the total sample size.

`=SUM`(number1,number2, ...) returns the sum of the specified numbers.

`=T.DIST`(x,deg_freedom,cumulative) returns the value of the left-tailed Student's t distribution of the random variable X .

`=T.DIST.2T`(x,deg_freedom) returns the two-tailed probability associated with Student's t distribution of the random variable X .

`=T.DIST.RT`(x,deg_freedom) returns the value of the right-tailed Student's t distribution of the random variable X .

`=T.TEST`(array1,array2,tails,type) returns the result of a t -test used to assess whether the means of two samples are significantly different. The argument "tails" specifies whether a one-tailed or two-tailed test is performed. The argument "type" specifies the type of t -test.

`=TRIMMEAN`(array,percent) returns the mean of the interior portion of a dataset obtained by excluding a specified percentage of data points from both the high and low ends of the dataset. The function is useful when calculating a trimmed mean.

`=VAR.P`(number1,number2, ...) returns the variance of a population.

`=VAR.S`(number1;number2, ...) returns the sample variance of a set of specified numbers.

`=WEIBULL.DIST`(x,alpha,beta,cumulative) returns the value of the probability density function or, respectively, the cumulative distribution function of the Weibull distribution of the random variable X .

`=Z.TEST`(array,x,sigma) returns the one-tailed p -value of a Z -test, where "sigma" denotes the standard deviation of the population.

Statistical Tables

Table 21.29: Critical Values for the Sign Test

Number of Pairs	Significance Level α		
n	0.01	0.05	0.10
5	-	-	0
6	-	0	0
7	-	0	0
8	0	0	1
9	0	1	1
10	0	1	1
11	0	1	2
12	1	2	2
13	1	2	3
14	1	2	3
15	2	3	3
16	2	3	4
17	2	4	4
18	3	4	5
19	3	4	5
20	3	5	5

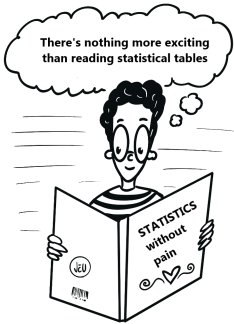


Table 21.30: Critical Values k_1 and k_2 for the Sign Test,
 $\mathbb{P}(Y \leq k_1) \leq \frac{\alpha}{2}, \mathbb{P}(Y \geq k_2) \leq \frac{\alpha}{2}$

$\alpha = 0.05 \quad \alpha = 0.01$					$\alpha = 0.05 \quad \alpha = 0.01$					$\alpha = 0.05 \quad \alpha = 0.01$				
N	k_1	k_2	k_1	k_2	N	k_1	k_2	k_1	k_2	N	k_1	k_2	k_1	k_2
6	0	6	-	-	21	5	16	4	17	36	11	25	9	27
7	0	7	-	-	22	5	17	4	18	37	12	25	10	27
8	0	8	0	8	23	6	17	4	19	38	12	26	10	28
9	1	8	0	9	24	6	18	5	19	39	12	27	11	28
10	1	9	0	10	25	7	18	5	20	40	13	27	11	29
11	1	9	0	11	26	7	19	6	20	41	13	28	11	30
12	2	10	1	11	27	7	20	6	21	42	14	28	12	30
13	2	11	1	12	28	8	20	6	22	43	14	29	12	31
14	2	12	1	13	29	8	21	7	22	44	15	29	13	31
15	3	12	2	13	30	9	21	7	23	45	15	30	13	32
16	3	13	2	14	31	9	22	7	24	46	15	31	13	33
17	4	13	2	15	32	9	23	8	24	47	16	31	14	33
18	4	14	3	15	33	10	23	8	25	48	16	32	14	34
19	4	15	3	16	34	10	24	9	25	49	17	32	15	34
20	5	15	3	17	35	11	24	9	26	50	17	33	15	35

