

MATHEMATICS (4), A30  
(Preliminary version)  
February 16, 2005  
Compiled by  
Josef Diblík

# MATHEMATICS (4), A30

Compiled by  
**Josef Diblík**

# MATHEMATICS (4), A30

*guaranteed by:* Prof. RNDr. Josef Diblík, DrSc. *and* RNDr. Veronika Chrastinová, Ph.D.

Programme in 2004/2005:

## A Elements of Probability Theory and Statistics

- 1 Random events, their probability, simple event, sample space. Probability, event, union and intersection of events. Additive rule, mutually exclusive events, complementary events, conditional probability
- 2 Probability addition and multiplication rules. Total probability formula. Baye's formula. Some counting rules (multiplicative, permutations, partitions)
- 3 Discrete and continuous random variables and probability distributions. Random variable and the law of its distribution. Mean, variance and standard deviation.
- 4,5 Chebyshev's theorem. Some special distributions: Uniform, Binomial, Poisson and Normal. Transformed random variables.
- 6,7 Statistics: basic concepts, point and interval estimator.

## B Elements of Numerical Methods

- 8 Interpolating (Newton and Lagrange polynomials, cubic splines, Least squares method).
- 9 Numerical differentiation and integration.
- 10,11 Solution of the equation  $f(x) = 0$ . Bisection method, Slope method, General iteration method. Iteration methods for the system of  $n$  linear equations and two nonlinear equations.
- 12,13 Numerical solution of Differential equations (e.g. Eulers method, finite - difference method) and partial differential equations. Eigenvalues and eigenvectors.



# Chapter 1

## ELEMENTS OF PROBABILITY THEORY AND STATISTICS

### 1.1 Sample Space

Suppose a coin is tossed once and the up face is recorded. This is an **observation**, or **measurement**. Any process of making an observation is called an **experiment**.

**Definition 1.** An **experiment** (trial) is an act or process that leads to a single outcome (event) that cannot be predicted with certainty.

The events that may occur and may not occur when the set of conditions connected with the possibility of their occurrence is realized are called **random** or **stochastic**. Random events are denoted, e.g., by the letters or numbers.

**Example 1.** Consider simple experiment consisting of tossing a die and observing the number on the up face. The six basic possible outcomes to this experiment are:

1. Observe a 1.
2. Observe a 2.
3. Observe a 3.
4. Observe a 4.
5. Observe a 5.

6. Observe a 6.

Note that if this experiment is conducted once, *you can observe one and only one of these six basic outcomes, and the outcome cannot be predicted with certainty*. Also these possibilities cannot be decomposed into more basic outcomes. The basic possible outcomes to an experiment are called **simple events**.

**Definition 2.** A **simple event** is the most basic outcome of an experiment.

**Example 2.** Two coins are tossed, and their up faces are recorded. List all the simple events for this experiment.

**Solution.** Even for a seemingly trivial experiment, we must be careful when listing the simple events. At first glance the basic outcomes seem to be:

1. Observe two heads;
2. Observe two tails;
3. Observe one head and one tail.

However, further reflection reveals that the last of these: *Observe one head and one tail*, can be decomposed into:

1. Head on coin 1, Tail on coin 2; and
2. Tail on coin 1, Head on coin 2.

Thus, the simple events are as follows:

1. Observe HH;
2. Observe HT;
3. Observe TH;
4. Observe TT,

where H in the first position means *Head on coin 1*, H in the second position means *Head on coin 2*, etc.

We often wish to refer to the collection of all simple events of an experiment. This collection is called the *sample space* of the experiment. For example, there are six simple events in the sample space associated with the die-toss experiment.

Let us discuss some examples.

**Example 3.** Observe the up face on a coin. Sample space:

1. Observe a head.
2. Observe a tail.

This sample space can be represented in set notation as a set containing two simple events:

$$S : \{H, T\}$$

where  $H$  represents the simple event *Observe a head* and  $T$  represents the simple event *Observe a tail*.

**Example 4.** Observe the up face on a die. Sample space:

1. Observe a 1.
2. Observe a 2.
3. Observe a 3.
4. Observe a 4.
5. Observe a 5.
6. Observe a 6.

This sample space can be represented in set notation as a set of six simple events:

$$S : \{1, 2, 3, 4, 5, 6\}.$$

**Example 5.** Observe the up faces on two coins. Sample space:

1. Observe HH.

2. Observe HT.
3. Observe TH.
4. Observe TT.

This sample space can be represented in set notation as a set of four simple events:

$$S : \{HH, HT, TH, TT\}.$$

**Definition 3.** The **sample space** of an experiment is the collection of all its simple events.

**Definition 4.** An **event** is a specific collection of simple events.

## 1.2 Probability

Now that we have defined simple events as the basic outcomes of the experiment and the sample space as the collection of all simple events, we are prepared to discuss the probabilities of simple events. You have undoubtedly used the term *probability* and have some intuitive idea about its meaning. Probability is generally used synonymously with “chance,” “odds,” and similar concepts. We will begin our treatment of probability using these informal concepts and then solidify what we mean later. For example, if a fair coin is tossed, we might reason that both the simple events, *Observe a head* and *Observe a tail*, have the same chance of occurring. Thus, we might state that *the probability of observing a head is 50%* or *the odds of seeing a head are 50 – 50*. Both these statements are based on an informal knowledge of probability. **The probability of a simple event is a number between 0 and 1 that measures the likelihood that event will occur when the experiment is performed.** Usually, except the simple events, we involve so called a **certain (sure)** event (i.e. an event which is sure to occur in each trial) and an **impossible event** (i.e. an event which cannot occur whatever the trial). Sometimes for impossible event  $A$  used the notation  $A = \emptyset$  and for the certain event  $A$  the notation  $A = I$ .

No matter how you assign the probabilities to simple events, the probabilities must obey two rules:

1. The probability of a certain event  $A$  is  $p(A) = 1$ .
2. The probability of an impossible event  $A$  is  $p(A) = 0$ .
3. All simple event probabilities *must* lie between 0 and 1.
4. The probabilities of all the simple events within a sample space *must* sum to 1.

Assigning probabilities to simple events is easy for some experiments. For example, if the experiment is to toss a fair coin and observe the up face, we would probably all agree to assign a probability of  $1/2$  to the simple events. *Observe a head* and *Observe a tail*. However, many experiments have simple events whose probabilities are more difficult to assign.

**Example 6.** Consider the experiment of tossing two coins. Suppose the coins are *not* balanced and the correct probabilities associated with the simple events are given in the table:

<i>Simple event</i>	<i>Probability</i>
$HH$	$4/9$
$HT$	$2/9$
$TH$	$2/9$
$TT$	$1/9$ .

Consider the events:

$A$ : {Observe exactly one head}

$B$ : {Observe at least one head}.

Calculate the probability of  $A$  and the probability of  $B$ .

**Solution.** Event  $A$  contains the simple events  $HT$  and  $TH$ . Since two or more simple events cannot occur at the same time, we can easily calculate the probability of event  $A$  by summing the probabilities of the two simple events. Thus, the probability of observing exactly one head (event  $A$ ), denoted by the symbol  $P(A)$ , is

$$P(A) = P(\text{Observe HT}) + P(\text{Observe TH}) = \frac{2}{9} + \frac{2}{9} = \frac{4}{9}.$$

Similarly, since  $B$  contains the simple events  $HH$ ,  $HT$ , and  $TH$ ,

$$P(B) = \frac{4}{9} + \frac{2}{9} + \frac{2}{9} = \frac{8}{9}.$$

**Definition 5.** The probability of an event  $A$  is calculated by summing the probabilities of the simple events in  $A$ .

The ratio of the number  $m$  of the occurrences of some random event  $A$  in a given series of trials to the total number  $n$  of the trials of that series is called the **frequency** of occurrence of the event  $A$  in the given series of trials (or simply the frequency of the event  $A$ ) and coincides with probability  $p(A) = m/n$ .

### 1.3 Unions and Intersections

An event can often be viewed as a composition of two or more events. Such events are called **compound events**; they can be formed (composed) in two ways.

**Definition 6.** The **union** of two events  $A$  and  $B$  is the event that occurs if either  $A$  or  $B$  or both occur on a single performance of the experiment. We denote the union of events  $A$  and  $B$  by the symbol  $A \cup B$ .

**Definition 7.** The **intersection** of two events  $A$  and  $B$  is the event that occurs if both  $A$  and  $B$  occur on a single performance of the experiment. We write  $A \cap B$  for the intersection of events  $A$  and  $B$ .

**Example 7.** Consider the die-toss experiment. Define the following events:

**A:** {Toss an even number},

**B:** {Toss a number less than or equal to 3}.

- a. Describe  $A \cup B$  for this experiment.
- b. Describe  $A \cap B$  for this experiment.
- c. Calculate  $P(A \cup B)$  and  $P(A \cap B)$  assuming the die is fair.

**Solution.**

- a. The union of  $A$  and  $B$  is the event that occurs if we observe either an even number, a number less than or equal to 3, or both on a single throw of the die. Consequently, the simple events in the event  $A \cup B$

are those for which  $A$  occurs,  $B$  occurs, or both  $A$  and  $B$  occur. Testing the simple events in the entire sample space, we find that the collection of simple events in the union of  $A$  and  $B$  is

$$A \cup B = \{1, 2, 3, 4, 6\}.$$

- b. The intersection of  $A$  and  $B$  is the event that occurs if we observe *both* an even number and a number less than or equal to 3 on a single throw of the die. Testing the simple events to see which imply the occurrence of *both* events  $A$  and  $B$ , we see that the intersection contains only one simple event:

$$A \cap B = \{2\}.$$

In other words, the intersection of  $A$  and  $B$  is the simple event “Observe a 2”.

- c. Recalling that the probability of an event is the sum of the probabilities of the simple events of which the event is composed, we have

$$P(A \cup B) = P(1) + P(2) + P(3) + P(4) + P(6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{5}{6},$$

and

$$P(A \cap B) = P(2) = \frac{1}{6}.$$

□

Unions and intersections can be defined for more than two events. For example, the event  $A \cup B \cup C$  represents the union of three events,  $A$ ,  $B$ , and  $C$ . This event, which includes the set of simple events in  $A$ ,  $B$ , or  $C$ , will occur if any one or more of the events  $A$ ,  $B$ , or  $C$  occurs. Similarly, the intersection  $A \cap B \cap C$  is the event that all three of the events  $A$ ,  $B$ , and  $C$  occur. Therefore,  $A \cap B \cap C$  is the set of simple events that are in all three of the events  $A$ ,  $B$ , and  $C$ .

## 1.4 The Additive Rule and Mutually Exclusive Events

**Theorem 1. (Additive rule of probability)** *The probability of the union of events  $A$  and  $B$  is the sum of the probability of events  $A$  and  $B$  minus*

the probability of the intersection of events  $A$  and  $B$ ; i.e.,

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

**Example 8.** Hospital records show that 12% of all patients are admitted for surgical treatment, 16% are admitted for obstetrics, and 2% receive both obstetrics and surgical treatment. If a new patient is admitted to the hospital, what is the probability that the patient will be admitted either for surgery, obstetrics, or both?

**Solution.** Consider the following events:

**A:** {A patient admitted to the hospital receives surgical treatment},

**B:** {A patient admitted to the hospital receives obstetrics treatment}.

Then, from the given information,

$$P(A) = 0,12, \quad P(B) = 0,16$$

and the probability of the event that a patient receives both obstetrics and surgical treatment is

$$P(A \cap B) = 0,02.$$

The event that a patient admitted to the hospital receives either surgical treatment, obstetrics treatment, or both is the union  $A \cup B$ . The probability of  $A \cup B$  is given by the additive rule of probability:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0,12 + 0,16 = 0,26.$$

Thus, 26% of all patients admitted to the hospital receive either surgical treatment, obstetrics treatment, or both.

A very special relationship exists between events  $A$  and  $B$  when  $A \cap B$  contains no simple events. In this case, we call the events  $A$  and  $B$  **mutually exclusive** events.

**Definition 8.** Events  $A$  and  $B$  are **mutually exclusive** if  $A \cap B$  contains no simple events.

**Theorem 2.** If two events  $A$  and  $B$  are mutually exclusive, then

$$P(A \cup B) = P(A) + P(B).$$

Above definitions can be generalized for the case of any finite number of events.

**Definition 9.** Events  $A_1, A_2, \dots, A_n$  are **mutually exclusive** if  $A_i \cap A_j$  contains no simple events for every  $i, j = 1, 2, \dots, n, i \neq j$ , i.e.,

$$A_i \cap A_j = \emptyset, \quad i, j = 1, 2, \dots, n, \quad i \neq j.$$

**Theorem 3.** *If events  $A_1, A_2, \dots, A_n$  are mutually exclusive then*

$$P(A_1 \cup A_2 \cup A_3 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n).$$

**Example 9.** Consider the experiment of tossing two balanced coins. Find the probability of observing at least one head.

**Solution.** Define the events:

**A:** {Observe at least one head},

**B:** {Observe exactly one head},

**C:** {Observe exactly two heads}.

Note that

$$A = B \cup C$$

and that  $B \cap C$  contains no simple events. Thus,  $B$  and  $C$  are mutually exclusive, so that

$$P(A) = P(B \cup C) = P(B) + P(C) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}.$$

## 1.5 Complementary Events

**Definition 10.** The **complement** of an event  $A$  (or opposite event, or contrary event) is the event that  $A$  does not occur - i.e., the event consisting of all simple events that are not in event  $A$ . We denote the complement of  $A$  by  $A'$ .

The sum of the probabilities of complementary events equals 1; i.e.,

$$P(A) + P(A') = 1.$$

In many probability problems it is easier to calculate the probability of the complement of the event of interest rather than the event itself. Then, since

$$P(A) + P(A') = 1$$

we can calculate  $P(A)$  by using the relationship

$$P(A) = 1 - P(A').$$

**Example 10.** Consider the experiment of tossing two fair coins. Calculate the probability of event  $A$ : {Observing at least one head} by using the complementary relationship.

**Solution.** We know that the event  $A$ : {Observing at least one head} consists of the simple events

$$A : \{HH, HT, TH\}.$$

The complement of  $A$  is defined as the event that occurs when  $A$  does not occur. Therefore,

$$A' : \{\text{Observe no heads}\} = \{TT\}.$$

Assuming the coins are balanced,

$$P(A') = P(TT) = \frac{1}{4}$$

and

$$P(A) = 1 - P(A') = 1 - \frac{1}{4} = \frac{3}{4}.$$

## 1.6 Conditional Probability

Sometimes we may wish to alter the probability of an event when we have additional knowledge that might affect its outcome. This probability is called the **conditional probability** of the event. For example, we have shown that the probability of observing an even number (event  $A$ ) on a toss of a fair die is  $1/2$ . However, suppose you are given the information that on a particular throw of the die the result was a number less than or equal to 3 (event  $B$ ). Would you still believe that the probability of observing

an even number on that throw of the die is equal to  $1/2$ ? If you reason that making the assumption that  $B$  has occurred reduces the sample space from six simple events to three simple events (namely, those contained in event  $B$ ), we get the reduced sample space. Because the simple events for the die-toss experiment are equally likely, each of the three simple events in the reduced sample space is assigned an equal *conditional probability* of  $1/3$ . Since the only even number of the three in the reduced sample space  $B$  is the number 2 and the die is fair, we conclude that the probability that  $A$  occurs *given that  $B$  occurs* is  $1/3$ . We use the symbol  $P(A | B)$  to represent the probability of event  $A$  given that event  $B$  occurs. For the die-toss example

$$P(A | B) = 1/3.$$

To get the probability of event  $A$  given that event  $B$  occurs, we proceed as follows. We divide the probability of the part of  $A$  that falls within the reduced sample space  $B$ , namely,  $P(A \cap B)$ , by the total probability of the reduced sample space, namely,  $P(B)$ . Thus, for the die-toss example with event  $A$ : {Observe a number less than or equal to 3}, we find

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{P(2)}{P(1) + P(2) + P(3)} = \frac{1/6}{3/6} = \frac{1}{3}.$$

The formula for  $P(A | B)$  is true in general:

*To find the **conditional probability** than event  $A$  occurs given that event  $B$  occurs, divide the probability that **both**  $A$  and  $B$  occur by the probability that  $B$  occurs, that is*

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

(We assume that  $P(B) \neq 0$ .) The formula adjust the probability  $A \cap B$  from its original value in the complete sample space  $S$  to a conditional probability in the reduced sample space  $B$ . If the simple events in the complete sample space are equally likely, then the formula will assign equal probabilities to the simple events in the reduced sample space, as in the die-toss experiment. If, on the other hand, the simple events have unequal probabilities, the formula will assign conditional probabilities proportional to the probabilities in the complete sample space. This is illustrated by the following example.

**Example 11.** Many medical researches have conducted experiments to examine the relationship between cigarette smoking and cancer. Let  $A$  represent the event that an individual smokes, and let  $C$  represent the event that an individual develops cancer. Therefore  $A \cap C$  is the simple event that an individual smokes and develops cancer;  $A \cap C'$  is the simple event that an individual smokes and does not develop cancer, etc. Assume that the probabilities associated with the four simple events are as shown in the table:

$A \cap C$	0,15
$A \cap C'$	0,25
$A' \cap C$	0,10
$A' \cap C'$	0,50

How can these simple event probabilities be used to examine the relationship between smoking and cancer?

**Solution.** One method of determining whether these probabilities indicate that smoking and cancer are related is to compare the conditional probability that an individual acquires cancer given that he or she smokes with the conditional probability that an individual acquires cancer given that he or she does not smoke.

First, we consider the reduced sample space  $A$  corresponding to smokers. The two simple events  $A \cap C$  and  $A \cap C'$  are contained in this reduced sample space, and the adjusted probabilities of these two simple events are the two conditional probabilities:

$$P(C | A) = \frac{P(A \cap C)}{P(A)} \quad \text{and} \quad P(C' | A) = \frac{P(A \cap C')}{P(A)} .$$

The probability of event  $A$  is the sum of the probabilities of the simple events in  $A$ :

$$P(A) = P(A \cap C) + P(A \cap C') = 0,15 + 0,25 = 0,40.$$

The the values of the two conditional probabilities in the reduced sample space  $A$  are

$$P(C | A) = \frac{0,15}{0,40} = 0,375 \quad \text{and} \quad P(C' | A) = \frac{0,25}{0,40} = 0,625.$$

These two numbers represent the probabilities that a smoker develops cancer and does not develop cancer, respectively. Notice that the conditional probabilities 0,625 and 0,375 are in the same 5 to 3 ratio as the original (unconditional) probabilities, 0,25 and 0,15. The conditional probability formula simply adjust the unconditional probabilities so that they add to 1 in the reduced sample space,  $A$ , of smokers.

In a like manner, the conditional probabilities of a nonsmoker developing cancer and not developing cancer are:

$$P(C | A') = \frac{P(A' \cap C)}{P(A')} = \frac{0,10}{0,60} = 0,167,$$

$$P(C' | A') = \frac{P(A' \cap C')}{P(A')} = \frac{0,50}{0,60} = 0,833,$$

Notice that the conditional probabilities 0,833 and 0,167 are in the same 5 to 1 ratio as the unconditional probabilities 0,5 and 0,1.

Two of the conditional probabilities give some insight into the relationship between cancer and smoking: the probability of developing cancer given that the individual is a smoker, and the probability of developing cancer given that the individual is not a smoker. The conditional probability that a smoker develops cancer (0,375) is more than twice the probability that a nonsmoker develops cancer (0,167). This does not imply that smoking *causes*, but it does suggest a pronounced link between smoking and cancer.

## 1.7 The Multiplicative Rule and Independent Events

The probability of an intersection of two events can be calculated using the **multiplicative rule**, which employs the conditional probabilities we defined in the Section 1.6 (page 12), as shown in the following example.

