University of Applied Sciences BFI Vienna
Economics Management Finance

Welcome to the admission procedure
of the master's programme Quantitative Asset and Risk Management!

We are pleased about your interest in our degree programme and would like to provide you with the following materials as a basis for preparing for our admission procedure. It consists of two parts: an **online multiple choice test** and a **personal admission interview**.

We hope you enjoy studying the literature, wish you all the best for the admission procedure and look forward to meeting you in person!

Silvia Helmreich
Degree programme director

Veronika Hallwirth
Coordinator of the degree programme

Quantitative Asset and Risk Management
Admission test and interview

Admission test

The admission test is carried out as an **electronic multiple choice test** (duration: 60 minutes).

The following literature is recommended to help applicants prepare for the Mathematics and Statistics test: *Wirth, M., Preparation text for the admission test: Mathematics, Returns and Statistics, Version 2026.* Please find it attached.

In addition, English language skills are assessed in the multiple-choice test. If necessary supplementary exams in English may be prescribed.

Permitted aids for the admission test:

- Electronic calculator (type Standard TR) integrated in the MC test
- a pen or pencil
- blank paper

Admission interview

After the admission test, all applicants will be invited to an admission interview. The interview is scheduled for 15 minutes. Applicants are provided with relevant information about the structure of the programme. In addition, the applicants' motivation, English communication skills, and expectations are assessed.

Detailed information regarding the exact time slot and procedure for your admission interview will be sent to you by email a few days after you took the admission test.

What happens after the admission interview?

Please refer to our Guide MA Application and admission for the further procedure.

Quantitative Asset and Risk Management (ARIMA)

Preparation text for the admission test

Mathematics, Returns and Statistics

Version 2026

Preamble

This text is the sole document needed to prepare for the admission test to the study program *Quantitative Asset and Risk Management (ARIMA)* at the *University of Applied Sciences BFI Vienna*. This admission test will be conducted as a PC-based multiple choice test. During the test, applicants are either asked to perform similar calculations as shown in the text or to use the definitions included for a plausibility check on the presented answers.

Overall, this text contains six chapters, of which the first three discuss important mathematical concepts such as algebra, the definition and properties of functions and the calculation of the derivative of a function. Chapter 4 presents the fundamental concepts used for calculating the rate of return. Returns and returns vectors are one of the most important inputs to models in mathematical finance and are therefore part of almost every course covered during the ARIMA study program.

Chapters 5 and 6 highlight basic statistical concepts. Chapter 5 discusses the possibilities of descriptive statistical methods used for analyzing and presenting given data sets. In the closing chapter the theoretical concept of random variables is briefly introduced. Additionally some of the most important theoretical probability distributions such as the Binomial distribution and the Gaussian distribution are covered in some detail.

The preparation text was written by the members of the ARIMA-Team. Although we review it in regular intervals, it may still include some typos and literal mistakes. Feedback is always welcomed! We wish everyone reading this text a good time and hope that we will you at our admission test.

Best regards,
The ARIMA-Team

December 2025 in Vienna, Austria

Table of Contents

1 Algebra

The major difference between algebra and arithmetics is the inclusion of variables in algebra. While in arithmetics only numbers and their arithmetical operations (such as $+, -, \cdot, \div$) occur, in algebra variables such as x and y or a and b are used to replace numbers.

The purpose of using variables is to allow generalisations in mathematics. This is useful because it allows

- arithmetical equations to be stated as laws (e.g. $a + b = b + a$ for all a and b).
- reference to values which are not known. In context of a problem, a variable may represent a certain value which is not yet known, but which may be found through the formulation and manipulation of equations.
- the exploration of mathematical relationships between quantities.

1.1 Elementary algebra

In elementary algebra, an *expression* may contain numbers, variables and arithmetical operations. These are conventionally written with ≪higher-power≫ terms on the left.

Examples for algebraic expressions

1. $x + 2$
2. $x^2 - 2y + 7$
3. $z^8 + x^3(b + 2x) + a^4b - \pi$
4. $\frac{a^3-4ab+3}{a^2-1}$

In mathematics it is important that the value of an expression is always computed the same way. Therefore, it is necessary to compute the parts of an expression in a particular order, known as the *order of operations*. In elementary algebra three mathematical operations are used. Algebraic expressions can be linked via *addition*, *multiplication* and *applying exponents*. Additionally, algebraic expressions can contain parenthesis and absolute value symbols.

The standard order of operations is expressed in the following list.

1. *parenthesis* and other grouping symbols including *brackets*
2. *absolute value symbols* and the *fraction bar*
3. *exponents* and *roots*
4. *multiplication* and *division*
5. *addition* and *substraction*

The three basic algebraic operations have certain properties. These are briefly described in the following list.

Addition The symbol for addition is $+$ and for its inverse operation substraction it is $-$. While addition is associative[1] and commutative[2], substraction is neither of them.

Multiplication The standard symbol for multiplication is $\cdot$. Sometimes $\times$ is used as well. By convention, the symbol for multiplication can be left out if the interpretation is clear. The symbol for its inverse operation division is $\div$ or the fraction bar. Similar to addition, multiplication is associative and commutative, while division is not. Furthermore multiplication is distributive[3] over addition.

Exponentiation Usually exponentiation is represented by raising the exponent, e.g. a^b. Like the two inverse operations substraction and division, exponentiation is neither commutative nor associative. But it distributes over multiplication, as

$$(ab)^c = a^c b^c$$

holds for all a, b and c.

---

[1] An operation $\diamond$ is *associative* if and only if

$$(a \diamond b) \diamond c = a \diamond (b \diamond c)$$

holds for all a, b and c.
For example:

$$(3 + 4) + 5 = 7 + 5 = 12 = 3 + 9 = 3 + (4 + 5).$$

[2] An operation $\diamond$ is *commutative* if and only if

$$a \diamond b = b \diamond a$$

holds for all a and b.
For example:

$$3 + 4 = 7 = 4 + 3.$$

[3] An operation $\star$ *distributes* over an operation $\diamond$, if and only if

$$(a \diamond b) \star c = a \star c \diamond b \star c$$

holds for all a, b and c.
For example:

$$(3 + 4) \cdot 5 = 7 \cdot 5 = 35 = 15 + 20 = (3 \cdot 5) + (4 \cdot 5).$$

Exponentiation has two inverse operations, the logarithm and the n^{th} root. For the logarithm

$$a^{\log_a b} = b = \log_a a^b$$

holds for all a and b whereas for the root

$$\left(\sqrt[b]{a}\right)^b = a$$

holds for all a and b.

Additionally the following equations hold

$$\begin{aligned} a^{\frac{m}{n}} &= \left(\sqrt[n]{a}\right)^m = \sqrt[n]{a^m}, \\ a^b a^c &= a^{b+c}, \\ \left(a^b\right)^c &= a^{bc}. \end{aligned}$$

Examples for elementary algebraic operations

1. Applying the multiplicative distribution rules for $(x^2+2x-1)(x+2)$ we have

$$\begin{aligned} (x^2+2x-1)(x+2) &= x^3+2x^2-x+2x^2+4x-2 \\ &= x^3+4x^2+3x-2. \end{aligned}$$

2. For $(a^2-3b)^2$ we have

$$(a^2-3b)^2 = a^4 - 6a^2b + 9b^2.$$

3. A fraction like $\frac{x+4}{x^2-1} + \frac{3x}{x^2-1} + 4x$ can be simplified as

$$\begin{aligned} \frac{x+4}{x^2-1} + \frac{3x}{x^2-1} + 4x &= \frac{(x+4)+(3x)+4x(x^2-1)}{x^2-1} \\ &= \frac{x+4+3x+4x^3-4x}{x^2-1} = \frac{4x^3+4}{x^2-1}. \end{aligned}$$

   For $\frac{4x}{7} + \frac{8}{7} - x$ we have

$$\frac{4x}{7} + \frac{8}{7} - x = \frac{4x-7x+8}{7} = \frac{-3x+8}{7}.$$

   Finally, $\frac{x^2-1}{x+1}$ is simplified as

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

4. Calculating with exponents simplifies $(3a)^4 \cdot (3a)^7 \cdot a$ to

$$(3a)^4 \cdot (3a)^7 \cdot a = (3a)^{4+7}\frac{(3a)^1}{3} = \frac{3^{12}a^{12}}{3} = 3^{11}a^{12}$$

   and $\sqrt{x} \cdot x^2$ to

$$\sqrt{x} \cdot x^2 = x^{\frac{1}{2}}x^2 = x^{\frac{5}{2}} = \sqrt{x^5}.$$

5. A standard multiplicative rule is that

$$(x+1)(x-1) = x^2 - x + x - 1 = x^2 - 1$$

holds.

1.2 Equations

In general, an equation is a mathematical statement that asserts the equality of two expressions. It is written by placing the expressions on either side of an equals sign =, for example

$$x + 2 = 6$$

asserts that $x + 2$ is equal to 6.

Equations often express relationships between given quantities, the *knowns*, and quantities yet to be determined, the *unknowns*. Usually unknowns are denoted by letters at the end of the alphabet $(x, y, z, w, \ldots)$ while knowns are denoted by letters at the beginning $(a, b, c, \ldots)$. The process of expressing the unknowns in terms of the knowns is called solving the equation.

The example stated above shows an equation with a single unknown. A value of that unknown for which the equation is true is called a *solution* or *root* of the equation. In the example above, 4 is the solution.

If both sides of the equations are polynomials the equation is called an algebraic equation.

1.2.1 Linear equations

The simplest equations to solve are linear equations that have only one variable. They contain only constant numbers and a single variable without an exponent.

The central technique is add, subtract, multiply or divide both sides of the equation by the same number in order to isolate the variable on one side of the equation. Once the variable is isolated, the other side of the equation is the value of the variable.

In the general case a linear equation looks as follows

$$ax + b = c.$$

The solution is given by

$$x = \frac{c-b}{a}.$$

1.2.2 Quadratic equations

All quadratic equations can be expressed in the form

$$ax^2 + bx + c = 0,$$

where a is not zero (if it were so, the equation would be a linear equation). As $a \neq 0$ we may divide by a and rearrange the equation into the standard form

$$x^2 + px + q = 0.$$

Then the solution of the equation can be calculated by using the following formula:

$$x_{1,2} = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q}. \tag{1.1}$$

A quadratic equation with real coefficients ($a, b, c \in \mathbb{R}$) can have either zero, one or two distinct real roots, or two distinct complex roots. The discriminant determines the number and nature of the roots. The discriminant is the expression underneath the square root sign. If it is negative, the solutions are complex ($\in \mathbb{C}$) and no real root exists. If it is positive, the solutions are real ($\in \mathbb{R}$) and if it is 0, there is only one real solution (formally, the two real solutions coincide).

1.2.3 Exponential and logarithmic equations

An exponential equation is an equation of the form

$$a^x = b \qquad \text{for } a > 0,$$

which has the solution

$$x = \log_a b = \frac{\log b}{\log a}$$

when $b > 0$.

Elementary algebraic techniques are used to rewrite a given equation. As mentioned above this is used to isolate the unknown and thus calculate the solution.

For example, if

$$3 \cdot 2^{x-1} + 1 = 10$$

then, by subtracting 1 from both sides, and then dividing both sides of the equation by 3 we obtain

$$2^{x-1} = 3.$$

Applying the natural logarithm (see section 2.6.6 for properties of the logarithmic functions) on both sides we get

$$(x - 1) \log 2 = \log 3$$

or

$$x = \frac{\log 3}{\log 2} + 1 \approx 2.5850.$$

A logarithmic equation is an equation of the form

$$\log_a x = b \qquad \text{for } a > 0,$$

which has the solution

$$x = a^b.$$

1.2.4 Radical equations

A radical equation is an equation of the form

$$x^{\frac{m}{n}} = a \qquad m, n \in \mathbb{N},\ n \neq 0.$$

For these types of equations two cases have to be considered separately.

- If the exponent m is *odd* the equation has the solution

$$x = \sqrt[m]{a^n} = \left(\sqrt[m]{a}\right)^n.$$

- If the exponent m is *even* and $a \geq 0$ the equation has the two solutions

$$x = \pm\sqrt[m]{a^n} = \pm\left(\sqrt[m]{a}\right)^n.$$

Examples for solving equations

1. The equation $4x = 8$ can be solved by dividing both sides by 4. The solution to the equation is $x = 2$.

2. The equation $x^2 + 10x - 11 = 0$ can be solved directly using the solution given in equation 1.1. As $p = 10$ and $q = -11$ we get:

$$\begin{aligned} x_{1,2} &= -5 \pm \sqrt{25 - (-11)} = -5 \pm \sqrt{36} \\ x_1 &= -5 + 6 = 1 \\ x_2 &= -5 - 6 = -11. \end{aligned}$$

3. Let us consider[4]

$$4\log_{10}(x - 3) - 2 = 6.$$

By adding 2 to both sides, followed by dividing both sides of the equation by 4 we get

$$\log_{10}(x - 3) = 2.$$

[4]For the definition of the logarithm to the base 10, denoted as $\log_{10}(\cdot)$ see section 2.6.6.

By applying the function

$$f(x) = 10^x$$

on both sides of the equation we obtain $x - 3 = 10^2 = 100$ and $x = 103$ respectively.

4. The equation

$$(x+5)^{\frac{2}{3}} = 4$$

has an even exponent $m = 2$ and positive $a = 4$. Therefore there exist two solutions. We get

$$x + 5 = \pm\left(\sqrt{4}\right)^3 = \pm 8$$
$$x_{1,2} = -5 \pm 8$$

and $x \in \{3, -13\}$.

2 Functions

The first formalization of functions was created in the 17^{th} century by Gottfried Wilhelm Leibniz. He coined the term *function* to indicate the dependence of one quantity on another.

A function can be described as a ≪machine≫, a ≪black box≫ or a ≪rule≫ that, for each input, returns a corresponding output. Figure 2.1 shows a schematic picture of the relationships of a function $f(x)$.

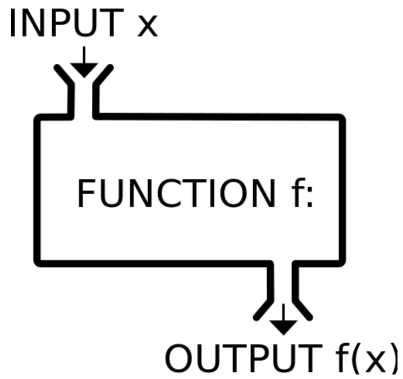

Figure 2.1: A function f takes an input x and returns an output $f(x)$.

2.1 Definition of functions

In formal terms a function f is given by

$$\begin{aligned} f : D &\to R \\ x &\mapsto f(x). \end{aligned}$$

To distinguish the different sets involved in the definition of a function the following terms are used:

Domain The input x to a function is called the *argument* of the function. The set of all possible inputs D is called the *domain*.

Range The output $f(x)$ of a function is called the *value*. The set R of all possible outputs is called the *range*.

Image The image of a subset A of the domain *under* the function can be defined as well. For $A \subseteq D$ the image of A under f is denoted $f(A) \subseteq R$ and is a subset of the functions range.

The image of the whole domain D under f is a subset of the range R as well. We have $f(D) \subseteq R$. $f(D)$ is sometimes simply referred to as the image of f.

Preimage The preimage of a subset B of the range of the function ($B \subseteq R$) under the function f is the set $A \subseteq D$ of all elements x for which $f(x)$ lies within B, formally $A = f^{-1}(B)$ or $A = \{x \in A : f(x) \in B\}$.

Graph The set $\{(x, f(x)) : x \in D\}$ of all paired inputs and outputs is called the *graph* of the function. A function is fully defined by its graph. In case of a real-valued function f of real values, the graph of f lies in the xy-plane.

2.2 Well defined functions

Generally spoken, a function is *well defined* if there is a *unique* $f(x)$ for every x in the domain of f. This means that a function is well defined if and only if all arguments x have only one possible function value $f(x)$ (or none).

For functions where the domain and the range consist of real numbers, there is a general rule to determine if the function is *well defined*. This rule is called the *vertical line test*.

> A curve in the xy-plane is the graph of some function f with $\mathbb{R}$ as domain and range, if and only if no vertical line intersects the curve more than once.

Examples for different functions

1. The function of a simple parabola in the plane is defined by

$$\begin{aligned} f : \mathbb{R} &\to \mathbb{R} \\ x &\mapsto x^2 \end{aligned}$$

   and shown in figure 2.2.

2. The function which aligns every integer number to the constant number 4 is defined by

$$\begin{aligned} f : \mathbb{Z} &\to \mathbb{Z} \\ x &\mapsto 4 \end{aligned}$$

   and shown in figure 2.3.

3. A function which aligns every number to its non-negative number is defined by

$$\begin{aligned} f : \mathbb{R} &\to \mathbb{R} \\ x &\mapsto |x| \end{aligned}$$

   and shown in figure 2.4.

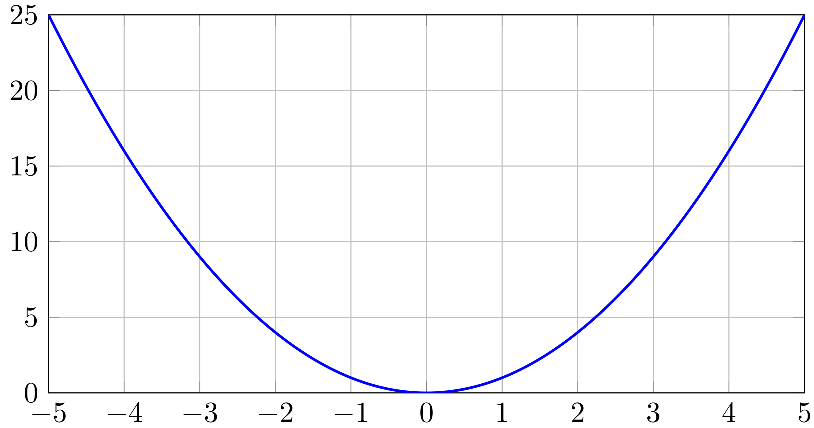

Figure 2.2: The polynomial function $f(x) = x^2$ from -5 to 5.

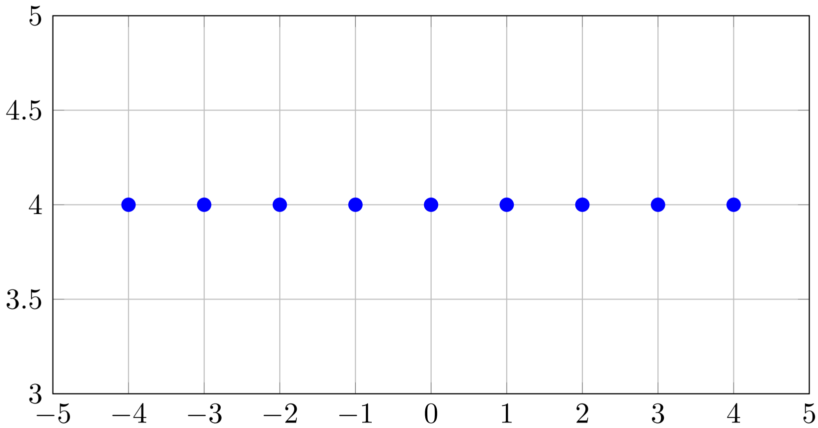

Figure 2.3: The constant function $f(x) = 4$ from -5 to 5.

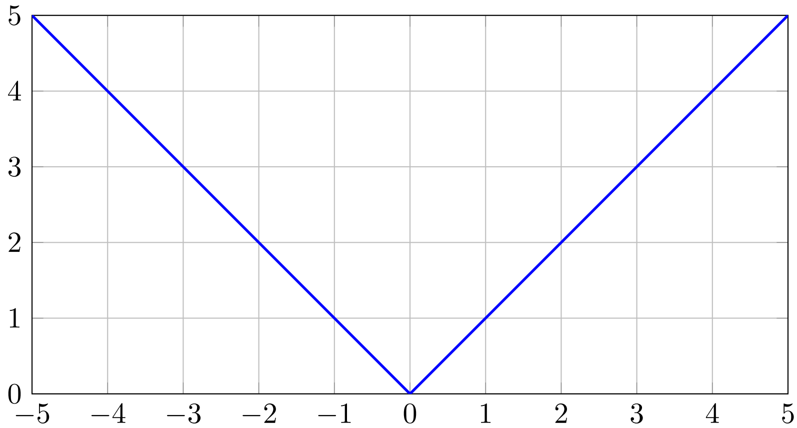

Figure 2.4: The absolute value function $f(x) = |x|$ from -5 to 5.

4. Furthermore, a function can be defined piecewise. That is a function which is defined by multiple subfunctions, each subfunction applying to a certain interval of the main functions domain (a

sub-domain):

$$f : \mathbb{R} \to \mathbb{R}$$
$$x \mapsto \begin{cases} 0, & x \leq -1 \\ +\sqrt{1-x^2}, & -1 < x < 1 \\ x, & x \geq 1. \end{cases}$$

Figure 2.5 shows the graph of the function.

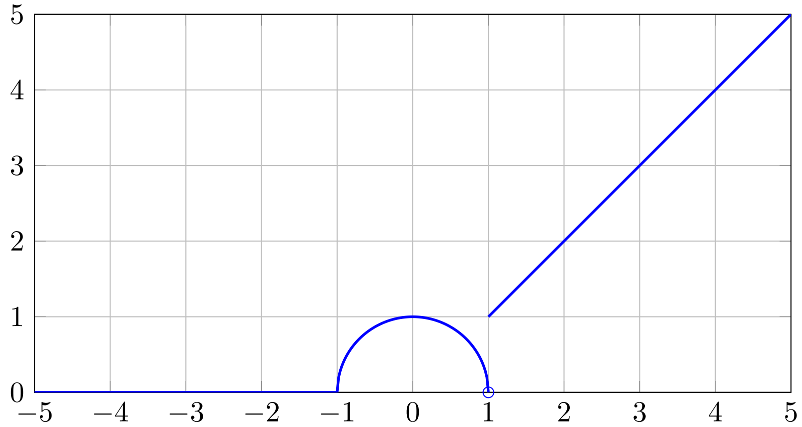

Figure 2.5: A piecewise defined function from -5 to 5 with a jump at $x = 1$.