Table 21.31: Critical Values w_n for the One-Sample Two-Sided and One-Sided Wilcoxon Signed-Rank Test, $\mathbb{P}(\min(S^+, S^-) \leq W_n(\alpha)) \leq \alpha$

n	Two-Sided Test		One-Sided Test	
	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
5	—	—	0	—
6	0	—	2	—
7	2	—	3	0
8	3	0	5	1
9	5	1	8	3
10	8	3	10	5
11	10	5	13	7
12	13	7	17	9
13	17	9	21	12
14	21	12	25	15
15	25	15	30	19
16	29	19	35	23
17	34	23	41	27
18	40	27	47	32
19	46	32	53	37
20	52	37	60	43
21	58	42	67	49
22	65	48	75	55
23	73	54	83	62
24	81	61	91	69
25	89	68	100	76
26	98	75	110	84
27	107	83	119	92
28	116	91	130	101
29	126	100	140	110
30	137	109	151	120

Table 21.32: Critical Values $W(0.05)$ for the Two-Sample Wilcoxon Rank-Sum Test, $\mathbb{P}(\min(U_1, U_2) \leq W(0.05)) \leq 0.05$

m/n	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
4	-	-	0														
5	-	0	1	2													
6	-	1	2	3	5												
7	-	1	3	5	6	8											
8	0	2	4	6	8	10	13										
9	0	2	4	7	10	12	15	17									
10	0	3	5	8	11	14	17	20	23								
11	0	3	6	9	13	16	19	23	26	30							
12	1	4	7	11	14	18	22	26	29	33	37						
13	1	4	8	12	16	20	24	28	33	37	41	45					
14	1	5	9	13	17	22	26	31	36	40	45	50	55				
15	1	5	10	14	19	24	29	34	39	44	49	54	59	64			
16	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75		
17	2	6	11	17	22	28	34	39	45	51	57	63	69	75	81	87	
18	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99
19	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106
20	2	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112
21	3	8	15	22	29	36	43	50	58	65	73	80	88	96	103	111	119
22	3	9	16	23	30	38	45	53	61	69	77	85	93	101	109	117	125
23	3	9	17	24	32	40	48	56	64	73	81	89	98	106	115	123	132
24	3	10	17	25	33	42	50	59	67	76	85	94	102	111	120	129	138
25	3	10	18	27	35	44	53	62	71	80	89	98	107	117	126	135	145
26	4	11	19	28	37	44	53	62	71	80	89	98	107	117	126	135	145
27	4	11	20	29	38	48	57	67	77	87	97	107	117	127	137	147	158
28	4	12	21	30	40	50	60	70	80	90	101	111	122	132	143	154	164
29	4	13	22	32	42	52	62	73	83	94	105	116	127	138	149	160	171
30	5	13	23	33	43	54	65	76	87	98	109	120	131	143	154	166	177

Table 21.33: Critical Values $W(0.01)$ for the Two-Sample Wilcoxon Rank-Sum Test, $\mathbb{P}(\min(U_1, U_2) \leq W(0.01)) \leq 0.01$

m/n	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
5	-	-	-	0													
6	-	-	0	1	2												
7	-	-	0	1	3	4											
8	-	-	1	2	4	6	7										
9	-	0	1	3	5	7	9	11									
10	-	0	2	4	6	9	11	13	16								
11	-	0	2	5	7	10	13	16	18	21							
12	-	1	3	6	9	12	15	18	21	24	27						
13	-	1	3	7	10	13	17	20	24	27	31	34					
14	-	1	4	7	11	15	18	22	26	30	34	38	42				
15	-	2	5	8	12	16	20	24	29	33	37	42	46				
16	-	2	5	9	13	18	22	27	31	36	41	45	50				
17	-	2	6	10	15	19	24	29	34	39	44	49	54	60			
18	-	2	6	11	16	21	26	31	37	42	47	53	58	64	70		
19	0	3	7	12	17	22	28	33	39	45	51	57	63	69	74	81	
20	0	3	8	13	18	24	30	36	42	48	54	60	67	73	79	86	92
21	0	3	8	14	19	25	32	38	44	51	58	64	71	78	84	91	98
22	0	4	9	14	21	27	34	40	47	54	61	68	75	82	89	96	104
23	0	4	9	15	22	29	35	43	50	57	64	72	79	87	94	102	109
24	0	4	10	16	23	30	37	45	52	60	68	75	83	91	99	107	115
25	0	5	10	17	24	32	39	47	55	63	71	79	87	96	104	112	121
26	0	5	11	18	25	33	41	49	58	66	74	83	92	100	109	118	127
27	1	5	12	19	27	35	43	52	60	69	78	87	96	105	114	123	132
28	1	5	12	20	28	36	45	54	63	72	81	91	100	109	119	128	138
29	1	6	13	21	29	38	47	56	66	75	85	94	104	114	124	134	144
30	1	6	13	22	30	40	49	58	68	78	88	98	108	119	129	139	150