**Example 12.** An agriculturist, who is interested in planting wheat next year, is concerned with the following events:

$$B : \{\text{The production of wheat will be profitable}\} \quad (1.1)$$

$$A : \{\text{A serious drought will occur}\}. \quad (1.2)$$

Based on available information, the agriculturist believes that the probability is 0,01 that production of wheat will be profitable **assuming a serious drought will occur** in the same year and that the probability is 0,05 that a serious drought will occur. That is,

$$P(B | A) = 0,01 \text{ and } P(A) = 0,05.$$

Based on the information provided, what is the probability that a serious drought will occur **and** that a profit will be made? That is, find  $P(A \cap B)$ , the probability of the intersection of events  $A$  and  $B$ .

**Solution.** As you will see, we have already developed a formula for finding the probability of an intersection of two events. Recall that the conditional probability of  $B$  given  $A$  is

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

Multiplying both sides of this equation by  $P(A)$ , we obtain a formula for the probability of the intersection of events  $A$  and  $B$ . This is often called the **multiplicative rule of probability** and is given by

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

Thus

$$P(A \cap B) = 0,05 \cdot 0,01 = 0,0005.$$

The probability that a serious drought occurs *and* the production of wheat is profitable is only 0,0005. As we might expect, this intersection is a very rare event.

### Multiplicative Rule of Probability:

$$P(A \cap B) = P(A) \cdot P(B | A) = P(B) \cdot P(A | B). \quad (1.3)$$

Intersections often contain only a few simple events. In this case, the probability of an intersection is easy to calculate by summing the appropriate

simple event probabilities. However, the formula for calculating intersection probabilities plays a very important role, particularly in an area of statistics known as **Bayesian statistics** (see part 1.11).

**Example 13.** Consider the experiment of tossing a fair coin twice and recording the up face on each toss. The following events are defined:

$$A : \{\text{First toss is a head}\} \quad (1.4)$$

$$B : \{\text{Second toss is a head}\}. \quad (1.5)$$

Does **knowing** that event  $A$  has occurred affect the probability that  $B$  will occur?

**Solution.** Intuitively the answer should be **no**, since what occurs on the first toss should in no way affect what occurs on the second toss. Let us check our intuition. Recall the sample space for this experiment:

$$1. \{\text{Observe } HH\}, \quad 2. \{\text{Observe } HT\}, \quad (1.6)$$

$$3. \{\text{Observe } TH\}, \quad 4. \{\text{Observe } TT\}. \quad (1.7)$$

Each of these simple events has a probability of  $1/4$ . Thus

$$P(B) = P(HH) + P(TH) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

and

$$P(A) = P(HH) + P(HT) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

Now

$$P(B | A) = \frac{P(A \cap B)}{P(A)} = \frac{P(HH)}{P(A)} = \frac{1/4}{1/2} = \frac{1}{2}.$$

We can now see that  $P(B) = 1/2$  and  $P(B | A) = 1/2$ . Knowing that the first toss resulted in a head does not affect the probability that the second toss will be a head. The probability is  $1/2$  whether or not we know the results of the first toss. When this occurs, we say that the two events  $A$  and  $B$  are **independent**.

**Definition 11.** Events  $A$  and  $B$  are **independent** if the occurrence of  $B$  does not alter the probability that  $A$  has occurred; i.e., events  $A$  and  $B$  are independent if

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

When events  $A$  and  $B$  are independent, it is also true that

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

Events that are not independent are said to be **dependent**.

In **events**  $A$  and  $B$  are **independent**, the probability of the intersection of  $A$  and  $B$  equals the product of the probabilities of  $A$  and  $B$ ; that is,

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

The converse is also true: If

$$P(A \cap B) = P(A)P(B),$$

then events  $A$  and  $B$  are independent.

By the probability multiplication rule, the probability of intersection of  $n$  events is specified by the formula

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_2 \cap A_1) \dots P(A_n | A_1 \cap A_2 \cap \dots \cap A_{n-1}).$$

## 1.8 Collectively Independent Events

**Definition 12.** The  $n$  events  $A_1, A_2, \dots, A_n$  are said to be **collectively independent** if the probability of occurrence of each of them is not affected by the occurrence of any other events taken in arbitrary combination.

**Theorem 4.** *The probability of intersection of  $n$  collectively independent events is equal to the product of their probabilities:*

$$P(A_1 \cap A_2 \cap A_3 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2) \cdots P(A_n).$$

**Example 14.** Three marksmen do target practise. For the first marksman the probability of hitting the target is 0,75, for the second, 0,8, for the third, 0,9. Determine the probability of the three marksmen hitting the target.

**Solution.** We have

$$P(A) = 0,75, \quad P(B) = 0,8, \quad P(C) = 0,9.$$

Events  $A$ ,  $B$  and  $C$  are collectively independent. Therefore

$$P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C) = 0,75 \cdot 0,8 \cdot 0,9 = 0,54.$$

**Example 15.** Under the conditions of the previous problem determine the probability of at least one marksman hitting the target.

**Solution.** Here  $P(\bar{A}) = 1 - 0,75 = 0,25$  (the probability of the first marksman missing the target);  $P(\bar{B}) = 1 - 0,8 = 0,2$  (the probability of the second marksman missing the target);  $P(\bar{C}) = 1 - 0,9 = 0,1$  (the probability of the third marksman missing the target). Then  $P(\bar{A} \cap \bar{B} \cap \bar{C})$ , the probability of all three marksman missing the target, is determined as follows:

$$P(\bar{A} \cap \bar{B} \cap \bar{C}) = P(\bar{A}) \cdot P(\bar{B}) \cdot P(\bar{C}) = 0,25 \cdot 0,2 \cdot 0,1 = 0,005.$$

But the event contrary to the event  $\bar{A} \cap \bar{B} \cap \bar{C}$  consists in the target being hit by at least one marksman. Consequently, the sought-for probability is

$$P = 1 - P(\bar{A} \cap \bar{B} \cap \bar{C}) = 1 - 0,005 = 0,995.$$

## 1.9 Some Counting Rules

### 1. The Multiplicative Rule

Suppose that we have  $k$  elements,  $n_1$  in the first set,  $n_2$  in the second set,  $\dots$ , and  $n_k$  in the  $k$ th set. Suppose we wish to form a sample of  $k$  elements by **taking one element from each** of the  $k$  sets. The number of different samples that can be formed is the product

$$n_1 n_2 n_3 \cdots n_k.$$

**Example 16.** There are 20 candidates for three different positions,  $E_1$ ,  $E_2$ , and  $E_3$ . How many different ways could you fill this positions?

**Solution.** For this example, there are  $k = 3$  sets of elements corresponding to:

*Set 1:* Candidates available to fill position  $E_1$ .

*Set 2:* Candidates remaining (after filling  $E_1$ ) that are available to fill  $E_2$ .

*Set 3:* Candidates remaining (after filling  $E_1$  and  $E_2$ ) that are available to fill  $E_3$ .

The numbers of elements in the sets are  $n_1 = 20$ ,  $n_2 = 19$ ,  $n_3 = 18$ . Therefore, the number of different ways of filling the three positions is given by the multiplicative rule as

$$n_1 \cdot n_2 \cdot n_3 = 20 \cdot 19 \cdot 18 = 6840.$$

## 2. The Permutations Rule

Given a single set of  $N$  distinctly different elements, you wish to select  $n$  elements from the  $N$  and **arrange** them within  $n$  positions. The number of different **permutations** of the  $N$  elements taken  $n$  at a time is denoted by  $P_n^N$  and is equal to

$$P_n^N = N \cdot (N - 1) \cdot (N - 2) \cdots (N - n + 1) = \frac{N!}{(N - n)!} \quad (1.8)$$

where  $n! = n \cdot (n - 1) \cdot (n - 2) \cdots 3 \cdot 2 \cdot 1$  and is called  **$n$  factorial**. The quantity  $0!$  is defined to be equal to 1.

**Example 17.** Consider the following transportation problem. You wish to drive, in sequence, from a starting point to each of five cities, and you wish to compare the distances - and ultimately the costs - of the different routes. How many different routes would have to be compared?

**Solution.** Denote the cities as  $C_1, C_2, \dots, C_5$ . Then a route moving from the starting point to  $C_2$  to  $C_1$  to  $C_3$  to  $C_4$  to  $C_5$  would be represented as  $C_2C_1C_3C_4C_5$ . The total number of routes would equal the number of ways you could rearrange the  $N = 5$  cities in  $n = 5$  positions. This number is

$$P_n^N = P_5^5 = \frac{5!}{(5 - 5)!} = \frac{5!}{0!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{1} = 120.$$

(Recall that  $0! = 1$ .)

## 3. The Partitions Rule

There exists a single set of  $N$  distinctly different elements. You wish to partition them into  $k$  sets, with the first set containing  $n_1$  elements, the

second containing  $n_2$  elements,  $\dots$ , and the  $k$ th set containing  $n_k$  elements. The number of different partitions is

$$\frac{N!}{n_1!n_2!\dots n_k!}, \quad (1.9)$$

where

$$n_1 + n_2 + n_3 + \dots + n_k = N.$$

**Example 18.** You have 12 construction workers. You wish to assign three to job site 1, four to job site 2, and five to job site 3. In how many different ways can you make this assignment?

**Solution.** For this example,  $k = 3$  (corresponding to the  $k = 3$  different job sites),  $N = 12$ , and  $n_1 = 3$ ,  $n_2 = 4$ , and  $n_3 = 5$ . Then the number of different ways to assign the workers to the job sites is

$$\frac{N!}{n_1!n_2!n_3!} = \frac{12!}{3!4!5!} \doteq 27720.$$

#### 4. The Combinations Rule

A special application of the partitions rule - partitioning a set of  $N$  elements into  $k = 2$  groups (the elements that appear in a sample and those that do not) - is very common. Therefore, we give a different name to the rule for counting the number of different ways of partitioning a set of elements into two parts: the combinations rule.

A sample of  $n$  elements is to be chosen from a set of  $N$  elements. Then the number of different samples of  $n$  elements that can be selected from  $N$  is denoted by  $\binom{N}{n}$  and is equal to

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}. \quad (1.10)$$

Note that the order in which the  $n$  elements are drawn is not important.

**Example 19.** Five soldiers from a squadron of 100 are to be chosen for a dangerous mission. In how many ways can groups of five be formed?

**Solution.** This problem is equivalent to sampling  $n = 5$  elements from a set of  $N = 100$  elements. Thus, the number of ways is the number of possible combinations of five soldiers selected from 100, or

$$\binom{100}{5} = \frac{100!}{5!95!} \doteq 75287520.$$

**Example 20.** Explain the formula (1.9) with the aid of the combinations rule (1.10) and the permutations rule (1.8).

**Solution.** A sample of  $n_1$  elements is chosen from a set of  $N$  elements and the number of different samples of  $n_1$  elements that can be selected from  $N$  is equal, by the combinations rule (1.10), to

$$\binom{N}{n_1}.$$

A sample of  $n_2$  elements is chosen from remaining set of  $N - n_1$  elements and the number of different samples of  $n_2$  elements that can be selected from  $N - n_1$  is equal, by the combinations rule (1.10) to

$$\binom{N - n_1}{n_2},$$

etc. At the end the sample of  $n_{k-1}$  elements is chosen from a set of

$$N - n_1 - n_2 - \cdots - n_{k-2}$$

elements and is equal, by the combinations rule (1.10), to

$$\binom{N - n_1 - n_2 - \cdots - n_{k-2}}{n_{k-1}}.$$

Finally, the sample of  $n_k$  elements is chosen from a set of

$$N - n_1 - n_2 - \cdots - n_{k-1} = n_k$$

elements and is equal, by the combinations rule (1.10), to

$$\binom{N - n_1 - n_2 - \cdots - n_{k-1}}{n_k}.$$

Now, in accordance with the permutations rule (1.8) and with the combinations rule (1.10) we get for number of different partitions

$$\begin{aligned} & \binom{N}{n_1} \times \binom{N-n_1}{n_2} \times \cdots \times \\ & \binom{N-n_1-n_2-\cdots-n_{k-2}}{n_{k-1}} \times \binom{N-n_1-n_2-\cdots-n_{k-1}}{n_k} = \\ & \frac{N!}{n_1!(N-n_1)!} \times \frac{(N-n_1)!}{n_2!(N-n_1-n_2)!} \times \cdots \times \\ & \frac{(N-n_1-n_2-\cdots-n_{k-2})!}{n_{k-1}!(N-n_1-n_2-\cdots-n_{k-1})!} \times \\ & \frac{(N-n_1-n_2-\cdots-n_{k-1})!}{n_k!(N-n_1-n_2-\cdots-n_k)!} = \frac{N!}{n_1!n_2!\cdots n_k!}. \quad \square \end{aligned}$$

## 1.10 Summary of Counting Rules

### 1. Multiplicative rule

If you are drawing one element from each of  $k$  sets of elements, with the sizes of the sets  $n_1, n_2, \dots, n_k$ , the number of different results is

$$n_1 n_2 n_3 \dots n_k.$$

### 2. Permutations rule

If you are drawing  $n$  elements from a set of  $N$  elements and arranging the  $n$  elements in a distinct order, the number of different results is

$$P_n^N = \frac{N!}{(N-n)!}.$$

### 3. Partitions rule

If you are partitioning the elements of a set of  $N$  elements into  $k$  groups consisting of  $n_1, n_2, n_3, \dots, n_k$  elements ( $n_1 + n_2 + n_3 + \cdots + n_k = N$ ), the number of different results is

$$\frac{N!}{n_1!n_2!\cdots n_k!}.$$

## 4. Combinations rule

If you are drawing  $n$  elements from a set of  $N$  elements without regard to the order of the  $n$  elements, the number of different results is

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}.$$

The combinations rule is a special case of the partitions rule when  $k = 2$ .

### 1.11 Total Probability Formula, Bayes's Formula

#### 1. Total Probability Formula

forming a complete group of mutually exclusive events, then the event  $A$  can be represented as a union of the events  $A \cap H_1, A \cap H_2, \dots, A \cap H_n$ , i.e.,

$$A = (A \cap H_1) \cup (A \cap H_2) \cup \dots \cup (A \cap H_n).$$

The probability of the event  $A$  can be found from the formula

$$P(A) = P(A \cap H_1) + P(A \cap H_2) + \dots + P(A \cap H_n)$$

or (after using multiplicative rule of probability (1.3))

$$P(A) = P(H_1) \cdot P(A | H_1) + P(H_2) \cdot P(A | H_2) + \dots + P(H_n) \cdot P(A | H_n),$$

i.e.,

$$P(A) = \sum_{i=1}^n P(H_i) \cdot P(A | H_i). \quad (1.11)$$

Formula (1.11) is known as the **total** (or **composite**) **probability formula**.

**Example 21.** The first urn contains 5 white and 10 black balls, the second 3 white and 7 black balls. We draw one ball from the second urn and place it into the first one. Then we randomly draw one ball from the first urn. Determine the probability of the drawn ball being white.

**Solution.** Denote as an event  $A$  appearance of the white ball and determine  $P(A)$ . After a ball was drawn from the second urn and placed into the first, two collections of balls have turned out in the first urn:

(1) 5 white and 10 black balls it contained prior to replacement;

(2) 1 ball replaced from the second urn.

Suppose that

$H_1$  is the event that the randomly drawn ball belongs (previously) to the first urn and

$H_2$  is the event that the randomly drawn ball belongs (previously) to the second urn.

The probability of appearance of a white ball belonging to the first collection is

$$P(A | H_1) = \frac{5}{15} = \frac{1}{3},$$

and the probability of appearance of a white ball belonging to the second collection is

$$P(A | H_2) = \frac{3}{10}.$$

The probability that the randomly drawn ball belongs to the first collection is

$$P(H_1) = \frac{15}{16},$$

and to the second,

$$P(H_2) = \frac{1}{16}.$$

Using the total probability formula (1.11), we obtain

$$P(A) = P(H_1) \cdot P(A | H_1) + P(H_2) \cdot P(A | H_2) = \frac{15}{16} \cdot \frac{1}{3} + \frac{1}{16} \cdot \frac{3}{10} = \frac{53}{160}.$$

## 2. Bayes's Formula

Provided the event  $A$  has occurred, the conditional probability of the event  $H_i$  can be determined by Bayes' formula

$$P(H_i | A) = \frac{P(A \cap H_i)}{P(A)} = \frac{P(H_i) \cdot P(A | H_i)}{\sum_{i=1}^n P(H_i) \cdot P(A | H_i)}, \quad (1.12)$$

where  $i = 1, 2, \dots, n$ .

**Example 22.** We have three boxes identical in appearance. The first box contains 20 white balls, the second box contains 10 white and 10 black balls and the third box contains 20 black balls. We draw a white ball from a randomly selected box. Calculate the probability of the ball being drawn from the first box.

**Solution.** Suppose  $H_1$ ,  $H_2$  and  $H_3$  are the hypotheses consisting in selecting the first, the second and the third box, respectively. Let the event  $A$  be the drawing of a white ball. Then

$$P(H_1) = P(H_2) = P(H_3) = \frac{1}{3}$$

since the selection of any of the boxes is equally probable. Moreover, the probabilities of drawing a white ball from the first box, from the second box and from the third box are:

$$P(A | H_1) = 1, \quad P(A | H_2) = \frac{10}{20} = \frac{1}{2}, \quad P(A | H_3) = 0.$$

The desired probability  $P(H_1 | A)$  can be found from Bayes' formula (1.12):

$$\begin{aligned} P(H_1 | A) &= \\ &= \frac{P(H_1) \cdot P(A | H_1)}{P(H_1) \cdot P(A | H_1) + P(H_2) \cdot P(A | H_2) + P(H_3) \cdot P(A | H_3)} = \\ &= \frac{1 \cdot (1/3)}{1 \cdot (1/3) + (1/2) \cdot (1/3) + 0 \cdot (1/3)} = \frac{2}{3}. \end{aligned}$$

## 1.12 Random Variable and the Law of Its Distribution

### 1. Definition of a random variable

Suppose an experiment has an outcome sample space  $S$ . A real variable  $X$  that is defined for all possible outcomes in  $S$  (so that a real number - not necessary unique - is assigned to each possible outcome) is called a **random variable**. The outcome of the experiment may already be a real number and hence a random variable, e.g. the number of heads obtained in 10 throws of a coin, or the sum of the values if two dice are thrown. However,

more arbitrary assignments are possible, e.g. the assignment of a “quality” rating to each successive item produced by a manufacturing process. Furthermore, assuming that a probability can be assigned to all possible outcomes in a sample space  $S$ , it is possible to assign a **probability distribution** to any random variable. Random variables can be divided into two classes, **discrete** and **continuous**. The relationship establishing in one way or another the connection between the possible values of a random variable and their probabilities is called the **law of distribution** of a random variable.

## 2. Discrete random variable

A random variable  $X$  that takes only discrete values  $x_1, x_2, \dots, x_n$ , with probabilities  $p_1, p_2, \dots, p_n$ , and  $\sum_{i=1}^n p_i = 1$  is called a discrete random variable. The number of values  $n$  for which  $X$  has a non-zero probability is finite or at most countably infinite. If  $X$  is a discrete random variable, we can define a **probability function**  $f(x)$  that assigns probabilities to all the distinct values that  $X$  can take, such that

$$f(x) = \Pr(X = x) = \begin{cases} p_i & \text{if } x = x_i, \\ 0 & \text{otherwise.} \end{cases} \quad (1.13)$$

By formula (1.13) is given the law of distribution of a discrete random variable. A typical probability function thus consists of spikes, at **valid values** of  $X$ , whose height at  $x$  corresponds to the probability that  $X = x$ . Since the probabilities must sum to unity, we require

$$\sum_{i=1}^n f(x_i) = 1. \quad (1.14)$$

We may also define the **cumulative probability function** (or **probability distribution function**) of  $X$ ,  $F(x)$ , whose value gives the probability that  $X \leq x$ , so that

$$F(x) := \Pr(X \leq x) = \sum_{x_i \leq x} f(x_i). \quad (1.15)$$

Sometimes is the definition modified as follows:

$$F(x) := \Pr(X < x) = \sum_{x_i < x} f(x_i). \quad (1.16)$$

Hence  $F(x)$  is a step function that has upward jumps of  $p_i$  at  $x = x_i$ ,  $i = 1, 2, \dots, n$ , and is constant between possible values of  $X$ . We may also calculate the probability that  $X$  lies between two limits  $a$  and  $b$  ( $a < b$ ); this is given by

$$\Pr(a < X \leq b) = \sum_{a < x_i \leq b} f(x_i) = F(b) - F(a), \quad (1.17)$$

i.e. it is the sum of all the probabilities for which  $x_i$  lies within the relevant interval.