There are many ways to describe or represent a function. Some functions may be described by a formula or an algorithm that defines how to compute the output for a given input. A function can also be presented graphically. Sometimes functions are given by a table that gives the outputs for selected inputs. Furthermore a function can be described through its relationship with other functions, for example as an inverse function or as a solution of a differential equation.

2.3 Properties of functions

It is not enough to say ≪f is a function≫ without specifying the domain and the range, unless these are known from the context. For example, a formula such as

$$f(x) = \sqrt{x^2 - 5x + 6}$$

is not a properly defined function on its own. However, it is standard to take the largest possible subset of $\mathbb{R}$ as the domain (in this case $x \leq 2$ or $x \geq 3$) and $\mathbb{R}$ as range.

2.3.1 Equations of functions

Well defined functions, which lie in the xy-plane can be represented by their equations. For a function $f(x)$ the equation can be obtained by replacing $f(x)$ with y.

For example, the function

$$\begin{aligned} f : \mathbb{R} &\to \mathbb{R} \\ x &\mapsto x^2 - 4x + 3 \end{aligned}$$

lies in the xy-plane. The equation for this function is $y = x^2 - 4x + 3$.

2.3.2 Monotone functions

A function whose graph is always rising as it is traversed from left to right is said to be an *increasing function.*
Definition Let x_1 and x_2 be arbitrary points in the domain of a function f. Then f is *strictly increasing* if

$$f(x_1) < f(x_2) \quad \text{whenever } x_1 < x_2.$$

A function whose graph is always falling as it is traversed from left to right is said to be a *decreasing* function.
Definition Again, let x_1 and x_2 be arbitrary points in the domain of a function f. Then f is *strictly decreasing* if

$$f(x_1) > f(x_2) \quad \text{whenever } x_1 < x_2.$$

Note that strictly increasing and strictly decreasing functions are always injective and therefore invertible (see section 2.5).

2.3.3 Continuity

Another important property of functions is the property of *continuity.* A continuous function is a function for which «small» changes in the input result in «small» changes in the output. Otherwise, a function is said to be *discontinuous.*

If the function is a real-valued function of real values and therefore its graph lies in the xy-plane, then the continuity of the function can be recognized in the graph. The function is continuous if, roughly speaking, the graph is a single unbroken curve with no *gaps* or *jumps.*

There are several ways to make this intuition mathematically rigorous. For example:

> The function $f : D \to R$ is *continuous* at $c \in D$, if and only if for every $\varepsilon > 0$ there is a $\delta > 0$ such that
>
> $$\text{for all } x \in D \text{ for which it holds that } |x - c| < \delta \Longrightarrow |f(x) - f(c)| < \varepsilon.$$
>
> More intuitively, we can say that if we want to get all the $f(x)$ values to stay in some small neighborhood around $f(c)$, we simply need to choose a small enough neighborhood for the x values around c. If that is possible, no matter of how small the $f(c)$-neighborhood is, then f is continuous at c.

A function f is said to be *continuous* if it is continuous at all points $c \in D$.

Examples of continuous and discontinuous functions

1. All polynomials are continuous, for example $f(x) = x^2 - 1$.

2. The absolute value function $f(x) = |x|$ is continuous.

3. The *sign*-function
$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$
is not continuous, as it jumps around $x = 0$.

2.4 Operations with functions

In analogy to arithmetics, it is possible to define addition, subtraction, multiplication and division of functions. Another important operation on functions is composition. When composing functions, the output from one function becomes the input to another function.

2.4.1 Arithmetic operations

For this section we consider the two function $f : D \to R$ and $g : D' \to R'$.

Addition of functions The addition of two functions f and g is defined as

$$(f + g)(x) = f(x) + g(x).$$

The domain of the new function $f + g$ is the intersection of the domains of f and g. Formally, we get $D \cap D'$ as the domain of the function $f + g$.

Subtraction of functions Similarly, the subtraction of two functions f and g is defined as

$$(f - g)(x) = f(x) - g(x).$$

Again, the domain of the new function $f - g$ is the intersection of the domains of f and g. Formally, we get $D \cap D'$ as the domain of the function $f - g$.

Multiplication of functions Additionally, the multiplication of two functions f and g is defined as

$$(f \cdot g)(x) = f(x) \cdot g(x),$$

where the domain of the new function $f \cdot g$ is the intersection of the domains of f and g. Again, we get $D \cap D'$ as the domain of the function $f \cdot g$.

Division of functions Finally, the division of two functions f and g is defined as

$$\left(\frac{f}{g}\right)(x) = \frac{f(x)}{g(x)}.$$

In this case the domain of the new function f/g is the intersection of the domains of f and g excluding the points where $g(x) = 0$ (to avoid division by zero). Note that the points x where $g(x) = 0$ can be written as the preimage of the set containing 0 under the function g. We get $(D \cap D') \setminus g^{-1}(\{0\})$ as the new domain.

2.4.2 Composition of functions

Now we consider a new operation on functions, called *composition*. This operation has no direct analogy in arithmetics. Informally stated, the operation of composition is performed by substituting some function for the variable of another function.

Definition Given functions $f : D \to R$ and $g : D' \to R'$, the *composition* of f and g, denoted by $f \circ g$, is the function defined by

$$(f \circ g)(x) = f(g(x)).$$

The domain of $f \circ g$ is defined as all x in the domain of g for which $g(x)$ is in the domain of f. If we denote the new domain with E, this means $E = \{x \in D' : g(x) \in D\}$.

Examples

1. Let $f(x) = x^2 - 3x + 4$ and $g(x) = x^2 - 1$. Then we have

   a. $(f + g)(x) = x^2 - 3x + 4 + x^2 - 1 = 2x^2 - 3x + 3$

   b. $(f - x)(x) = x^2 - 3x + 4 - (x^2 - 1) = -3x + 5$

   c. $(f \cdot g)(x) = (x^2 - 3x + 4)(x^2 - 1) = x^4 - 3x^3 + 3x^2 + 3x - 4$

   d. $\left(\frac{f}{g}\right)(x) = \frac{x^2-3x+4}{x^2-1}$, which is defined for $x \in \mathbb{R} \setminus \{-1, 1\}$.

2. Let $f(x) = x^3$ and $g(x) = x^2 - 4$. Then we have

   a. $(f + g)(x) = x^3 + x^2 - 4$

   b. $(f - g)(x) = x^3 - x^2 + 4$

   c. $(f \cdot g)(x) = x^5 - 4x^3$

   d. $\left(\frac{f}{g}\right)(x) = \frac{x^3}{x^2-4}$, which is defined for $x \in \mathbb{R} \setminus \{-2, 2\}$.

3. Let $f(x) = x^2 - 3x + 4$ and $g(x) = x^2 - 1$. Lets calculate $f \circ g(x)$ and $g \circ f(x)$.

a. For $f \circ g(x)$ we get

$$f(g(x)) = (x^2 - 1)^2 - 3(x^2 - 1) + 4 = x^4 - 5x^2 + 8.$$

b. For $g \circ f(x)$ we get

$$g(f(x)) = (x^2 - 3x + 4)^2 - 1 = x^4 - 6x^3 + 17x^2 - 24x + 15.$$

4. Let $f(x) = \frac{1}{x^2}$ and $g(x) = x^2 + 4$. Lets calculate $f \circ g(x)$ and $g \circ f(x)$.

a. For $f \circ g(x)$ we get

$$f(g(x)) = \frac{1}{(x^2 + 4)^2}.$$

b. For $g \circ f(x)$ we get

$$g(f(x)) = \left(\frac{1}{x^2}\right)^2 + 4 = \frac{4x^4 + 1}{x^4}.$$

2.5 Inverse functions

In mathematics, an *inverse function* is a function that undoes another function. Two functions f and g are said to be inverse to each other, if an input x into the function f produces an output y and plugging this y into the function g produces the output x, and vice versa.

Example
Suppose g is the inverse function of f. Then $f(x) = y$ and $g(y) = x$. Furthermore $f(g(x)) = x$, meaning $g(x)$ composed with $f(x)$ leaves x unchanged.

A function f that has an inverse is called *invertible*. The inverse function is unique for every f and is denoted by f^{-1}.

Definition and determination of the inverse function

Let f be a function whose domain is the set D and whose range is the set R. Then f is *invertible* if there exists a function g with domain R and range D with the property:

$$f(x) = y \text{ if and only if } g(y) = x.$$

If f is invertible, the function g is unique and is called the *inverse* of f, denoted by f^{-1}.

Not all functions have an inverse. For this rule to be applicable, each element $y \in R$ must correspond to no more than one $x \in D$. A function f with this property is called injective.

The most powerful approach to find a formula for f^{-1} , if it exists, is to solve the equation $x = f(y)$ for y.

Graph of the inverse function

If f and f^{-1} are inverses, then the graph of the function $y = f^{-1}(x)$ is the same as the graph of the equation $x = f(y)$. Thus the graph of f^{-1} can be obtained from the graph of f by switching the positions of the x and y axes. This is equivalent to *reflecting* the graph across the line $y = x$.

In figure 2.6 the graphs of the functions $f(x)$ and $f^{-1}(x)$ of example 5 below are shown.

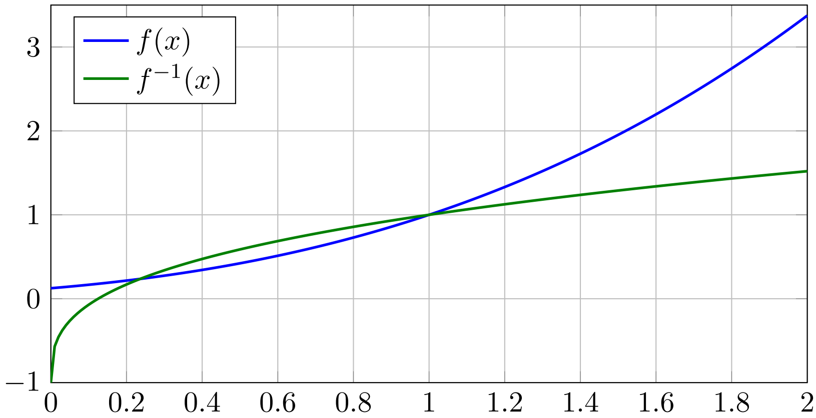

Figure 2.6: $f(x) = \left(\frac{x+1}{2}\right)^3$ in blue and its inverse $f^{-1}(x) = 2\sqrt[3]{x} - 1$ in green from 0 to 2.

Examples

1. The function $f(x) = x^2$ is only invertible if the domain is restricted to positive numbers (or alternative restricted to negative numbers). If so, the inverse function is $f^{-1}(x) = \sqrt{x}$.
2. For a function $f(x) = x - a$, where a is a constant, the inverse function is $f^{-1}(x) = x + a$.
3. For a function $f(x) = mx$, where m is a constant, the inverse function is $f^{-1}(x) = \frac{x}{m}$. Thereby $m \neq 0$ must hold.
4. For a function $f(x) = \frac{1}{x}$ the inverse function is $f^{-1}(x) = \frac{1}{x}$, where $x \neq 0$ must hold.
5. f is the function
$$f(x) = \left(\frac{x}{2} + \frac{1}{2}\right)^3.$$
To find the inverse we must solve the equation $x = \left(\frac{y}{2} + \frac{1}{2}\right)^3$ for y:
$$\begin{aligned} x &= \left(\frac{y}{2} + \frac{1}{2}\right)^3 \\ \sqrt[3]{x} &= \frac{y+1}{2} \\ 2\sqrt[3]{x} - 1 &= y \end{aligned}$$

Thus the inverse function f^{-1} is given by the formula

$$f^{-1}(x) = 2\sqrt[3]{x} - 1.$$

2.6 Special functions

There exists an overwhelming number of different functions. Nevertheless there are several classes of functions, which should be considered separately.

2.6.1 Polynomial functions

A *polynomial function* is a function that can be defined by evaluating a polynomial. A function f of one argument is called a polynomial function if it satisfies

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_2 x^2 + a_1 x + a_0 = \sum_{i=0}^{n} a_i x^i \tag{2.1}$$

for all arguments x, where n is a non-negative integer and $a_0, a_1, a_2, \ldots, a_n$ are constant coefficients. A polynomial function is either zero, or can be written as the sum of one or more non-zero terms as denoted in equation 2.1. The number of terms is always finite.

The exponent of a variable in a term is called the *degree* of that variable in that term. The degree of a polynomial is the largest degree of any term. Since $x = x^1$, the degree of a variable without exponent is one. A term with no variable is called a constant term, or just a constant. The degree of a constant term is 0, since $x^0 = 1$.

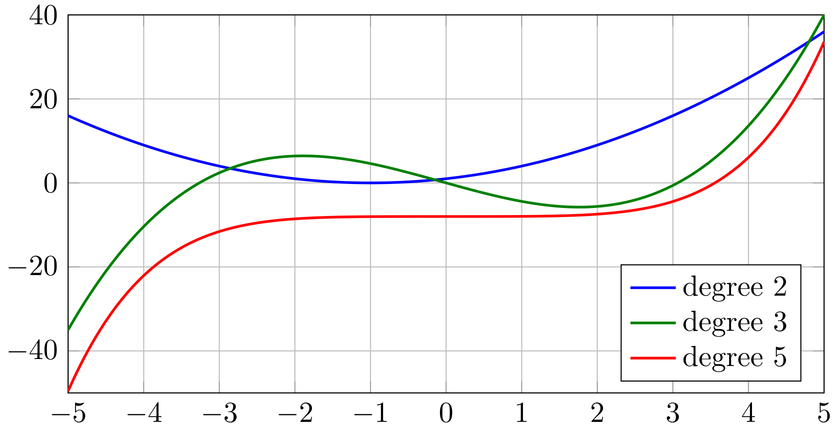

Figure 2.7: Three polynomial functions of degree $2, 3$ and 5 from -5 to 5.

Polynomial functions are widely used in different applications in physics, economics, calculus and numerical analysis. Furthermore they can be used to approximate other functions. The application of polynomial functions in a wide range of topics is due to some elementary *properties of polynomials*:

1. A *sum* of polynomials is a polynomial. The degree of the resulting polynomial equals the higher degree of the two added polynomials. For example $(x^2 + 3x + 1) + (x - 1) = x^2 + 4x$.

2. A *product* of polynomials is a polynomial. The degree of the resulting polynomial equals the sum of the degrees of the two multiplied polynomials. For example $(x^2 + 3x + 1)(x - 1) = x^3 + 2x^2 - 2x - 1$.

3. A *composition* of two polynomials is a polynomial. It is obtained by substituting a variable of the first polynomial by the second polynomial. The degree of the resulting polynomial equals the product of the degrees of the two polynomials of the composition.
   For example, for $f(x) = x^2 + 3x + 1$ and $g(x) = x - 1$ the composition $f \circ g$ results in
   $$f \circ g(x) = x^2 + x - 1.$$
   Note that the composition $g \circ f$ results in a different polynomial, namely
   $$g \circ f(x) = x^2 + 3x.$$

4. All polynomials are continuous functions.

5. The *derivative*[5] of the polynomial
   $$a_n x^n + a_{n-1}x^{n-1} + \ldots + a_2 x^2 + a_1 x + a_0$$
   is the polynomial
   $$n a_n x^{n-1} + (n-1)a_{n-1}x^{n-2} + \ldots + 2a_2 x + a_1.$$
   Furthermore polynomials are *smooth* functions, which means that all orders of derivatives of polynomials exist. The resulting derivatives of any order are always polynomials as well.

6. Any polynomial function $P(x)$ of n^{th} order has n roots (roots are values for x where $P(x) = 0$). These roots may be complex numbers ($\in \mathbb{C}$).

   Furthermore every polynomial function can be written as product of simpler polynomials, using its roots. Let $(x_i)_{i=1,\ldots,n}$ denote the roots of the polynomial function $P(x)$ of degree n. Then the following holds
   $$\begin{aligned} P(x) &= a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \\ &= a_n (x - x_1) \cdot \ldots \cdot (x - x_n). \end{aligned}$$

Examples of different polynomials

1. A constant polynomial would be $f(x) = 3$.

2. Linear polynomials look like $f(x) = 2x - \frac{1}{2}$ and $f(x) = \frac{3x}{2} - 2$.

3. A polynomial of degree 2 is $f(x) = x^2 + 3x + 20$.

[5] For more information on derivatives, see chapter 3.

4. A polynomial of degree 3 is $f(x) = \frac{7x^3}{12} + x^2 - 5x$.

5. A polynomial of degree 5 looks like $f(x) = \frac{x^5 - 3x^3}{100} - 8$.

6. In figure 2.7 the graphs of the polynomials given in examples 3-5 are shown.

2.6.2 Power functions

Power functions are elementary mathematical functions of the form

$$f : x \mapsto a \cdot x^r \qquad a, r \in \mathbb{R}.$$

The possible *domain*[6] depends on the exponent. If roots of negative numbers are not allowed (which is mostly the case, except if the range includes complex numbers $\mathbb{C}$), then the maximum size of the domain D of the power function depends on the exponent $r \in \mathbb{R}$. Table 2.1 lists the possible domains, depending on the nature of the exponent r.

	$r > 0$	$r < 0$
$r \in \mathbb{Z}$	$D \subseteq \mathbb{R}$	$D \subseteq \mathbb{R} \setminus \{0\}$
$r \notin \mathbb{Z}$	$D \subseteq \mathbb{R}_0^+$	$D \subseteq \mathbb{R}^+$

Table 2.1: The possible domains D for power functions, depending on $r \in \mathbb{R}$.

For the maximum *range* of a power function, the sign of a has to be considered as well. If $r \in \mathbb{Z}$, the range differs depending on whether r is an odd or even number. Table 2.2 lists the possibilities for the maximum range of the power function, depending on the nature of the coefficient $a \in \mathbb{R}$ and the exponent $r \in \mathbb{R}$ of the power function.

	$r > 0$		$r < 0$	
	r even or $r \notin \mathbb{Z}$	r odd	r even or $r \notin \mathbb{Z}$	r odd
$a > 0$	$R \subseteq \mathbb{R}_0^+$	$R \subseteq \mathbb{R}$	$R \subseteq \mathbb{R}^+$	$R \subseteq \mathbb{R} \setminus \{0\}$
$a < 0$	$R \subseteq \mathbb{R}_0^-$	$R \subseteq \mathbb{R}$	$R \subseteq \mathbb{R}^-$	$R \subseteq \mathbb{R} \setminus \{0\}$

Table 2.2: The possible maximum ranges R for power functions, depending on $a, r \in \mathbb{R}$.

Examples of power functions

1. $f(x) = x^4$

2. $f(x) = \sqrt[7]{x^3} = x^{\frac{3}{7}}$

3. $f(x) = \frac{5}{\sqrt{x}} = 5x^{-\frac{1}{2}}$

4. $f(x) = \frac{1}{7} \cdot \frac{1}{x^5} = \frac{1}{7}x^{-5}$

[6]See section 2.1.

2.6.3 Rational functions

A *rational function* is a function which can be written as ratio of two polynomial functions[7]. Neither the coefficients of the polynomials nor the values taken by the function are necessarily rational numbers.

In case of one variable x, a function is called a rational function if and only if it can be written in the form

$$f(x) = \frac{P(x)}{Q(x)}$$

where $P(x)$ and $Q(x)$ are polynomial functions and Q is not the zero polynomial. The *domain* of f is the set of all points x for which the denominator $Q(x)$ is not zero.

Examples for different rational functions

1. An example for a rational function of degree 2 is

$$f(x) = \frac{x^2 - 3x - 2}{x^2 - 4}.$$

The graph of this function is shown in figure 2.8. It can be seen that the function has a horizontal asymptote at $y = 1$ and two vertical asymptotes at $x = \pm 2$. For $x = \pm 2$ the function is not defined.

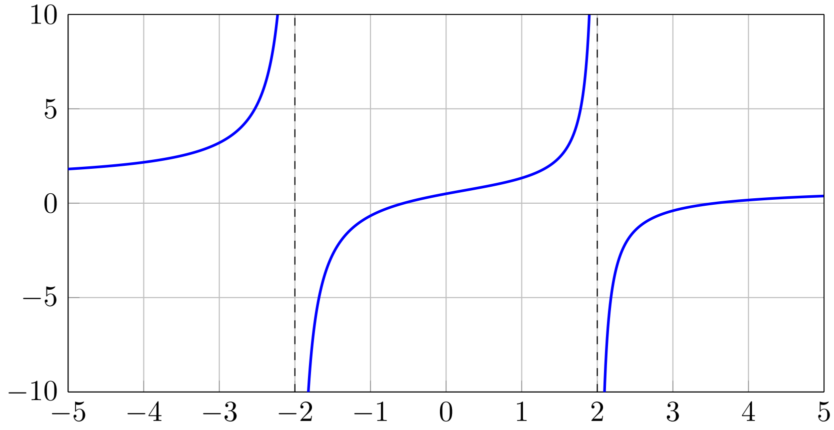

Figure 2.8: $f(x) = \frac{x^2-3x-2}{x^2-4}$ from -5 to 5 with vertical asymptotes at $x \pm 2$.

2. A rational function of degree 3 is

$$f(x) = \frac{x^3 - 2x}{2(x^2 - 5)}.$$

The graph of this function is shown in figure 2.9. It can be seen that the function has two vertical asymptotes at $x = \pm\sqrt{5}$. Furthermore the linear function $g(x) = \frac{x}{2}$ is an asymptote. For $x = \pm\sqrt{5}$ the function is not defined.

[7] See section 2.6.1.