Table 21.34: Critical Values $q_{m,\nu}(0.05)$ for Tukey's Multiple Comparison Procedure, $X \sim q_{m,\nu}$, $\mathbb{P}(X \geq q_{m,\nu}(0.05)) = 0.05$

ν/m	2	3	4	5	6	7	8	9	10	11	12	13	14
1	18.0	27.0	32.8	37.1	40.4	43.1	45.4	47.4	49.1	50.6	52.0	53.2	54
2	6.08	8.33	9.80	10.9	11.7	12.4	13.0	13.5	14.0	14.4	14.7	15.1	15
3	4.50	5.91	6.82	7.50	8.04	8.48	8.85	9.18	9.46	9.72	9.95	10.2	10
4	3.93	5.04	5.76	6.29	6.71	7.05	7.35	7.60	7.83	8.03	8.21	8.37	8.5
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17	7.32	7.47	7.6
6	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49	6.65	6.79	6.92	7.0
7	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30	6.43	6.55	6.6
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05	6.18	6.29	6.3
9	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74	5.87	5.98	6.09	6.1
10	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72	5.83	5.93	6.0
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61	5.71	5.81	5.9
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39	5.51	5.61	5.71	5.8
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43	5.53	5.63	5.7
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36	5.46	5.55	5.6
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.31	5.40	5.49	5.5
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26	5.35	5.44	5.5
17	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11	5.21	5.31	5.39	5.4
18	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17	5.27	5.35	5.4
19	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.14	5.23	5.31	5.3
20	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11	5.20	5.28	5.3
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01	5.10	5.18	5.2
30	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82	4.92	5.00	5.08	5.1
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73	4.82	4.90	4.98	5.0
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73	4.81	4.88	4.9
120	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56	4.64	4.71	4.78	4.8
∞	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55	4.62	4.68	4.7

Table 21.35: Critical Values $q_{m,\nu}(0.01)$ for Tukey's Multiple Comparison Procedure, $X \sim q_{m,\nu}$, $\mathbb{P}(X \geq q_{m,\nu}(0.01)) = 0.01$

ν/m	2	3	4	5	6	7	8	9	10	11	12	13	14
1	90.0	135	164	186	202	216	227	237	246	253	260	266	272
2	14.0	19.0	22.3	24.7	26.6	28.2	29.5	30.7	31.7	32.6	33.4	34.1	34.8
3	8.26	10.6	12.2	13.3	14.2	15.0	15.6	16.2	16.7	17.1	17.5	17.9	18.2
4	6.51	8.12	9.17	9.96	10.6	11.1	11.5	11.9	12.3	12.6	12.8	13.1	13.3
5	5.70	6.97	7.80	8.42	8.91	9.32	9.67	9.97	10.2	10.5	10.7	10.9	11.1
6	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30	9.49	9.65	9.81
7	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37	8.55	8.71	8.86	9.00
8	4.74	5.63	6.20	6.63	6.96	7.24	7.47	7.68	7.87	8.03	8.18	8.31	8.44
9	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.32	7.49	7.65	7.78	7.91	8.03
10	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	7.36	7.48	7.60	7.71
11	4.39	5.14	5.62	5.97	6.25	6.48	6.67	6.84	6.99	7.13	7.25	7.36	7.46
12	4.32	5.04	5.50	5.84	6.10	6.32	6.51	6.67	6.81	6.94	7.06	7.17	7.26
13	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	6.79	6.90	7.01	7.10
14	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	6.66	6.77	6.87	6.96
15	4.17	4.83	5.25	5.56	5.80	5.99	6.16	6.31	6.44	6.55	6.66	6.76	6.84
16	4.13	4.78	5.19	5.49	5.72	5.92	6.08	6.22	6.35	6.46	6.56	6.66	6.74
17	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27	6.38	6.48	6.57	6.66
18	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.31	6.41	6.50	6.58
19	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14	6.25	6.34	6.43	6.51
20	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.19	6.29	6.37	6.45
25	3.96	4.54	4.91	5.17	5.37	5.54	5.69	5.81	5.92	6.02	6.11	6.19	6.26
30	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	5.85	5.93	6.01	6.08
40	3.82	4.37	4.70	4.93	5.11	5.27	5.39	5.50	5.60	5.69	5.77	5.84	5.90
60	3.76	4.28	4.60	4.82	4.99	5.13	5.25	5.36	5.45	5.53	5.60	5.67	5.73
120	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	5.38	5.44	5.51	5.56
∞	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.23	5.29	5.35	5.40