**Example 23.** A bag contains seven red balls and three white balls. Three balls are drawn at random and not replaced. Find the probability function for the number of red balls drawn.

**Solution.** Let  $X$  be the number of red balls drawn. Then

$$\begin{aligned} \Pr(X = 0) &= f(0) = \frac{3}{10} \times \frac{2}{9} \times \frac{1}{8} = \frac{1}{120}, \\ \Pr(X = 1) &= f(1) = \frac{3}{10} \times \frac{2}{9} \times \frac{7}{8} \times 3 = \frac{7}{40}, \\ \Pr(X = 2) &= f(2) = \frac{3}{10} \times \frac{7}{9} \times \frac{6}{8} \times 3 = \frac{21}{40}, \\ \Pr(X = 3) &= f(3) = \frac{7}{10} \times \frac{6}{9} \times \frac{5}{8} = \frac{7}{24}, \end{aligned}$$

It should be noted that  $\sum_{i=0}^3 f(i) = 1$ , as expected.

### 3. Continuous random variable

A random variable  $X$  is said to have a **continuous** distribution if  $X$  is defined for a continuous range of values between given limits. It is convenient to represent the law of distribution of a continuous random variable with the aid of so-called **probability density function**  $f(x)$ . The probability  $\Pr(a < X \leq b)$  of the fact that the value assumed by the random variable  $X$  will fall in the interval  $(a, b]$  is defined by the equality

$$\Pr(a < X \leq b) = \int_a^b f(x) dx.$$

The graph of the function  $f(x)$  is called a **distribution curve**. In terms of geometry, the probability that the random variable will fall in the interval  $(a, b]$  is equal to the area of the corresponding curvilinear trapezoid

bounded by the distribution curve, the  $Ox$  axis and the straight lines  $x = a$ ,  $x = b$ . The probability density function  $f(x)$  possesses the following properties:

1.  $f(x) \geq 0$ .
2.  $\int_{-\infty}^{\infty} f(x)dx = 1$ .

(If all the values of the random variable  $X$  belong to the interval  $(a, b)$ , the last property can be written as  $\int_a^b f(x)dx = 1$ .)

Let us now consider the function

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

It follows from the last equality that

$$f(x) = F'(x).$$

The function  $f(x)$  is sometimes called a **probability distribution differential function**, and the function  $F(x)$ , a **probability distribution integral function**.

Note the most significant properties of a probability distribution function:

1.  $F(x)$  is a non-decreasing function.
2.  $F(-\infty) = 0$ .
3.  $F(+\infty) = 1$ .

**Example 24.** Given the distribution series for the random variable  $X$ :

$x_i$	10	20	30	40	50
$p_i$	0,2	0,3	0,35	0,1	0,005

Construct the probability distribution function  $F(x) := \Pr(X < x)$  for that variable.

**Solution.**

- |                       |      |   |
|-----------------------|------|---|
| if $x \leq 10$ ,      | then | $F(x) = \Pr(X < x) = 0$ ;                 |
| if $10 < x \leq 20$ , | then | $F(x) = \Pr(X < x) = 0,2$ ;               |
| if $20 < x \leq 30$ , | then | $F(x) = \Pr(X < x) = 0,2 + 0,3 = 0,5$ ;   |
| if $30 < x \leq 40$ , | then | $F(x) = \Pr(X < x) = 0,5 + 0,35 = 0,85$ ; |
| if $40 < x \leq 50$ , | then | $F(x) = \Pr(X < x) = 0,85 + 0,1 = 0,95$ ; |
| if $x > 50$ ,         | then | $F(x) = \Pr(X < x) = 0,95 + 0,05 = 1$ .   |

**Example 25.** The random variable  $X$  is defined by the distribution function (integral function):

$$F(x) = \begin{cases} 0 & \text{if } x < 1; \\ (x - 1)/2 & \text{if } 1 \leq x \leq 3; \\ 1 & \text{if } x > 3. \end{cases}$$

Calculate the probabilities of the random variable  $X$  falling in the intervals  $(1, 5; 2, 5)$  and  $(2, 5; 3, 5)$ .

**Solution.** We have

$$P_1 = F(2, 5) - F(1, 5) = (2, 5 - 1)/2 - (1, 5 - 1)/2 = 0, 75 - 0, 25 = 0, 5, \\ P_2 = F(3, 5) - F(2, 5) = 1 - (2, 5 - 1)/2 = 1 - 0, 75 = 0, 25.$$

#### 4. The Mean Value, the Variance and Standard Deviation of the Random variable

##### The Mean Value

The **mean value** (or simply **mean**, or **mathematical expectation**) is the property most commonly used to characterize a probability distribution. Several abbreviations of mean value are used, e.g.:  $E(\mathbf{X})$  or  $M(X)$ .

For the **discrete random** variable mean value is defined as the sum of the products of the values of the random variable by the probabilities of these values. If the random variable  $X$  is characterized by the finite distribution series

$x_i$	$x_1$	$x_2$	$x_3$	$\dots$	$x_n$
$p_i$	$p_1$	$p_2$	$p_3$	$\dots$	$p_n$

then the mean value  $E(X)$  can be determined from the formula

$$E(X) = x_1p_1 + x_2p_2 + \dots + x_np_n, \quad (1.18)$$

i.e.,

$$E(X) = \sum_{i=1}^n x_i p_i.$$

Since  $p_1 + p_2 + \dots + p_n = 1$ , it follows that

$$E(X) = \frac{x_1p_1 + x_2p_2 + \dots + x_np_n}{p_1 + p_2 + \dots + p_n}.$$

Thus,  $E(X)$  is the weighted arithmetic mean of the values  $x_1, x_2, \dots, x_n$  of the random variable for the weights  $p_1, p_2, \dots, p_n$ . If  $n = \infty$ , then

$$E(X) = \sum_{i=1}^{\infty} x_i p_i$$

provided the sum of the series is finite.

The concept of the mean can be extended to a **continuous random variable**. Suppose  $f(x)$  is the probability density function of the random variable  $X$ . Then the mean value of the continuous random variable  $X$  is specified by the equality

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

provided the value of the integral is finite.

### The Variance

The **variance** (or the **dispersion**) of a random variable is the mean value of the square of the deviation of the random variable from its mean value:

$$D(X) = E[(X - E(X))^2].$$

Here is used the notation  $D(X)$ . Nevertheless, we can meet, e.g., the following ones:  $V(R)$  or  $\text{var}(R)$ , too. If we introduce the notation  $E(X) = m$ , then the formulas for variance of the discrete random variable  $X$  will be written in the form

$$D(X) = \begin{cases} \sum_{i=1}^n p_i (x_i - m)^2, \\ \sum_{i=1}^{\infty} p_i (x_i - m)^2 \quad (\text{for } n = \infty), \end{cases}$$

and for the continuous random variable  $X$ , in the form

$$D(X) = \int_{-\infty}^{+\infty} (x - m)^2 f(x) dx.$$

From the definitions we may easily derive the following useful properties of  $D(X)$ :

1.

$$D(X) = E[(X - a)^2] - [E(X) - a]^2 \quad (1.19)$$

or

$$D(X) = E[(X - a)^2] - [m - a]^2$$

is valid for the variance of a random variable, where  $a$  is an arbitrary number. This formula is often used to calculate the variance of a random variable since the calculations performed by this formula are simpler than previous formulas.

2.

$$D(a) = 0$$

where where  $a$  is an arbitrary number.

3.

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

where where  $a, b$  are arbitrary constants.

### The Standard Deviation

The variance of a distribution is always positive; its positive square root is known as the standard deviation of the distribution and is often denoted by  $\sigma$ , i.e.

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

Roughly speaking,  $\sigma$  measures the spread (about  $x = m$ ) of the values that  $X$  can assume.

**Example 26.** Given the function:

$$f(x) = \begin{cases} 0 & \text{if } x < 0; \\ (1/2) \sin x & \text{if } 0 \leq x \leq \pi; \\ 0 & \text{if } x > \pi. \end{cases}$$

Show that  $f(x)$  can serve as the probability density function of some random variable  $X$ . Find the mathematical expectation, the variance and the standard deviation of the random variable  $X$ .

**Solution.** We have

$$\int_{-\infty}^{\infty} f(x)dx = \int_0^{\pi} f(x)dx = \frac{1}{2} \int_0^{\pi} \sin x dx = -\frac{1}{2} \cos x \Big|_0^{\pi} = 1.$$

In addition,  $f(x) \geq 0$ . Consequently,  $f(x)$  can serve as the probability density function of some random variable. Since the straight line  $x = \pi/2$  is the symmetry axis of the corresponding arc of the curve  $y = (1/2) \sin x$ , the expectation of the random variable  $X$  is  $\pi/2$ , i.e.  $E(X) = \pi/2$ . Let us find the variance. We put  $a = 0$ ,  $E(X) = \pi/2$  in formula (1.19) and calculate the integral determining  $E(X^2)$ . We have

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x) dx = \frac{1}{2} \int_0^{\pi} x^2 \sin x dx = \\ &= \frac{1}{2} \left[ -x^2 \cos x + 2x \sin x + 2 \cos x \right] \Big|_0^{\pi} = \frac{1}{2} (\pi^2 - 4). \end{aligned}$$

Therefore

$$D(X) = \frac{1}{2} (\pi^2 - 4) - \left( \frac{\pi}{2} \right)^2 = \frac{\pi^2}{4} - 2,$$

and

$$\sigma = \sqrt{\frac{\pi^2}{4} - 2} \approx 0,69.$$

**Example 27.** The random variable  $X$  is characterized by the distribution series:

$x_i$	0	1	2	3	4
$p_i$	0,2	0,4	0,3	0,08	0,02

Determine the mathematical expectation, the variance and the standard deviation of the random variable  $X$ .

**Solution.** We find the mean by the formula (1.18):

$$E(X) = 0 \cdot 0,2 + 1 \cdot 0,4 + 2 \cdot 0,3 + 3 \cdot 0,08 + 4 \cdot 0,02 = 1,32.$$

We find the variance by formula (1.19) putting  $a = 2$ ; hence  $E(X) - a = 1,32 - 2 = -0,68$ . We compile a table:

$x_i$	0	1	2	3	4
$x_i - a$	-2	-1	0	1	2
$(x_i - a)^2$	4	1	0	1	4
$p_i$	0,2	0,4	0,3	0,08	0,02
$p_i(x_i - a)^2$	0,8	0,4	0	0,08	0,008

Now we find

$$E[(X - a)^2] = \sum_{i=0}^4 p_i (x_i - a)^2 = 1,36;$$

$$D(X) = 1,36 - (-0,68)^2 = 1,36 - 0,4634 = 0,8966;$$

$$\sigma = \sqrt{0,8966} = 0,95.$$

## 5. The Mode and the Median

The **mode of the discrete random variable**  $X$  is its most frequent value. The **mode of the continuous random variable**  $X$  is the point at which the probability density function has its greatest value. The mode is designated as  $\overline{M}$ . The **median** of the continuous random variable  $X$  is its value  $\mu$  for which is equally probable that the random variable turns out to be less or greater than  $\mu$ , i.e.

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

**Example 28.** Given the probability density function of the random variable  $f(x) = ae^{2x-x^2}$  with  $a > 0$ . Find the mode of this random variable.

**Solution.** To find the maximum of the function  $y = f(x)$ , we find the derivatives of the first and the second orders:

$$f'(x) = 2a(1-x)e^{2x-x^2}, \quad f''(x) = -2ae^{2x-x^2} + 4a(1-x^2)e^{2x-x^2}.$$

From the equation  $f'(x) = 0$  we get  $x = 1$ . Since  $f''(1) = -2ae < 0$ , it follows that for  $x = 1$  the function  $f(x)$  possesses a maximum, i.e., for mode we have:  $\overline{M} = 1$ . Note that the maximum of  $f(x)$  does not depend on the numerical value of  $a$ .

**Example 29.** Given the probability density function of the random variable  $X$ :

$$f(x) = \begin{cases} 0 & \text{if } x < 0; \\ x - x^3/4 & \text{if } 0 \leq x \leq 2; \\ 0 & \text{if } x > 2. \end{cases}$$

Find the median of this random variable.

**Solution.** We find the median  $\mu$  from the condition  $\Pr(X < \mu) = 0,5$ . In the given case

$$\Pr(X < \mu) = \int_0^\mu \left(x - \frac{x^3}{4}\right) dx = \frac{\mu^2}{2} - \frac{\mu^4}{16}.$$

Thus we arrive the equation

$$\frac{\mu^2}{2} - \frac{\mu^4}{16} = 0,5, \quad \text{or} \quad \mu^4 - 8\mu^2 + 8 = 0,$$

whence

$$\mu = \pm\sqrt{4 \pm \sqrt{8}}.$$

From the four roots of the equation we should choose the root contained between 0 and 2. Hence,  $\mu = \sqrt{4 - \sqrt{8}} \approx 1,09$ .

## 1.13 Some Special Distributions

### 1. Discrete Uniform Distribution

If the random variable  $X$  assumes the values  $x_1, x_2, \dots, x_k$ , with equal probabilities, then the discrete uniform distribution is given by

$$f(x) = \frac{1}{k}, \quad \text{if } x = x_i, i = 1, 2, \dots, k.$$

**Example 30.** When a light bulb is selected at random from a box that contains a 40-watt bulb, a 60-watt bulb, a 75-watt bulb, and a 100-watt bulb, each element of the sample space  $S = \{40, 60, 75, 100\}$  occurs with probability  $1/4$ . Therefore we have a uniform discrete distribution, with

$$f(x) = \frac{1}{4}, \quad \text{if } x = 40, 60, 75, 100.$$

**Theorem 5.** *The mean and variance of the discrete uniform distribution  $f(x)$  are*

$$E(X) = \frac{1}{k} \sum_{i=1}^k x_i, \quad D(X) = \frac{1}{k} \sum_{i=1}^k (x_i - E(X))^2.$$

**Proof.** By definition

$$E(X) = \frac{1}{k} \sum_{i=1}^k x_i f(x_i) = \frac{1}{k} \sum_{i=1}^k x_i.$$

Also by definition

$$\begin{aligned} D(X) = E[(X - E(X))^2] &= \sum_{i=1}^k (x_i - E(X))^2 f(x_i) = \\ &= \sum_{i=1}^k (x_i - E(X))^2 \frac{1}{k} = \frac{1}{k} \sum_{i=1}^k (x_i - E(X))^2. \end{aligned}$$

## 2. Continuous Uniform Distribution

The distribution of random variables whose all values lie in an interval  $[a, b]$  and possess a constant probability density  $h > 0$  on that interval is known as **uniform** distribution. Thus

$$f(x) = \begin{cases} 0 & \text{if } x < a; \\ h & \text{if } a \leq x \leq b; \\ 0 & \text{if } x > b. \end{cases}$$

Since  $h(b - a) = 1$ , we have  $h = 1/(b - a)$  and, consequently,

$$f(x) = \begin{cases} 0 & \text{if } x < a; \\ 1/(b - a) & \text{if } a \leq x \leq b; \\ 0 & \text{if } x > b. \end{cases}$$

**Example 31.** Determine the mean of a random variable with uniform distribution.

**Solution.** We have

$$E(X) = \int_a^b x f(x) dx = \int_a^b \frac{x}{b - a} dx = \frac{1}{b - a} \cdot \frac{x^2}{2} \Big|_a^b = \frac{1}{b - a} \cdot \frac{b^2 - a^2}{2} = \frac{b + a}{2},$$

i.e.

$$E(X) = \frac{b + a}{2}.$$

**Example 32.** Calculate the variance and the standard deviation for a random variable with uniform distribution.

**Solution.** We use the formula

$$D(X) = E(X^2) - [E(X)]^2,$$

taking into account the value

$$E(X) = \frac{b+a}{2}$$

found in the preceding problem. Thus, it remains to calculate  $E(X^2)$ . We have

$$E(X^2) = \int_a^b \frac{x^2}{b-a} dx = \frac{x^3}{3(b-a)} \Big|_a^b = \frac{b^3 - a^3}{3(b-a)} = \frac{b^2 + 2ab + a^2}{3}.$$

It follows that

$$D(X) = \frac{b^2 + 2ab + a^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}.$$

Consequently,

$$\sigma_x = \sqrt{D(X)} = \frac{(b-a)}{2\sqrt{3}}.$$

### 3. Binomial Distribution. Poisson's Distribution

#### Bernoulli's Formula

If  $n$  independent trials are performed in each of which the probability of occurrence of the event  $A$  is the same and is equal to  $p$ , then the probability of the event  $A$  occurring  $m$  times in these  $n$  trials is expressed by **Bernoulli's formula**

$$\Pr = P_{m,n} = \binom{n}{m} p^m q^{n-m}, \quad (1.20)$$

where  $q = 1 - p$ .

**Example 33.** There are 20 white and 10 black balls in the urn. Four balls are drawn successively with replacement, the urn being shaken before every new drawing. What is the probability of two of the four drawn balls being white?

**Solution.** The probability of a white ball being drawn,

$$p = \frac{20}{30} = \frac{2}{3},$$

may be assumed to be the same in the four trials and  $q = 1 - p = 1/3$ . Using Bernoulli's formula we get

$$P_{2,4} = \binom{4}{2} p^2 q^2 = \frac{4 \cdot 3}{1 \cdot 2} \cdot \left(\frac{2}{3}\right)^2 \cdot \left(\frac{1}{3}\right)^2 = \frac{8}{27}.$$

### Binomial Distribution

The distribution of the random variable  $X$  which can assume  $n + 1$  values  $(0, 1, \dots, n)$  described by Bernoulli's formula (1.20) is known as a **binomial distribution**. The mean and the variance of binomial distribution is

$$E(X) = np, \quad D(X) = npq.$$

**Example 34.** The probability of the marksman hitting the target is  $2/3$ . The marksman fired 15 shots. The random variable  $X$  is the number of hits. Find the mean value and the variance of the random variable  $X$ .

**Solution.** We must use here the mean values and the variances of the binomial distribution:

$$E(X) = np = 15 \cdot (2/3) = 10, \quad D(X) = npq = 15 \cdot (2/3) \cdot (1/3) = 10/3.$$

### Poisson's Distribution

The distribution of the random variable  $X$  which can assume any integral nonnegative values  $(0, 1, \dots, n, \dots)$  described by the formula

$$\Pr(X = n) = \frac{a^n}{n!} e^{-a}, \quad (1.21)$$

is known as **Poisson's distribution**.

The following random variables have a Poisson's distribution:

**a)** Suppose  $n$  points are randomly distributed over the interval  $(0, N)$  of the  $Ox$  axis and the events consisting in the falling of one point in any preassigned segment of constant (say, unit) length are equally probable. If

$N \rightarrow \infty$ ,  $n \rightarrow \infty$  and  $a = \lim_{N \rightarrow \infty} n/N$ , then the random variable  $X$  equal to the number of points falling in the preassigned segment of unit length (which can assume the values  $0, 1, \dots, m, \dots$ ) has Poisson's distribution.