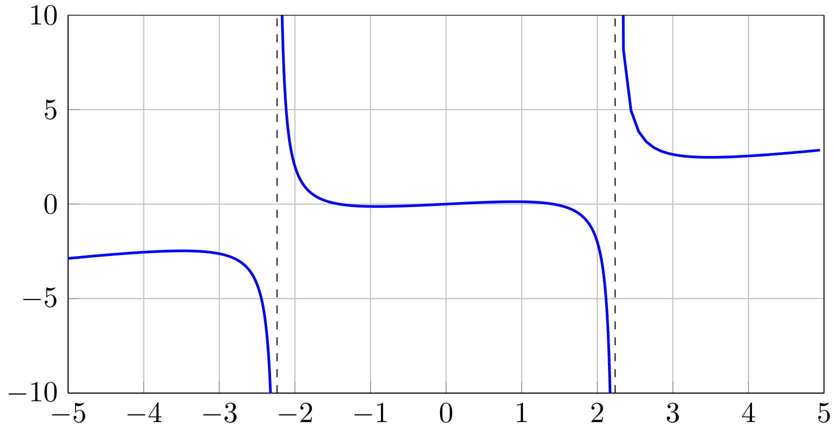

Figure 2.9: $f(x) = \frac{x^3-2x}{2(x^2-5)}$ from -5 to 5 with vertical asymptotes at $x \pm \sqrt{2}$.

2.6.4 Trigonometric functions

Trigonometric functions are functions of an angle. They are used to relate angles of a triangle to the length of the sides of that triangle. The most familiar trigonometric functions are *sine, cosine* and *tangent*.

The definition of trigonometric functions works best in the context of the standard unit circle with radius 1. In such a circle a triangle is formed by a ray originating at the origin and making some angle θ with the x-axis. The *sine* of the angle θ gives the length of the y-component (rise) of the triangle, the *cosine* gives the length of the x-component (run) and the *tangent* function gives the slope (y-component divided by the x-component). Thus it holds that $\tan(\theta) = \frac{\sin(\theta)}{\cos(\theta)}$. These relationships are shown in figure 2.10.

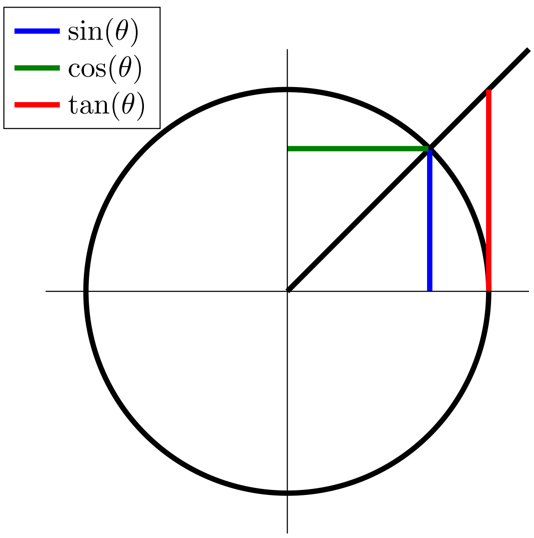

Figure 2.10: The unit circle, a ray ($\theta = 45°$) and the functions $\sin(\theta)$, $\cos(\theta)$ and $\tan(\theta)$.

Trigonometric functions have a wide range of applications including computing unknown lengths and angles in triangles or modeling periodic phenomena. In figure 2.11 the sine function $\sin(\theta)$ within the model of the unit circle and its function values are presented. It is easy to see that sine is a periodic function oscillating between -1 and 1.

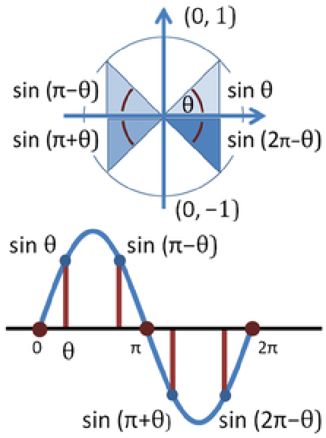

Figure 2.11: The sine function $\sin(\theta)$, the unit circle and the graph of the function $\sin(\theta)$.

For angles greater than 2π or less than -2π the circle is rotated more than once and therefore sine and cosine are periodic functions with periodicity 2π. To be more exact, the following holds for any angle θ and integer k:

$$\sin\theta = \sin(\theta + 2\pi k)$$
$$\cos\theta = \cos(\theta + 2\pi k).$$

The trigonometric functions satisfy a range of identities which make them very powerful in algebraic calculations. Here is a list of the most important identities.

- Most frequently used is the *Pythagorean identity*, which states that for any angle, the square of the sine plus the square of the cosine is 1. In symbolic form, the Pythagorean identity is written as
$$\sin^2 x + \cos^2 x = 1.$$

- Other key relationships are the *sum and difference formulas*, which give the sine and cosine of the sum and difference of two angles in terms of sines and cosines of the angles themselves. For the sine the formulas state
$$\sin(x + y) = \sin x \cos y + \cos x \sin y$$
$$\sin(x - y) = \sin x \cos y - \cos x \sin y.$$

  Similarly for the cosine the following holds
$$\cos(x + y) = \cos x \cos y - \sin x \sin y$$
$$\cos(x - y) = \cos x \cos y + \sin x \sin y.$$

- When the two angles are equal, the sum formulas reduce to simpler equations known as the *double-angle formulae.*

$$\sin 2x = 2\sin x\cos x$$
$$\cos 2x = \cos^2 x - \sin^2 x = 2\cos^2 x - 1$$

2.6.5 Exponential function e^x

The exponential function is the function e^x, where e is the number

$$2.71828\,18284\,59045\,23536\,02874\,71352\,66249\,77572\,47093\,69995\ldots$$

such that the function e^x is its own derivative. The exponential function can be used to model a relationship in which a constant change in the independent variable gives the same proportional change (e.g. percentage increase or decrease) in the dependent variable.

The function is often written as $\exp(x)$, especially when it is impractical to write the independent variable as superscript. The graph of the exponential function for real numbers is illustrated in figure 2.12.

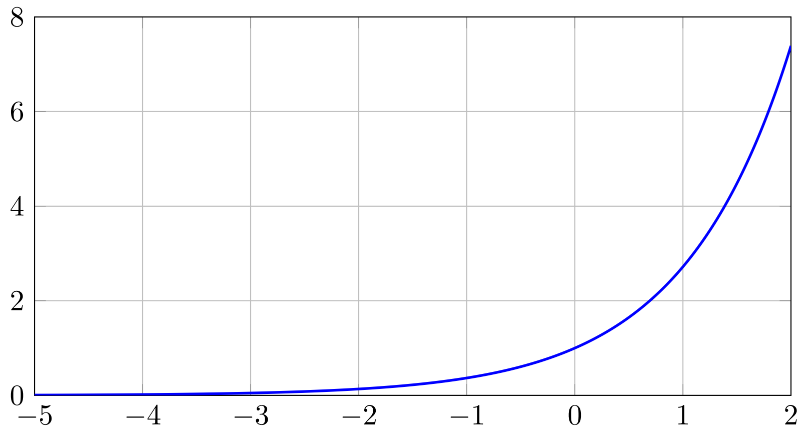

Figure 2.12: The exponential function $\exp(x)$ from -5 to 2.

The exponential function arises whenever a quantity grows or decays at a rate proportional to its current value. One such situation is *continuously compounded interest.* Leonhard Euler has proven in the 18^{th} century that the number

$$\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n$$

actually exists. It is now known as e.

If a principal amount of 1 earns interest at an annual rate of x compounded monthly, then the interest earned each month is $\frac{x}{12}$ times the current value, so each month the total value is multiplied by $(1+\frac{x}{12})$, and the value at the end of the year is

$$\left(1+\frac{x}{12}\right)^{12}.$$

If instead interest is compounded daily this becomes

$$\left(1 + \frac{x}{365}\right)^{365}.$$

Letting the number of time intervals per year grow without bound leads to the limit definition of the exponential function

$$e^x = \exp(x) = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n,$$

first given by Euler. Another very important application of the exponential function stands in the center of mathematical finance which is the core of our study program, *Quantitative Asset and Risk Management (ARIMA)*. This is the calculation of continuously compounded rates of return. For a brief introduction to return calculation, see chapter 4.

The importance of the exponential function in mathematics and sciences stems mainly from the properties of its derivative (see section 3). In particular,

$$\frac{\mathrm{d}}{\mathrm{d}x} e^x = e^x.$$

That is, e^x is its own derivative.

2.6.6 Logarithmic functions

The logarithm of a number is the *exponent* by which another fixed value, the *base*, has to be raised to produce that number. For example, the logarithm of 100 to base 10 is 2, because $100 = 10^2$.

More generally, if $x = b^y$, then y is the logarithm of x to base b, and is written

$$y = \log_b(x).$$

For example $\log_{10}(100) = 2$.

Logarithms were introduced by John Napier in the early 17^{th} century as a means to simplify calculations. They were rapidly adopted by scientists, engineers and others to perform computations more easily using slide rules and logarithm tables. These devices rely on the fact that the logarithm of a product is the sum of the logarithms of the factors:

$$\log_b(xy) = \log_b(x) + \log_b(y).$$

There are three distinct bases for which the logarithm is used often:

- The logarithm to base $b = 10$ is called the *common logarithm* and has many applications in science and engineering. In this text the notation $\log_{10}(x)$ for the logarithm to base 10 will be used. For example $\log_{10}(10000) = \log_{10}(10^4) = 4$.

- The logarithm where the base is the Euler constant e is called the *natural logarithm*. It is mostly used in mathematics and for the definition of continuously compounded rates of return in mathematical finance (see section 4.2). In this text the notation $\log(x)$ for the natural logarithm will be used. For example $\log(e^3) = 3$.

- The logarithm to base $b = 2$ is called the *binary logarithm* and is used primarily in computer science. In this text, we will not use the binary logarithm. In literature the notation $\mathrm{lb}(x)$ is most commonly used. For example $\mathrm{lb}(8) = 3$.

The graph of the natural logarithm function for real numbers is shown in figure 2.13.

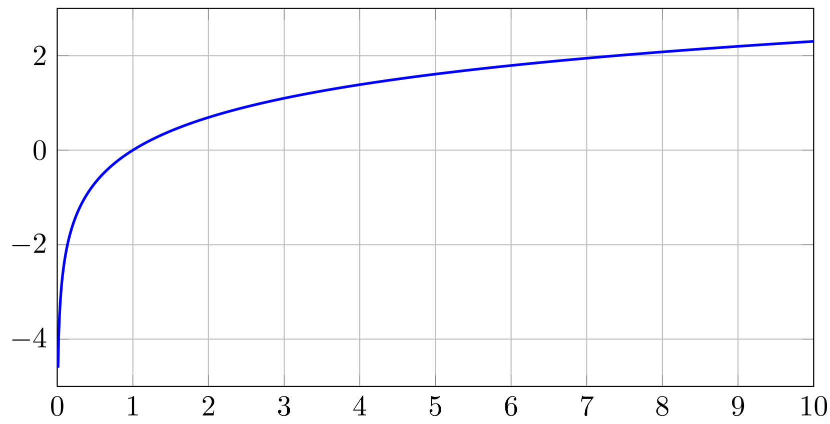

Figure 2.13: The natural logarithm function $\log(x)$ for $x \in]0, 10]$.

It has been shown above that the logarithm of a product is the sum of the logarithms of the numbers being multiplied. Analogical the logarithm of the ratio of two numbers is the difference of the logarithms. The following listing shows all properties of logarithms regarding to arithmetic operations.

product The logarithm of the product is the sum of the logarithms. In formula:

$$\log_b(xy) = \log_b(x) + \log_b(y)$$

quotient The logarithm of the ratio is the difference of the logarithms. In formula:

$$\log_b\left(\frac{x}{y}\right) = \log_b(x) - \log_b(y)$$

power The logarithm of x to the power of p is p times the logarithm of x. In formula:

$$\log_b(x^p) = p \cdot \log_b(x)$$

root The logarithm of the p-root of x is the logarithm of x divided by p. In formula:

$$\log_b \sqrt[p]{x} = \frac{\log_b(x)}{p}$$

Finally the base of the logarithm can be changed. The logarithm $\log_b(x)$ can be computed from the logarithms of x and b with respect to an arbitrary base c using the formula

$$\log_b(x) = \frac{\log_c(x)}{\log_c(b)}.$$

3 Differentiation

In mathematics the derivative of a function f is a measure of how the function changes as its input changes. Loosely speaking, a derivative can be thought of as how much one quantity is changing in response to changes in some other quantity.

The derivative of a function at a chosen input value describes the best linear approximation to the function near that input value. For a real-valued function with a single variable, the derivative at a point equals the *slope* of the tangent line to the function at that point (see figure 3.1).

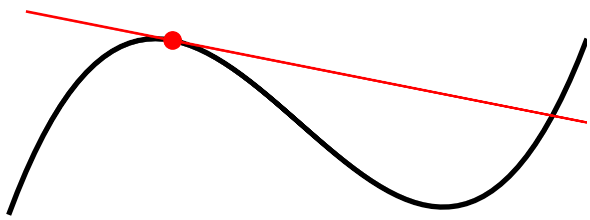

Figure 3.1: The graph of a function and its tangent line.

The process of finding a derivative is called *differentiation*. As stated above, differentiation is a method to compute the rate at which a dependent output y changes with respect to the change in the independent input x. This rate of change is called the *derivative* of y with respect to x.

The simplest case is when y is a linear function of x, meaning that the graph $y = f(x)$ in the xy-plane is a straight line. In this case,

$$y = f(x) = kx + d$$

for $k, d \in \mathbb{R}$ and the slope k is given by

$$k = \frac{\text{change in } y}{\text{change in } x} = \frac{\Delta y}{\Delta x},$$

where the symbol Δ (the uppercase form of the Greek letter Delta) is an abbreviation for «change in». This gives the exact value k for the slope of the straight line and subsequently for its derivative.

If the function f is not linear, the change in y divided by the change in x varies as x varies. Differentiation is a method to find an exact value for this rate of change at any given value of x. The idea is to compute the rate of change as the limit of the ratio of the differences

$$\frac{\Delta y}{\Delta x} = \frac{f(x+h) - f(x)}{x + h - x} = \frac{f(x+h) - f(x)}{h},$$

as Δx (or h respectively) becomes infinitely small.

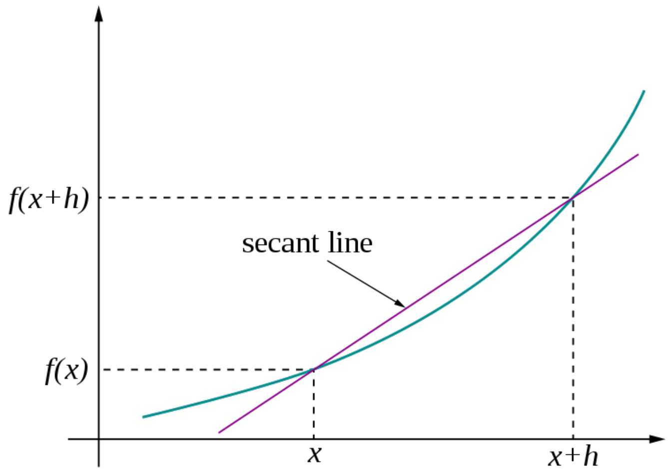

Figure 3.2: The graph of a function and a secant line to that function.

In figure 3.2 the graph of a function and its secant line are shown. As h becomes infinitely small, the secant line becomes the tangent line. Its slope is the derivative of the function at x.

In *Leibniz's notation*, such an infinitesimal change in x is denoted by dx, and the derivative of y with respect to x is written

$$\frac{dy}{dx},$$

suggesting the ratio of two infinitesimal quantities. The expression is read as ≪the derivative of y with respect to x≫ or ≪dy over dx≫.

Simultaneously with Leibniz, the English mathematician and physicist Newton has developed the concept of differentiation in the 17$^{\text{th}}$ century. He developed a slightly different notation where the difference quotient is defined as

$$\frac{\Delta f(x)}{\Delta x} = \frac{f(x+h) - f(x)}{h}.$$

The derivative is the value of the difference quotient as h becomes infinitesimally small.

3.1 Continuity and differentiability

For a function f to have a derivative at point a, it is necessary for the function f to be continuous at a (see section 2.3.3), but continuity alone is not sufficient. For example, the absolute value function $f(x) = |x|$ is continuous at $x = 0$, but is not differentiable there (see figure 3.3).

In the graph of the function in the xy-plane this can be seen as a ≪kink≫ or a ≪cusp≫ in the graph at $x = 0$. The derivative does not exist at $x = 0$ as the tangent line at this point is not unique.

Let f be a differentiable function, and let $f'(x)$ be its derivative. The derivative of $f'(x)$ - if it has one - is written $f''(x)$ and is called the *second derivative of f*. Similarly the derivative of a second

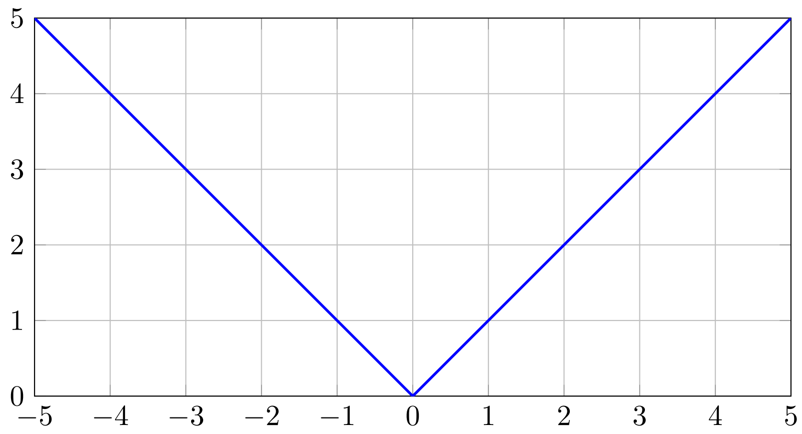

Figure 3.3: The function $f(x) = |x|$. It is not differentiable at $x = 0$.

derivative is written $f'''(x)$ and is called the third derivative of f. These repeated derivatives are called *higher-order derivatives*.

A function that has infinitely many derivatives is called *infinitely differentiable* or *smooth*. For example, every polynomial function is infinitely differentiable (see section 2.6.1).

3.2 Computing the derivative

In theory the derivative of a function can be computed from the definition by considering the difference quotient and computing its limit. In practice the derivatives are computed using *rules* for obtaining derivatives of complicated functions from simpler ones.

3.2.1 Differentiation is linear

For any functions $f(x)$ and $g(x)$ and any real numbers a, b and c, the derivative of the function $h(x) = a \cdot f(x) + b \cdot g(x) + c$ with respect to x is

$$h'(x) = a \cdot f'(x) + b \cdot g'(x).$$

Special cases include

- the constant multiple rule
$$(a \cdot f(x))' = a \cdot f'(x)$$

- the sum rule
$$(f(x) + g(x))' = f'(x) + g'(x)$$

- the subtraction rule
$$(f(x) - g(x))' = f'(x) - g'(x)$$

- the constant rule
$$(f(x) + c)' = f'(x)$$

3.2.2 The product rule

For the functions $f(x)$ and $g(x)$, the derivative of the function $h(x) = f(x) \cdot g(x)$ with respect to x is
$$h'(x) = f'(x) \cdot g(x) + f(x) \cdot g'(x).$$

3.2.3 The quotient rule

For the functions $f(x)$ and $g(x)$, the derivative of the function $h(x) = \frac{f(x)}{g(x)}$ with respect to x is

$$h'(x) = \left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x) \cdot g(x) - g'(x) \cdot f(x)}{g(x)^2}$$

whenever $g(x)$ is nonzero.

3.2.4 The chain rule

The derivative of the function of a function $h(x) = f(g(x))$ (or $h(x) = f \circ g(x)$ respectively, see section 2.4.2) with respect to x is
$$h'(x) = f'(g(x)) \cdot g'(x).$$

3.2.5 The elementary and the generalized power rule

If $f(x) = x^r$ and $r \in \mathbb{R} \setminus \{0\}$
$$f'(x) = rx^{r-1}.$$
If $r = 0$ then the function f is a constant function and its derivative equals 0.

The most general power rule is the functional power rule. For any functions $f(x)$ and $g(x)$,

$$\left(f(x)^{g(x)}\right)' = \left(e^{g(x) \cdot \log(f(x))}\right)' = f(x)^{g(x)} \cdot \left(f'(x)\frac{g(x)}{f(x)} + g'(x) \cdot \log(f(x))\right),$$

where both sides are well defined.