Table 21.36: Table of Critical Values for Nemenyi's Multiple Comparison Procedure Based on Ranks, $\alpha = 0.05$

m/I	3	4	5	6	7	8	9	10
1	3.3	4.7	6.1	7.5	9.0	10.5	12.0	13.5
2	8.8	12.6	16.5	20.5	24.7	28.9	33.1	37.4
3	15.7	22.7	29.9	37.3	44.8	52.5	60.3	68.2
4	23.9	34.6	45.6	57.0	68.6	80.4	92.4	104.6
5	33.1	48.1	63.5	79.3	95.5	112.0	128.8	145.8
6	43.3	62.9	83.2	104.0	125.3	147.0	169.1	191.4
7	54.4	79.1	104.6	130.8	157.6	184.9	212.8	240.9
8	66.3	96.4	127.6	159.6	192.4	225.7	259.7	294.1
9	78.9	114.8	152.0	190.2	229.3	269.1	309.6	350.6
10	92.3	134.3	177.8	222.6	268.4	315.0	362.4	410.5
11	106.3	154.8	205.0	256.6	309.4	363.2	417.9	473.3
12	120.9	176.2	233.4	292.2	352.4	413.6	476.0	539.1
13	136.2	198.5	263.0	329.3	397.1	466.2	536.5	607.7
14	152.1	221.7	293.8	367.8	443.6	520.8	599.4	679.0
15	168.6	245.7	325.7	407.8	491.9	577.4	664.6	752.8
16	185.6	270.6	358.6	449.1	541.7	635.9	732.0	829.2

Table 21.37: Critical Values for Nemenyi’s Multiple Rank Comparison Procedure, $\alpha = 0.01$

m/I	3	4	5	6	7	8	9	10
1	4.1	5.7	7.3	8.9	10.5	12.2	13.9	15.6
2	10.9	15.3	19.7	24.3	28.9	33.6	38.3	43.1
3	19.5	27.5	35.7	44.0	52.5	61.1	69.8	78.6
4	29.7	41.9	54.5	67.3	80.3	93.6	107.0	120.6
5	41.2	58.2	75.8	93.6	111.9	130.4	149.1	168.1
6	53.9	76.3	99.3	122.8	146.7	171.0	195.7	220.6
7	67.6	95.8	124.8	154.4	184.6	215.2	246.3	277.7
8	82.4	116.8	152.2	188.4	225.2	262.6	300.6	339.0
9	98.1	139.2	181.4	224.5	268.5	313.1	358.4	404.2
10	114.7	162.8	212.2	262.7	314.2	366.5	419.5	473.1
11	132.1	187.6	244.6	302.9	362.2	422.6	483.7	545.6
12	150.4	213.5	278.5	344.9	412.5	481.2	551.0	621.4
13	169.4	240.6	313.8	388.7	464.9	542.4	621.0	700.5
14	189.1	268.7	350.5	434.2	519.4	606.0	693.8	782.6
15	209.6	297.8	388.5	481.3	575.8	671.9	769.3	867.7
16	230.7	327.9	427.9	530.1	634.2	740.0	847.3	955.7