**b)** If  $n$  is an average number of calls received by a given telephone exchange during one hour, then the number of calls received during one minute is approximatively distributed by Poisson's law, with  $a = n/60$ . The mean and the variance of Poisson's distribution is

$$E(X) = a, \quad D(X) = a.$$

**Example 35.** A person receives on average one e-mail per half-hour interval. Assuming that the e-mails are received randomly in time, find the probabilities that in any particular hour 0, 1, 2, 3, 4, 5 messages are received.

**Solution.** Let  $X =$  number of e-mails received per hour. Clearly the mean number of e-mails per hour is two, and so  $X$  follows a Poisson distribution with  $\lambda = 2$ , i.e.

$$\Pr(X = x) = \frac{2^x}{x!} e^{-2}.$$

Thus

$$\begin{aligned} \Pr(X = 0) &= e^{-2} = 0,135, \\ \Pr(X = 1) &= 2e^{-2} = 0,271, \\ \Pr(X = 2) &= 2^2 e^{-2}/2! = 0,271, \\ \Pr(X = 3) &= 2^3 e^{-2}/3! = 0,180, \\ \Pr(X = 4) &= 2^4 e^{-2}/4! = 0,090, \\ \Pr(X = 5) &= 2^5 e^{-2}/5! = 0,036. \end{aligned}$$

**Example 36.** Show that the binomial distribution approaches Poisson's distribution in the limit if  $n \rightarrow \infty$ ,  $p \rightarrow 0$ , but  $np = a$ .

**Solution.** Let us rewrite the binomial probabilities (1.20) in the form

$$P_{m,n} = \binom{n}{m} p^m q^{n-m} = \frac{(np)^m \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{m-1}{n}\right)}{m!} \cdot (1-p)^{\frac{np}{p} - m}.$$

Due to conditions indicated

$$P_{m,n} = \frac{a^m \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{m-1}{n}\right)}{m!} \cdot (1-p)^{\frac{a}{p} - m}.$$

Since it is easy to see that

$$\lim_{p \rightarrow 0} (1-p)^{a/p} = e^{-a}$$

we conclude

$$\begin{aligned} \lim_{p \rightarrow 0, n \rightarrow \infty} P_{m,n} &= \frac{a^m}{m!} \times \\ \lim_{p \rightarrow 0, n \rightarrow \infty} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{m-1}{n}\right) (1-p)^{a/p} (1-p)^{-m} &= \\ &= \frac{a^m e^{-a}}{m!}. \end{aligned}$$

So,

$$P_{m,n} \approx \frac{a^m e^{-a}}{m!}$$

if  $np = a$  and  $n \rightarrow \infty, p \rightarrow 0$ .

#### 4. Normal Distribution (Gaussian normal distribution) and Laplace Function

##### Normal Distribution

We say that the continuous random variable  $X$  has **normal distribution** (or Gaussian normal distribution) if it is characterized by the density (Gaussian curve)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[ -\frac{1}{2} \cdot \left( \frac{x - \mu}{\sigma} \right)^2 \right],$$

where  $\mu$  and  $\sigma$  are constants. The probability density function  $f(x)$  satisfies two conditions of a density function:

$$f(x) > 0, \quad \int_{-\infty}^{\infty} f(x) dx = 1.$$

For normal distribution with these parameters we use notation  $\text{No}(\mu, \sigma^2)$ . The mean and the variance of normal distribution is

$$E(X) = \mu, \quad D(X) = \sigma^2.$$

The probability distribution integral function has the form

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left[-\frac{1}{2} \cdot \left(\frac{t - \mu}{\sigma}\right)^2\right] dt.$$

The random variable

$$Z = \frac{X - \mu}{\sigma}$$

is called the **standard normal variable**. For it the probability density function takes the form

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

which is called the **standard normal distribution** with

$$E(Z) = 0, \quad D(Z) = 1.$$

In this case we use notation  $\text{No}(0, 1)$ .

### Laplace Function

The function

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{t^2}{2}\right) dt$$

is called the **Laplace function** (or error function). The following relation holds:

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

The probability that the normally distributed random variable  $X$  will fall in the interval  $(a, b)$  is determined from the values of the Laplace function by the formula

$$\Pr(a < X < b) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right).$$

Between basic properties of Laplace function belong the following:

$$\Phi(0) = \frac{1}{2}, \quad \Phi(-t) = 1 - \Phi(t).$$

**Example 37.** Sawmill  $A$  produces boards whose length are Gaussian distributed with mean 209,4 cm and standard deviation 5,0 cm. A board is accepted if it is longer than 200 cm but it is rejected otherwise. Show that 3% of boards are rejected.

**Solution.** If  $X$  is length of boards from  $A$ , so that  $X$  has distribution  $No(209,4; (5,0)^2)$ . Then

$$\Pr(X < 200) = \Phi\left(\frac{200 - \mu}{\sigma}\right) = \Phi\left(\frac{200 - 209,4}{5,0}\right) = \Phi(-1,88).$$

Since  $\Phi(-t) = 1 - \Phi(t)$  we have

$$\Pr(X < 200) = 1 - \Phi(1,88) = \text{by the table} = 1 - 0,9699 = 0,0301,$$

i.e. 3,0% of boards are rejected.

**Example 38.** Sawmill  $B$  produces boards of the same standard deviation but of mean length 210,1 cm. Find the proportion of boards rejected if they are drawn at random from the outputs of  $A$  and  $B$  in the ratio 3 : 1.

**Solution.** Now let  $Y$  is length of boards from  $B$ , so that  $Y$  has distribution  $No(210,1; (5,0)^2)$ . Then

$$\Pr(Y < 200) = \Phi\left(\frac{200 - \mu}{\sigma}\right) = \Phi\left(\frac{200 - 210,1}{5,0}\right) = \Phi(-2,02).$$

Since  $\Phi(-t) = 1 - \Phi(t)$  we have

$$\Pr(Y < 200) = 1 - \Phi(2,02) = \text{by the table} = 1 - 0,9783 = 0,0217.$$

Therefore, when taken alone, only 2,2% of boards from  $B$  are rejected. If, however, boards are drawn at random from  $A$  and  $B$  in the ratio 3 : 1 then the proportion rejected is

$$\frac{1}{4}(3 \times 0,030 + 1 \times 0,022) = 0,028 = 2,8\%.$$

## 1.14 Chebyshev's Theorem

If a random variable has a small variance or standard deviation, we would expect most the values to be grouped around the mean. Therefore, the

probability that random variable assumes a value within a certain interval about the mean greater than for a similar random variable with a larger standard deviation, we think of probability in terms of area, we would expect a continuous contribution with a small standard deviation to have most of its area close to  $\mu$ . However, a large value of  $\sigma$  indicates a greater variability, therefore we should expect the area to be more spread out.

The following theorem (named after the Russian mathematician P.L. Chebyshev, 1821–1894) gives an estimate of the probability that a random variable (continuous or discrete) assumes a value within  $k$  standard deviations of its mean for any real number  $k$ .

**Theorem 6 (Chebyshev)** *The probability that any random variable  $X$  will assume a value within  $k$  standard deviations of the mean is **at least**  $1 - 1/k^2$ . That is,*

$$\Pr(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}.$$

**Proof.** Let us prove this theorem for the case of continuous random variable. By definition of the variance of  $X$  we can write

$$\begin{aligned} \sigma^2 &= E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \\ &\int_{-\infty}^{\mu - k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu - k\sigma}^{\mu + k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu + k\sigma}^{\infty} (x - \mu)^2 f(x) dx \geq \\ &\int_{-\infty}^{\mu - k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu + k\sigma}^{\infty} (x - \mu)^2 f(x) dx, \end{aligned}$$

since the middle of the three integrals is nonnegative. Now since  $|x - \mu| \geq k\sigma$  in both remaining integrals, we have

$$\sigma^2 \geq \int_{-\infty}^{\mu - k\sigma} k^2 \sigma^2 f(x) dx + \int_{\mu + k\sigma}^{\infty} k^2 \sigma^2 f(x) dx,$$

or

$$\int_{-\infty}^{\mu - k\sigma} f(x) dx + \int_{\mu + k\sigma}^{\infty} f(x) dx \leq \frac{1}{k^2}.$$

Hence

$$\Pr(\mu - k\sigma < X < \mu + k\sigma) = \int_{\mu - k\sigma}^{\mu + k\sigma} f(x) dx \geq 1 - \frac{1}{k^2}.$$

□ For  $k = 2$  the theorem states that the random variable  $X$  has probability of at least  $1 - 1/2^2 = 3/4$  of falling within two standard deviations of the

mean. That is, three-fourths or more of the observations of any distribution lie in the interval  $\mu \pm 2\sigma$ . Similarly, for  $k = 3$  the theorem says that the random variable  $X$  has probability of at least  $1 - 1/3^2 = 8/9$  of falling within two standard deviations of the mean. That is, eight-ninths of the observations of any distribution lie in the interval  $\mu \pm 3\sigma$ .

**Example 39.** A random variable  $X$  has a mean  $\mu = 8$ , a variance  $\sigma^2 = 9$ , and an unknown probability distribution. Find  $\Pr(-4 < X < 20)$ , and  $\Pr(|X - 8| \geq 6)$ .

**Solution.** In accordance with Chebyshev's theorem we get

$$\Pr(-4 < X < 20) = \Pr(8 - 4 \cdot 3 < X < 8 + 4 \cdot 3) \geq \frac{15}{16}$$

in the first case and

$$\begin{aligned} \Pr(|X - 8| \geq 6) &= 1 - \Pr(|X - 8| < 6) = 1 - \Pr(-6 < X - 8 < 6) = \\ &= 1 - \Pr(8 - 2 \cdot 3 < X < 8 + 2 \cdot 3) \geq \frac{1}{4} \end{aligned}$$

in the second one.

**Remark 1.** Chebyshev's theorem holds for **any distribution** of observations and, for this reason, the results are usually weak. The values given by the theorem is a lower bound only. That is, we know, e.g. that the probability of a random variable falling within two standard deviations of the mean can be **no less** than  $3/4$ , but we never know how much more it might actually be. Only when the probability distribution is known we can determine exact probabilities.

## 1.15 Normal Approximation to the Binomial

**Theorem 7.** *If  $X$  is a binomial random variable with mean  $\mu = np$  and variance  $\sigma^2 = npq$ , then the limiting form of the distribution of*

$$Z = \frac{X - np}{\sqrt{npq}}.$$

*as  $n \rightarrow \infty$ , is the standard normal distribution  $\text{No}(0, 1)$ .*

## 1.16 The Central Limit Theorem

**Theorem 8.** *Suppose that  $X_1, X_2, \dots, X_n$  are independent random variables, each of which is described by a probability density function  $f_i(x)$  with a mean  $\mu_i$  and a variance  $\sigma_i^2$ . The random variable*

$$Z = \frac{\sum_{i=1}^n X_i}{n},$$

*i.e. the “mean” of the  $X_i$ , has the following properties:*

1) *its expectation value*

$$E(Z) = \frac{\sum_{i=1}^n \mu_i}{n},$$

2) *its variance*

$$V(Z) = \frac{\sum_{i=1}^n \sigma_i^2}{n^2},$$

3) *as  $n \rightarrow \infty$ , the probability of  $Z$  tends to a Gaussian with corresponding mean and variance*

**Proof.** First two properties are easily proved.

$$E(Z) = \frac{E(X_1) + E(X_2) + \dots + E(X_n)}{n} = \frac{\mu_1 + \mu_2 + \dots + \mu_n}{n} = \frac{\sum_{i=1}^n \mu_i}{n}.$$

Let us note that this result does not require that  $X_i$  are independent random variables. If  $\mu_i = \mu$  for all  $i$  then this becomes

$$E(Z) = \frac{n\mu}{n} = \mu.$$

If  $X_i$  are independent random variables, then

$$V(Z) = V\left[\frac{X_1 + X_2 + \dots + X_n}{n}\right] = \frac{V(X_1) + V(X_2) + \dots + V(X_n)}{n^2} = \frac{\sum_{i=1}^n \sigma_i^2}{n^2}.$$

## 1.17 Transformed Random Variables

Frequently we need to derive the probability distribution of a function of random variable. For example, suppose that  $X$  is a **discrete** random variable with probability distribution  $f(x)$  and suppose further that  $Y = u(X)$  defines a one-to-one transformation between the values  $X$  and  $Y$ . We wish to find the probability distribution of  $Y$ . It is important to note that the one-to-one transformation implies that each value  $x$  is related to one, and only one, value  $y = u(x)$  and that each value  $y$  is related to one, and only one, value  $x = w(y)$ , where  $w(y)$  is obtained by solving  $y = u(x)$  for  $x$  in terms of  $y$ . The random variable  $Y$  assumes the value  $y$  when  $X$  assumes the value  $w(y)$ . Consequently, the probability distribution of  $Y$  is given by

$$g(y) = \Pr(Y = y) = \Pr[X = w(y)] = f[w(y)].$$

**Theorem 9.** *Suppose that  $X$  is a **discrete** random variable with probability distribution  $f(x)$ . Let  $Y = u(X)$  define one-to-one transformation between the values of  $X$  and  $Y$  so that the equation  $y = u(x)$  can be uniquely solved for  $x$  in terms of  $y$ , say  $x = w(y)$ . Then the probability distribution of  $Y$  is*

$$g(y) = f[w(y)].$$

**Example 40.** Let  $X$  be a geometric random variable with probability distribution

$$f(x) = \frac{3}{4} \cdot \left(\frac{1}{4}\right)^{x-1},$$

where  $x = 1, 2, 3, \dots$ . Find the probability distribution of the random variable  $Y = X^2$ .

**Solution.** Since the value of  $X$  are all positive, the transformation defines a one-to-one correspondence between the  $x$  and  $y$  values,  $y = x^2$  and  $x = \sqrt{y}$ . Hence

$$g(y) = \begin{cases} f(\sqrt{y}) = \frac{3}{4} \cdot \left(\frac{1}{4}\right)^{\sqrt{y}-1}, & y = 1, 2, 3, \dots, \\ 0 & \text{elsewhere.} \end{cases}$$

□

To find the probability distribution of the random variable  $Y = u(X)$  when  $X$  is a **continuous** random variable and the transformation is one-to-one, we shall need the following

**Theorem 10.** *Suppose that  $X$  is a **continuous** random variable with probability distribution  $f(x)$ . Let  $Y = u(X)$  define one-to-one correspondence between the values of  $X$  and  $Y$  so that the equation  $y = u(x)$  can be uniquely solved for  $x$  in terms of  $y$ , say  $x = w(y)$ . Then the probability distribution of  $Y$  is*

$$g(y) = f[w(y)] \cdot |J|,$$

where  $J = w'(y)$  and is called the **Jacobian** of the transformation.

**Proof.** Suppose that  $y = u(x)$  is an increasing function. Then whenever  $Y$  falls between  $a$  and  $b$ , the random variable  $X$  must fall between  $w(a)$  and  $w(b)$ . Hence

$$\Pr(a < Y < b) = \Pr[w(a) < X < w(b)] = \int_{w(a)}^{w(b)} f(x)dx.$$

Changing the variable of integration from  $x$  to  $y$  by the relation  $x = w(y)$ , we obtain  $dx = w'(y)dy$ , and hence

$$\Pr(a < Y < b) = \int_a^b f[w(y)]w'(y)dy.$$

Since the integral gives the desired probability for every  $a < b$  within the permissible set of  $y$ , then the probability distribution of  $Y$  is

$$g(y) = f[w(y)]w'(y) = f[w(y)] \cdot J = f[w(y)] \cdot |J|.$$

Suppose that  $y = u(x)$  is a decreasing function. Then

$$\Pr(a < Y < b) = \Pr[w(b) < X < w(a)] = \int_{w(b)}^{w(a)} f(x)dx.$$

After changing the variable of integration from  $x$  to  $y$  by the relation  $x = w(y)$ , we obtain  $dx = w'(y)dy$ , and hence

$$\Pr(a < Y < b) = \int_b^a f[w(y)]w'(y)dy = - \int_a^b f[w(y)]w'(y)dy,$$

and, consequently,

$$g(y) = -f[w(y)]w'(y) = -f[w(y)] \cdot J = f[w(y)] \cdot |J|.$$

□

**Example 41.** Let  $X$  be a continuous random variable with probability distribution

$$f(x) = \begin{cases} \frac{x}{12}, & 1 < x < 5, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the probability distribution of the random variable  $Y = 2X - 3$ .

**Solution.** The inverse solution of  $y = 2x - 3$  yields  $x = (y + 3)/2$ , from which we obtain

$$L = w'(y) = \frac{dx}{dy} = \frac{1}{2}.$$

Therefore, by previous theorem, we find the density function of  $Y$  to be

$$g(y) = \begin{cases} \frac{(y+3)/2}{12} \cdot \left(\frac{1}{2}\right) = \frac{y+3}{48}, & -1 < y < 7, \\ 0 & \text{elsewhere.} \end{cases}$$

□

## 1.18 Statistics

Statistics is concerned with the analysis of experimental data. We may regard the product of any experiment as a set on  $n$  measurements of some quantity. This set constitutes the **data**. Each measurement (or **data item**) consists accordingly of a single number or a set of numbers. We will assume that each data item is a single number. As a result of inaccuracies in the measurement process, or because of intrinsic variability in the quantity being measured, one would expect the  $n$  measured values  $x_1, x_2, \dots, x_n$  to be different each time the experiment is performed. In other words, an experiment consisting of  $n$  measurements is considered as a **random sample** of size  $n$  from a **population**  $f(x)$ , where  $x$  denotes a point in the  $n$ -dimensional data space having coordinates  $(x_1, x_2, \dots, x_n)$ . In selecting a random sample of size  $n$  from a population  $f(x)$ , let us define the random variable  $X_i$ ,  $i = 1, 2, \dots, n$ , to represent the  $i$ th measurement or sample value that we observe. The random variables  $X_1, X_2, \dots, X_n$  will then constitute a random sample from the population  $f(x)$  with numerical values  $x_1, x_2, \dots, x_n$  if the measurements are obtained by repeating the experiment  $n$  independent times under essentially the same conditions.

Because of the identical conditions under which the elements are selected, it is reasonable to assume that the  $n$  random variables  $X_1, X_2, \dots, X_n$  are independent and that each has the same probability distribution  $f(x)$ . That is, the probability distributions of  $X_1, X_2, \dots, X_n$  are, respectively  $f(x_1), f(x_2), \dots, f(x_n)$  and their joint probability distribution is

$$f(x_1, x_2, \dots, x_n) = f(x_1)f(x_2) \dots f(x_n).$$

**Definition 13.** Let  $X_1, X_2, \dots, X_n$  be  $n$  independent random variables each having the same probability distribution  $f(x)$ . We then define

$$X_1, X_2, \dots, X_n$$

to be a random sample of size  $n$  from the population  $f(x)$  and write its joint probability distribution as

$$f(x_1, x_2, \dots, x_n) = f(x_1)f(x_2) \dots f(x_n).$$

## 1. Sample Statistics

Any function of the random variables constituting a random sample is called a **statistic**, or **sample statistic**.

### Sample Mean

**Definition 14.** If  $X_1, X_2, \dots, X_n$  represent a random sample of size  $n$ , then the **sample mean** is defined by the statistic

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}.$$

Note that the statistic  $\bar{X}$  assumes the value

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

when  $X_1$  assumes the value  $x_1$ ,  $X_2$  assumes the value  $x_2$ , and so forth. In practice the value of a statistic is usually given the same name as the statistic. For instance, the term **sample mean** is applied to both the statistic  $\bar{X}$  and its computed value  $\bar{x}$ .

### Sample Median

**Definition 15.** If  $X_1, X_2, \dots, X_n$  represent a random sample of size  $n$ , arranged in increasing order of magnitude, then the **sample median** is defined by the statistic

$$\tilde{X} = \begin{cases} X_{(n+1)/2}, & \text{if } n \text{ is odd,} \\ \frac{X_{n/2} + X_{(n/2)+1}}{2}, & \text{if } n \text{ is even.} \end{cases}$$

**Example 42.** The number of foreign ships arriving at an east coast port on 7 randomly selected days were 8, 3, 9, 5, 6, 8 and 5. Find the sample median.