3.2.6 Derivatives of exponential and logarithmic functions

The exponential function is its own derivative (see section 2.6.5), therefore

$$(e^x)' = e^x.$$

For the logarithmic function the following holds

$$(\log_c x)' = \frac{1}{x \log c}, \qquad c > 0, c \neq 1, x \neq 0$$

and

$$(\log x)' = \frac{1}{x}, \qquad x \neq 0.$$

Furthermore

$$(\log |x|)' = \frac{1}{x} \quad \text{and} \quad (x^x)' = x^x(1 + \log x)$$

holds as well. By applying the chain rule it can be stated that

$$(\log f)' = \frac{f'}{f} \qquad \text{wherever } f \text{ is positive.}$$

3.2.7 Derivatives of trigonometric functions

The most important trigonometric functions were introduced in section 2.6.4. Their derivatives are

$$\begin{aligned}
(\sin x)' &= \cos x, \\
(\cos x)' &= -\sin x, \\
(\tan x)' &= \frac{1}{\cos^2 x} = 1 + \tan^2 x.
\end{aligned}$$

Examples

1. The derivative of

   $$f(x) = 4x^7$$

   is

   $$f'(x) = 4 \cdot 7x^6 = 28x^6.$$

2. The derivative of

   $$f(x) = 3x^8 + 2x + 1$$

   is

   $$f'(x) = 3 \cdot 8x^7 + 2x^0 = 24x^7 + 2.$$

3. The derivative of

$$f(x) = 2\sqrt{x} = 2x^{\frac{1}{2}}$$

is

$$f'(x) = 2 \cdot \frac{1}{2}x^{-\frac{1}{2}} = \frac{1}{\sqrt{x}}.$$

4. The derivative of

$$f(x) = (x+1)(x^2 - x + 1)$$

using the product rule is

$$\begin{aligned} f'(x) &= x^0 \cdot (x^2 - x + 1) + (x+1)(2x - 1x^0) \\ &= x^2 - x + 1 + (x+1)(2x-1) = 3x^2. \end{aligned}$$

5. The derivative of

$$f(x) = \frac{2x+4}{x^2+1}$$

using the quotient rule is

$$f'(x) = \frac{2(x^2+1) - 2x(2x+4)}{(x^2+1)^2} = -2\frac{x^2+4x-1}{(x^2+1)^2}.$$

6. The derivative of

$$f(x) = \sin\left(\frac{1}{x^2}\right)$$

using the chain rule is

$$f'(x) = \cos\left(\frac{1}{x^2}\right) \cdot (-2x^{-3}) = -\cos\left(\frac{1}{x^2}\right)\frac{2}{x^3}.$$

7. The derivative of

$$f(x) = \log 5x$$

is

$$f'(x) = \frac{5}{5x} = \frac{1}{x}.$$

8. The derivative of

$$f(x) = x^3 e^{\cos x}$$

using the product rule and the chain rule is

$$f'(x) = 3x^2 e^{\cos x} - x^3 \sin x e^{\cos x}.$$

4 Rate of return calculation

The rate of return expresses the profit, respectively loss, as a percentage of the original value of the asset. It measures by what percentage the original value changed.

4.1 Discrete rate of return

The *discrete rate of return* R on an asset (alternatively called simple rate of return or periodically compounded rate of return) is calculated as

$$R = \frac{S_{t_2} - S_{t_1}}{S_{t_1}} = \frac{S_{t_2}}{S_{t_1}} - 1,$$

where S_{t_1} and S_{t_2} are prices of an asset at different points in time t_1 and t_2, where $t_1 < t_2$. Hence R represents a T-period discrete rate of return, where $T = t_2 - t_1$.

In order to be able to compare rates of return that are calculated over different time periods, the rates of return are typically annualised. The annualised discrete rate of return R^{ann} is calculated from the T-period rate of return R as

$$R^{ann} = (1 + R)^{1/T} - 1.$$

A T-period rate of return R is calculated from an annualised rate of return R^{ann} as

$$R = (1 + R^{ann})^T - 1.$$

The value at t_2 of an investment in a specific asset at time t_1 is

$$V_{t_2} = V_{t_1} \cdot (1 + R) = V_{t_1} \cdot (1 + R^{ann})^T,$$

where V_{t_1} and V_{t_2} represent the value of the investment at time t_1 and t_2, respectively.

Examples

Example 4.1 Consider the price information for four stocks, given in table 4.1.

Assume we want to calculate the T-period discrete rate of return R, the annualized discrete rate of return R^{ann} and the value of an original investment of €10 000 in 2019 for all assets as of January 1st, 2021.

For Stock 1 we get:

Asset	Stock 1	Stock 2	Stock 3	Stock 4
Price on January 1st, 2019 (S_{t_1})	100€	67€	3 486€	587€
Price on January 1st, 2021 (S_{t_2})	110€	69€	3 674€	203€

Table 4.1: Prices for four stocks on two dates for example 4.1.

- a 2-year discrete rate of return:
$$R_1 = \frac{110}{100} - 1 = 10\%.$$

- an annualized discrete rate of return:
$$R_1^{ann} = 1.1^{1/2} - 1 = 4.881\%.$$

- a value on January 1st, 2021:
$${}_1V_{2016} = 10\,000 \cdot 1.04881^2 = 11\,000€.$$

For Stock 2 we get:

- a 2-year discrete rate of return:
$$R_2 = \frac{69}{67} - 1 = 2.985\%.$$

- an annualized discrete rate of return:
$$R_2^{ann} = 1.02985^{1/2} - 1 = 1.482\%.$$

- a value on January 1st, 2021:
$${}_2V_{2016} = 10\,000 \cdot 1.01482^2 = 10\,298.51€.$$

For Stock 3 we get:

- a 2-year discrete rate of return:
$$R_3 = \frac{3\,674}{3\,486} - 1 = 5.393\%.$$

- an annualized discrete rate of return:
$$R_3^{ann} = 1.05393^{1/2} - 1 = 2.661\%.$$

- a value on January 1st, 2021:
$${}_3V_{2016} = 10\,000 \cdot 1.02661^2 = 10\,539.30€.$$

For Stock 4 we get:

- a 2-year discrete rate of return:

$$R_4 = \frac{203}{587} - 1 = -65.417\%.$$

- an annualized discrete rate of return:

$$R_4^{ann} = 0.34583^{1/2} - 1 = -41.193\%.$$

- a value on January 1st, 2021:

$${}_4V_{2016} = 10\,000 \cdot 0.58807^2 = 3\,458.26€.$$

Example 4.2 Consider the price information for four stocks, given in table 4.2.

Asset	Stock 1	Stock 2	Stock 3	Stock 4
Price on November 30th, 2020 (S_{t_1})	107€	68€	3 681€	242€
Price on December 17th, 2020 (S_{t_2})	109€	70€	3 682€	199€

Table 4.2: Prices for four stocks on two dates for example 4.2.

Assume we want to calculate the T-period discrete rate of return R, the annualized discrete rate of return R^{ann} and the value of an original investment of € 10 000 on November 30th, 2020 for all four assets as of December 17th, 2020.

The time period T is 17 days, corresponding to $T = 17/365 = 0.04657534$ years.

For Stock 1 we get:

- a $17/365$-year discrete rate of return:

$$R_1 = \frac{109}{107} - 1 = 1.869\%.$$

- an annualized discrete rate of return[8]:

$$R_1^{ann} = 1.01869^{\frac{365}{17}} - 1 = 48.827\%.$$

- a value on December 17th, 2020:

$${}_1V_{t_2} = 10\,000 \cdot 1.48827^{\frac{17}{365}} = 10\,186.92€.$$

[8]Please note that

$$\frac{1}{\frac{17}{365}} = \frac{365}{17}.$$

For Stock 2 we get:

- a $^{17}/_{365}$-year discrete rate of return:

$$R_2 = \frac{70}{68} - 1 = 2.941\%.$$

- an annualized discrete rate of return:

$$R_2^{ann} = 1.02941^{\frac{365}{17}} - 1 = 86.336\%.$$

- a value on December 17$^{\text{th}}$, 2020:

$$_2V_{t_2} = 10\,000 \cdot 1.86336^{\frac{17}{365}} = 10\,294.12€.$$

For Stock 3 we get:

- a $^{17}/_{365}$-year discrete rate of return:

$$R_3 = \frac{3\,682}{3\,681} - 1 = 0.027\%.$$

- an annualized discrete rate of return:

$$R_3^{ann} = 1.00027^{\frac{365}{17}} - 1 = 0.585\%.$$

- a value on December 17$^{\text{th}}$, 2020:

$$_3V_{t_2} = 10\,000 \cdot 1.00585^{\frac{17}{365}} = 10\,002.72€.$$

For Stock 4 we get:

- a $^{17}/_{365}$-year discrete rate of return:

$$R_4 = \frac{199}{242} - 1 = -17.769\%.$$

- an annualized discrete rate of return:

$$R_4^{ann} = 0.82231^{\frac{365}{17}} - 1 = -98.501\%.$$

- a value on December 17$^{\text{th}}$, 2020:

$$_4V_{t_2} = 10\,000 \cdot 0.01499^{\frac{17}{365}} = 8\,223.14€.$$

4.2 Continuously compounded rate of return

The continuously compounded rate of return r on an investment (alternatively called log-return) is mostly used in financial models. It is calculated as

$$r = \log(S_{t_2}) - \log(S_{t_1}) = \log\left(\frac{S_{t_2}}{S_{t_1}}\right),$$

where log(·) is the natural logarithm (compare section 2.6.6) and S_{t_1} and S_{t_2} are prices of an asset at different points in time t_1 and t_2, where $t_1 < t_2$. Hence r represents a T-period continuously compounded rate of return, where $T = t_2 - t_1$.

The annualized continuously compounded rate of return r^{ann} is calculated from the T-period rate of return r as

$$r^{ann} = \frac{r}{T}.$$

A T-period rate of return r is calculated from an annualized rate of return r^{ann} as

$$r = r^{ann} \cdot T.$$

The value at t_2 of an investment in a specific asset at time t_1 is

$$V_{t_2} = V_{t_1} \cdot e^{r} = V_{t_1} \cdot e^{r^{ann} \cdot T},$$

where V_{t_1} and V_{t_2} represent the value of the investment at time t_1 and t_2, respectively. e is the Euler's number, which was introduced in section 2.6.5.

Examples

Example 4.3 Consider the price information for four stocks, given in table 4.3.

Asset	Stock 1	Stock 2	Stock 3	Stock 4
Price on January 1st, 2019 (S_{t_1})	100€	67€	3 486€	587€
Price on January 1st, 2021 (S_{t_2})	110€	69€	3 674€	203€

Table 4.3: Prices for four stocks on two dates for example 4.3.

Assume we want to calculate the T-period continuously compounded rate of return r, the annualized continuously compounded rate of return r^{ann} and the value of an original investment of € 10 000 in 2019 for all four assets as of January 1st, 2021.

For Stock 1 we get:

- a 2-year continuously compounded rate of return:

$$r_1 = \log\left(\frac{110}{100}\right) = 9.531\%.$$

- an annualized continuously compounded rate of return:

$$r_1^{ann} = 0.09531 \cdot \frac{1}{2} = 4.766\%.$$

- a value on January 1st, 2021:

$$_1V_{2016} = 10\,000 \cdot e^{0.04766 \cdot 2} = 11\,000€.$$

For Stock 2 we get:

- a 2-year continuously compounded rate of return:

$$r_2 = \log\left(\frac{69}{67}\right) = 2.941\%.$$

- an annualized continuously compounded rate of return:

$$r_2^{ann} = 0.02941 \cdot \frac{1}{2} = 1.471\%.$$

- a value on January 1st, 2021:

$$_2V_{2016} = 10\,000 \cdot e^{0.01471 \cdot 2} = 10\,298.51€.$$

For Stock 3 we get:

- a 2-year continuously compounded rate of return:

$$r_3 = \log\left(\frac{3\,674}{3\,486}\right) = 5.253\%.$$

- an annualized continuously compounded rate of return:

$$r_3^{ann} = 0.05253 \cdot \frac{1}{2} = 2.626\%.$$

- a value on January 1st, 2021:

$$_3V_{2016} = 10\,000 \cdot e^{0.02626 \cdot 2} = 10\,539.30€.$$

For Stock 4 we get:

- a 2-year continuously compounded rate of return:

$$r_4 = \log\left(\frac{203}{587}\right) = -106.182\%.$$

- an annualized continuously compounded rate of return:

$$r_4^{ann} = -1.06182 \cdot \frac{1}{2} = -53.091\%.$$

- a value on January 1st, 2021:

$${}_4V_{2016} = 10\,000 \cdot e^{-0.53091 \cdot 2} = 3\,458.26€.$$

Example 4.4 Consider the price information for for stocks, given in table 4.4.

Asset	Stock 1	Stock 2	Stock 3	Stock 4
Price on November 30th, 2020 (S_{t_1})	107€	68€	3 681€	242€
Price on December 17th, 2020 (S_{t_2})	109€	70€	3 682€	199€

Table 4.4: Prices for four stocks on two dates for example 4.4.

Assume we want to calculate the T-period continuously compounded rate of return r, the annualized continuously compounded rate of return r^{ann} and the value of an original investment of €10 000 on November 30th, 2020 for all assets as of December 17th, 2020.

The time period T is 17 days, corresponding to $T = 17/365 = 0.04657534$ years.

For Stock 1 we get:

- a $17/365$-year continuously compounded rate of return:

$$r_1 = \log\left(\frac{109}{107}\right) = 1.852\%.$$

- an annualized continuously compounded rate of return:

$$r_1^{ann} = 0.01852 \cdot \frac{365}{17} = 39.761\%.$$

- a value on December 17th, 2020:

$${}_1V_{t_2} = 10\,000 \cdot e^{0.39761 \cdot \frac{17}{365}} = 10\,186.92€.$$

For Stock 2 we get:

- a $17/365$-year continuously compounded rate of return:

$$r_2 = \log\left(\frac{70}{68}\right) = 2.899\%.$$

- an annualized continuously compounded rate of return:

$$r_2^{ann} = 0.02899 \cdot \frac{365}{17} = 62.238\%.$$

- a value on December 17^{th}, 2020:

$$_2V_{t_2} = 10\,000 \cdot e^{0.62238 \cdot \frac{17}{365}} = 10\,294.12€.$$

For Stock 3 we get:

- a $^{17}/_{365}$-year continuously compounded rate of return:

$$r_3 = \log\left(\frac{3\,682}{3\,681}\right) = 0.027\%.$$

- an annualized continuously compounded rate of return:

$$r_3^{ann} = 0.00027 \cdot \frac{365}{17} = 0.583\%.$$

- a value on December 17^{th}, 2020:

$$_3V_{t_2} = 10\,000 \cdot e^{0.00583 \cdot \frac{17}{365}} = 10\,002.72€.$$

For Stock 4 we get:

- a $^{17}/_{365}$-year continuously compounded rate of return:

$$r_4 = \log\left(\frac{199}{242}\right) = -19.563\%.$$

- an annualized continuously compounded rate of return:

$$r_4^{ann} = -0.19536 \cdot \frac{365}{17} = -420.035\%.$$

- a value on December 17^{th}, 2020:

$$_4V_{t_2} = 10\,000 \cdot e^{-4.20035 \cdot \frac{17}{365}} = 8\,223.14€.$$

4.3 The relationship between discrete and continuously compounded rate of return

A continuously compounded rate of return is calculated from a discrete rate of returns as

$$r = \log(1 + R)$$

and a discrete rate of return is calculated from a continuously compounded rate of return as

$$R = e^r - 1.$$

If, for example, the discrete rate of return is 10%, $R = 10\%$, the corresponding continuously compounded rate of return is $r = \log(1.1) = 9.531\%$. We may verify that $e^{0.09531} - 1 = 10\%$. As a matter of fact $r \leq R$ holds in all cases.

A total loss of the investment (if the value of the invested amount corresponds to zero, $V_{t_2} = 0$) corresponds to $R = -100\%$ and $r = -\infty$.

Figure 4.1 shows the relationship between these two definitions of returns. The figure shows the resulting returns for an investment with $V_{t_1} = 100$ and different end-values V_{t_2}.

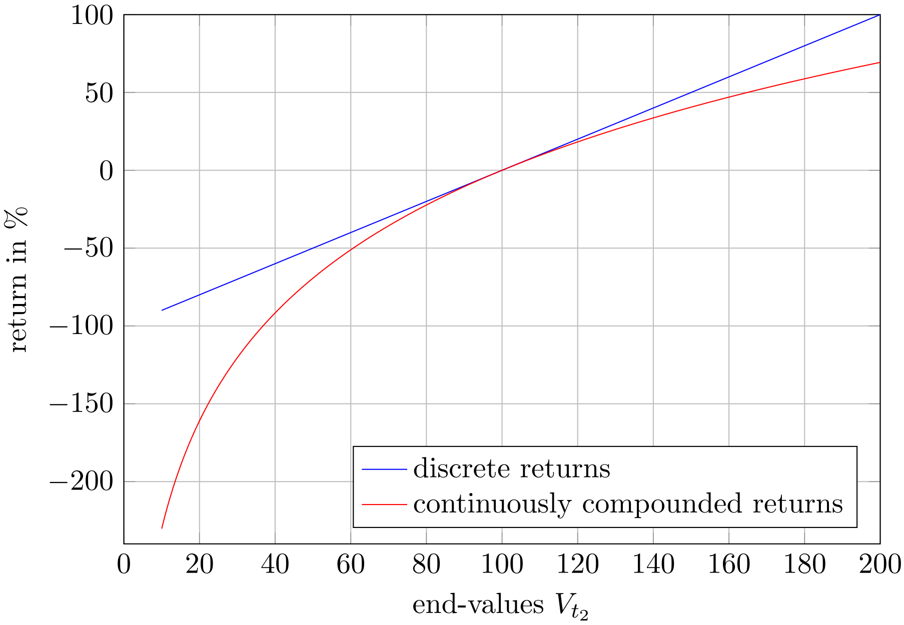

Figure 4.1: The relationship between continuously compounded and discrete returns.

5 Descriptive statistics

Descriptive statistics is the study of collection, organization, analysis and interpretation of data. It deals with all aspects of this, including the planning of data collection in terms of surveys and experiments. Statistical methods can be used to summarize and describe a collection of data.

5.1 Basic concepts of descriptive statistics

The set on which descriptive statistics is usually applied is called the population or the sample, depending on the size of the set. The term statistical characteristics denotes the entities we actually want to measure and analyse. The following sections provide a short overview on the most important basic concepts of descriptive statistics.

5.1.1 Population

For applying statistics to a problem it is necessary to begin with a *population* to be studied. In general, a population consists of all elements - individuals, items or objects - whose characteristics we wish to study.

Examples for populations

- The ballots of *every* person who voted in a given election.
- The sex of *every* persons of a country.
- The electric charge of *every* atom composing a crystal.
- Price of gasoline at *every* petrol station in a country.

5.1.2 Sample

A *sample* is a selection or a subset of entities we can actually study from within a population. This may be used to estimate characteristics of the whole population. A sample can also be composed of observations of a process at various points in time, with data from each observation serving as a different member of the overall group. Data collected this way constitutes what is called a *time series*.

Usually, an entire population cannot be surveyed because the cost of a census is too high. The main advantages of sampling are that

1. the costs are lower,
2. data collection is faster,
3. studying the population is not practical,
4. if the object of interest is defined as a time series, data for the whole population cannot be collected as we would need observations for *all* points in time.

Examples for samples

- The ballots of 5000 voters in a given election.
- The returns of a given stock during 2008 - 2018.
- Prices of 50 petrol stations in a country.

5.1.3 Statistical characteristics and their attributes

When the population and the sampling process are specified, the entities for statistical investigation are fixed. The set of all entities in the population is usually denoted as Ω. Examples are voters, manufactured products or stocks.

Next the *statistical characteristic*, the object of statistical investigation, has to be specified. For this script we denote the set of statistical characteristics as $\mathcal{S}$. Examples are the vote, exact dimension of the product or prices of a stock at certain points in time.

Finally the *statistical attributes*, the individual values of a statistical characteristic, can be described. The statistical attributes comprise of all elements of the set $\mathcal{S}$. Examples are $\mathcal{S} = \{$SPÖ, ÖVP, $\ldots\}$, $\mathcal{S} = \{$2.1243cm, 2.1245cm $,\ldots\}$ and $\mathcal{S} = \{€75, €77, \ldots\}$.

Summarizing, the model used for statistical investigations can be formally described as a function where every element of the population $\omega \in \Omega$ is mapped to its measurable statistical attribute s.

$$X : \Omega \to \mathcal{S}$$
$$\omega \mapsto s.$$

The set of statistical characteristics can be classified as qualitative or quantitative characteristics.

Qualitative statistical characteristics

If a sample is measured using qualitative characteristics, the statistical sample cannot directly be measured with numbers. Within qualitative statistical characteristics two concepts can be distinguished, namely nominal and ordinal characteristics.

1. A *nominal characteristic* is measured on a nominal scale. Within such a scale, there is *no* hierarchy of attributes and the distance between the attributes *cannot* be interpreted.

   Therefore nominal characteristics only allow qualitative classification. The elements under investigation are measured in terms of whether the individual item belongs to some distinctive category, but these categories cannot be quantified or ordered.

   Examples for nominal characteristics
   Eye color, occupation, religious affiliation, name of the stock exchange of an equity share.

2. An *ordinal characteristic* is measured on a so-called ordinal scale. For this scale, there *is* a hierarchy of attributes, but the distance between the attributes *cannot* be interpreted.

   Therefore an ordinal characteristic allows one to rank measured items in terms of which has less and which has more of the quality represented by the characteristic, but there is no reliable information about ≪how much more≫.

   Examples for ordinal characteristics
   Drinking habit, level of education, grades, rating of a company.

Quantitative statistical characteristics

The attributes of a quantitative characteristic are characterized by their magnitude. In these cases there is a hierarchy of the attributes (in this case values) and the distance between them *can* be interpreted meaningfully. Depending on the values an attribute can have, two types can be distinguished.

1. A *discrete characteristic* may possess only certain *distinct* attributes. The number of values can be finite or countably infinite. Most of the time, the set of a discrete quantitative characteristic is a subset of the natural numbers, formally $\mathcal{S} \subseteq \mathbb{N}$.

   Examples for discrete characteristics
   Number of bad items in a sample, monthly income in €, number of employees of a company.

2. A *continuous characteristic* may possess at least theoretically *any* value within an interval. In general the set of such a statistical characteristic $\mathcal{S}$ is a subset of the real numbers, formally $\mathcal{S} \subseteq \mathbb{R}$.

   Examples for continuous characteristics
   Temperature, length, daily return of a stock.

5.2 Distribution analysis

For this section we concentrate on a statistical model with a finite set of statistical characteristics. We have a sample (or population) that contains n observations (or elements of the population). The set of all statistical characteristics $\mathcal{S}$ is a finite set containing k attributes a_j, $j = 1, 2, \ldots, k$.