Table 21.38: Critical Values for Friedman’s Test, $\alpha = 0.05$

I/J	3	4	5	6	7	8	9	10	11	12
3	6.000	7.4	8.53	9.86	11.24	12.57	13.88	15.19	16.48	17.76
4	6.500	7.8	8.8	10.24	11.63	12.99	14.34	15.67	16.98	18.3
5	6.400	7.8	8.99	10.43	11.84	13.23	14.59	15.93	17.27	18.6
6	7.000	7.6	9.08	10.54	11.97	13.38	14.76	16.12	17.4	18.8
7	7.143	7.8	9.11	10.62	12.07	13.48	14.87	16.23	17.6	18.9
8	6.250	7.65	9.19	10.68	12.14	13.56	14.95	16.32	17.7	19.0
9	6.222	7.66	9.22	10.73	12.19	13.61	15.02	16.40	17.7	19.1
10	6.200	7.67	9.25	10.76	12.23	13.66	15.07	16.44	17.8	19.2
11	6.545	7.68	9.27	10.79	12.27	13.70	15.11	16.48	17.9	19.2
12	6.167	7.70	9.29	10.81	12.29	13.73	15.15	16.53	17.9	19.3
13	6.000	7.70	9.30	10.83	12.32	13.76	15.17	16.56	17.9	19.3
14	6.143	7.71	9.32	10.85	12.34	13.78	15.19	16.58	17.9	19.3
15	6.400	7.72	9.33	10.87	12.35	13.80	15.20	16.6	18.0	19.3
16	5.99	7.73	9.34	10.88	12.37	13.81	15.23	16.6	18.0	19.3
20	5.99	7.74	9.37	10.92	12.41	13.8	15.3	16.7	18.0	19.4
∞	5.99	7.82	9.49	11.07	12.59	14.07	15.51	16.92	18.31	19.68

Table 21.39: Critical Values for Friedman's Test, $\alpha = 0.01$

I/J	3	4	5	6	7	8	9	10	11	12
3	-	9.000	10.13	11.76	13.26	14.78	16.28	17.74	19.19	20.61
4	8.000	9.600	11.20	12.59	14.19	15.75	17.28	18.77	20.24	21.7
5	8.400	9.96	11.43	13.11	14.74	16.32	17.86	19.37	20.86	22.3
6	9.000	10.200	11.75	13.45	15.10	16.69	18.25	19.77	21.3	22.7
7	8.857	10.371	11.97	13.69	15.35	16.95	18.51	20.04	21.5	23.0
8	9.000	10.35	12.14	13.87	15.53	17.15	18.71	20.24	21.8	23.2
9	8.667	10.44	12.27	14.01	15.68	17.29	18.87	20.42	21.9	23.4
10	9.600	10.53	12.38	14.12	15.79	17.41	19.00	20.53	22.0	23.5
11	9.455	10.60	12.46	14.21	15.89	17.52	19.10	20.64	22.1	23.6
12	9.500	10.68	12.53	14.28	15.96	17.59	19.19	20.73	22.2	23.7
13	9.385	10.72	12.58	14.34	16.03	17.67	19.25	20.80	22.3	23.8
14	9.000	10.76	12.64	14.40	16.09	17.72	19.31	20.86	22.4	23.9
15	8.933	10.80	12.68	14.44	16.14	17.78	19.35	20.9	22.4	23.9
16	8.79	10.84	12.72	14.48	16.18	17.81	19.40	20.9	22.5	24.0
20	8.87	10.94	12.83	14.60	16.30	18.00	19.5	21.1	22.6	24.1
∞	9.21	11.35	13.28	15.09	16.81	18.48	20.09	21.67	23.21	24.73

Table 21.40: Critical Values for Multiple Comparisons in Friedman's Test, $\alpha = 0.05$