**Solution.** Arranging the observations in increasing order of magnitude, we get

$$3, 5, 5, 6, 8, 8, 9$$

and hence  $\tilde{x} = 6$ .

### Sample Mode

**Definition 16.** If  $X_1, X_2, \dots, X_n$ , not necessarily all different, represent a random sample of size  $n$ , the the **mode**  $M$  is that value of the sample that occurs most often or with the greatest frequency. The mode may not exist, and when it does it is not necessarily unique.

The mode does not always exist. This is true when e.g. all observations occur with the same frequency.

**Example 43.** The number of movies attended last month by a random sample of 12 students were recorded as follows: 2, 0, 3, 1, 2, 4, 2, 5, 4, 0, 1, 4. In this case, there are two modes, 2 and 4, since both 2 and 4 occur with the greatest frequency. (The distribution is said to be bimodal.)

### Sample Variance

**Definition 17.** If  $X_1, X_2, \dots, X_n$  represent a random sample of size  $n$ , then the **sample variance** is defined by the statistic

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}.$$

The computed value of  $S^2$  for a given sample is denoted by  $s^2$ . In the denominator is used  $n - 1$  as a divisor. In some books the number  $n$  is used.

**Example 44.** A comparison of coffee prices at 4 randomly selected grocery stores showed increases from the previous month of 12, 15, 17, and 20 cents for a 200-gram jar. Find the variance of this random sample of price increases.

**Solution.** Calculating the sample mean, we get

$$\tilde{x} = \frac{12 + 15 + 17 + 20}{4} = 16 \text{ cents.}$$

Therefore

$$s^2 = \frac{\sum_{i=1}^4 (x_i - 16)^2}{3} = \frac{(12 - 16)^2 + (15 - 16)^2 + (17 - 16)^2 + (20 - 16)^2}{3} = \frac{(-4)^2 + (-1)^2 + (1)^2 + (4)^2}{3} = \frac{34}{3}.$$

**Theorem 11.** If  $S^2$  is the variance of a random sample of size  $n$ , we may write

$$S^2 = \frac{n \sum_{i=1}^n X_i^2 - \left( \sum_{i=1}^n X_i \right)^2}{n(n - 1)}.$$

**Proof.** By definition,

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} = \frac{\sum_{i=1}^n (X_i^2 - 2\bar{X}X_i + \bar{X}^2)}{n - 1} = \frac{\sum_{i=1}^n X_i^2 - 2\bar{X} \sum_{i=1}^n X_i + n\bar{X}^2}{n - 1}.$$

Replacing  $\bar{X}$  by  $\sum_{i=1}^n X_i/n$  and multiplying numerator and denominator by  $n$ , we obtain the more useful computational formula

$$S^2 = \frac{n \sum_{i=1}^n X_i^2 - \left( \sum_{i=1}^n X_i \right)^2}{n(n-1)}.$$

**Example 45.** Find the variance of the data 3, 4, 5, 6, 6, 7.

**Solution.** We find that  $n = 6$ ,  $\sum_{i=1}^6 x_i = 31$ ,  $\sum_{i=1}^6 x_i^2 = 171$ . Hence

$$s^2 = \frac{6 \cdot 171 - 31^2}{6 \cdot 5} = \frac{13}{6}.$$

### Sample Standard Deviation

**Definition 18.** The **sample standard deviation**, denoted by  $S$ , is the positive square root of the sample variance.

## 2. Point Estimator

Suppose a random sample  $x_1, x_2, \dots, x_n$  of a unknown random variable  $X$  is given. The central aim of statistics is to use the sample values  $x_1, x_2, \dots, x_n$  to infer certain properties of the unknown random variable, such as its mean or variance. Suppose, we wish to estimate the value of a parameter  $a$  (e.g. mean or variance). Since the sample values  $x_1, x_2, \dots, x_n$  provide our only source of information, any estimate of  $a$  must be some function of the  $x_1, x_2, \dots, x_n$ , i.e. some sample statistic (we denote it  $\hat{a}(x)$ ). Such a statistic is called an **estimator** or **decision function**. Hence the decision function  $S^2$ , which is a function of the random sample, is an estimator of  $\sigma^2$  and the estimate  $s^2$  is the “action” taken. If a number of random samples, each of the same size  $n$ , are taken from the one-dimensional random variable  $X$  then the value of the estimator  $\hat{a}(x)$  will vary from one sample to the next and in general will not be equal to the true value  $a$ . This variation in the estimator is described by its **sampling distribution**. For any particular quantity  $a$  we may define any number of different estimators, each of it will have its own sampling distribution. The quality of a given estimator  $\hat{a}$  may be assessed by investigating certain properties of its sampling distribution. In particular, an estimator  $\hat{a}(x)$  is usually judged

on the three criteria of **consistency**, **bias** and **efficiency**. A **point estimate** of some population parameter  $\theta$  is a single value  $\hat{\theta}$  of a statistic  $\hat{\Theta}$ . For example, the value  $\bar{x}$  of the statistics  $\bar{X}$ , computed from a sample of size  $n$ , is a point estimate of the population parameter  $\mu$ .

An estimator is not expected to estimate the population parameter without error. We do not expect  $\bar{X}$  to estimate  $\mu$  exactly, but we certainly hope that it is not too far off. For a particular sample it is possible to obtain a closer estimate of  $\mu$  by using the sample median  $\tilde{X}$  as an estimator. Consider, for instance, a sample consisting of the values 2, 5, and 11 from a population whose mean is 4 but supposedly unknown. We would estimate  $\mu$  to be  $\bar{x} = 6$ , using the sample mean as our estimate, or  $\tilde{x}$ , using the sample median as our estimate. In this case the estimator  $\tilde{X}$  produces an estimate closer to the true parameter than that of the estimator  $\bar{X}$ . On the other hand, if our random sample contains the values 2, 6, and 7, then  $\bar{x} = 5$  and  $\tilde{x} = 6$ , so that  $\bar{X}$  is now the better estimator. Not knowing the true value of  $\mu$ , we must decide in advance to use  $\bar{X}$  or  $\tilde{X}$  as our estimator.

## Consistency

An estimator  $\hat{a}$  is consistent if its value tends to the true value  $a$  in the large-sample limit, i.e.

$$\lim_{n \rightarrow \infty} \hat{a} = a.$$

Consistency is usually a minimum requirement for a useful estimator.

## Bias

The bias of an estimator is defined as

$$b(a) = E[\hat{a}] - a.$$

If  $b(a) = 0$ , i.e. if  $a = E[\hat{a}]$  then  $\hat{a}$  is called an **unbiased** estimator of the parameter  $a$ .

**Example 46.** An estimator  $\hat{a}$  is biased in such a way that

$$E[\hat{a}] = a + b(a),$$

where the bias  $b(a)$  is given by  $(b_1 - 1)a + b_2$  and  $b_1$  and  $b_2$  are known constants. Construct an unbiased estimator of  $a$ .

**Solution.** Let us first write  $E[\hat{a}]$  in the clearer form

$$E[\hat{a}] = a + (b_1 - 1)a + b_2 = b_1a + b_2.$$

The task of constructing an unbiased estimator is now trivial, an appropriate choice is

$$\hat{a}' = (\hat{a} - b_2)/b_1,$$

which (as required) has the expectation value

$$E[\hat{a}'] = \frac{E[\hat{a}] - b_2}{b_1} = a.$$

**Example 47.** Show that  $S^2$  is an unbiased estimator of the parameter  $\sigma^2$ .

**Solution.** Let us write

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X})^2 &= \sum_{i=1}^n [(X_i - \mu) - (\bar{X} - \mu)]^2 = \\ &= \sum_{i=1}^n (X_i - \mu)^2 - 2(\bar{X} - \mu) \sum_{i=1}^n (X_i - \mu) + n(\bar{X} - \mu)^2 = \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2. \end{aligned}$$

Now

$$\begin{aligned} E(S^2) &= E \left[ \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \right] = \\ &= \frac{1}{n-1} \left[ \sum_{i=1}^n E(X_i - \mu)^2 - nE(\bar{X} - \mu)^2 \right] = \frac{1}{n-1} \left( \sum_{i=1}^n \sigma_{X_i}^2 - n\sigma_{\bar{X}}^2 \right). \end{aligned}$$

Since  $\sigma_{X_i}^2 = \sigma^2$  for  $i = 1, 2, \dots, n$  and  $\sigma_{\bar{X}}^2 = \sigma^2/n$ , we get

$$E(S^2) = \frac{1}{n-1} \left( n\sigma^2 - n\frac{\sigma^2}{n} \right) = \sigma^2.$$

## Efficiency

If  $\hat{\Theta}_1$  and  $\hat{\Theta}_2$  are two unbiased estimators of the same population parameter  $\theta$ , we would choose the estimator whose sampling distribution has the smaller variance. Hence if  $\sigma_{\hat{\Theta}_1}^2 < \sigma_{\hat{\Theta}_2}^2$ , we say that  $\hat{\Theta}_1$  is a **more efficient estimator** of  $\theta$  than  $\hat{\Theta}_2$ . If we consider all possible unbiased estimators of some parameter  $\theta$ , the one with the smallest variance is called the **most efficient estimator** of  $\theta$ .

The variance of an estimator describes the spread of values  $\hat{a}$  about  $E[\hat{a}]$ . An estimator with a smaller variance is said to be more **efficient** than one with a larger variance. For any given quantity  $a$  there exists a theoretical **lower limit**  $V_{\min}$  of the variance of **any** estimator  $\hat{a}$ . The efficiency  $e$  of an estimator is defined as

$$e = V_{\min}/V[\hat{a}].$$

An estimator for which  $e = 1$  is called a **minimum-variance of efficient** estimator. Otherwise, if  $e < 1$ ,  $\hat{a}$  is called an **inefficient** estimator.

Note that some qualities of estimators are related. For example, suppose that  $\hat{a}$  is an unbiased estimator, so that  $E[\hat{a}] = a$  and  $V[\hat{a}] \rightarrow 0$  as  $n \rightarrow \infty$ . It can be proved that  $\hat{a}$  is also a consistent estimator.

### Point estimator

**Theorem 12.** *The point estimation of  $\mu$  is the value  $\mu = \bar{x}$ ; the point estimation of  $\sigma^2$  is the value  $\sigma^2 = s^2$ .*

### 3. Interval Estimation

An interval estimate of a population parameter  $\theta$  is an interval of the form  $\hat{\theta}_L < \theta < \hat{\theta}_U$  where  $\hat{\theta}_L$  and  $\hat{\theta}_U$  depend on the value of the statistic  $\hat{\Theta}$  for a particular sample and also on the sampling distribution on  $\hat{\Theta}$ .

Since different samples will generally yield different values of  $\hat{\Theta}$  and, therefore, different values of  $\hat{\theta}_L$  and  $\hat{\theta}_U$ , these end points of the interval are values of corresponding random variables  $\hat{\Theta}_L$  and  $\hat{\Theta}_U$ . From the sampling distribution of  $\hat{\Theta}$  we shall be able to determine  $\hat{\theta}_L$  and  $\hat{\theta}_U$  such that the

$$\Pr(\hat{\Theta}_L < \theta < \hat{\Theta}_U)$$

is equal to any positive fractional value we care to specify. If, for instance, we find  $\hat{\theta}_L$  and  $\hat{\theta}_U$  such that

$$\Pr(\hat{\Theta}_L < \theta < \hat{\Theta}_U) = 1 - \alpha,$$

for  $0 < \alpha < 1$ , then we have a probability of  $1 - \alpha$  of selecting a random sample that will produce an interval containing  $\theta$ . The interval

$$\hat{\theta}_L < \theta < \hat{\theta}_U,$$

computed from the selected sample, is then called a  $(1 - \alpha)100\%$  **confidence interval**, the fraction  $1 - \alpha$  is called the **confidence coefficient** or the **degree of confidence**, and the end points,  $\hat{\theta}_L$  and  $\hat{\theta}_U$ , are called the lower and upper **confidence limits**. Thus, when  $\alpha = 0,05$ , we have a 95% confidence interval, and when  $\alpha = 0,01$ , we obtain a wider 99% confidence interval.

Of course, it is better to be 95% confident that the average life of a certain car is between 6 and 7 years than to be 99% confident that it is between 3 and 10 years. Ideally we prefer a short interval with a high degree of confidence.

### Estimating the Mean

The sampling distribution of  $\bar{X}$  is centered at  $\mu$  and in most applications the variance is smaller than that of any other estimators of  $\mu$ . Thus the sample mean  $\bar{x}$  will be used as a point estimate for the population mean  $\mu$ . Recall that  $\sigma_{\bar{X}}^2 = \sigma^2/n$ , so that a large sample will yield a value of  $\bar{X}$  that comes from a sampling distribution with a small variance. Hence  $\bar{x}$  is likely to be a very accurate estimate of  $\mu$  when  $n$  is large.

Let us now consider the interval estimate of  $\mu$ . If our sample is selected from a normal population or, failing this, if  $n$  is sufficiently large, we can establish a confidence interval for  $\mu$  by considering the sampling distribution of  $\bar{X}$ . According to the Central Limit Theorem, we can expect the sampling distribution of  $\bar{X}$  to be approximately normally distributed with mean  $\mu_{\bar{X}} = \mu$  and standard deviation  $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ . Writing  $z_{\alpha/2}$  for the  $z$ -value above which we find an area of  $\alpha/2$ , we can see that

$$\Pr(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha,$$

where

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Hence

$$\Pr\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha.$$

Multiplying each term in the inequality by  $\sigma/\sqrt{n}$ , and then subtracting  $\bar{X}$

from each term and multiplying by  $-1$  we obtain

$$\Pr \left( \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha.$$

A random sample of size  $n$  is selected from a population whose variance  $\sigma^2$  is known and the mean  $\bar{x}$  is computed to give the following  $(1 - \alpha)100\%$  confidence interval.

### Confidence Interval of $\mu$ ; $\sigma$ Known

If  $\bar{x}$  is the mean of a random sample of size  $n$  from a population with known variance  $\sigma^2$ , a  $(1 - \alpha)100\%$  confidence interval for  $\mu$  is given by

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$

where  $z_{\alpha/2}$  is the  $z$ -value leaving an area of  $\alpha/2$  to the right.

For small samples selected from nonnormal populations, we cannot expect our degree of confidence to be accurate. However, for samples of size  $n \geq 30$ , regardless of the shape of most populations, sampling theory guarantees good results.

**Example 48.** The mean of the quality point averages of a random sample of 36 college seniors is calculated to be 2,6. Find the 95% and 99% confidence intervals for the mean of the entire senior class. Assume that the population standard deviation is 0,3.

**Solution.** The point estimate of  $\mu$  is  $\bar{x} = 2,6$ . The  $z$ -value, leaving an area of 0,025 to the right and therefore an area of 0,975 to the left, is (using table)  $z_{0,025} = 1,96$ . Hence the 95% confidence interval is

$$2,6 - (1,96) \cdot \left( \frac{0,3}{\sqrt{36}} \right) < \mu < 2,6 + (1,96) \cdot \left( \frac{0,3}{\sqrt{36}} \right),$$

which reduces to

$$2,50 < \mu < 2,70.$$

To find a 99% confidence interval, we find the  $z$ -value leaving an area of 0,005 to the right and 0,995 to the left. Therefore (using table again),

$z_{0,005} = 2.575$ , and the 99% confidence interval is

$$2,6 - (2,575) \cdot \left( \frac{0,3}{\sqrt{36}} \right) < \mu < 2,6 + (2,575) \cdot \left( \frac{0,3}{\sqrt{36}} \right),$$

or, simply

$$2,47 < \mu < 2,73.$$

We now see that a longer interval is required to estimate  $\mu$  with a higher degree of accuracy.

The  $(1 - \alpha)100\%$  confidence interval provides an estimate of the accuracy of our point estimate. If  $\mu$  is actually the center value of the interval, then  $\bar{x}$  estimates  $\mu$  without error. Most of the time, however,  $\bar{x}$  will not be exactly equal to  $\mu$  and the point estimator is in error. The size of this error will be the absolute value of the difference between  $\mu$  and  $\bar{x}$ , and we can be  $(1 - \alpha)100\%$  confident that this difference will not exceed  $z_{\alpha/2}\sigma/\sqrt{n}$ .

**Theorem 13.** *If  $\bar{x}$  is used as an estimate of  $\mu$ , we can then be  $(1 - \alpha)100\%$  confident that the error will not exceed  $z_{\alpha/2}\sigma/\sqrt{n}$ .*

In Example 48 we are 95% confident that the sample mean  $\bar{x} = 2,6$  differs from the true mean  $\mu$  by an amount less than 0,1 and 99% confident that the difference is less than 0,13.

Frequently we wish to know how large a sample is necessary to ensure that the error in estimating  $\mu$  will be less than a specified amount  $e$ . By previous theorem this means that we must choose  $n$  such that  $z_{\alpha/2}\sigma/\sqrt{n} = e$ . Solving this equation gives the following formula for  $n$ .

**Theorem 14.** *If  $\bar{x}$  is used as an estimate of  $\mu$ , we can be  $(1 - \alpha)100\%$  confident that the error will not exceed a specified amount  $e$  when the sample size is*

$$n = \left( \frac{z_{\alpha/2}\sigma}{e} \right)^2.$$

When solving for the sample size,  $n$ , all fractional values are rounded up to the next whole number. By adhering to this principle, we can be sure that our degree of confidence never falls below  $(1 - \alpha)100\%$ . Strictly speaking, the formula in Theorem 13 is applicable only if we know the variance of the population from which we are to select our sample. Lacking this information, we could take a preliminary sample of size  $n \geq 30$  to provide

an estimate of  $\sigma$ . Then, using  $s$  as an approximation for  $\sigma$  in Theorem 13, we could determine approximately how many observations are needed to provide the desired degree of accuracy.

**Example 49.** How large a sample is required in Example 48 if we want to be 95% confident that our estimate of  $\mu$  is off by less than 0,05?

**Solution.** The population standard deviation is  $\sigma = 0,3$ . Then, by Theorem 13,

$$n = \left[ \frac{(1,96) \cdot (0,3)^2}{0,05} \right]^2 = 138,3.$$

Therefore, we can be 95% confident that a random sample of size 139 will be provide an estimate  $\bar{x}$  differing from  $\mu$  by an amount less than 0,05.



## Chapter 2

# Polynomial Approximation to Functions

### 2.1 Lagrange's Interpolation Polynomial

Let there be known the values of some function  $f$  at  $(n + 1)$  distinct points  $x_0, x_1, \dots, x_n$  which will be denoted as follows:

$$f_i = f(x_i), \quad i = 0, 1, 2, \dots, n.$$

There arises the problem of an approximate reconstruction of the function  $f$  at an arbitrary point  $x$ . Frequently, in order to solve this problem, we construct an algebraic polynomial  $L_n(x)$  of degree  $n$  which attains the assigned values at the points  $x_i$ , that is,

$$L_n(x_i) = f_i, \quad i = 0, 1, 2, \dots, n \quad (2.1)$$

which is called the *interpolation polynomial*. The points  $x_i, i = 0, 1, \dots, n$ , are called the *interpolation points*. For the sake of convenience, here and elsewhere by a polynomial of degree  $n$  we shall understand a polynomial of degree not higher than  $n$ . For instance, if  $f_i = 0, i = 0, 1, \dots, n$ , then the interpolation polynomial  $L_n(x) \equiv 0$  actually has a zero degree, but is will also be called the polynomial of degree  $n$ .

An approximate reconstruction of the function  $f$  by the formula

$$f(x) \approx L_n(x) \quad (2.2)$$

is called the *interpolation* of the function  $f$  (with the aid of algebraic polynomial). If  $x$  is situated outside the minimum interval containing all the

interpolation points  $x_0, x_1, \dots, x_n$ , then the replacement of the function  $f$  by (2.2) is also called *extrapolation*.