We get

$$\begin{aligned} X : \Omega &\to \mathcal{S} = \{a_1, a_2, \ldots, a_k\} \\ \omega &\mapsto a_j \end{aligned}$$

5.2.1 Absolute and relative frequencies

The **absolute frequency** of an attribute a_j, $j = 1, \ldots, k$ is defined as

$$\mathbf{F}(a_j) = \ll\text{Number of cases in which } a_j \text{ occurs}\gg = n_{a_j}.$$

The **relative frequency** of an attribute a_j, $j = 1, \ldots, k$ is defined as

$$f(a_j) = \frac{1}{n} \cdot \mathbf{F}(a_j).$$

Important properties of absolute and relative frequencies are

1. The absolute frequency of any attribute is less than or equal to the number of observations,
$$0 \leq \mathbf{F}(a_j) \leq n \text{ for all } j = 1, \ldots, k.$$
2. The relative frequency of any attribute is less than or equal to 1 (or 100%),
$$0 \leq f(a_j) \leq 1 \text{ for all } j = 1, \ldots, k.$$
3. The sum of all absolute frequencies of all attributes equals the number of observations n,
$$\sum_{j=1}^{k} \mathbf{F}(a_j) = n.$$
4. The sum of all relative frequencies of all attributes equals 1 (or 100%),
$$\sum_{j=1}^{k} f(a_j) = 1.$$

Based on the calculation of frequencies, a **frequency distribution** can be defined. The sequence

$$(\mathbf{F}(a_1), \mathbf{F}(a_2), \ldots, \mathbf{F}(a_k))$$

is called the absolute frequency distribution.

Analogously, the sequence

$$(f(a_1), f(a_2), \ldots, f(a_k))$$

is called the relative frequency distribution.

Example for the calculation of frequencies of grades
Let $X : \Omega \to \mathcal{S}$ be the statistical model which assigns a grade in a given exam to all students. Therefore the set Ω is the set of all students, while the set $\mathcal{S}$ consists of all possible grades[9], $\mathcal{S} = \{1, 2, 3, 4, 5\}$.

Clearly, each grade can occur several times within a given sample. The number of students who received a certain grade, is the absolute frequency of this grade. The relative proportion of students who received a certain grade is the relative frequency of this grade.

Example for absolute and relative frequencies
Let $X : \Omega \to \mathcal{S}$ be a statistical model for the Austrian real GDP growth[10]. For the sake of simplicity we consider rounded values. The data is presented in table 5.1.

year	*Austria real GDP growth in % (rounded values)*
2007	4
2008	2
2009	-4
2010	2
2011	3
2012	1
2013	0
2014	1
2015	1
2016	2
2017	3

Table 5.1: Rounded values for the Austrian real GDP growth since 2007 (in %).

The data starts 2007, contains 11 data points with 6 different values. Therefore

$$n = 11 \quad \text{and} \quad k = 6.$$

The absolute and relative frequencies for this data set are presented in table 5.2.

[9]The Austrian grade system is used for this example.
[10]Source of the data: Statistik Austria. *Wachstum des Bruttoinlandsprodukts (BIP) in Österreich in den Jahren 2007 bis 2017 (gegenüber dem Vorjahr).* Statista, `de.statista.com/statistik/daten/studie/14530/umfrage/wachstum-des-bruttoinlandsprodukts-in-oesterreich/`

a_j	$\mathbf{F}(a_j)$	$f(a_j)$
-4	1	0.091
0	1	0.091
1	3	0.273
2	3	0.273
3	2	0.182
4	1	0.091
$\sum$	11	1.000

Table 5.2: Absolute and relative frequencies for the Austrian real GDP growth given in table 5.1.

5.2.2 Visualization of a frequency distribution

There are several ways to visualize frequency distributions. In this section the most important methods are presented.

Bar charts

A bar chart depicts the frequencies by a series of bars. Absolute or relative frequencies can be depicted (see figure 5.1 for a bar chart of the absolute frequencies $\mathbf{F}(a_j)$).

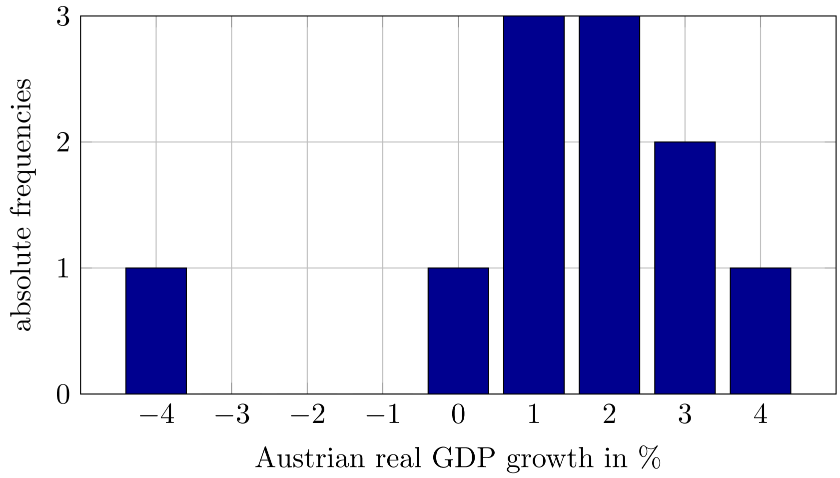

Figure 5.1: Bar chart of the absolute frequencies for the Austrian real GDP growth rates.

A common term for a bar chart of the frequency distribution is **histogram**. For many applications a histogram combines a bar chart of the relative frequency distribution with the probability density function of a normal distribution[11] with same mean and standard deviation as the given data.

Figure 5.2 shows the relative frequencies of the data given in table 5.1 combined with a normal distribution with the same mean and standard deviation as red line.

[11]For the definition of normal distributions see chapter 6.6.4.

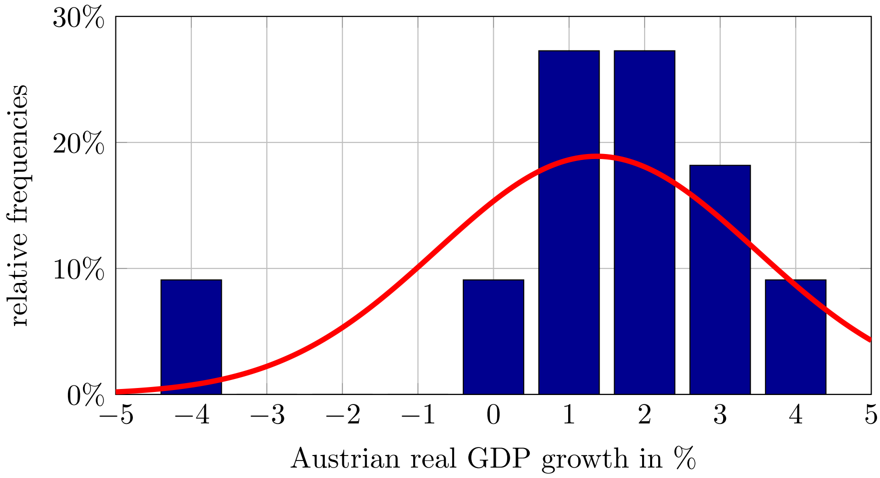

Figure 5.2: Histogram for Austrian real GDP growth combined with a normal distribution.

Polygons

A frequency polygon is a simple line graph of the relative or absolute frequency distribution. See figure 5.3 for a relative frequency line chart the data given in table 5.1.

Figure 5.3: Line chart of the relative frequencies for the Austrian real GDP growth rates.

The frequency distribution charts with bars or polygons allow us to describe the frequency distribution in terms of the symmetry of the data (skewness, see section 5.2.3) and dispersion of the data (kurtosis, see section 5.2.4).

Pie charts

A pie chart is a pie-shaped figure in which pieces of the pie represent the relative frequencies. See figure 5.4 for a relative frequency pie chart for the data given in table 5.1.

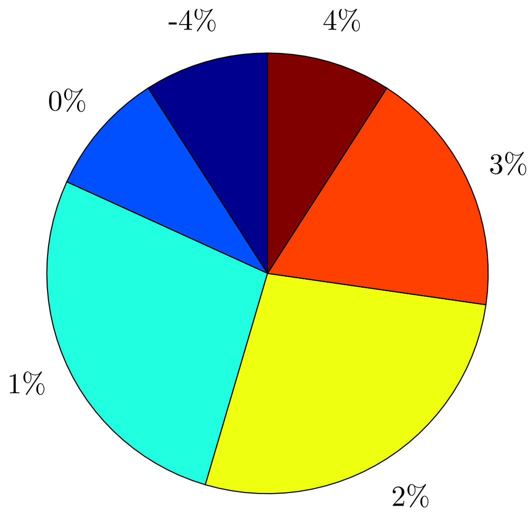

Figure 5.4: Pie chart for the relative frequencies of the Austrian real GDP growth since 2007.

5.2.3 Skewness

In terms of *skewness*, a frequency curve can be

positively skewed The curve is nonsymmetric with the ≪tail≫ to the right. See the blue distribution curve on the right of figure 5.5.

symmetrical The curve is symmetric. See the black distribution curve at the center of figure 5.5.

negatively skewed The curve is nonsymmetric with the ≪tail≫ to the left. See the red distribution curve on the left of figure 5.5.

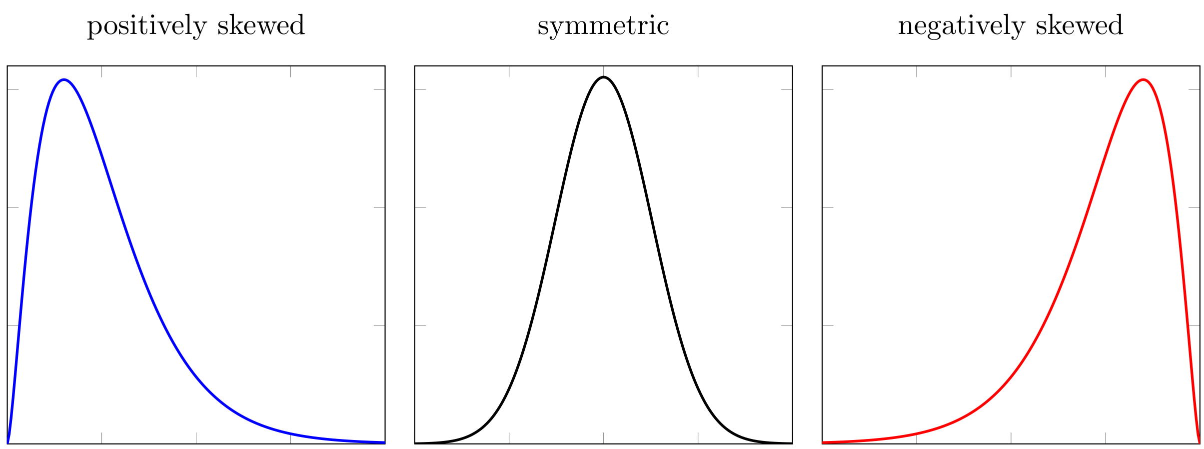

Figure 5.5: The different forms of frequency curves in terms of skewness.

Karl Pearson suggested a simple calculation as a measure of skewness (sometimes denoted as Pearson's median). He defined the skewness coefficient S as[12]

$$S = \frac{3 \cdot \left(\overline{X} - \text{Me}\right)}{s}, \qquad \text{where } -3 \leq S \leq 3.$$

Example for the calculation of the skewness coefficient
The skewness coefficient S of the Austrian real GDP growth data presented in table 5.1 is

$$S = \frac{3 \cdot (1.364 - 2.0)}{2.111} = -0.905.$$

As can be seen easily, the data given in table 5.1 is moderately negatively skewed.

5.2.4 Kurtosis

In terms of *kurtosis*, a frequency curve can be

mesokurtic The curve is neither flat nor peaked in terms of the distribution of observed values. The frequency distribution curve is similar to the probability density function of a normal distributed random variable (see chapter 6.6.4). Compare type A in figure 5.6.

leptokurtic The curve is peaked with the observations concentrated within a narrow range of values. See type B in figure 5.6.

platykurtic The curve is flat as the observations in the data are scattered heavily. An example is shown in type C in figure 5.6.

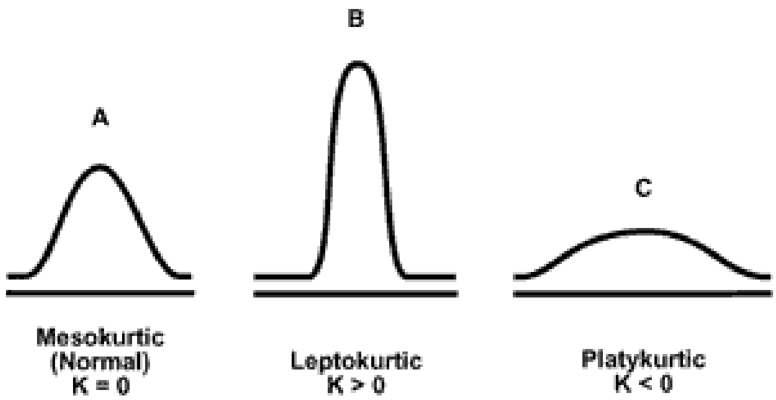

Figure 5.6: The different forms of frequency curves in terms of kurtosis.

[12]The measures of location *arithmetic mean* $\overline{X}$ and *median* Me are defined in section 5.5. The definition of the standard deviation s can be found in section 5.6.

5.3 Cumulative absolute and relative frequencies

Additionally to the absolute and relative frequency distribution the cumulative distribution can be of interest. The cumulative frequency distribution visualises the number of values or percentage of values equal to or below a distinct attribute a_j.

Formally, the *cumulative absolute frequency* of the attribute a_j for $j = 1, \ldots, k$ is defined as

$$\mathbf{F}(a_1) + \mathbf{F}(a_2) + \ldots + \mathbf{F}(a_j) = \sum_{i=1}^{j} \mathbf{F}(a_i), \qquad j = 1, \ldots, k.$$

Similarly, the *cumulative relative frequency* of the attribute a_j for $j = 1, \ldots, k$ is defined as

$$f(a_1) + f(a_2) + \ldots + f(a_j) = \sum_{i=1}^{j} f(a_i), \qquad j = 1, \ldots, k.$$

Continued example with data given in table 5.1
The cumulative absolute and relative frequencies for the rounded values of the Austrian real GDP growth data given in table 5.3.

j	a_j	$\mathbf{F}(a_j)$	$f(a_j)$	$\sum_{i=1}^{j} \mathbf{F}(a_i)$	$\sum_{i=1}^{j} f(a_i)$
1	-4	1	0.091	1	0.091
2	0	1	0.091	2	0.182
3	1	3	0.273	5	0.455
4	2	3	0.273	8	0.727
5	3	2	0.182	10	0.909
6	4	1	0.091	11	1.000
	Sum	*11*	*1.000*		

Table 5.3: Absolute and relative frequencies combined with their cumulative frequencies calculated for the Austrian real GDP growth rates since 2007.

A polygon plot of the cumulative absolute or relative frequencies is denoted as *ogive*. An ogive shows the values of the data set on the horizontal axis and either the cumulative absolute frequencies or the cumulative relative frequencies on the vertical axis. In figure 5.7 an ogive with cumulative relative frequencies is shown.

5.4 Empirical distribution function

The *empirical distribution function*, or *empirical CDF*, is the cumulative distribution function associated with the empirical measure derived from the sample. This function is a step function that jumps up by $1/n$ at each of the n data points.

Figure 5.7: An ogive for cumulative relative frequencies for the rounded values if Austrian real GDP growth rates since 2007.

The empirical cumulative distribution function can be used as an estimate for the true underlying cumulative distribution function (CDF) of the stochastic data set (for the definition of a cumulative distribution function F see section 6.1).

The empirical CDF $\hat{F}$ is defined as follows

$$\hat{F}(x) = \frac{\text{number of elements in the sample with attributes } a_j \text{ smaller or equal to } x}{n}$$

$$= \begin{cases} 0 & \text{for} \quad x < a_1 \\ \sum_{i=1}^{j} f(a_i) & \text{for} \quad a_j \leq x < a_{j+1} \\ 1 & \text{for} \quad x \geq a_k. \end{cases}$$

The empirical cumulative distribution function fulfills some important properties.

1. The empirical CDF describes the proportion of values in the sample which are smaller than a distinct value x. Formally

$$\hat{F}(x) = \mathbf{P}(X \leq x) \quad (\mathbf{P}: \text{proportion})$$

2. The empirical CDF always lies between 0 and 1. Formally

$$0 \leq \hat{F}(x) \leq 1 \qquad \text{for all } x$$

3. The empirical CDF is a monotonically increasing function. Formally

$$\text{for all } x_1, x_2: \quad x_1 < x_2 \Rightarrow \hat{F}(x_1) \leq \hat{F}(x_2)$$

4. The empirical CDF can be used to calculate the proportion of values within a certain half-open interval $]x_1, x_2]$.

$$\mathbf{P}(x_1 < X \leq x_2) = \hat{F}(x_2) - \hat{F}(x_1)$$

5. $\hat{F}(x)$ is at least right-sided continuous and has at most a finite number of jump discontinuities.

6. All empirical CDFs approach 0 as $x \to -\infty$ and 1 as $x \to \infty$.

Example for an empirical CDF as chart and table
The empirical CDF can be represented in different ways. In figure 5.8 the empirical cumulative distribution function for the Austrian real GDP growth rates given in table 5.1 is shown. The circles denote that the corresponding points are excluded and the filled points are included in the graph.

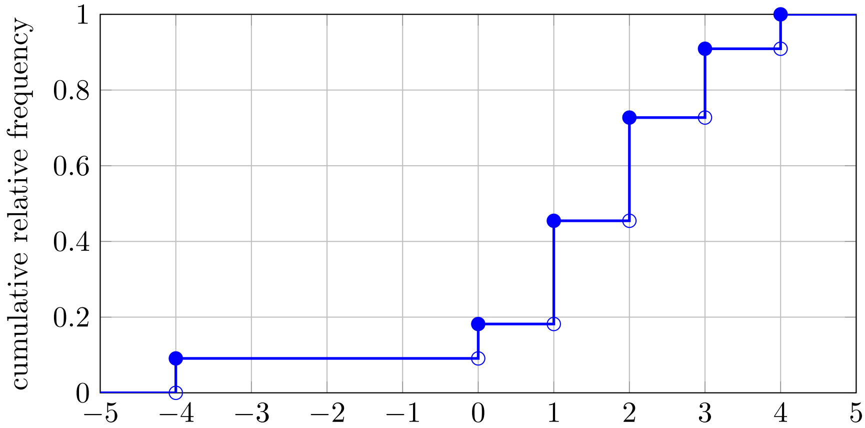

Figure 5.8: Empirical cumulative distribution function $\hat{F}(x)$ for the rounded values of Austrian real GDP growth rates since 2007.

Additionally, the empirical CDF can be represented in tabular form. Thereby the domain of the function is divided in half-open intervals and the function value can be specified for each interval. Table 5.4 shows the empirical CDF for the Austrian real GDP growth rates.

Interval for x	$\hat{F}(x)$
$]-\infty, -4[$	0.000
$[-4, 0[$	0.091
$[0, 1[$	0.182
$[1, 2[$	0.455
$[2, 3[$	0.727
$[3, 4[$	0.909
$[4, +\infty[$	1.000

Table 5.4: Empirical cumulative distribution function $\hat{F}$ for Austrian real GDP growth rates since 2007 in tabular form.

Example for the interpretation of the values of the empirical CDF
The empirical CDF allows the following calculations and interpretations.

1. The value of the empirical CDF at 2.2 is

$$\hat{F}(2.2) = 0.727.$$

This can be interpreted as 72.7% of the rounded values of Austrian real GDP growth rates since 2007 are less than or equal to 2.2.

2. The value of the difference of the empirical CDF at 3 and 0.5 can be used to calculate the probability of the real GDP growth rates lying between 0.5 and 3.0. We get

$$\hat{F}(3) - \hat{F}(0.5) = 0.909 - 0.182 = 0.727.$$

Therefore about 73% of the Austrian real GDP growth rates lie between 0.5 and 3.0.

5.5 Measures of location

5.5.1 Arithmetic mean

The *arithmetic mean*, or simply the *mean* when the context is clear, is a measure for the central tendency of a collection of values. It is calculated as the sum of the values divided by the size of the collection. Let

$$x_i, \quad i = 1, \ldots, n$$

be the observed values of a variate having the attributes

$$a_j, \quad j = 1, \ldots k.$$

If values for the whole population of size N are known, the mean is called the *population mean*. It is denoted as μ and is calculated as

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i = \frac{1}{N} \sum_{j=1}^{k} a_j \cdot \mathbf{F}(a_j)$$

and therefore is a simple arithmetic mean of population data.

If observed values consist of a sample, the mean is called the *sample mean*. We denote it as $\hat{\mu}$ (sometimes denoted as $\overline{X}$) and calculate it as arithmetic mean of the sample data as in

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \sum_{j=1}^{k} a_j \cdot \mathbf{F}(a_j).$$

The arithmetic mean fulfills the following properties:

1. The arithmetic mean of data which has been linearly transformed by its arithmetic mean is always 0. Formally

$$\sum_{i=1}^{n} (x_i - \hat{\mu}) = 0.$$

2. The arithmetic mean is the best measurement of the central tendency of the data in terms of least squares. Formally

$$\text{for all } a \in \mathbb{R} \text{ it holds that} \qquad \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \leq \sum_{i=1}^{n} (x_i - a)^2 .$$

3. Given a linear transformation of the data set as in

$$x_i^* = \alpha + \beta \cdot x_i \quad \text{for } i = 1, \dots, n \text{ and } \alpha, \beta \in \mathbb{R},$$

we obtain

$$\hat{\mu}^* = \alpha + \beta \cdot \hat{\mu}.$$

Example for the calculation of the sample mean
The sample mean for the Austrian real GDP growth data presented in table 5.1 is

$$\hat{\mu} = \frac{1}{11} \sum_{i=1}^{11} x_i = \frac{15}{11} = 1.364.$$

5.5.2 Median

The *median* is the numerical value separating the higher half of a sample or a population from the lower half. The median of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one.