I/J	3	4	5	6	7	8	9	10
1	3.3	4.7	6.1	7.5	9.0	10.5	12.0	13.5
2	4.7	6.6	8.6	10.7	12.7	14.8	17.0	19.2
3	5.7	8.1	10.6	13.1	15.6	18.2	20.8	23.5
4	6.6	9.4	12.2	15.1	18.0	21.0	24.0	27.1
5	7.4	10.5	13.6	16.9	20.1	23.5	26.9	30.3
6	8.1	11.5	14.9	18.5	22.1	25.7	29.4	33.2
7	8.8	12.4	16.1	19.9	23.9	27.8	31.8	35.8
8	9.4	13.3	17.3	21.3	25.5	29.7	34.0	38.3
9	9.9	14.1	18.3	22.6	27.0	31.5	36.0	40.6
10	10.5	14.8	19.3	23.8	28.5	33.2	38.0	42.8
11	11.0	15.6	20.2	25.0	29.9	34.8	39.8	44.9
12	11.5	16.2	21.1	26.1	31.2	36.4	41.6	46.9
13	11.9	16.9	22.0	27.2	32.5	37.9	43.3	48.8
14	12.4	17.5	22.8	28.2	33.7	39.3	45.0	50.7
15	12.8	18.2	23.6	29.2	34.9	40.7	46.5	52.5
16	13.3	18.8	24.4	30.2	36.0	42.0	48.1	54.2

Table 21.41: Critical Values for Multiple Comparisons Following Friedman's Test, $\alpha = 0.01$

I/J	3	4	5	6	7	8	9	10
1	4.1	5.7	7.3	8.9	10.5	12.2	13.9	15.6
2	5.8	8.0	10.3	12.6	14.9	17.3	19.7	22.1
3	7.1	9.8	12.6	15.4	18.3	21.2	24.1	27.0
4	8.2	11.4	14.6	17.8	21.1	24.4	27.8	31.2
5	9.2	12.7	16.3	19.9	23.6	27.3	31.1	34.9
6	10.1	13.9	17.8	21.8	25.8	29.9	34.1	38.2
7	10.9	15.0	19.3	23.5	27.9	32.3	36.8	41.3
8	11.7	16.1	20.6	25.2	29.8	34.6	39.3	44.2
9	12.4	17.1	21.8	26.7	31.6	36.6	41.7	46.8
10	13.0	18.0	23.0	28.1	33.4	38.6	44.0	49.4
11	13.7	18.9	24.1	29.5	35.0	40.5	46.1	51.8
12	14.3	19.7	25.2	30.8	36.5	42.3	48.2	54.1
13	14.9	20.5	26.2	32.1	38.0	44.0	50.1	56.3
14	15.4	21.3	27.2	33.3	39.5	45.7	52.0	58.4
15	16.0	22.0	28.2	34.5	40.8	47.3	53.9	60.5
16	16.5	22.7	29.1	35.6	42.2	48.9	55.6	62.5

Table 21.42: Critical Values $D_n(\alpha)$ for the One-Sample Kolmogorov–Smirnov Test