**Theorem 15.** *There exists the unique interpolation polynomial of degree  $n$  satisfying the conditions (2.1).*

**Proof.** Let us establish the existence of the interpolation polynomial directly by writing out its expressions. Let  $n = 1$ , then

$$L_1(x) = \frac{x - x_1}{x_0 - x_1} \cdot f_0 + \frac{x - x_0}{x_1 - x_0} \cdot f_1. \quad (2.3)$$

For  $n = 2$

$$L_2(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} \cdot f_0 + \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} \cdot f_1 + \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} \cdot f_2 \quad (2.4)$$

and, finally, in the general case for any natural  $n$

$$L_n(x) = \sum_{i=0}^n p_{ni}(x) \cdot f_i, \quad (2.5)$$

where

$$p_{ni}(x) = \frac{(x - x_0) \dots (x - x_{i-1})(x - x_{i+1}) \dots (x - x_n)}{(x_i - x_0) \dots (x_i - x_{i-1})(x_i - x_{i+1}) \dots (x_i - x_n)}, \quad i = 0, 1, 2, \dots, n. \quad (2.6)$$

The expression (2.3) represents a linear function (a polynomial of first degree) and  $L_1(x_0) = f_0$  and  $L_1(x_1) = f_1$ . The formula (2.4) specifies a second-degree polynomial  $L_2(x)$  which satisfies the conditions (2.1) for  $n = 2$ . For arbitrary  $n$  (2.5) is also an algebraic polynomial of degree  $n$ , and since  $p_{ni}(x_i) = 1$  and  $p_{ni}(x_j) = 0$  for  $j \neq i$ ,  $0 \leq j \leq n$ , the requirements (2.1) are fulfilled.

It remains to prove the uniqueness of the interpolation polynomial. Suppose that along with the interpolation polynomial (2.5) there is also some algebraic polynomial  $\tilde{L}_n(x)$  of degree  $n$  satisfying the conditions

$$\tilde{L}_n(x_i) = f_i, \quad i = 0, 1, \dots, n.$$

Then

$$\tilde{L}_n(x_i) - L_n(x_i) = 0, \quad i = 0, 1, 2, \dots, n. \quad (2.7)$$

If

$$\tilde{L}_n(x) - L_n(x) \neq 0,$$

then this difference, being an algebraic polynomial of degree not higher than  $n$ , by virtue of the fundamental theorem of higher algebra, has at most  $n$  roots, which contradicts the equalities (2.7) whose number is equal to  $n + 1$ . Consequently,

$$\tilde{L}_n(x) \equiv L_n(x).$$

The interpolation polynomial (2.5) is called *Lagrange's interpolation polynomial*, and the polynomials (2.6) - the *Lagrangian coefficients*.

**Example 50.** Construct Lagrange's interpolation polynomial given by the data:

$i$	0	1	2	3
$x_i$	0	2	3	5
$f_i$	1	3	2	5

**Solution.** According to (2.5), for  $n = 3$  we have

$$\begin{aligned} L_3(x) &= \frac{(x-2)(x-3)(x-5)}{(0-2)(0-3)(0-5)} \cdot 1 + \frac{x(x-3)(x-5)}{2(2-3)(2-5)} \cdot 3 + \\ &\quad \frac{x(x-2)(x-5)}{3(3-2)(3-5)} \cdot 2 + \frac{x(x-2)(x-3)}{5(5-2)(5-3)} \cdot 5 = \\ &= \dots = 1 + \frac{62}{15} \cdot x - \frac{13}{6} \cdot x^2 + \frac{3}{10} \cdot x^3. \end{aligned}$$

## 2.2 Interpolation Error

We shall obtain some expression for the remainder  $R_n(x) = f(x) - L_n(x)$  in the supposition that  $f \in C^{n+1}([a, b], \mathbb{R})$ , where  $[a, b]$  is the interval containing all the interpolation points  $x_i$ ,  $i = 0, 1, \dots, n$  and the point  $x$ . We seek  $R_n(x)$  in the form:

$$R_n(x) = \omega(x)r_n(x)$$

where

$$\omega_n(x) = (x - x_0)(x - x_1) \dots (x - x_n),$$

and  $r_n(x)$  is some function.

The function

$$\varphi(t) = L_n(t) + \omega_n(t)r_n(t) - f(t) \quad (2.8)$$

vanishes for  $t = x_i$ ,  $i = 0, 1, \dots, n$  and  $t = x$ , that is, at least at  $n + 2$  points of the interval  $[a, b]$ . By Rolle's theorem,  $\varphi'(t)$  vanishes at least at the  $(n + 1)$ st point of the interval  $(a, b)$ ,  $\varphi''(t)$  is equal to zero at least at  $n$  points and so forth. Thus, there is at least one point  $\xi \in (a, b)$  at which  $\varphi^{(n+1)}(\xi) = 0$ . Whence from (2.8), bearing in mind that  $L_n^{(n+1)}(\xi) = 0$ ,  $\omega_n^{(n+1)}(\xi) = (n + 1)!$ , we obtain

$$(n + 1)!r_n(x) - f^{(n+1)}(\xi) = 0.$$

Consequently

$$r_n(x) = \frac{f^{(n+1)}(\xi)}{(n + 1)!}$$

and

$$R_n(x) = \omega_n(x) \cdot \frac{f^{(n+1)}(\xi)}{(n + 1)!};$$

$$f(x) = L_n(x) + \omega_n(x) \cdot \frac{f^{(n+1)}(\xi)}{(n + 1)!},$$

where  $\xi = \xi(x) \in (a, b)$  is some unknown point. Moreover, the estimate of the interpolation error at the running point  $x \in [a, b]$ :

$$|f(x) - L_n(x)| \leq \frac{M_{n+1}}{(n + 1)!} \cdot |\omega_n(x)| \quad (2.9)$$

and

$$\max_{[a,b]} |f(x) - L_n(x)| \leq \frac{M_{n+1}}{(n + 1)!} \max_{[a,b]} |\omega_n(x)| \quad (2.10)$$

where

$$M_{n+1} = \max_{[a,b]} |f^{(n+1)}(x)|.$$

**Example 51.** Estimate the error of approximating the function  $f(x) = \sqrt{x}$  at the point  $x = 116$  and on the entire interval  $[a, b]$ , where  $a = 100$  and  $b = 144$ , with the aid of Lagrange's interpolation polynomial  $L_2(x)$  of second degree constructed with the interpolation points  $x_0 = 100$ ,  $x_1 = 121$ , and  $x_2 = 144$ .

**Solution.** Let us compute:

$$f'(x) = \frac{1}{2\sqrt{x}}, \quad f''(x) = -\frac{1}{4\sqrt{x^3}}, \quad f'''(x) = \frac{3}{8\sqrt{x^5}}$$

and

$$M_3 = \max_{[100,144]} |f'''(x)| = \frac{3}{8\sqrt{100^5}} = \frac{3 \cdot 10^{-5}}{8}.$$

on the basis of (2.9), we obtain

$$\begin{aligned} |\sqrt{116} - L_2(116)| &\leq \frac{3 \cdot 10^{-5}}{8} \cdot \frac{1}{3!} |(116 - 100) \cdot (116 - 121) \cdot (116 - 144)| = \\ &= \frac{1}{16} \cdot 10^{-5} \cdot 16 \cdot 5 \cdot 28 = 1 \cdot 41 \cdot 10^{-3}. \end{aligned}$$

By virtue of (2.10)

$$\max_{[100,144]} |\sqrt{x} - L_2(x)| \leq \frac{10^{-5}}{16} \max_{[100,144]} |(x-100) \cdot (x-121) \cdot (x-144)| \approx 2 \cdot 5 \cdot 10^{-3}.$$

## 2.3 Linear Interpolation

The interpolation with the aid of the linear function

$$L_1(x) = \frac{x - x_1}{x_0 - x_1} \cdot f_0 + \frac{x - x_0}{x_1 - x_0} \cdot f_1$$

is known as *linear interpolation*. Setting  $h = x_1 - x_0$  and  $q = (x - x_0)/h$ , we can write the formula for linear interpolation as

$$f(x) \approx L_1(x) = L_1(x_0 + qh) = (1 - q)f_0 + qf_1.$$

The quantity  $q$  is called the *interpolation phase*. The latter changes within the limits from 0 to 1 as  $x$  runs through the values from  $x_0$  to  $x_1$ . Geometrically, linear interpolations means the replacement of the graph of the

function on the interval  $[x_0, x_1]$  by the chord joining the points  $(x_0, f_0)$  and  $(x_1, f_1)$ . Since

$$\omega_2(x) = (x - x_0)(x - x_1)$$

and, consequently,

$$\max_{[x_0, x_1]} |\omega_2(x)| = \max_{[x_0, x_1]} |(x - x_0)(x - x_1)| = \frac{h^2}{4},$$

the estimate of the maximum error of the linear interpolation on  $[x_0, x_1]$  has the form

$$\max_{[x_0, x_1]} |f(x) - L_1(x)| \leq \frac{M_2 \cdot h^2}{8},$$

where

$$M_2 = \max_{[x_0, x_1]} |f''(x)|.$$

Frequently, a table is given containing a great number of values of some function  $f$  with constant step  $h$  of argument variation. Then for a given  $x$  two nearest to it interpolation points are chosen. The left-hand point is taken for  $x_0$  and the right-hand point for  $x_1$  and linear interpolation by above formula is realized.

## 2.4 Finite and Divided Differences

**Finite differences.** Let  $x_k = x_0 + kh$ , where  $k$  is an integer,  $h$  is the step,  $h > 0$  and  $f_k = f(x_k)$ . The quantity

$$\Delta f_k = f(x_k + h) - f(x_k) = f(x_{k+1}) - f(x_k) = f_{k+1} - f_k$$

is called the *finite difference of the first order* of the function  $f$  at the point  $x_k$  (with step  $h$ ), and

$$\begin{aligned} \Delta^2 f_k &= \Delta(\Delta f_k) = \Delta f_{k+1} - \Delta f_k = \\ &= (f_{k+2} - f_{k+1}) - (f_{k+1} - f_k) = f_k - 2f_{k+1} + f_{k+2} \end{aligned}$$

is the *finite difference of the second order* at the point  $x_k$ . In general, the  $n$ th - order finite difference of the function  $f$  at the point  $x_k$  is determined by the recurrence formula

$$\Delta^n f_k = \Delta(\Delta^{n-1} f_k) = \Delta^{n-1} f_{k+1} - \Delta^{n-1} f_k,$$

where  $n \geq 1$  and  $\Delta^0 f_k \equiv f_k$ . When carrying out computations, it is convenient to write finite differences in tabular form:

$x_0$	$f_0$				
		$\Delta f_0$			
$x_1$	$f_1$		$\Delta^2 f_0$		
		$\Delta f_1$		$\Delta^3 f_0$	
$x_2$	$f_2$		$\Delta^2 f_1$		$\Delta^4 f_0$
		$\Delta f_2$		$\Delta^3 f_1$	
$x_3$	$f_3$		$\Delta^2 f_2$		
		$\Delta f_3$			
$x_4$	$f_4$				

**Theorem 16.** *The finite difference of order  $n$  of an algebraic polynomial of degree  $n$  is constant, i.e. is independent of  $k$ , while the finite differences of any higher order are equal to zero.*

**Divided differences.** Let now  $x_0, x_1, \dots, x_n$  be arbitrary points (nodes) on the  $x$ -axis, and  $x_i \neq x_j$  for  $i \neq j$ . The values  $f(x_0), f(x_1), \dots, f(x_n)$  are called the *divided differences of zeroth order*. The number

$$f(x_0; x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

is called the *divided difference of the first order* of the function  $f$ . The *divided difference of the order  $n$*  of the function  $f$  is defined via the divided differences of order  $n - 1$  by the recurrence formula

$$f(x_0; x_1; \dots; x_n) = \frac{f(x_1, x_2, \dots, x_n) - f(x_0, x_1, \dots, x_{n-1})}{x_n - x_0}.$$

The divided differences are written in the form of array:

$x_0$	$f(x_0)$				
		$f(x_0; x_1)$			
$x_1$	$f(x_1)$		$f(x_0; x_1; x_2)$		
		$f(x_1; x_2)$		$f(x_0; x_1; x_2; x_3)$	
$x_2$	$f(x_2)$		$f(x_1; x_2; x_3)$		$f(x_0; x_1; x_2; x_3; x_4)$
		$f(x_2; x_3)$		$f(x_1; x_2; x_3; x_4)$	
$x_3$	$f(x_3)$		$f(x_2; x_3; x_4)$		
		$f(x_3; x_4)$			
$x_4$	$f(x_4)$				

**Theorem 17.** *The  $n$ th-order divided difference is expressed in terms of the nodal values of the function by the formula*

$$f(x_0; x_1; \dots; x_n) = \sum_{i=0}^n \frac{f(x_i)}{(x_i - x_0) \dots (x_i - x_{i-1})(x_i - x_{i+1}) \dots (x_i - x_n)},$$

*i.e. it is a symmetric function of its arguments.*

**Theorem 18.** *If  $x_k = x_0 + kh$ ,  $k = 0, 1, \dots$  then there is the following relation between the  $n$ th-order divided difference and the  $n$ th-order finite difference:*

$$f(x_0; x_1; \dots; x_n) = \frac{\Delta^n f_0}{n!h^n}.$$

**Theorem 19.** *Let  $[\alpha, \beta]$  be the minimum interval containing the points  $x_0, x_1, \dots, x_n$ ,  $f \in C^n([\alpha, \beta], \mathbb{R})$ . Then there is a point  $\eta \in (\alpha, \beta)$  such that*

$$f(x_0; x_1; \dots; x_n) = \frac{f^n(\eta)}{n!}.$$

## 2.5 Newton's Interpolation Formula

Let  $x_0, x_1, \dots, x_n$  be arbitrary pairwise noncoinciding interpolation points at which the values of the function  $f$  are known.

**Theorem 20.** *The  $n$ th-degree algebraic polynomial*

$$\begin{aligned} l_n(x) = & f(x_0) + (x - x_0)f(x_0; x_1) + \\ & (x - x_0)(x - x_1)f(x_0; x_1; x_2) + \dots + \\ & (x - x_0)(x - x_1) \dots (x - x_{n-1})f(x_0; x_1; \dots; x_n) \end{aligned} \quad (2.11)$$

*is an interpolation one, that is,*

$$l_n(x_i) = f(x_i), \quad i = 0, 1, \dots, n.$$

Let us prove this equalities for  $n = 2$ . We have

$$l_2(x) = f(x_0) + (x - x_0)f(x_0; x_1) + (x - x_0)(x - x_1)f(x_0; x_1; x_2).$$

Obviously,  $l_2(x_0) = f(x_0)$ . Further

$$\begin{aligned} l_2(x_1) = & f(x_0) + (x_1 - x_0)f(x_0; x_1) = \\ & f(x_0) + (x_1 - x_0) \cdot \frac{f(x_1) - f(x_0)}{x_1 - x_0} = f(x_1). \end{aligned}$$

Finally

$$\begin{aligned}
 l_2(x_2) &= f(x_0) + (x_2 - x_0)f(x_0; x_1) + \\
 & (x_2 - x_0)(x_2 - x_1)f(x_0; x_1; x_2) = f(x_0) + \frac{x_2 - x_0}{x_1 - x_0} \cdot (f(x_1) - f(x_0)) + \\
 & \frac{(x_2 - x_0)(x_2 - x_1)}{x_2 - x_0} \cdot \left( \frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0} \right) = \\
 & f(x_0) + \frac{x_2 - x_0}{x_1 - x_0} \cdot (f(x_1) - f(x_0)) + f(x_2) - f(x_1) - \\
 & \frac{x_2 - x_1}{x_1 - x_0} \cdot (f(x_1) - f(x_0)) = f(x_2).
 \end{aligned}$$

For arbitrary natural  $n > 2$  equalities can be proved by induction.

## 2.6 The case of equally spaced interpolation points

Let  $x_k = x_0 + kh$ ,  $h > 0$ ,  $k = 0, 1, \dots, n$ , and  $f_k = f(x_k)$ . We may rewrite the interpolation polynomial (2.11) in the form (where  $q = (x - x_0)/h$ ):

$$\begin{aligned}
 l_n(x) &= l_n(x_0 + qh) = \\
 & f_0 + q \cdot \frac{\Delta f_0}{1!} + q(q - 1) \cdot \frac{\Delta^2 f_0}{2!} + \dots + q(q - 1) \dots (q - (n - 1)) \cdot \frac{\Delta^n f_0}{n!}.
 \end{aligned} \tag{2.12}$$

It is so called *Newton's interpolation polynomial for forward interpolation*. The point  $q$  is situated at the leftmost interpolation point  $x_0$ . This polynomial may be conveniently used for interpolation at the beginning of the table and for extrapolation to the left of the point  $x_0$ , that is, for  $q < 0$ .

**Example 52.** Let there be given a table of the values of the function

$f(x) = \sin x$  and its finite differences:

$x$	$f(x)$	$\Delta f(x)$	$\Delta^2 f(x)$	$\Delta^3 f(x)$
$5^\circ$	0,087156			
		34.713		
$7^\circ$	0,121869		-148	
		34.565		-42
$9^\circ$	0,156434		-190	
		34.375		-43
$11^\circ$	0,190809		-233	
		34.142		-41
$13^\circ$	0,224951		-274	
		33.868		
$15^\circ$	0,258819			

(For the sake of simplicity, the finite differences of a function are customarily written without an explicit indication of the position of the decimal point.) Suppose that it is required to find  $\sin 6^\circ$ . In (2.12) we set  $n = 3$ ,  $x_0 = 5^\circ$ ,  $h = 2$ ,  $q = (6^\circ - 5^\circ)/2^\circ = 1/2$ . The relevant computations have the form:

$f_0 =$	0.087156
$q\Delta f_0 = 0.5 \cdot 0.034713 =$	0.0073565
$q(q-1)\Delta^2 f_0/2! = (1/8) \cdot 0.000148 =$	0.0000185
$q(q-1)(q-2)\Delta^3 f_0/3! = (-1/16) \cdot 0.000042 =$	- 0.0000026
$l_3(6^\circ) =$	0.104528

Here, the intermediate quantities are found with seven digits after decimal point. The seventh digit is a spare one and in the final result it is rounded off. The exact value of  $\sin 6^\circ$  rounded off to six decimal places is equal to 0.104528, that is, all digits of  $l_3(6^\circ)$  turn out to be correct. (Let us note that the exact value:  $\sin 6^\circ = 0,1045284$ .)

The interpolation polynomial with points  $x_0, x_{-1}, \dots, x_{-n}$ , where  $x_{-k} = x_0 - k \cdot h$ , has the form

$$l_n(x) = l_n(x_0 + qh) = f_0 + q \cdot \frac{\Delta f_{-1}}{1!} + q(q+1) \cdot \frac{\Delta^2 f_{-2}}{2!} + \dots + q(q+1) \dots (q+n-1) \cdot \frac{\Delta^n f_{-n}}{n!}. \quad (2.13)$$

It is called *Newton's interpolation polynomial for backward interpolation*. In it the reference point  $q$  is situated at the rightmost interpolation point  $x_0$ , and the finite differences used are contained in the table from  $f_0$  to the right upwards:

$x_{-4}$	$f_{-4}$				
		$\Delta f_{-4}$			
$x_{-3}$	$f_{-3}$		$\Delta^2 f_{-4}$		
		$\Delta f_{-3}$		$\Delta^3 f_{-4}$	
$x_{-2}$	$f_{-2}$		$\Delta^2 f_{-3}$		$\Delta^4 f_{-4}$
		$\Delta f_{-2}$		$\Delta^3 f_{-3}$	
$x_{-1}$	$f_{-1}$		$\Delta^2 f_{-2}$		
		$\Delta f_{-1}$			
$x_0$	$f_0$				

The interpolation polynomial (2.13) may be conveniently used for interpolation at the end of the table and for extrapolation to the right of the point  $x_0$ , that is, for  $q > 0$ .