If there is an even number of observations, then there is no single value in the middle. The median is then usually defined to be the mean of the two middle values. Formally, the *median* of a data set is defined by

$$\text{Me} = \begin{cases} x_{\left[\frac{n+1}{2}\right]} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n}{2}+1\right]}\right) & \text{if } n \text{ is even.} \end{cases}$$

Here $[\cdot]$ denotes the position of the observation in the *arranged* (ascending or descending) data set.

Example for the calculation of the median
The median for the Austrian real GDP growth rate data presented in table 5.1 is

$$\text{Me} = 2.$$

5.5.3 α-quantile

Quantiles are points taken at regular intervals from the cumulative distribution function (CDF) of a random variable (see section 6) or the corresponding empirical CDF of a data set. Dividing ordered data into $\frac{1}{\alpha}$ essentially equal-sized data subsets is the motivation for α-quantiles. The quantiles are the data values marking the boundaries between consecutive subsets.

Put in another way, the k^{th} α-quantile for a random variable is the value x such that the probability that the random variable will be *less* than x is at most

$$k \cdot \alpha$$

and the probability that the random variable will be *more* than x is at most

$$1 - k \cdot \alpha.$$

For a *ranked* data set, the α-quantile ($0 < \alpha < 1$) is defined by

$$\tilde{x}_\alpha = \begin{cases} x_{[k]} & \text{if } n \cdot \alpha \text{ is not an integer (then } k \text{ is the following integer)} \\ \frac{1}{2}\left(x_{[k]} + x_{[k+1]}\right) & \text{if } n \cdot \alpha \text{ is integer } (k = n\alpha) \end{cases}$$

In many applications the *lower* and *upper* quartiles are used. Usually the 25% quantile is called *lower (first) quartile* and the 75% quartile is called *upper (third) quartile* respectively.

Example for the calculation of a 25% quantile
When calculating the 25% quantile for the Austrian real GDP growth rates presented in table 5.1 we get:

$$11 \cdot 25\% = 11 \cdot 0.25 = 2.75.$$

As $n \cdot \alpha$ is not an integer we search for the data point in the *sorted* data set at the position $\tilde{x}_{[k]} = \tilde{x}_{[3]}$. This is

$$\tilde{x}_{[3]} = 1.$$

5.5.4 Mode

The *mode* is the number that appears the most often in a set of numbers. In other words, the mode of a data set is the statistical attribute with the highest absolute frequency. Usually it is denoted as Mo. Like the mean (see section 5.5.1) and median (see section 5.5.2), the mode is a way of expressing, in a single number, a measurement of the central tendency of the data set.

The numerical value of the mode is usually different from those of the mean and median and may be very different for strongly skewed distributions. The mode of a *discrete probability distribution* (compare section 6.2) is the value x at which its probability mass function has its maximum value. Informally speaking, the mode is at the peak.

The mode is not necessarily unique, since the same maximum frequency may be attained at different values. The most extreme case occurs in uniform distributions, where all values occur equally frequently.

Example for the determination of the mode
The mode for the Austrian real GDP growth rates given in table 5.1 is not unique as the values 1 and 2 both occur 3 times. We get

$$\text{Mo} = 1 \quad \text{or} \quad \text{Mo}' = 2.$$

5.5.5 Relationship among mean, median and mode

The information about mean, median and mode allows us to determine some properties of the sample distribution or population distribution respectively. Among other properties the skewness (see section 5.2.3) of the distribution can be identified.

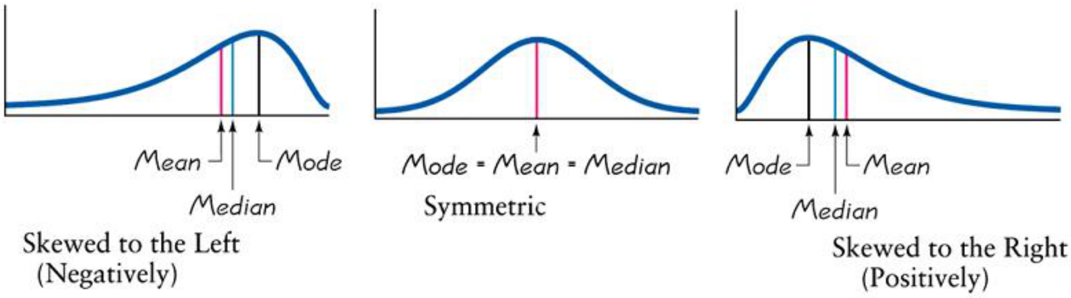

Figure 5.9: The values of mean, median and mode for three possibilities the distribution of a sample or a population.

In figure 5.9 three possibilities are presented:

1. For a symmetric frequency distribution curve *with one peak*, the values of the mean, median and mode are identical. They lie in the centre of the distribution. This is displayed in the image in the center of figure 5.9.

2. For a frequency distribution curve skewed to the right, the value of the mean is the largest, that of the mode is the smallest and the value of the median lies between the two. This is displayed in the right image of figure 5.9.

3. If a frequency distribution curve is skewed to the left, the value of the mean is the smallest and that of the mode is the largest with the value of the median lying between these two. This is displayed in the left image of figure 5.9.

5.5.6 Geometric mean

The geometric mean is a type of mean or average, which indicates the central tendency or typical value of a set of numbers. A geometric mean is often used when comparing different items - finding a single ≪figure of merit≫ for these items - when each item has multiple properties that have different numeric ranges.

The geometric mean is similar to the arithmetic mean (see section 5.5.1), except that the numbers are multiplied and then the n^{th} root (n is the size of the data set) of the resulting product is taken. It only works for data sets containing only positive numbers and no zeros.

Formally, for the sample of size n the geometric mean is defined as

$$\overline{X}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}, \qquad x_i > 0 \text{ for all } i = 1, \ldots, n.$$

Example of usage of the geometric mean
The geometric mean can give a meaningful «average» to compare two companies which are each rated at 1 to 5 for their environmental sustainability and are rated at 1 to 100 for their financial viability.

If an arithmetic mean is used instead of a geometric mean, the financial viability is given more weight because its numeric range is larger. A small percentage change in the financial rating (e.g. going from 80 to 90) makes a much larger difference in the arithmetic mean than a large percentage change in environmental sustainability (e.g. going from 2 to 5).

The use of a geometric mean «normalizes» the ranges being averaged, so that no range dominates the weighting, and a given percentage change in any of the properties has the same effect on the geometric mean. So, a 20% change in environmental sustainability from 4 to 4.8 has the same effect on the geometric mean as a 20% change in financial viability from 60 to 72.

Examples for measures of location

Summarizing, table 5.5 shows various measures of location for the data set for Austrian real GDP growth given in table 5.1. The geometric mean does not work for the given data set as its usage is

Mean $\hat{\mu}$	1.36
Median Me	2.00
25%-quantile	1.00
75%-quantile	3.00

Table 5.5: Measures of location for Austrian real GDP growth rates

limited to data sets containing only positive numbers and no zeros. The data set given in table 5.1 contains a negative observation as well as zero.

5.6 Measures of dispersion

Measures of location indicate the general magnitude of the data and locate only the center of a distribution. They do not establish the degree of variability or the spread of the individual items and the deviation from the average.

Two distributions of statistical data may be symmetrical and have common arithmetic mean, median and modes. Yet with these points in common they may differ widely in the scatter or in their values in measures of dispersion.

Statistical dispersion (also called *statistical variability* or *variation*) is variability or spread in a variable or a probability distribution. Common examples of measures for statistical dispersion are the variance, standard deviation and interquartile range.

5.6.1 Range

In descriptive statistics, the *range* of a set of data is the difference between the largest and smallest values. It is the smallest interval which contains all the data and provides an indication of statistical dispersion. Formally, the *range* is defined by

$$R = x_{\text{max}} - x_{\text{min}}$$

where x_{min} denotes the smallest value in the data set and x_{max} denotes the largest value in the data set.

The absolute range of a set of data has some disadvantages for interpretation. Firstly it is strongly influenced by outliers and secondly its calculation is based on two values only. Thus, the range is not always a satisfactory measure of dispersion.

The *interquartile range* counterbalances some of these disadvantages. It is defined as

$$R_Q = \tilde{x}_{75\%} - \tilde{x}_{25\%}$$

where $\tilde{x}_{75\%}$ is the 75% quantile and $\tilde{x}_{25\%}$ is the 25% quantile.

Example for the range of a distribution
The 25% quantile of the Austrian real GDP growth data presented in table 5.1 is $\tilde{x}_{25\%} = 1$, the 75% quantile is $\tilde{x}_{75\%} = 3$. The lowest value is $x_{\text{min}} = -4$ and the highest value equals $x_{\text{max}} = 4$.

Therefore we get

$$R = 4 - (-4) = 8$$

as range R and

$$R_Q = 3 - 1 = 2$$

as interquartile range R_Q.

5.6.2 Average absolute deviation

The *average absolute deviation* of a data set is the average of the absolute deviation and is a summary statistic of statistical dispersion or variability. The point from which the deviations are measured is a measure of central tendency, mostly the median or the mean of the data set.

1. Absolute average deviation from the median is defined as

$$d_{\text{Me}} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \text{Me}| \, .$$

2. Absolute average deviation from the mean is defined as

$$d_{\hat{\mu}} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \hat{\mu}| \,.$$

Example of the absolute average deviation from the median and the mean
The absolute average deviation from the median for the Austrian real GDP growth data presented in table 5.1 is $d_{\text{Me}} = 1.3636$.

The absolute average deviation from the mean of the same data set is $d_{\hat{\mu}} = 1.421$.

5.6.3 Average squared deviation

The *average squared deviation* of a data set is the average of the squared deviations and measures statistical dispersion or variability. The average squared deviation from the arithmetic average is the most important measure of dispersion and is called *variance*. There are two definitions of variance, depending on the data set.

1. If the data set contains the whole population, the *population variance* σ^2 is defined as

   $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$$

   where N is the size of the population and μ is the arithmetic mean of the population.

2. If the data set contains just a sample, the *sample variance* $\hat{\sigma}^2$ is defined as

   $$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

   where n is the sample size and $\hat{\mu}$ denotes the sample mean.

For manual calculations the formulae for the population variance and the sample variance can be reformulated in order to obtain the following short cut formulas:

1. For the population variance we get:

   $$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2 - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{N}}{N} = \frac{\sum_{i=1}^{N} x_i^2 - N\mu^2}{N}.$$

2. The sample variance can be reformulated into:

   $$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\hat{\mu}^2}{n-1}.$$

The standard deviation is then defined as the positive square root of the variance. Hence

$$\sigma = +\sqrt{\sigma^2}$$

is the *population standard deviation* and

$$\hat{\sigma} = +\sqrt{\hat{\sigma}^2}$$

is the *sample standard deviation.*

If the mean and the standard deviation of a data set or a distribution are known, the *coefficient of variation* can be calculated.

1. For a population data set the coefficient of variation is defined as

$$v = \frac{\sigma}{\mu}.$$

2. For a sample data set the coefficient of variation is defined as

$$v = \frac{\hat{\sigma}}{\hat{\mu}}.$$

Example for the calculation of the sample variance and sample standard deviation
The sample variance s^2 of the Austrian real GDP growth data presented in table 5.1 is

$$\hat{\sigma}^2 = 4.4545.$$

Its sample standard deviation $\hat{\sigma}$ is

$$\hat{\sigma} = 2.1106.$$

The coefficient of variation equals

$$v = \frac{2.1106}{1.3636} = 1.548.$$

5.6.4 Box-and-whisker plot

A *box-and-whisker plot* (or simply *box plot*) is a convenient way of graphically depicting a data set through its five number summaries. It is constructed by drawing a box and two whiskers. The five values used to construct a box plot are:

1. a measure for the smallest observation or a low quantile for the lower whisker

2. the lower quartile ($\tilde{x}_{25\%}$) for the lower end of the box

3. the median for the horizontal line within the box

4. the upper quartile ($\tilde{x}_{75\%}$) for the upper end of the box

5. a measure for the largest observation or an upper quantile for the upper whisker

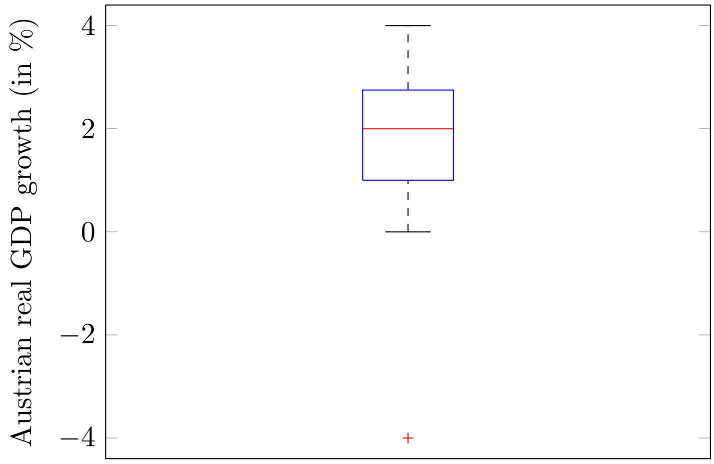

Figure 5.10: A box plot of the Austrian GDP data presented in table 5.1.

Depending on the data, a box plot may also indicate which observations might be considered outliers. This is the case if the whiskers are defined alternatively. Possible definitions for the whiskers are

- The minimum and the maximum.
- One standard deviation above and below the mean or the median.
- Some other multiplier of standard deviations, for figure 5.10 all data points outside of $\text{Me} \pm 2.7 \cdot \sigma$ are considered outliers.
- The 9^{th} percentile and the 91^{th} percentile.
- The 2^{nd} percentile and the 98^{th} percentile.

Figure 5.10 shows the box plot for the Austrian real GDP growth data presented in table 5.1.

Examples for measures of dispersion

Summarizing, table 5.6 shows various measures of dispersion for the data set for Austrian real GDP growth given in table 5.1.

Range R	8
Interquartile Range R_Q	2
Average absolute deviation from the median d_{Me}	1.3636
Average absolute deviation from the mean $d_{\overline{X}}$	1.4215
Population variance σ^2	4.0496
Sample variance $\hat{\sigma}^2$	4.4545

Table 5.6: Measures of dispersion for Austrian real GDP growth rates

5.7 Correlation and regression

5.7.1 Correlation

Correlation refers to any of a broad class of statistical relationships between two random variables or two sets of data. Correlations can (but does not necessarily) indicate a predictive relationship that can be exploited in practice. For example, an electrical utility may produce less power on a mild day based on the correlation between electricity demand and weather. In this example there is a causal relationship, because extreme weather causes people to use more electricity for heating or cooling. However it is necessary to note that statistical dependence is *not* sufficient to demonstrate the presence of such a causal relationship.

The following possible forms of correlation should be distinguished:

1. *Linear* and *non-linear* correlations
   Most methods for measuring correlations implicitly assume that the correlation is of a linear form. An exception to that is the Spearman correlation coefficient which will be introduced later in this chapter.

2. *Positive* and *negative* correlations
   If two variables change in the same direction, then this is called a *positive correlation.* (e.g. stock returns of two companies in similar markets)

   If two variables change in the opposite direction, then this is called a *negative correlation.* (e.g. returns of stocks and bonds)

The value of a correlation coefficient always lies within the interval $[-1, 1]$. In table 5.7 different degrees of correlation are distinguished.

degrees	**positive**	**negative**
absence of correlation	0	0
perfect correlation	+1	-1
high degree	[0.75,1[	]-1,-0.75]
moderate degree	[0.25,0.75[	]-0.75,-0.25]
low degree	]0,0.25[	]-0.25,0[

Table 5.7: Depending on the value of the Pearson correlation coefficient different degrees of correlation can be distinguished.

Data for the following examples
In the following sections we will compare a time series for the real Austrian GDP growth in % and the change of unemployment in % in Austria in the years from 2007 up to 2017[13]. The data is presented in table 5.8.

[13]Source of the data: AMS Österreich and Statistik Austria. *Arbeitslosenquote in Österreich von 2007 bis 2017.* Statista, `de.statista.com/statistik/daten/studie/17304/umfrage/arbeitslosenquote-in-oesterreich/`

year	real Austrian GDP growth (in %)	change in Austrian unemployment (ILO standard, in %)
2007	3.7	-0.4
2008	1.5	-0.8
2009	-3.8	1.2
2010	1.8	-0.5
2011	2.9	-0.2
2012	0.7	0.3
2013	0.0	0.5
2014	0.7	0.2
2015	1.1	0.1
2016	2.0	0.3
2017	2.6	-0.5

Table 5.8: %-changes in the Austrian GDP and the Austrian unemployment rate according to ILO standards from 2007 to 2017.

Pearson correlation coefficient

There are several *correlation coefficients*, often denoted ρ or r, measuring the degree of correlation. The most common of these is the *Pearson correlation coefficient*, which is sensitive only to a linear relationship between two variables. Furthermore it should be used only when both variables are quantitative.

The Pearson correlation coefficient, when applied to a sample, is commonly represented by the letter r and may be referred to as the *sample correlation coefficient*.

For the two variables X and Y the Pearson correlation coefficient is defined as

$$r = \frac{\sum_{i=1}^{n} (x_i - \hat{\mu}_X) \cdot (y_i - \hat{\mu}_Y)}{\sqrt{\sum_{i=1}^{n} (x_i - \hat{\mu}_X)^2 \cdot \sum_{i=1}^{n} (y_i - \hat{\mu}_Y)^2}}.$$

If the sample means ($\hat{\mu}_X$ and $\hat{\mu}_Y$) and the sample standard deviations ($\hat{\sigma}_X$ and $\hat{\sigma}_Y$) are already known, the Pearson correlation coefficient can alternatively be calculated as

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \hat{\mu}_X}{\hat{\sigma}_X} \right) \left(\frac{y_i - \hat{\mu}_Y}{\hat{\sigma}_Y} \right).$$

Example for the Pearson correlation coefficient
The Pearson correlation coefficient of the data presented in table 5.8 is

$$r = -0.815.$$

Therefore the data is highly negatively correlated. This result is plausible, it means that unemployment tends to go up if GDP growth gets very low (or even negative) and vice versa.

Spearman correlation coefficient

If the variables are at least ordinal, the *Spearman's rank correlation coefficient* can be used. It is usually denoted by r_S and is a non-parametric measure of statistical dependence between two variables. It is applicable to linear and non-linear relationships between two variables.

It is defined using the ranks of the values of the variables. For assigning ranks, the variables must be ordinal or quantitative. For a sample of size n we have

$$\begin{aligned} R_i, \quad i = 1, \ldots n: &\quad \text{ranks of the characteristic } X \\ R'_i, \quad i = 1, \ldots n: &\quad \text{ranks of the characteristic } Y. \end{aligned}$$

Then the Spearman correlation coefficient is defined as

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{n} (R_i - R'_i)^2}{(n-1) \cdot n \cdot (n+1)}.$$

Example for the calculation of the Spearman correlation coefficient

year	real Austrian GDP growth (in %)	change in Austrian unemployment (ILO standard, in %)	R_i	R'_i	$(R_i - R'_i)^2$
2007	3.7	-0.4	11	4	49
2008	1.5	-0.8	6	1	25
2009	-3.8	1.2	1	11	100
2010	1.8	-0.5	7	2	25
2011	2.9	-0.2	10	5	25
2012	0.7	0.3	3	8	25
2013	0.0	0.5	2	10	64
2014	0.7	0.2	3	7	16
2015	1.1	0.1	5	6	1
2016	2.0	0.3	8	8	0
2017	2.6	-0.5	9	2	49
					$\sum 379$

Table 5.9: Calculation of the Spearman correlation coefficient for the Austrian GDP growth and the change in Austrian unemployment.

In table 5.9 the given data and their corresponding ranks R_i and R'_i are shown. The lowest rank is given to the lowest data point in each column, the second lowest the the following data points. If several data points coincide, all get the same rank but the following ranks have to be skipped in order to balance the effect.

For example, the years 2012 and 2014 saw a real GDP growth of 0.7. Therefore both pints for 2012 and 2014 get the same rank 3, but the rank 4 is skipped in the consecutive ranking process. The next value for real GDP growth observed in 2015, which equals 1.1, gets rank 5.

Additionally the table includes values for $(R_i - R'_i)^2$, which are used to calculate the Spearman correlation coefficient r_S.

The Spearman correlation coefficient of the data presented in table 5.8 equals

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{11}(R_i - R_i')^2}{(11-1)\cdot 11 \cdot (11+1)} = 1 - \frac{6 \cdot 379}{1320} = -0.723.$$

We see again that the data is highly negatively correlated.

Scatter plot

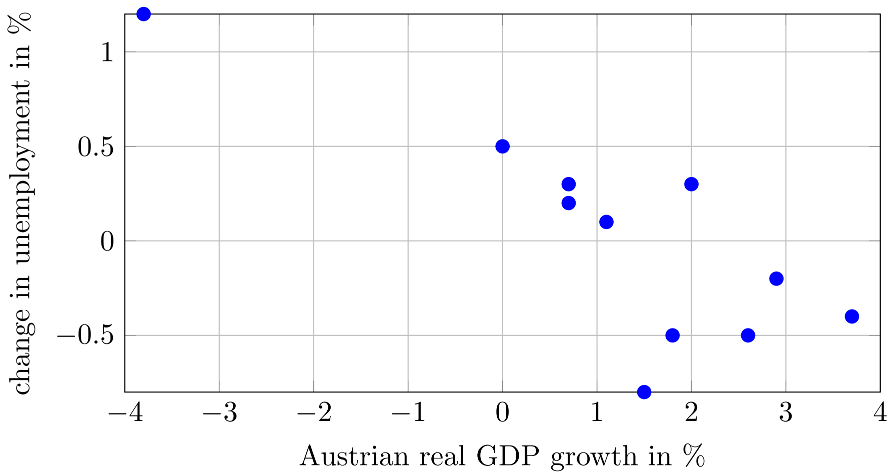

Figure 5.11: A scatter plot of the real Austrian GDP growth (x-axis) and the change in the Austrian unemployment rate (y-axis) presented in table 5.8.