n	$\alpha = 0.05$	$\alpha = 0.01$	n	$\alpha = 0.05$	$\alpha = 0.01$	n	$\alpha = 0.05$	$\alpha = 0.01$
1	.97500	.99500	31	.23788	.28530	61	.17091	.20506
2	.84189	.92929	32	.23424	.28094	62	.16956	.20343
3	.70760	.82900	33	.23076	.27677	63	.16823	.20184
4	.62394	.73424	34	.22743	.27279	64	.16693	.20029
5	.56328	.66853	35	.22425	.26897	65	.16567	.19877
6	.51926	.61661	36	.22119	.26532	66	.16443	.19729
7	.48342	.57581	37	.21826	.26180	67	.16322	.19584
8	.45427	.54179	38	.21544	.25843	68	.16204	.19442
9	.43001	.51332	39	.21273	.25205	69	.16088	.19303
10	.40925	.48893	40	.21012	.25205	70	.15975	.19167
11	.39122	.46770	41	.20760	.24904	71	.15864	.19034
12	.37543	.44905	42	.20517	.24613	72	.15755	.18903
13	.36143	.43247	43	.20283	.24332	73	.15649	.18776
14	.34890	.41762	44	.20056	.24060	74	.15544	.18650
15	.33760	.40420	45	.19837	.23798	75	.15442	.18528
16	.32733	.39201	46	.19625	.23544	76	.15342	.18408
17	.31796	.38086	47	.19420	.23298	77	.15244	.18290
18	.30936	.37062	48	.19221	.23059	78	.15147	.18174
19	.30143	.36117	49	.19028	.22828	79	.15052	.18060
20	.29408	.35241	50	.18841	.22604	80	.14960	.17949
21	.28724	.34427	51	.18659	.22386	81	.14868	.17840
22	.28087	.33666	52	.18482	.22174	82	.14779	.17732
23	.27490	.32954	53	.18311	.21968	83	.14691	.17627
24	.26931	.32286	54	.18144	.21768	84	.14605	.17523
25	.26404	.31657	55	.17981	.21574	85	.14520	.17421
26	.25907	.31064	56	.17823	.21384	86	.14437	.17321
27	.25438	.30502	57	.17669	.21199	87	.14355	.17223
28	.24993	.29971	58	.17519	.21019	90	.14117	.16938
29	.24571	.29466	59	.17373	.20844	95	.13746	.16493
30	.24170	.28987	60	.17231	.20673	100	.13403	.16081

Table 21.43: Critical Values for the Correlation Coefficient r

n	$\alpha = 0.05$	$\alpha = 0.01$	n	$\alpha = 0.05$	$\alpha = 0.01$	n	$\alpha = 0.05$	$\alpha = 0.01$
3	0.9969	0.9999	14	0.5324	0.6614	25	0.3961	0.5052
4	0.9500	0.9900	15	0.5140	0.6411	30	0.3610	0.4629
5	0.8783	0.9587	16	0.4973	0.6226	35	0.3338	0.4296
6	0.8114	0.9172	17	0.4822	0.6055	40	0.3120	0.4026
7	0.7545	0.8745	18	0.4683	0.5897	45	0.2940	0.3801
8	0.7067	0.8343	19	0.4555	0.5751	50	0.2787	0.3610
9	0.6664	0.7977	20	0.4438	0.5614	60	0.2542	0.3301
10	0.6319	0.7646	21	0.4329	0.5487	70	0.2352	0.3060
11	0.6021	0.7348	22	0.4227	0.5368	80	0.2352	0.2864
12	0.5760	0.7079	23	0.4123	0.5256	90	0.2072	0.2702
13	0.5529	0.6835	24	0.4044	0.5151	100	0.1966	0.2565

Table 21.44: Critical Values for Spearman's Rank Correlation Coefficient

n	$\alpha = 0.05$	$\alpha = 0.01$	n	$\alpha = 0.05$	$\alpha = 0.01$	n	$\alpha = 0.05$	$\alpha = 0.01$
5	0.9000	-	15	0.5179	0.6536	25	0.3977	0.5100
6	0.8286	0.9429	16	0.5000	0.6324	26	0.3894	0.5002
7	0.7450	0.8929	17	0.4853	0.6152	27	0.3822	0.4915
8	0.6905	0.7571	18	0.4716	0.5975	28	0.3749	0.4828
9	0.6833	0.8167	19	0.4579	0.5825	29	0.3685	0.4744
10	0.6364	0.7818	20	0.4451	0.5684	30	0.3620	0.4665
11	0.6091	0.7545	21	0.4351	0.5545			
12	0.5804	0.7273	22	0.4241	0.5426			
13	0.5549	0.6978	23	0.4150	0.5306			
14	0.5341	0.6747	24	0.4061	0.5200			