If for a given  $x$ , the table of the values has a sufficient number of interpolation points on either side of  $x$ , then, it is convenient to choose the interpolation points  $x_0, x_1, \dots, x_n$  so that the point  $x$  is found as close as possible to the middle of the minimum interval these points. In this case, the interpolation polynomial can be constructed in different ways. The most natural way to specify the interpolation polynomial in the form (2.11) where the nearest to  $x$  point is taken as  $x_0$  (on one side of  $x$ ) and then the nearest to  $x$  point is taken as  $x_1$  (on the opposite side). The subsequent interpolation points are taken by turns on opposite sides of  $x$  and are situated as close to  $x$  as possible. With such a choice of the interpolation points, the subsequent terms in (2.11) are usually decreasing if  $h$  is small and  $n$  is not large.

The remainder of the interpolation polynomial (2.12) may be written (if use the remainder of  $L_n(x)$ ) in the form

$$R_n(x) = R_n(x_0 + qh) = h^{n+1} \bar{\omega}_n(q) \cdot \frac{f^{n+1}(\xi)}{(n+1)!},$$

where

$$\bar{\omega}_n(q) = q(q-1) \dots (q-n)$$

and the remainder of the interpolation polynomial (2.13) may be written in the form

$$R_n(x) = R_n(x_0 + qh) = h^{n+1}q(q+1)\dots(q+n) \cdot \frac{f^{n+1}(\xi)}{(n+1)!},$$

where  $f^{n+1}$  is the derivative with respect to  $x$  and  $\xi$  is some point of the minimum interval containing the interpolation points  $x_0, x_{-1}, \dots, x_{-n}$  and the point  $x$ . If  $h$  is small and the function  $f$  is sufficiently smooth, then the current term in (2.12) is approximately equal to the error of interpolation by the polynomial made up of all the previous terms. This remark also refers to the interpolation polynomial (2.13) for backward interpolation.

# Chapter 3

## Numerical Differentiation

### 3.1 Simplest Formulas of Numerical Differentiation

Let us assume that at some point  $x$  the function  $f$  has the derivative

$$f'(x) = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$

Then it is natural to set

$$f'(x) \approx \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$

There arises the question: What is the error (i.e. what is the difference between the left-hand and right-hand members) of this approximate equality? To obtain the quantitative estimates of the error, the sole fact of the existence of  $f'(x)$  is insufficient. Therefore, when analyzing the error of approximate formulas of numerical differentiation, we usually demand that the given function have some derivative of a higher order than the desired derivative.

Let  $x_i = x_0 + i \cdot h$ ,  $i = 0, \pm 1, \pm 2, \dots$ , where  $h > 0$  is the step. Let us set  $f_i = f(x_i)$ ,  $f'_i = f'(x_i)$ , and so forth. Suppose that  $f \in C^2([x_0, x_1], \mathbb{R})$ . Then there is a point  $\xi$  such that

$$f'_0 = \frac{f_1 - f_0}{h} - \frac{h}{2} \cdot f''(\xi), \quad x_0 < \xi < x_1. \quad (3.1)$$

If  $f \in C^3([x_{-1}, x_1], \mathbb{R})$ , then, in addition,

$$f'_0 = \frac{f_1 - f_{-1}}{2h} - \frac{h^2}{6} \cdot f'''(\xi), \quad x_{-1} < \xi < x_1. \quad (3.2)$$

Under the condition that  $f \in C^{(4)}[x_{-1}, x_1]$ , we have

$$f_0'' = \frac{f_{-1} - 2f_0 + f_1}{h^2} - \frac{h^4}{12} \cdot f^{(4)}(\xi), \quad x_{-1} < \xi < x_1. \quad (3.3)$$

The point  $\xi$  in each of the formulas is unknown.

Let us prove the relationships (3.1) and (3.3). According to Taylor's formula, we have

$$f_1 = f_0 + hf_0' + \frac{h^2}{2}f_0''(\xi),$$

where  $\xi$  is some point of the interval  $(x_0, x_1)$ , that is, (3.1) holds true. Analogously, if  $f \in C^{(4)}$  on  $[x_{-1}, x_1]$ , then

$$f_{\pm 1} = f_0 \pm hf_0' + \frac{h^2}{2}f_0'' + \frac{h^4}{24}f^{(4)}(\xi_{\pm}),$$

where  $\pm$  may be replaced either everywhere by  $+$  or everywhere by  $-$  and  $x_{-1} < \xi_- < \xi_+ < x_1$ . Then

$$f_{-1} + f_1 = 2f_0 + h^2f_0'' + \frac{h^4}{24} \cdot (f^{(4)}(\xi_+) + f^{(4)}(\xi_-)).$$

It can be proved that

$$f^{(4)}(\xi_+) + f^{(4)}(\xi_-) = 2f^{(4)}(\xi),$$

where  $\xi \in [\xi_-, \xi_+]$ . Then

$$f_0'' = \frac{f_{-1} - 2f_0 + f_1}{h^2} - \frac{h^4}{12} \cdot f^{(4)}(\xi),$$

that is, we arrive at the relationship (3.3). The formulas (3.1) - (3.3) are called the *formulas of numerical differentiation with remainders*, and the formulas

$$f_0' \approx \frac{f_1 - f_0}{h}, \quad f_0' \approx \frac{f_1 - f_{-1}}{2h}, \quad f_0'' \approx \frac{f_{-1} - 2f_0 + f_1}{h^2}$$

simply the *formulas of numerical differentiation*. Errors of these formulas are

$$\left| f_0' - \frac{f_1 - f_0}{h} \right| \leq \frac{h}{2} \max_{[x_0, x_1]} |f''(x)|,$$

(error is the first order with respect to  $h$  (or to be of order  $h$ ));

$$\left| f'_0 - \frac{f_1 - f_{-1}}{2h} \right| \leq \frac{h^2}{6} \max_{[x_{-1}, x_1]} |f'''(x)|,$$

(error here and in the next relation is said to have the second order with respect to  $h$  (or to be of order  $h^2$ )),

$$\left| f''_0 - \frac{f_{-1} - 2f_0 + f_1}{h^2} \right| \leq \frac{h^2}{12} \max_{[x_{-1}, x_1]} |f^{(4)}(x)|.$$

### 3.2 Applying Lagrange's Interpolation Polynomial

One of the general-purpose methods of constructing the formulas of numerical differentiation consists in the following: using the values of the function  $f$  at some points  $x_0, x_1, \dots, x_n$ , we construct the interpolation polynomial  $L_n(x)$  and approximately set:

$$f^{(m)} \approx L_n^{(m)}, \quad 0 \leq m \leq n.$$

If e.g.  $n = 2$  (the case of three points) then

$$L'_2(x) = \frac{2x - x_1 - x_2}{(x_0 - x_1)(x_0 - x_2)} \cdot f_0 + \frac{2x - x_0 - x_2}{(x_1 - x_0)(x_1 - x_2)} \cdot f_1 + \frac{2x - x_0 - x_1}{(x_2 - x_0)(x_2 - x_1)} \cdot f_2,$$

$$L''_2(x) = \frac{2f_0}{(x_0 - x_1)(x_0 - x_2)} + \frac{2f_1}{(x_1 - x_0)(x_1 - x_2)} + \frac{2f_2}{(x_2 - x_0)(x_2 - x_1)}$$

and

$$f'_0 = \frac{1}{2h} \cdot (-3f_0 + 4f_1 - f_2) + \frac{h^3}{3} \cdot f'''(\xi),$$

$$f'_1 = \frac{1}{2h} \cdot (f_2 - f_0) - \frac{h^2}{6} \cdot f'''(\xi),$$

$$f'_2 = \frac{1}{2h} \cdot (f_0 - 4f_1 + 3f_2) + \frac{h^2}{3} \cdot f'''(\xi),$$

$$f''_0 = \frac{1}{h^2} \cdot (f_0 - 2f_1 + f_2) - h \cdot f'''(\xi),$$

$$f_1'' = \frac{1}{h^2} \cdot (f_0 - 2f_1 + f_2) - \frac{h^2}{12} \cdot f^{(4)}(\xi),$$

$$f_2'' = \frac{1}{h^2} \cdot (f_0 - 2f_1 + f_2) + h \cdot f'''(\xi).$$

If  $m = 1$  and  $n = 3$  (the case of four points), then:

$$f_0' = \frac{1}{6h} \cdot (-11f_0 + 18f_1 - 9f_2 + 2f_3) - \frac{h^3}{4} \cdot f^{(4)}(\xi),$$

$$f_1' = \frac{1}{6h} \cdot (-2f_0 - 3f_1 + 6f_2 - f_3) + \frac{h^3}{12} \cdot f^{(4)}(\xi),$$

$$f_2' = \frac{1}{6h} \cdot (f_0 - 6f_1 + 3f_2 + 2f_3) - \frac{h^3}{12} \cdot f^{(4)}(\xi),$$

$$f_3' = \frac{1}{6h} \cdot (-2f_0 + 9f_1 - 18f_2 + 11f_3) + \frac{h^3}{4} \cdot f^{(4)}(\xi).$$

If  $m = 2$  and  $n = 3$  then:

$$f_0'' = \frac{1}{h^2} \cdot (2f_0 - 5f_1 + 4f_2 - f_3) + \frac{11h^2}{12} \cdot f^{(4)}(\xi),$$

$$f_1'' = \frac{1}{h^2} \cdot (f_0 - 2f_1 + f_2) - \frac{h^2}{12} \cdot f^{(4)}(\xi),$$

$$f_2'' = \frac{1}{h^2} \cdot (f_1 - 2f_2 + f_3) - \frac{h^2}{12} \cdot f^{(4)}(\xi),$$

$$f_3'' = \frac{1}{h^2} \cdot (-f_0 + 4f_1 - 5f_2 + 2f_3) + \frac{11h^2}{12} \cdot f^{(4)}(\xi).$$

With an *increase* in  $n$  and an appropriate smoothness of the function  $f$ , the order of accuracy of the formulas is *increased*, and with an *increase* in  $m$  (that is, in the derivative number) the order of accuracy with respect to  $h$  *decreases*. The expressions of the derivatives at the points situated closer to the middle of interval  $[x_0, x_n]$  are simpler than those at its ends. For an even  $n$  the order of accuracy of the formula at the middle point for an even derivative is a unity higher than at the remaining points. Therefore it is recommended to use the formulas of numerical differentiation with points arranged symmetrically about the point at which the derivative is determined.

In formulas of numerical differentiation with constant step  $h$  the values of

the function  $f$  are divided by  $h^m$ , where  $m$  is the order of the computed derivative. Therefore for a small  $h$  the inherent errors in the values of the function  $f$  strongly affect the result of numerical differentiation. Thus, there arises the problem of the choice of an optimal step  $h$ , since the error of the method itself tends to zero as  $h \rightarrow 0$ , and the inherent errors increases.

### 3.3 Applications of Newton's Interpolation Polynomial

All the listed formulas are expressed in terms of the tabular values of the given function. If we differentiate Newton's interpolation polynomial  $l_n(x)$ , then we obtain the formula of numerical differentiation expressed in terms of finite differences of the function. Taking into consideration that

$$\frac{d}{dx} = \frac{1}{h} \frac{d}{dq}$$

we find (from (2.12), p. 69)

$$f'(x) \approx \frac{1}{h} \cdot \frac{d}{dq} l_n(x_0 + qh) = \frac{1}{h} \cdot \left( \Delta f_0 + (2q - 1) \cdot \frac{\Delta^2 f_0}{2!} + (3q^2 - 6q + 2) \cdot \frac{\Delta^3 f_0}{3!} + \dots \right).$$

This formula is convenient to be used for interpolation at the beginning of the table of the values of the function  $f$  with step  $h$ . A similar formula of numerical differentiation can be obtained from Newton's interpolation polynomial for backward interpolation.

### 3.4 General error estimate

It is possible to obtain the error estimate of the general formula of numerical differentiation which is expressed in the form of an inequality in terms of the modulus maximum of the derivatives. We shall confine ourselves to considering the case when the interpolation points are spaced at a constant step  $h$ .

**Theorem 21.** *Let  $x_i = x_0 + ih$ ,  $h > 0$ ,  $i = 0, 1, \dots, n$ ,  $0 \leq k \leq n$ , and  $f \in C^{k+1}([x_0, x_n], \mathbb{R})$ . Then there are constants  $a_{nkm}$  dependent only on  $n, k$  and  $m$  and independent of the step  $h$  and the function  $f$  such that*

$$\max_{[x_0, x_n]} |f^{(m)}(x) - L_n^{(m)}(x)| \leq h^{k+1-m} \cdot a_{nkm} \max_{[x_0, x_n]} |f^{(k+1)}(x)|,$$

where  $L_n(x)$  is Lagrange's interpolation polynomial for the function  $f$ ,  $0 \leq m \leq k \leq n$ .

# Chapter 4

## Splines

Let the interval  $[a, b]$  be divided into  $N$  equal subintervals  $[x_i, x_{i+1}]$ , where  $x_i = a + ih$ ,  $i = 0, 1, \dots, N - 1$ ,  $x_N = b$ , and  $h = (b - a)/N$ .

The *spline* is defined as a function which, together with its several derivatives, is continuous throughout the given interval  $[a, b]$  and on each separate subinterval  $[x_i, x_{i+1}]$  it is some algebraic polynomial. The maximum (over all the subintervals) degree of polynomials is called the *degree* of a spline, and the difference between the degree of a spline and the order of the highest derivative continuous on  $[a, b]$  is called the *defect* of a spline.

For instance, a continuous piecewise linear function (a broken line) is a first-degree spline with defect equal to unity, since only the function itself (or zero derivative) is continuous, while the first derivative is already discontinuous.

In practice, the most common are third-degree splines having a continuous, at least, first derivative on  $[a, b]$ . These are called *cubic splines* and are denoted by  $S_3(x)$ . The quantity  $m_i = S'_3(x_i)$  is termed the *inclination* of the spline at the point (knot)  $x_i$ . It is not difficult to make sure that the cubic spline  $S_3(x)$  which attains the respective values  $f_i$  and  $f_{i+1}$  at the knots  $x_i$  and  $x_{i+1}$  has the following expression on the subinterval  $[x_i, x_{i+1}]$ :

$$S_3(x) = \frac{1}{h^3} \cdot (x_{i+1} - x)(2(x - x_i) + h) \cdot f_i + \frac{1}{h^3} \cdot (x - x_i)^2(2(x_{i+1} - x) + h) \cdot f_{i+1} + \frac{1}{h^2} \cdot (x_{i+1} - x)^2(x - x_i) \cdot m_i + \frac{1}{h^2} \cdot (x - x_i)^2(x - x_{i+1}) \cdot m_{i+1}.$$

Indeed, it is readily seen that  $S_3(x_i) = f_i$  and  $S_3(x_{i+1}) = f_{i+1}$ . Besides, simple computations show that  $S'_3(x_i) = m_i$  and  $S'_3(x_{i+1}) = m_{i+1}$ . It is possible to prove that any third-degree algebraic polynomial, which attains

the values respectively equal to  $f_i$  and  $f_{i+1}$  at the knots  $x_i$  and  $x_{i+1}$  and has a derivative respectively equal to  $m_i$  and  $m_{i+1}$  at these points, coincides identically with given polynomial.

Thus, to define the cubic spline  $S_3(x)$  over the whole interval  $[a, b]$ , we have to specify its values  $f_i$  at  $N + 1$  knots  $x_i$  and its inclinations  $m_i$   $i = 0, 1, \dots, N$ . A cubic spline which attains the same values  $f_i$  at the knots  $x_i$  as a function  $f$  is called the *interpolation spline*. It serves for approximating the function  $f$  on the interval  $[a, b]$  together with several derivatives.

## 4.1 Methods of Specifying the Inclinations of an Interpolation Cubic Spline

### 1. Method I. (Simplified)

We set

$$m_0 = \frac{1}{2h}(4f_1 - f_2 - 3f_0),$$

$$m_i = \frac{1}{2h}(f_{i+1} - f_{i-1}), \quad i = 1, 2, \dots, N - 1,$$

$$m_N = \frac{1}{2h}(3f_N + f_{N-2} - 4f_{N-1}).$$

These formulas are formulas of numerical differentiation of the second order of accuracy with respect to the step  $h = (b - a)/N$ .

### 2. Method II

If the values  $f'_i$  of the derivative  $f'$  at the knots  $x_i$  are known, then we set

$$m_i = f'_i, \quad i = 0, 1, \dots, N.$$

Methods I and II are called *local* since with their aid the spline is constructed separately on each subinterval  $[x_i, x_{i+1}]$  (directly by above formula). In doing so, nevertheless, the continuity of the derivative  $S'_3(x)$  at the knots  $x_i$  is observed. But the continuity of the second derivative  $S''_3(x)$  at the knots of the spline constructed by these methods is not guaranteed. Therefore the defect of such a spline is usually equal to two.

### Method III (Global)

Compute

$$S_3''(x_i + 0) = -\frac{4m_i}{h} - \frac{2m_{i+1}}{h} + 6 \cdot \frac{f_{i+1} - f_i}{h^2},$$

$$S_3''(x_i - 0) = \frac{2m_{i-1}}{h} + \frac{4m_i}{h} - 6 \cdot \frac{f_i - f_{i-1}}{h^2}.$$

Then the continuity of  $S''(x)$  is required at the knots:

$$S_3''(x_i - 0) = S_3''(x_i + 0), \quad i = 1, 2, \dots, N - 1.$$

We arrive at the following system of linear algebraic equations with respect to inclinations:

$$m_{i-1} + 4m_i + m_{i+1} = \frac{3(f_{i+1} - f_{i-1})}{h}, \quad i = 1, 2, \dots, N - 1. \quad (4.1)$$

Since there are  $N + 1$  unknowns, it is necessary to specify two more conditions which are called the *boundary conditions*. Let us give three variants of boundary conditions:

(a) With  $f'_0 = f'(a)$  and  $f'_N = f'(b)$  known we specify:

$$m_0 = f'_0, \quad m_N = f'_N.$$

(b) The derivatives  $f'_0$  and  $f'_N$  are approximated by the formulas of numerical differentiation of the third order of accuracy. We set

$$m_0 = \frac{1}{6h} \cdot (-11f_0 + 18f_1 - 9f_2 + 2f_3),$$

$$m_N = \frac{1}{6h} \cdot (11f_N - 18f_{N-1} + 9f_{N-2} - 2f_{N-3}).$$

(b) In some cases the values  $f''$  at the end points of the interval  $[a, b]$  are known, that is, the quantities  $f''_0 = f''(a)$  and  $f''_N = f''(b)$ . Then the requirements  $S_3''(a) = f''_0$  and  $S_3''(b) = f''_N$  lead to the boundary conditions

$$m_0 = -\frac{m_1}{2} + \frac{3}{2h} \cdot (f_1 - f_0) - \frac{h}{4}f''_0,$$

$$m_N = -\frac{m_{N-1}}{2} + \frac{3}{2h} \cdot (f_N - f_{N-1}) + \frac{h}{4}f''_N.$$

The boundary conditions (a) - (c) may be combined. For all considered boundary conditions, the system (4.1) has a unique solution. Solving this system we find the inclination  $m_i$ ,  $i = 0, 1, \dots, N$ . Then we specify the spline on each subinterval  $[x_i, x_{i+1}]$ ,  $i = 0, 1, \dots, N - 1$ . This spline has a defect not exceeding unity since its second derivative is continuous on  $[a, b]$ .

## 4.2 The Error of Approximation by a Spline

**Theorem 22.** *If  $f \in C^{k+1}([a, b], \mathbb{R})$ ,  $0 \leq k \leq 3$ , then the interpolation spline  $S_3(x)$ , with the inclinations given by Method II or by Method III, satisfies the inequality*

$$\max_{[x_i, x_{i+1}]} |f^{(m)}(x) - S_3^{(m)}(x)| \leq Ch^{k+1-m} \cdot \max_{[a, b]} |f^{(k+1)}(x)|,$$

where  $i = 0, 1, \dots, N - 1$ ,  $m = 0, 1, \dots, k$ , and  $C$  is a constant independent of  $h, i$  and  $f$ .