For this method, the values of the two variables for every observation are plotted on a graph. One is taken along the horizontal axis (x-axis) and the other along the vertical axis (y-axis). By plotting the data we obtain points on the graph which are generally scattered, hence the name ≪scatter plot≫.

The manner in which these points are scattered suggests the degree and the direction of correlation. Let the degree of correlation be denoted by ρ. Its direction is given by the signs positive and negative.

i) If all points lie exactly on a rising straight line the correlation is perfectly positive and $\rho = +1$.

ii) If all points lie exactly on a falling straight line the correlation is perfectly negative and $\rho = -1$.

iii) If the points lie in a narrow rising strip, the correlation is highly positive.

iv) If the points lie in a narrow falling strip, the correlation is highly negative.

v) If the points are spread widely over a broad rising strip the correlation is weakly positive.

vi) If the points are spread widely over a broad falling strip, the correlation is weakly negative.

vii) If the points are scattered without any specific pattern, the correlation is absent, i.e. $\rho = 0$.

Example for a scatter plot
Figure 5.11 shows the scatter plot of the data presented in table 5.8. It can easily be seen that the data is highly negatively correlated.

5.7.2 Linear Regression

Linear Regression is an approach for modeling the relationship between a scalar *dependent variable* Y and one or more *explanatory variables* denoted X. The case of one explanatory variable is called *simple regression*. More than one explanatory variable is called *multiple regression*.

In linear regression, data is modeled using *linear predictor functions*. Such functions are linear functions with a set of coefficients and explanatory variables, whose value is used to predict the outcome of a dependent variable Y. They can be denoted as

$$\hat{y} = f(x).$$

Ordinary least squares

A large number of procedures have been developed for parameter estimation and inference in linear regression. In this text we focus on the basic and most important model of *ordinary least squares estimation* (also referred to as OLS). It is a simple and computationally straightforward method.

The ordinary least squares method minimizes the sum of squared residuals, formally

$$\sum_{i=1}^{n} (y_i - \hat{y})^2 \longrightarrow \text{min!}$$

If the data X consists of only one explanatory variable, then the linear predictor function is defined by a *constant* and a *scalar multiplier* for x_i. This is called the *simple regression model*.

The vector of parameters in such model is 2-dimensional, and is commonly denoted as (α, β):

$$\hat{y}_i = \alpha + \beta \cdot x_i.$$

The coefficients α and β can be obtained by solving the following system of linear equations

$$\begin{array}{rcrcr} n \cdot \alpha & + & \beta \sum_{i=1}^{n} x_i & = & \sum_{i=1}^{n} y_i \\ \alpha \sum_{i=1}^{n} x_i & + & \beta \sum_{i=1}^{n} x_i^2 & = & \sum_{i=1}^{n} x_i \cdot y_i \end{array}$$

Alternatively the coefficients can be calculated directly by using the formulas

$$\beta = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n} \left(\sum x_i\right)^2} = \frac{\text{Cov}(x, y)}{\text{Var}(x)} \quad \text{and} \quad \alpha = \overline{y} - \beta \overline{x}.$$

Coefficient of determination

We can further obtain the *coefficient of correlation* R of a linear regression model as

$$R = \frac{n \sum x_i y_i - \sum x_i \cdot \sum y_i}{\sqrt{\left(n \sum x_i^2 - (\sum x_i)^2\right) \cdot \left(n \sum y_i^2 - (\sum y_i)^2\right)}}, \quad -1 \leq R \leq 1.$$

The *coefficient of determination* is then

$$R^2, \quad 0 \leq R^2 \leq 1.$$

Formally, the coefficient of determination is defined as

$$R^2 = \frac{SSR}{SST},$$

where SSR denotes the *sum of square due to regression*

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

and SST denotes the *total sum of squares*

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

The coefficient of determination R^2 can be interpreted as a measure for the goodness of fit of the model. Values close to one indicate a good model fit.

Applications of linear regression

Linear regression has many practical uses. Most applications of linear regression fall into one of the following two broad categories:

1. If the goal is prediction, or forecasting, linear regression can be used to fit a predictive model to an observed data set of Y and X values. After developing such a model, if an additional value of X is then given without its accompanying value of Y, the fitted model can be used to make a prediction of the value of Y.

2. Given a variable Y and a number of variables $X_1, \ldots, X_p$ that may be related to Y, linear regression analysis can be applied to quantify the strength of the relationship between Y and the X_j, to assess which X_j may have no relationship with Y at all and to identify which subsets of the X_j contain redundant information about Y.

Example of an OLS calculation

The simple regression model can be used to explain the ≪change in Austrian unemployment rate≫ with the ≪change in the real Austrian GDP≫. Both time series are given in table 5.8. We search for α, β such that

$$\text{DeltaUnemployment}_i = \alpha + \beta \cdot \text{GrowthGDP}_i + \varepsilon_i \quad \sum_{i=1}^{n} \varepsilon_i^2 \to \min .$$

To use the formulas from the previous section we denote the change in the Austrian unemployment rate as y_i and the change in the real Austrian GDP as x_i. We get the sums

$$\sum_{i=1}^{n} x_i = 13.2 \quad \text{and} \quad \sum_{i=1}^{n} y_i = 0.2,$$

$$\sum_{i=1}^{n} x_i \cdot y_i = -8.96 \quad \text{and} \quad \sum_{i=1}^{n} x_i^2 = 54.98.$$

This leads to the following linear system to be solved

$$\begin{array}{rcrcr} 11.00 \cdot \alpha & + & 13.20 \cdot \beta & = & 0.20 \\ 13.20 \cdot \alpha & + & 54.98 \cdot \beta & = & -8.96 \end{array}$$

Solutions to the linear system are

$$\alpha = 0.3002 \quad \text{and} \quad \beta = -0.2351.$$

The linear predictor function is

$$\hat{y}_i = \alpha + \beta \cdot x_i = 0.3002 - 0.2351 \cdot x_i.$$

Based on this predictor function we see that unemployment rises for approximately 0.3 if the real GDP doesn't grow. Furthermore one can state that the unemployment rate gets reduced by 0.235 for every percent of real Austrian GDP growth.

Table 5.10 shows the calculated predictions $\hat{y}_i$ in the forth column and the two square sums needed to calculation the coefficient of determination in the last two columns. For the sum of squares due to regression we get

$$SSR = 2.162$$

and the total sum of squares equals

$$SST = 3.256.$$

Therefore the coefficient of determination is

$$R^2 = \frac{SSR}{SST} = \frac{2.162}{3.256} = 0.6641.$$

This can be interpreted as:

year	GrowthGDP as x_i	DeltaUnemployment as y_i	$\hat{y}_i$	$(\hat{y}-\overline{y})^2$	$(y-\overline{y})^2$
2007	3.7	-0.4	-0.569	0.345	0.175
2008	1.5	-0.8	-0.052	0.005	0.669
2009	-3.8	1.2	1.193	1.381	1.397
2010	1.8	-0.5	-0.123	0.020	0.269
2011	2.9	-0.2	-0.381	0.160	0.048
2012	0.7	0.3	0.136	0.014	0.079
2013	0.0	0.5	0.300	0.080	0.232
2014	0.7	0.2	0.136	0.014	0.033
2015	1.1	0.1	0.042	0.001	0.007
2016	2.0	0.3	-0.170	0.035	0.079
2017	2.6	-0.5	-0.311	0.108	0.269
$\sum$				2.162	3.256

Table 5.10: Calculation of SSR and SST for the growth in real Austrian GDP and the change in the Austrian unemployment rate.

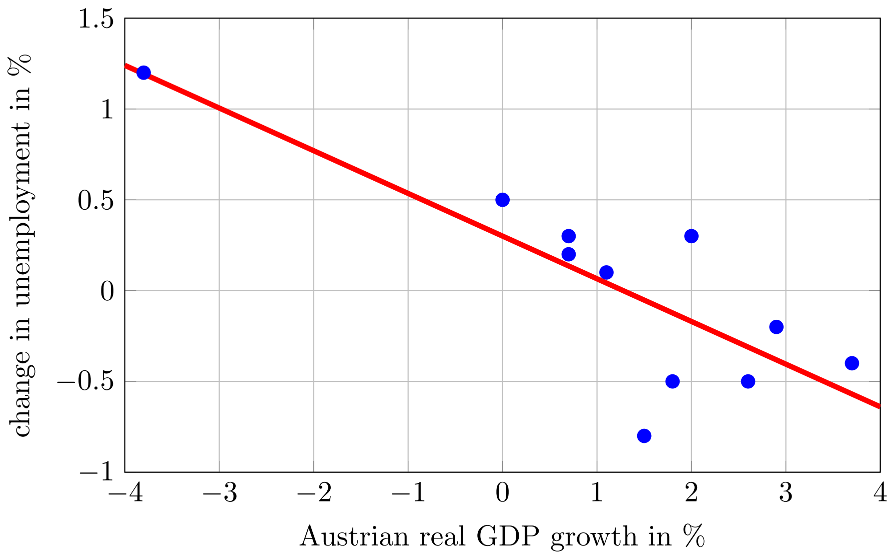

Figure 5.12: Simple linear regression explaining the change in the Austrian unemployment rate by the change of the real Austrian GDP.

> About 66% of the variance of the change in the Austrian unemployment rate can be explained by the variance of the real Austrian GDP growth.

Figure 5.12 shows a scatter plot of the change the Austrian unemployment rate according to ILO standards and the real Austrian GDP growth. The resulting linear regression line is shown in red.

6 Random variables

As opposed to other mathematical variables, a *random variable* conceptually does not have a single fixed value. Rather it can take on a set of possible different values, each with an associated probability. In probability theory, random variables are defined in terms of functions on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$.

A probability space consists of three parts.

1. Ω is the sample space, which is the set of all possible outcomes.

2. $\mathcal{F}$ is the sigma algebra on the set of possible outcomes. It is the set of all measurable subsets of Ω, thus all possible events. An event is a set containing zero or more possible outcomes.

    - If the sample space Ω is a finite set, the set $\mathcal{F}$ is a subset of the power set of Ω, $\mathcal{F} \subseteq \mathcal{P}(\Omega)$.

    - For an uncountable infinite sample space (for example $\mathbb{R}$) the sigma algebra $\mathcal{F}$ is just somewhat similar to the power set $\mathcal{P}(\Omega)$.

3. $\mathbf{P}$ assigns the probability to the events. Therefore $\mathbf{P}$ is a function from $\mathcal{F}$ to probability levels,

$$\begin{aligned} \mathbf{P} : \mathcal{F} &\to [0,1] \\ E &\mapsto \mathbf{P}(E) \end{aligned}$$

Example
Let Ω be the sample space of all humans. The random variable X assigns each human to the measurable attribute height. We have

$$\begin{aligned} X : \Omega &\to \mathbb{R} \\ \omega &\mapsto \ll\text{height of the person}\gg \end{aligned}$$

We see that a random variable is a function wit the set of all possible outcomes as domain (compare section 2.1). Random variables are typically classified as either *discrete random variables* or *continuous random variables*.

Discrete variables can take on either a finite or at most a countably infinite set of discrete values. Their probability distribution is given by a *probability mass function* which maps a value of the random variable to a probability.

Continuous variables take on values that vary continuously within one or more (possibly infinite) intervals. There are an uncountably infinite number of individual outcomes, and each has a probability 0. The probability distribution for a continuous random variable is defined using a *probability density function*. It indicates the «density» of the probability of the value of a random variable being in a small neighborhood around a given value. More technically, the probability that the value of the random value is in a particular range is derived from the definite integral of the probability density function for that range.

Both concepts can be united using a cumulative distribution function (CDF) $F(x)$, which describes the probability that the value of an outcome will be less than or equal to a specified value.

Examples for possible discrete and continuous random variables
Discrete random variables could be

- The number of cars passing a checkpoint in an hour.
- The number of home mortgages approved by a bank per year.
- The received points in an exam.

Continuous random variables could be

- The time it takes to make breakfast.
- The exact length of a car.
- The rate of return of a certain stock within one year.

6.1 Cumulative distribution function

In statistics the *cumulative distribution function (CDF)* $F_X(x)$ describes the probability that the measured outcome of a random variable X with a given probability distribution will be found at a value less or equal to x (similar to the empirical cumulative distribution function introduced in section 5.4).

Formally, for every number x, the cumulative distribution function of a random variable X is given by

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

where the right-hand side represents the *probability* that the random variable X takes on a value less or equal to x. The probability that X lies in the interval $]a, b]$, where $a < b$, is

$$\mathbf{P}(\{\omega \in \Omega : a < X(\omega) \leq b\}) = F_X(b) - F_X(a).$$

If treating several random variables $X, Y, \ldots$ etc. the corresponding letters are used as subscripts while, if treating only one, the subscript is omitted. For this script we use bold F for cumulative distribution functions, in contrast to lower-case f used for probability density functions and probability mass functions.

6.2 Probability mass functions for discrete random variables

If X is a *discrete random variable*, then the function

$$f_X(x) = \mathbf{P}(\{\omega \in \Omega : X(\omega) = x\})$$

defined on the measured outcomes of X is called the *probability mass function (PMF)* of the discrete random variable X.

Suppose that

$$X : \Omega \to A \qquad A \subseteq \mathbb{N}$$

is a discrete random variable defined on a sample space Ω and a natural number as measured outcome. Then the probability mass function

$$f_X : \mathbb{N} \to [0,1]$$

for X is defined as

$$f_X(x) = \mathbf{P}(X = x) = \mathbf{P}(\{\omega \in \Omega : X(\omega) = x\}).$$

Note that f_X is defined for all numbers, including those not in the image of X. It holds that $f_X(x) = 0$ for all $x \notin X(\Omega)$. The sum of probabilities across all x must equal 1,

$$\sum_{x \in \Omega} f_X(x) = 1.$$

For a a discrete random variable X which attains values $x_1, x_2, \ldots, x_n$ with probabilities $\mathbf{P}(\{\omega \in \Omega : X(\omega) = x_i\}) = p_i$ for all $i = 1, \ldots, n$, the cumulative distribution function of X will be discontinuous at points x_i and constant in between:

$$\begin{aligned} F_X(x) &= \mathbf{P}(\{\omega \in \Omega : X(\omega) \leq x\}) \\ &= \sum_{x_i \leq x} \mathbf{P}(\{\omega \in \Omega : X(\omega) = x_i\}) \\ &= \sum_{x_i \leq x} p_i. \end{aligned}$$

Example for a discrete random variable
Suppose that Ω is the sample space of all outcomes of a single toss of a fair coin and X is the random variable defined on Ω assigning 0 to ≪tails≫ and 1 to ≪heads≫. Since the coin is fair, the probability mass function is

$$f_X(x) = \begin{cases} \frac{1}{2} & x \in \{0,1\} \\ 0 & x \notin \{0,1\}. \end{cases}$$

Therefore the variable X has the outcomes $x_1 = 0$ and $x_2 = 1$ and we can write the probability mass function in tabular form as

	$x_1 = 0$	$x_2 = 1$	$x \notin \{0,1\}$
$\mathbf{P}(\{\omega \in \Omega : X(\omega) = x\}) = f_X(x)$	$1/2$	$1/2$	0

The cumulative distribution function of the random variable X is

$$F_X(x) = \begin{cases} 0 & \text{if} \quad -\infty < x < 0 \\ 1/2 & \text{if} \quad 0 \leq x < 1 \\ 1 & \text{if} \quad x \geq 1. \end{cases}$$

This is a special case of the binomial distribution (compare section 6.6.2).

6.3 Probability density functions for continuous random variables

A continuous random variable is one which can take a continuous range of values, as opposed to a discrete distribution (see previous section), where the set of possible values is at most countable. Note that for continuous variables the probability that the random variable attains any distinct value *always equals zero!* Merely the probability that the outcome will fall into an interval is nonzero.

For a continuous random variable X the *probability density function* $f_X(x)$ describes the behavior of the random variable. The cumulative distribution function of such a continuous random variable X, denoted $F_X(x)$, can be calculated using its probability density function f_X as follows:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,\mathrm{d}u.$$

Equivalently, if the continuous random variable X has a differentiable cumulative distribution function $F_X(x)$, then

$$F_X(x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u,$$

and the function

$$f_X(x) = \frac{\mathrm{d}F_X}{\mathrm{d}x}(x)$$

is the *probability density function* of X.

The probability of X falling into a given interval, say $[a, b]$ is given by the integral

$$\mathbf{P}(\{\omega \in \Omega : a \leq X(\omega) \leq b\}) = \int_{a}^{b} f_X(x)\,\mathrm{d}x.$$

Here we see that the probability for X taking any single value a, that is the stochastic event

$$\{\omega \in \Omega : a \leq X(\omega) \leq a\},$$

is zero, because an integral with coinciding upper and lower limits always equals zero.

6.4 Expected value of a random variable

The *expected value* of a random variable is the weighted average of all possible values that this random variable can take on. The weights used in computing this average correspond to the *probabilities* in case of discrete random variables, or *densities* in case of continuous variables.

6.4.1 Expected value for a discrete random variable

Suppose the random variable X takes value x_1 with probability p_1, value x_2 with probability p_2, and so on, up to value x_n with probability p_n. The *expected value* of this random variable X is defined as

$$\mathbf{E}(X) = x_1 \cdot p_1 + x_2 \cdot p_2 + \ldots + x_n \cdot p_n = \sum_{i=1}^{n} x_i \cdot p_i.$$

Since all probabilities p_i add up to 1 ($\sum_{i=1}^{n} p_i = 1$) the expected value can be viewed as the weighted average, with p_i's being the weights.

If all outcomes x_i are equally likely (that is $p_1 = p_2 = \ldots = p_n$[14]), then the weighted average becomes a simple arithmetic average. For a fair dice we have

$$\mathbf{E}(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

The expected value of a discrete random variable with countable infinite possible outcomes x_i for all $i \in \mathbb{N}$ only exists if

$$\sum_{i=1}^{\infty} |x_i| \cdot p_i < \infty.$$

The generalized expected value of the form $\mathbf{E}(g(X))$ for any function $g(X)$ can be calculated as

$$\mathbf{E}(g(X)) = \sum_{i=1}^{\infty} g(x_i) \cdot p_i.$$

6.4.2 Expected value for a continuous random variable

If the probability distribution of X has the probability density function $f_X(x)$, the expected value can be computed as

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} x \cdot f_X(x) \, \mathrm{d}x.$$

Similarly to the discrete case, the expected value of an continuous random variable only exists if

$$\int_{-\infty}^{\infty} |x| \cdot f(x) \, \mathrm{d}x < \infty.$$

If this holds, we can calculate the expected value of an arbitrary function of X, say $g(X)$. With respect to the probability density function $f_X(x)$ the expected value of $g(X)$ is given by the inner product of f_X and g

$$\mathbf{E}(g(X)) = \int_{-\infty}^{\infty} g(x) \cdot f_X(x) \, \mathrm{d}x.$$

Example for the calculation of the expected value of a continuous random variable with given probability density function

[14] E.g. probability a certain value when of tossing a fair dice always equals $1/6$.

Suppose, we have a continuous random variable X with a probability density function

$$f_X : \mathbb{R} \to \mathbb{R}$$
$$x \mapsto \begin{cases} 0, & x \notin [0,3] \\ -\frac{2}{9}x + \frac{2}{3}, & x \in [0,3]. \end{cases}$$

Then the expected value $\mathbf{E}(X)$ can be calculated as

$$\begin{aligned}
\mathbf{E}(X) &= \int_{-\infty}^{\infty} x \cdot f_X(x)\,\mathrm{d}x = \int_0^3 x\left(-\frac{2}{9}x + \frac{2}{3}\right)\mathrm{d}x \\
&= \int_0^3 \left(-\frac{2}{9}x^2 + \frac{2}{3}x\right)\mathrm{d}x \\
&= -\frac{2}{9} \cdot \frac{x^3}{3} + \frac{2}{3} \cdot \frac{x^2}{2}\bigg|_0^3 \\
&= -\frac{2}{9} \cdot \frac{27}{3} + \frac{9}{3} = 1.
\end{aligned}$$

6.4.3 Properties of the expected value

There are three major properties of the expected value.

Constants The expected value of a constant is equal to the constant itself. For example, if c is a constant, then $\mathbf{E}(c) = c$.

Monotonicity If X and Y are random variables such that $X \leq Y$ *almost surely*, then

$$\mathbf{E}(X) \leq \mathbf{E}(Y).$$

Linearity The expected value $\mathbf{E}$ is linear in the sense that

$$\begin{aligned}
\mathbf{E}(X + c) &= \mathbf{E}(X) + c \\
\mathbf{E}(X + Y) &= \mathbf{E}(X) + \mathbf{E}(Y) \\
\mathbf{E}(a \cdot X) &= a \cdot \mathbf{E}(X)
\end{aligned}$$

where X, Y are random variables defined on the same probability space and $a, c \in \mathbb{R}$.

6.5 Variance of a random variable

The *variance* of a probability distribution is a measure of how far a set of values is spread out (compare section 5.6.3). It is one of several descriptors of a probability distribution, describing how far the values lie away from the expected value (e.g. the mean).

The variance is a parameter describing in part either the actual probability distribution of an observed sample of numbers, or the theoretical probability distribution of a population of numbers. An estimate for the variance for the theoretical probability distribution can be calculated using a sample of data from such a distribution. In the simplest case this estimate can be the *sample variance*, defined in section 5.6.3.