Bibliography

- [1] Anděl, J. (1978). *Matematická statistika*. SNTL/ALFA.
- [2] Anděl, J. (2003). *Statistické metody*. Matfyzpress.
- [3] Anděl, J. (2005). *Základy matematické statistiky*. Matfyzpress.
- [4] Anděl, J. (2011). *Statistické metody*. Matfyzpress.
- [5] Budíková, M., Mikoláš, Š., & Osecký, P. (2007). *Teorie pravděpodobnosti a matematická statistika* (3. vyd.). Masarykova univerzita, Přírodovědecká fakulta.
- [6] Büchter, A., & Henn, H.-W. (2007). *Elementare Stochastik*. Eine Einführung in die Mathematik der Daten und des Zufalls. Springer.
- [7] Denny, M., & Gaines, S. (2000). *Chance in Biology, Using probability to Explore Nature*. Princeton University Press.
- [8] Dupač, V., & Hušková, M. (1999). *Pravděpodobnost a matematická statistika*. Karolinum.
- [9] Hindls, R., Hronová, S., & Seger, J. (2004). *Statistika pro ekonomy*. Professional Publishing.
- [10] Hendl, J. (2015). *Přehled statistických metod*. Portál.
- [11] Hindls, R., Arltová, M., Hronová, S., Malá, I., Marek, L., Pecáková, I., & Řezanková, H. (2018). *Statistika v ekonomii*. Professional publishing.
- [12] Kahounová, J., & Hebák, P. (2014). *Počet pravděpodobnosti v příkladech*. Informatorium.
- [13] Mrkvička, T., & Petrášková, V. (2006). *Úvod do statistiky*. Jihočeská univerzita v Českých Budějovicích.

- [14] Mrkvička, T., & Petrášková, V. (2008). *Úvod do teorie pravděpodobnosti*. Jihočeská univerzita v Českých Budějovicích.
- [15] Neubauer, J., Sedlačík, M., & Kříž, O. (2021). *Základy statistiky*. Grada.
- [16] Płocki, A., & Tlustý, P. (2007). *Pravděpodobnost a statistika pro začátečníky a mírně pokročilé*. Prometheus.
- [17] Potocký, R. (1986). *Zbierka úloh z pravdepodobnosti a matematickej štatistiky*. Alfa, vydavateľstvo technickej a ekonomickej literatúry.
- [18] Prášilová, M., & Svatošová, L. (1997). *Cvičení ze statistiky*. Česká zemědělská univerzita v Praze, Provozně ekonomická fakulta.
- [19] Prášilová, M., & Zeipelt, R. (2001). *Cvičení ze statistiky I*. Česká zemědělská univerzita v Praze, Provozně ekonomická fakulta.
- [20] Stocker, T. C., & Steinke, I. (2017). *Statistik*. Übungsbuch, DE GRUYTER Oldenbourg.

Probability and Statistics with Solved Examples in MS Excel

Prof. RNDr. Tomáš Mrkvička, Ph.D.

Doc. RNDr. Vladimíra Petrášková, Ph.D.

Mgr. Michal Houda, Ph.D.

RNDr. PhDr. Jan Fiala, Ph.D.

ISBN 978-80-7694-176-2





Pedagogická
fakulta
Faculty
of Education

Jihočeská univerzita
v Českých Budějovicích
University of South Bohemia
in České Budějovice

Probability theory and mathematical statistics are fundamental disciplines of modern mathematics. Together, they provide an essential foundation for the analysis and interpretation of data across a wide range of scientific, technical, economic, and social science disciplines. This textbook offers a systematic presentation of both foundational and more advanced concepts, methods, and techniques in probability theory and mathematical statistics, while emphasising their practical application to real-world problems.

The conceptual exposition is complemented by a comprehensive collection of worked examples, practice exercises, and methodological commentaries designed to foster understanding of the subject matter and support the development of analytical thinking. Particular attention is devoted to the use of Microsoft Excel for statistical computation, data analysis, and interpretation of results, enabling readers to acquire not only theoretical knowledge but also the applied skills required for effective data work.

The book is primarily intended for undergraduate and postgraduate students in technical, natural science, economic, and teacher education programmes. At the same time, it serves as a valuable study and reference resource for anyone interested in probability theory, mathematical statistics, data analysis, and statistical reasoning.

Its clear and accessible presentation, strong integration of theory and practice, and emphasis on the use of widely available software tools make this publication an effective guide to the study of probability and statistics, as well as a useful resource for professional practice.

ISBN 978-80-7694-176-2



9 788076 941762