If a random variable X has the expected value $\mu = \mathbf{E}(X)$, then the variance of X is given by

$$\mathrm{Var}(X) = \mathbf{E}\left((X-\mu)^2\right).$$

Here it can be seen that the variance is the expected value of the squared difference between the variable's realisation and the variable's expected value.

This definition can be expanded using the properties of variance and we get:

$$\begin{aligned}\mathrm{Var}(X) &= \mathbf{E}\left((X-\mu)^2\right) = \mathbf{E}(X^2 - 2\mu X + \mu^2) \\ &= \mathbf{E}(X^2) - 2\mu\mathbf{E}(X) + \mu^2 = \mathbf{E}(X^2) - 2\mu^2 + \mu^2 \\ &= \mathbf{E}(X^2) - \mu^2 = \mathbf{E}(X^2) - (\mathbf{E}(X))^2.\end{aligned}$$

A mnemonic for the expression above is «mean of the square minus square of the mean». The variance of a random variable X is typically designated as $\mathrm{Var}(X)$, σ_X^2, or simply σ^2. The *standard deviation*, denoted by σ, is defined as

$$\sigma = +\sqrt{\sigma^2}.$$

6.5.1 Variance of a discrete random variable

If the random variable X is discrete with *probability mass function* $x_1 \mapsto p_1, \ldots, x_n \mapsto p_n$, then

$$\mathrm{Var}(X) = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2$$

where μ is the expected value, i.e.

$$\mu = \sum_{i=1}^{n} p_i \cdot x_i.$$

That is the expected value of the square of the deviation of X from its own mean. It can be expressed as «the mean of the squares of the deviations of the data points from the average». It is thus the *mean squared deviation.*

Example for the calculation of the variance of rolling an unbiased die
The variance of the discrete random variable describing rolling an unbiased die is

$$\mathrm{Var}(X) = (1-3.5)^2 \cdot \frac{1}{6} + (2-3.5)^2 \cdot \frac{1}{6} + \ldots + (6-3.5)^2 \cdot \frac{1}{6} \approx 2.92$$

6.5.2 Variance of a continuous random variable

If the random variable X is continuous with *probability density function* $f_X(x)$, then the variance equals the second central moment, given by

$$\mathrm{Var}(X) = \int (x-\mu)^2 \cdot f_X(x)\,\mathrm{d}x,$$

where μ is the expected value,

$$\mu = \int x \cdot f_X(x)\,\mathrm{d}x,$$

and where the integrals are definite integrals for x ranging over the full range of X.

Note that not all distributions have a variance. If a continuous distribution does not have an expected value, as is the case for the Cauchy distribution, it does not have a variance either. Many other distributions for which the expected value does exist also do not have a variance because the integral in the variance definition diverges.

Example for the calculation of the variance of an arbitrary random variable with given probability density function
The variance of the random variable X with the probability density function

$$\begin{aligned} f_X &: \mathbb{R} \to \mathbb{R} \\ x &\mapsto \begin{cases} 0, & x \notin [0,3] \\ -\frac{2}{9}x + \frac{2}{3}, & x \in [0,3]. \end{cases} \end{aligned}$$

is

$$\begin{aligned} \mathrm{Var}(X) &= \int_0^3 (x-1)^2 \left(-\frac{2}{9}x + \frac{2}{3}\right) \mathrm{d}x \\ &= \int_0^3 \left(-\frac{2}{9}x^3 + \frac{10}{9}x^2 - \frac{14}{9}x + \frac{2}{3}\right) \mathrm{d}x \\ &= -\frac{2}{9} \cdot \frac{x^4}{4} + \frac{10}{9} \cdot \frac{x^3}{3} - \frac{14}{9} \cdot \frac{x^2}{2} + \frac{2}{3}x \Big|_0^3 \\ &= \frac{1}{2}. \end{aligned}$$

6.5.3 Properties of the variance

1. Variance is non-negative because the squares are positive or zero,

$$\mathrm{Var}(X) \geq 0.$$

2. The variance of a constant random variable is zero; and if the variance of a variable in a data set is 0, then all entries have the same value,

$$\mathbf{P}(\{\omega \in \Omega : X(\omega) = a\}) = 1 \quad \Longleftrightarrow \quad \mathrm{Var}(X) = 0.$$

3. Variance is invariant respect to changes in a location parameter. That is, if a constant is added to the values of the variable, the variance is unchanged,

$$\mathrm{Var}(X + a) = \mathrm{Var}(X).$$

4. If all values are scaled by a constant, the variance is scaled by the square of that constant,

$$\mathrm{Var}(a \cdot X) = a^2 \cdot \mathrm{Var}(X).$$

5. The variance of a sum of two random variables is given by

$$\begin{aligned}\mathrm{Var}(a \cdot X + b \cdot Y) &= a^2 \cdot \mathrm{Var}(X) + b^2 \cdot \mathrm{Var}(Y) + 2ab \cdot \mathrm{Cov}(X, Y)\\ \mathrm{Var}(X - Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X, Y)\end{aligned}$$

The variance of a finite sum of *uncorrelated* random variables is equal to the sum of their variances. This stems from the identity above and the fact that the covariance is zero for uncorrelated variables.

6.6 Important probability distributions

6.6.1 Hypergeometric distribution

In statistics, the *hypergeometric distribution* is a discrete probability distribution that describes the probability of k successes in n draws from a finite population of size N containing m successes *without replacement*. Formally, a random variable X follows the hypergeometric distribution if its probability mass function is given by

$$f(k; N, m, n) = \mathbf{P}(\{\omega \in \Omega : X(\omega) = k\}) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}},$$

where

- N is the population size
- m is the number of success states in the population
- n is the number of draws
- k is the number of successes in the draws
- $\binom{a}{b}$ is the binomial coefficient[15]

It is positive when

$$\max(0, n + m - N) \leq k \leq \min(m, n).$$

[15] The binomial coefficient is defined as

$$\binom{a}{b} = \frac{a!}{b!(a-b)!} \quad \text{for } 0 \leq b \leq a.$$

Expected value The expected value of the hypergeometric distribution is

$$\mathbf{E}(X) = n\frac{m}{N}.$$

Variance The variance of the hypergeometric distribution is

$$\mathrm{Var}(X) = n\frac{m}{N}\left(1 - \frac{m}{N}\right)\frac{N-n}{N-1}.$$

Example for a hypergeometric distribution
The classical application of the hypergeometric distribution is *sampling without replacement*. Think of an urn with two types of marbles, black ones and white ones. Define drawing a white marble as a success and drawing a black marble as a failure. If the variable N describes the number of all marbles in the urn and m describes the number of white marbles, then $N - m$ corresponds to the number of black marbles. In this example X is the random variable whose outcome is k, the number of white marbles actually drawn in the experiment.

Suppose, there are 5 white and 45 black marbles in the urn. Standing next to the urn, you close your eyes and draw 10 marbles without replacement. What is the probability that exactly 4 of the 10 are white?

We get

$$\mathbf{P}(\{\omega \in \Omega : X(\omega) = 4\}) = \frac{\binom{5}{4}\binom{50-5}{10-4}}{\binom{50}{10}} = \frac{5 \cdot 8145060}{10272278170} = 0.00396 = 0.396\%.$$

6.6.2 Binomial distribution

The *binomial distribution* is the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success with probability p. The binomial distribution is frequently used to model the number of successes in a sample of size n drawn *with replacement* from a population size N.

In general, if a random variable X follows the binomial distribution with the parameters n and p, we write $X \sim \mathrm{B}(n, p)$. The probability of getting exactly k successes in n is given by the probability mass function

$$f(k; n, p) = \mathbf{P}(\{\omega \in \Omega : X(\omega) = k\}) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{for } k = 0, 1, 2, \ldots, n.$$

Figure 6.1 shows the probability mass function for different values for p, $n = 100$ and $k = 0, \ldots, n$.

Expected value The expected value of a binomial distributed random variable $X \sim \mathrm{B}(n, p)$ is

$$\mathbf{E}(X) = n \cdot p.$$

Variance The variance is

$$\mathrm{Var}(X) = n \cdot p \cdot (1-p).$$

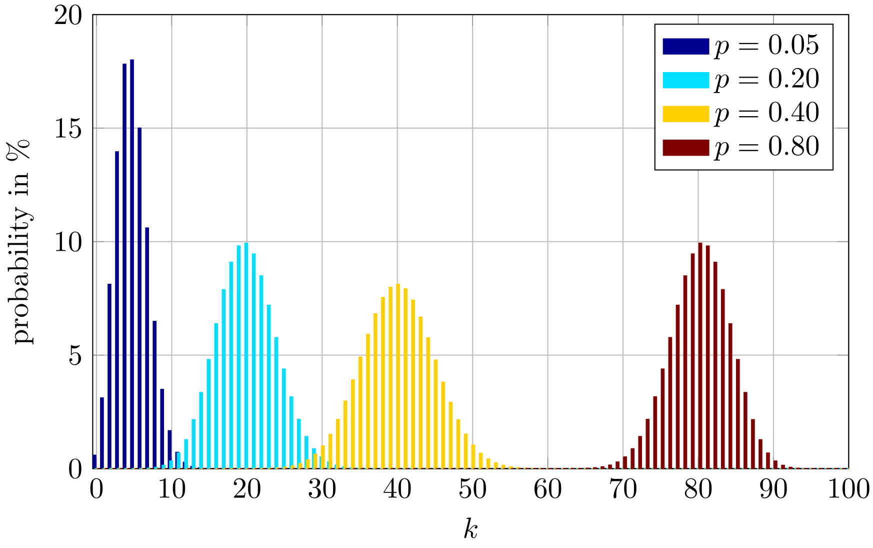

Figure 6.1: Probability mass functions of the binomial distribution with different values for p, $n = 100$ and $k = 0, \ldots, n$.

Example for a binomial distribution
A box contains 25 items, 10 of which are defective. A sample of two items *with replacement* will be taken. Then the probability, that one of the items are defective is

$$\mathbf{P}(\{\omega \in \Omega : X(\omega) = 1\}) = \binom{2}{1} \cdot \left(\frac{10}{25}\right)^1 \cdot \left(\frac{15}{25}\right)^{2-1} = \frac{12}{25} = 0.48 = 48\%.$$

6.6.3 Poisson distribution

The *Poisson distribution* is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and *independently* of the time since the last event. The distribution was first introduced by Siméon Denis Poisson (1781-1840). His work focused on certain random variables N that count the number of discrete occurrences that take place during a time-interval of given length.

A discrete stochastic variable X is said to have a Poisson distribution with parameter $\lambda > 0$ if the probability mass function of X for $k = 0, 1, 2, \ldots$ is given by

$$f(k, \lambda) = \mathbf{P}(\{\omega \in \Omega : X(\omega) = k\}) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}.$$

Figure 6.2 shows the probability mass function for different values for λ and $k = 0, \ldots, 20$.

For a Poisson distributed random variable $X \sim \text{Pois}(\lambda)$, the positive real number $\lambda \in \mathbb{R}$ is equal to the *expected value* of X and also to the variance

$$\lambda = \mathbf{E}(X) = \text{Var}(X).$$

The Poisson distribution can be applied to systems with a large (theoretically infinite) number of possible events, each of which is rare.

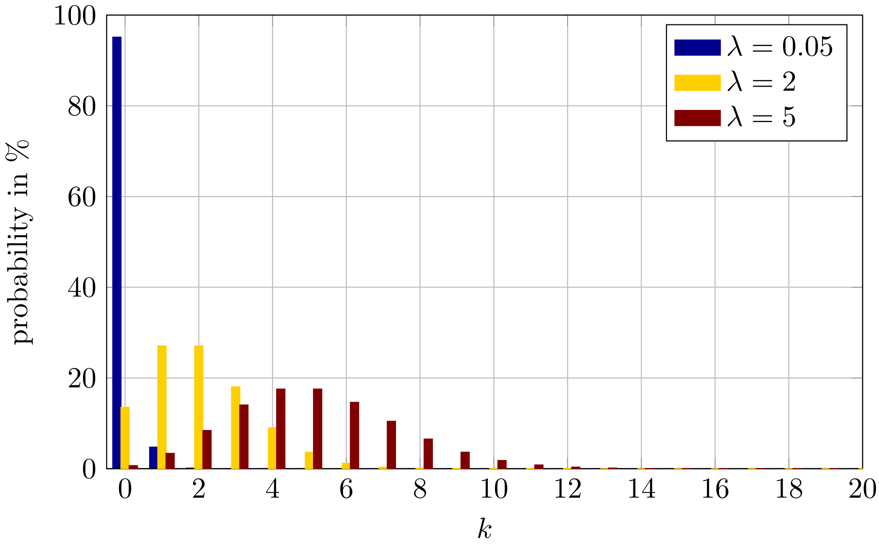

Figure 6.2: Probability mass functions of the Poisson distribution with different values for λ.

Example for a Poisson distribution
Assume that the number of defaults per year on a large set of loans will be measured as no defaults, one default, two defaults and so on. In average the number of defaults per year is 0.3. Then, the probability that there are two defaults in one year is

$$\mathbf{P}(\{\omega \in \Omega : X(\omega) = 2\}) = \frac{0.3^2 e^{-0.3}}{2!} = \frac{0.066673}{2} = 0.03333.$$

Therefore the probability of two defaults in one year is 3.33%.

6.6.4 Normal distribution

In 1809 Carl Friedrich Gauss published his monograph ≪Theoria motus corporum coelestium in sectionibus conicis solem ambientium≫ where among other things he introduced several important statistical concepts, such as the method of least squares, the method of maximum likelihood and the *normal distribution*.

In probability theory, the *normal distribution* (or *Gaussian distribution*) is a continuous probability distribution that has a bell-shaped probability density function, known as the Gaussian function. A continuous variable X is normally distributed if its probability density function is of the form

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Figure 6.3 shows the probability density function of the normal distribution for different values for μ and σ^2. Sometimes it is denoted $\varphi_{\mu,\sigma^2}(x)$.

For a normal distributed continuous random variable $X \sim \mathrm{N}(\mu, \sigma^2)$, the parameter μ is the mean or expected value,

$$\mathbf{E}(X) = \mu.$$

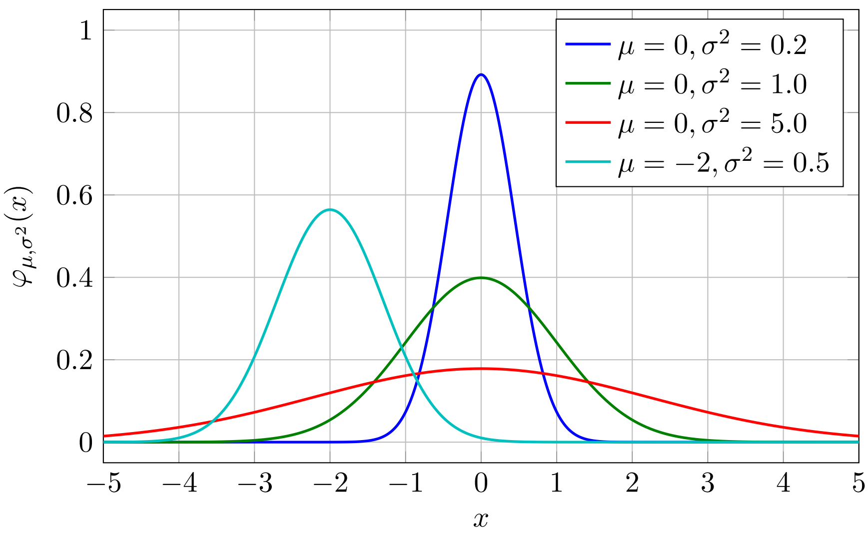

Figure 6.3: Probability density function $\varphi_{\mu,\sigma^2}(x)$ of normally distributed random variables with different values for μ and σ^2.

The second parameter σ^2 is its variance,

$$\text{Var}(X) = \sigma^2.$$

σ is known as the standard deviation. The distribution with $\mu = 0$ and $\sigma^2 = 1$ is called the *standard normal distribution*.

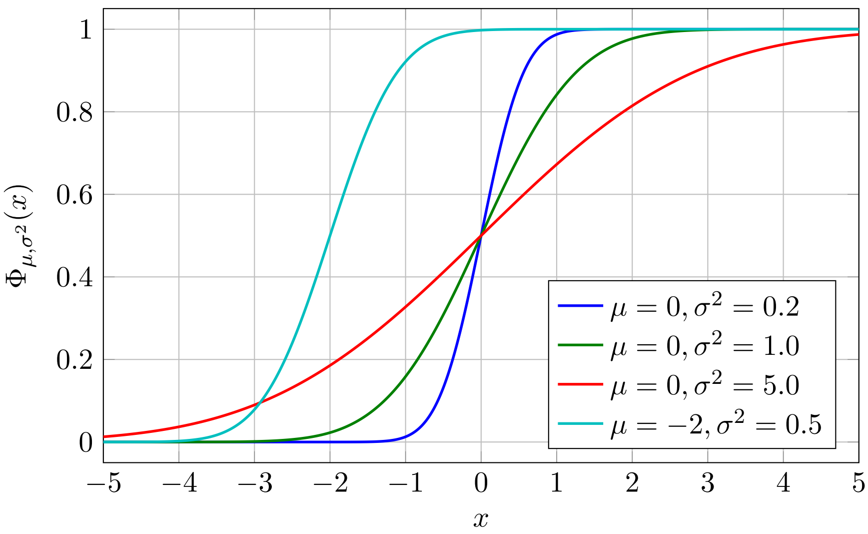

Figure 6.4: Cumulative distribution functions $\Phi_{\mu,\sigma^2}(x)$ of normally distributed random variables with different values for μ and σ^2.

The normal distribution is considered the most prominent probability distribution in statistics. There are several reasons for this:

1. The normal distribution arises from the *central limit theorem*, which states that under mild conditions the mean of a large number of random variables drawn from the same distribution is distributed approximately normally, *irrespective* of the form of the original distribution.

2. The normal distribution is very tractable analytically. This means, that a large number of results involving this distribution can be derived in *explicit* form.

There are several *rules of the thumb* describing properties of the normal distribution.

- About 68% of the values lie within 1 standard deviation of the mean.
- Similarly, about 95% of the values lie within 2 standard deviations of the mean.
- Nearly all (99.7%) of the values lie within 3 standard deviations. In figure 6.5 these levels are shown graphically.

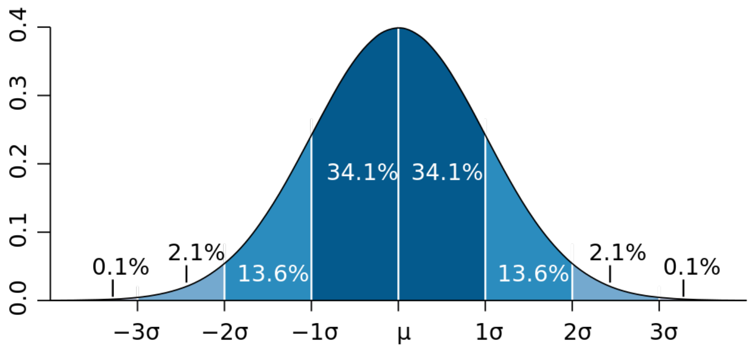

Figure 6.5: Empirical rules of a normally distributed random variable shown with its density function.

6.6.5 Standardized normal distribution

As stated above, the normal distribution with $\mu = 0$ and $\sigma^2 = 1$ is called the *standard normal distribution*. Every normally distributed random variable X with known parameters for μ and σ^2 can be transformed into a random variable Z with standard normal distribution by the linear transformation

$$Z = \frac{X - \mu}{\sigma},$$

where μ is the mean of the population and σ is the standard deviation of the population. Here, Z represents the distance between the raw score and the population mean in units of the standard deviation. Z is negative where the raw score is below the mean, positive when above.

Standardization can be used to calculate *prediction intervals* for a normally distributed random variable X with known mean and variance. A prediction interval is an estimate of an interval in which future observations will fall with a certain probability, given what has already been observed. A prediction interval $[l, u]$ for a future observation of the random variable $X \sim N(\mu, \sigma^2)$ with known mean and variance can be calculated from

$$\begin{aligned}\gamma &= \mathbf{P}(\{\omega \in \Omega : l < X(\omega) < u\}) \\ &= \mathbf{P}\left(\left\{\omega \in \Omega : \frac{l-\mu}{\sigma} < \frac{X(\omega)-\mu}{\sigma} < \frac{u-\mu}{\sigma}\right\}\right) \\ &= \mathbf{P}\left(\left\{\omega \in \Omega : \frac{l-\mu}{\sigma} < Z(\omega) < \frac{u-\mu}{\sigma}\right\}\right),\end{aligned}$$

where

$$Z = \frac{X-\mu}{\sigma},$$

is the standard score of X and is distributed according to the standard normal distribution, $Z \sim N(0,1)$.

significance γ	z *value*
50%	0.67
90%	1.64
95%	1.96
99%	2.58

Table 6.1: z-values for different levels of significance for symmetric intervals.

A prediction interval is conventionally written as

$$[\mu - z\sigma, \mu + z\sigma].$$

The most important values for z are shown in table 6.1 for symmetric intervals. For a single tail $\tilde{\gamma}$ has to be transformed as in

$$\tilde{\gamma} = 1 - \frac{1-\gamma}{2}.$$

Example for the calculation of a prediction interval for a normally distributed random variable
Assume, we want to calculate the 95% prediction interval for a normal distribution with a mean of $\mu = 5$ and a standard deviation of $\sigma = 2$.

Then z equals 1.96. Therefore, the lower limit of the prediction interval is

$$l = 5 - (2 \cdot 1.96) = 1.0801$$

and the upper limit equals

$$u = 5 + (2 \cdot 1.96) = 8.9199.$$

Thus we get the prediction interval of approximately 1 to 9, where 95% of all observations will lie in.

List of Tables

List of Figures