Splines are a more convenient means of approximation of functions on big intervals (for large  $N$ ) than, say interpolation polynomials. An approximation of a function on a big interval by one polynomial may require a considerable increase in its degree in order to achieve the assigned accuracy, which is unacceptable in practice. The subdivision of the given interval  $[a, b]$  into several parts with an interpolation polynomial constructed independently on each subinterval is inconvenient, since the first derivative of two neighbouring interpolation polynomials will have a discontinuity. It may even happen that the values themselves of two adjacent interpolation polynomials will not coincide at the joint if the joint point is not their common knot. The cubic spline  $S_3(x)$  whose inclinations are found by the global method is twice continuously differentiable on the entire interval  $[a, b]$ , that is, it has a continuous curvature. The accuracy of the approximation of the function  $f$  by the spline  $S_3(x)$  is controlled by the choice of  $N$ , that is, by the step  $h = (b - a)/N$ .

# Chapter 5

## The Method of Least Squares

### 5.1 Introduction

**Definition 19.** A set  $M$  is called a *linear space* if the operations of addition and multiplication by the real numbers within the bounds of  $M$  and satisfying the following conditions are defined in this set:

1. addition is associative:  $(f + g) + r = f + (g + r)$ ;
2. addition is commutative:  $f + g = g + f$ ;
3. there exists a zero element  $\theta \in M$ , i.e.  $f + \theta = f$  for every  $f \in M$ ;
4.  $0 \cdot f = \theta$  for every  $f \in M$ ;
5.  $(\alpha + \beta) \cdot f = \alpha f + \beta f$ ;
6.  $\alpha(f + g) = \alpha f + \alpha g$ ;
7.  $\alpha(\beta f) = (\alpha\beta)f$ ;
8.  $1 \cdot f = f$ .

Here  $\alpha$  and  $\beta$  are real numbers.

**Definition 20.** A *scalar product* is said to be introduced in the linear space  $F$  if to each pair of elements  $f, g \in F$  there corresponds a real number denoted by  $(f, g)$  and called the *scalar product* of the elements  $f$  and  $g$  which satisfies the following *axioms of scalar product*:

1.  $(f, g) = (g, f)$ ;

2.  $(f, f) \geq 0$ , where  $(f, f) = 0$  if and only if  $f = \theta$ , that is,  $f$  is a zero element in  $F$ ;
3.  $(\alpha f, g) = \alpha(f, g)$ , where  $\alpha$  is any real number;
4.  $(f_1 + f_2, g) = (f_1, g) + (f_2, g)$ .

The linear space  $F$  with scalar product  $(f, g)$  introduced in it is called the *Euclidean space* and is denoted by  $E$ .

We shall be interested in two concrete Euclidean spaces, namely, the space  $E = E_C$  of functions which are continuous on the interval  $[a, b]$  with the scalar product

$$(f, g) = \frac{1}{b-a} \cdot \int_a^b f(x)g(x) dx \quad (5.1)$$

and the linear space  $E = E_{n+1}$  of functions defined on a finite (discrete) set of points  $x_0, x_1, \dots, x_n$  of some interval  $[a, b]$  with the scalar product

$$(f, g) = \frac{1}{n+1} \cdot \sum_{i=0}^n f(x_i)g(x_i). \quad (5.2)$$

**Definition 21.** A set  $F$  is called a *linear normed space* if it is linear and each element  $f \in F$  is associated with a real number  $\|f\|$  which is called the *norm* of  $f$  and satisfies the *axioms of norm*:

1.  $\|f\| \geq 0$ , and  $\|f\| = 0$  if and only if  $f = \theta$ , that is,  $f$  is a zero element in  $F$ ;
2.  $\|\alpha f\| = |\alpha|\|f\|$  for any real  $\alpha$ ;
3.  $\|f + g\| \leq \|f\| + \|g\|$  for any  $f, g \in F$ .

Axiom 3 is called the *triangle inequality* for the norm.

**Example 53.** The class  $C[a, b]$  of all continuous functions specified on the interval  $[a, b]$  is, obviously, a linear space, since the sum of any two continuous functions is continuous, and a continuous function multiplied by any number is also continuous. The zero element in this space is represented by the only function which is identically equal to zero on  $[a, b]$ .

If we introduce the norm

$$\|f\| = \max_{[a,b]} |f(x)| \quad (5.3)$$

into the class  $C[a, b]$ , then it becomes a normed space.

**Definition 22.** A set  $M$  is called a *metric space* if any pair of its elements is associated with a nonnegative number  $\rho(f, g)$ , called the *distance* between the elements  $f$  and  $g$ , which satisfies the following *axioms of distance*:

1.  $\rho(f, g) = 0$  if and only if  $f = g$ ;
2.  $\rho(f, g) = \rho(g, f)$ , for every  $f, g \in M$ ;
3.  $\rho(f, r) \leq \rho(f, g) + \rho(g, r)$ , for every  $f, g, r \in M$ .

Axiom (3) is called the *triangle inequality*. The function  $\rho(f, g)$  will also be called the *metric* of the space  $M$ .

A metric space may be exemplified by any set  $M$  of the points  $x, y, z, \dots$  of the  $n$ -dimensional space  $\mathbb{R}^n$  with the distance

$$\rho(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

where  $x = (x_1, x_2, \dots, x_n)$  and  $y = (y_1, y_2, \dots, y_n)$ .

For the norm introduced via a scalar product by the method (5.5), all the three axioms of norm are fulfilled. Axiom (1) follows directly from Axiom (2) of definition of scalar product. Let us verify Axiom (2). Let  $\alpha$  be any real number. We have

$$\|\alpha f\|^2 = (\alpha f, \alpha f) = \alpha(f, \alpha f) = \alpha(\alpha f, f) = \alpha^2(f, f) = \alpha^2\|f\|^2.$$

Hence,

$$\|\alpha f\| = |\alpha|\|f\|,$$

since  $\|\alpha f\| \geq 0$ . Axiom (2) has been fulfilled. Axiom (3) can be established by means of Minkowski's inequality.

*Any linear normed space is at the same time a metric space with the distance (metric)*

$$\rho(f, g) = \|f - g\|.$$

In one and the same linear space, the norm can be introduced by a different methods. For instance, in the class of continuous functions  $C[0, 1]$  the norm can also be specified, besides (5.3), in the following form:

$$\|f\| = \left( \int_0^1 f^2(x) dx \right)^{\frac{1}{2}}. \quad (5.4)$$

In order to distinguish between various norms, we shall use appropriate subscripts. For example, for the norm (5.3) the following symbols are used:

$$\|f\|_{C[a,b]}, \quad \|f\|_C,$$

and for the norm (5.4) the symbol

$$\|f\|_{L_2}$$

is frequently utilized.

Note that *every Euclidean space is simultaneously a linear normed space with the norm*

$$\|f\| = \sqrt{(f, f)} \quad (5.5)$$

and, consequently, *a metric space with the distance*

$$\rho(f, g) = \|f - g\| = \sqrt{(f - g, f - g)}. \quad (5.6)$$

Let in the Euclidean space  $E$  there be given a system of functions

$$\varphi_0, \varphi_1, \dots, \varphi_m.$$

**Definition 23.** The determinant

$$\begin{vmatrix} (\varphi_0, \varphi_0) & (\varphi_1, \varphi_0) & \dots & (\varphi_m, \varphi_0) \\ (\varphi_0, \varphi_1) & (\varphi_1, \varphi_1) & \dots & (\varphi_m, \varphi_1) \\ \dots & \dots & \dots & \dots \\ (\varphi_0, \varphi_m) & (\varphi_1, \varphi_m) & \dots & (\varphi_m, \varphi_m) \end{vmatrix}$$

made up of scalar products is called **Gram determinant** (or *gramian*) of the system of functions  $\varphi_0, \varphi_1, \dots, \varphi_m$ .

**Lemma 1.** *The Gram determinant is equal to zero if and only if the system of functions  $\varphi_0, \varphi_1, \dots, \varphi_m$  is linearly dependent.*

**Definition 24.** A system of functions  $\varphi_0, \varphi_1, \dots, \varphi_m$  is said to be *orthogonal* if

$$(\varphi_j, \varphi_k) = 0, \quad j \neq k, \quad (\varphi_j, \varphi_j) > 0, \quad (5.7)$$

where  $0 \leq j, k \leq m$ .

If the system of functions  $\varphi_0, \varphi_1, \dots, \varphi_m$  is orthogonal, then it is linearly independent.

## 5.2 A Polynomial of Best Mean-square Approximation

**Definition 25.** The function

$$\Phi_m(x) = c_0\varphi_0(x) + c_1\varphi_1(x) + \dots + c_m\varphi_m(x), \quad (5.8)$$

where  $c_0, c_1, \dots, c_m$  are numerical coefficients, is called the *generalized polynomial* with respect to the system of functions  $\varphi_0, \varphi_1, \dots, \varphi_m$ .

The following problem is posed: find a polynomial of the form (5.8) such that the distance  $\rho(f, \Phi_m)$  is minimal where  $f$  is an arbitrary given function. This distance is called the *mean-square deviation of the polynomial  $\Phi_m$  from the function  $f$* . The polynomial  $\Phi_m$  possessing the indicated property is called the *polynomial of best mean-square approximation* of the function  $f$ .

Let us show that if the system of functions  $\varphi_0, \varphi_1, \dots, \varphi_m$  is linearly independent, then for any function  $f \in E$  the polynomial of best mean-square exists and is *unique*. In accordance with (5.6),

$$\begin{aligned} \rho^2(f, \Phi_m) &= \|f - \Phi_m\|^2 = (f - \Phi_m, f - \Phi_m) = \\ &= (f - c_0\varphi_0(x) - \dots - c_m\varphi_m(x), f - c_0\varphi_0(x) - \dots - c_m\varphi_m(x)) = \\ &= (f, f) + \sum_{j,k=1}^m c_j \cdot c_k \cdot (\varphi_j, \varphi_k) - 2 \sum_{j=0}^m c_j \cdot (f, \varphi_j). \end{aligned} \quad (5.9)$$

Thus, the quantity  $\rho^2(f, \Phi_m)$  is a quadratic form relative to the desired coefficients  $c_j$  of the polynomial (5.8). Since for any  $c_j$ ,  $j = 0, 1, \dots, m$ ,  $\rho^2(f, \Phi_m) \geq 0$ , it is known from the theory of quadratic forms that the

quadratic form (5.9) reaches its nonnegative minimum. Simultaneously with  $\rho^2(f, \Phi_m)$ , the distance

$$\rho(f, \Phi_m) = \sqrt{\rho^2(f, \Phi_m)}$$

also reaches its minimum.

Equating the partial derivatives of the form (5.9) with respect to  $c_i$ ,  $i = 0, 1, \dots, m$ , to zero, reducing the coefficient equal to 2, and transposing the constant terms to the right, we arrive at the following system of linear algebraic equations:

$$\begin{aligned} c_0(\varphi_0, \varphi_0) + c_1(\varphi_1, \varphi_0) + \dots + c_m(\varphi_m, \varphi_0) &= (f, \varphi_0), \\ c_0(\varphi_0, \varphi_1) + c_1(\varphi_1, \varphi_1) + \dots + c_m(\varphi_m, \varphi_1) &= (f, \varphi_1), \\ &\dots \\ c_0(\varphi_0, \varphi_m) + c_1(\varphi_1, \varphi_m) + \dots + c_m(\varphi_m, \varphi_m) &= (f, \varphi_m), \end{aligned} \tag{5.10}$$

called the *normal system*. By Lemma 1 its determinant, which is the Gram determinant, of the linearly independent system of functions  $\varphi_0, \varphi_1, \dots, \varphi_m$  is not equal to zero. Therefore for any function  $f \in E$  the system (5.10) has the unique solution  $c_0, c_1, \dots, c_m$  which corresponds to the unique stationary point of the quadratic form (5.9). This stationary point can be only the point of minimum, since the form reaches its minimum.

In the space  $E_C$  of continuous functions with the scalar product (5.1) the distance  $\rho(f, g)$ , called the *mean-square distance*, in accordance with (5.6), acquires the form

$$\rho(f, g) = \sqrt{\frac{1}{b-a} \int_a^b [f(x) - g(x)]^2 dx}, \tag{5.11}$$

and in the space  $E_{n+1}$  of functions defined on the discrete set  $D = \{x_i\}_{i=0}^n$  with the scalar product (5.2) the mean-square distance is given by the formula

$$\rho(f, g) = \sqrt{\frac{1}{n+1} \sum_{i=0}^n [f(x_i) - g(x_i)]^2}. \tag{5.12}$$

It should be borne in mind that the proximity of two continuous functions by the distance (5.11), that is, in the sense of mean-square deviation, does not guarantee the smallness of their maximum deviation from each other.

For instance, let  $g(t) \equiv 0$  for  $x \in [a, b]$  and let the function  $f(x)$  be given, differing from the former by a narrow tooth of altitude  $n$  and thickness at the base equal to  $1/n^3$ . In this case

$$\begin{aligned} \rho(f, g) &= \sqrt{\frac{1}{b-a} \int_a^b [f(x) - g(x)]^2 dx} = \\ &= \sqrt{\frac{1}{b-a} \int_a^b f^2(x) dx} \leq \sqrt{\frac{1}{b-a} \cdot n^2 \cdot \frac{1}{n^3}} = \frac{1}{\sqrt{(b-a)n}}, \end{aligned}$$

that is, by choosing  $n$ , we can make the mean-square distance  $\rho(f, g)$  arbitrarily small and the quantity

$$\max_{[a,b]} |f(x) - g(x)|$$

arbitrary large.

### 5.3 Mean-square Approximations by Algebraic Polynomials

We often use mean-square approximations of functions by algebraic polynomials, that is, instead of the system of functions  $\varphi_0, \varphi_1, \dots, \varphi_m$  we take the powers of  $x$ :  $1, x, x^2, \dots, x^m$ . The system of these functions is linearly independent in  $E_C$  (that is, on the given interval  $[a, b]$ ) for any  $m$ . In  $E_{n+1}$  it is linearly independent if  $m \leq n$ . For  $m \geq n + 1$  the system  $1, x, x^2, \dots, x^m$  is linearly dependent in  $E_{n+1}$ .

**Example 54.** On the interval  $[0, 1]$  construct the polynomial of the best mean-square approximation  $\Phi_1(x) = c_0 + c_1x$  for the function  $f(x) = \sqrt{x}$ .

**Solution.** We have  $\varphi_0(x) = 1, \varphi_1(x) = x$ . Therefore

$$\begin{aligned} (\varphi_0, \varphi_0) &= \int_0^1 1^2 dx = 1, & (\varphi_1, \varphi_1) &= \int_0^1 x^2 dx = \frac{1}{3}, \\ (\varphi_0, \varphi_1) &= (\varphi_1, \varphi_0) = \int_0^1 x dx = \frac{1}{2}, \\ (f, \varphi_0) &= \int_0^1 \sqrt{x} dx = \frac{2}{3}, & (f, \varphi_1) &= \int_0^1 \sqrt{x} \cdot x dx = \frac{2}{5}. \end{aligned}$$

Consequently, the normal system of equations (5.10) is the following:

$$\begin{aligned} c_0 + \frac{1}{2} \cdot c_1 &= \frac{2}{3}, \\ \frac{1}{2} \cdot c_0 + \frac{1}{3} \cdot c_1 &= \frac{2}{5}. \end{aligned}$$

Hence,

$$c_0 = \frac{4}{15}, \quad c_1 = \frac{4}{5},$$

and

$$\Phi_1(x) = \frac{4}{15} + \frac{4}{5} \cdot x.$$

In this case, the mean-square deviation  $\Phi_1$  from  $f$  has the value

$$\rho(f, \Phi_1) = \sqrt{\int_0^1 \left( \sqrt{x} - \frac{4}{15} - \frac{4}{5} \cdot x \right)^2 dx} = \frac{\sqrt{2}}{30}.$$

When finding the algebraic polynomial of best mean-square approximation in  $E_C$  for the function  $f$  of a continuous argument on the interval  $[a, b]$  we may encounter some difficulties in connection with the evaluation of the right-hand members of the equations (5.10), that is, the integrals

$$(f, \varphi_i) = \frac{1}{b-a} \int_a^b f(x) \cdot x^i dx, \quad i = 0, 1, \dots, m,$$

although the coefficients of the system, that is, the scalar products

$$(\varphi_j, \varphi_k) = \frac{1}{b-a} \int_a^b x^{j+k} dx,$$

are evaluated readily.

Therefore the method of least squares is used in the discrete variant, that is, in  $E_{n+1}$ . A set of points  $\{x_i\}_{i=0}^n$  from the interval  $[a, b]$  is given by experimental evaluations of the function  $f$ . It is desirable that the number of points exceeds the degree  $m$  of the polynomial at least by one and a half or two times. The points on the interval  $[a, b]$  are arranged as uniformly as possible or are somewhat concentrated on the portion of the interval where it is important to obtain a more exact approximation of the function. In the discrete variant, the computation of the coefficients and right-hand sides of the normal system of equations (5.10) in terms of the scalar product

(5.2) is not difficult. The found algebraic polynomial of best mean-square approximation of the function  $f$  on the discrete set  $\{x_i\}_{i=0}^n \subset [a, b]$  in sense of the distance (5.12) is usually taken for some approximating polynomial of the function  $f$  throughout the interval  $[a, b]$ .

If  $m = n$ , then the algebraic polynomial of degree  $n$  which is found in the discrete variant by the least-squares method coincides with the interpolation polynomial, since the deviation of the interpolation polynomial from the given function  $f$  on the set of points  $\{x_i\}_{i=0}^n$  in sense of the distance (5.12) is equal to zero.

The mean-square approximation of a function by an algebraic polynomial is used when the function to be approximated is not sufficiently smooth and one fails to construct a suitable interpolation polynomial for it, spline or a polynomial of uniform approximation, and also if the values of the function are known at a sufficiently large number of points, but with random errors.

## 5.4 Application of Orthogonal Polynomials

The solution of the normal system of equations (5.10) is found in a most simple way if the system of functions  $\varphi_0, \varphi_1, \dots, \varphi_m$  is orthogonal, that is, satisfies the condition (5.7). In this case, the matrix of the system becomes diagonal and coefficients

$$c_j = \frac{(f, \varphi_j)}{(\varphi_j, \varphi_j)}, j = 0, 1, \dots, m. \quad (5.13)$$

They are called the *Fourier coefficients* of the function  $f$  with respect to the orthogonal system  $\varphi_0, \varphi_1, \dots, \varphi_m$ .

Taking into account (5.13) we find

$$\rho^2(f, \Phi_m) = \|f\|^2 - \sum_{j=0}^m c_j^2 \cdot \|\varphi_j\|^2,$$

where  $\Phi_m$  is the polynomial of best mean-square approximation of the function  $f$  constructed with respect to the orthogonal system  $\varphi_0, \varphi_1, \dots, \varphi_m$ , and  $c_j$  are its coefficients. Hence it is clearly seen that with an increase in  $m$ , that is, with new functions added to the orthogonal system  $\varphi_0, \varphi_1, \dots, \varphi_m$  (without changing the old ones), the quantity  $\rho^2(f, \Phi_m)$ , generally speaking, decreases (does not increase).

## 5.5 Method of least squares - continuation (a practical approach)

Consider now an elementary approach to the method of least squares. The following problem is often encountered in practical applications. Suppose two functionally related quantities  $x$  and  $y$  are associated with  $n$  pairs of known values  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ . It is required to determine, in the preassigned formula

$$y = f(x, \alpha_1, \alpha_2, \dots, \alpha_m)$$

$m$  parameters  $\alpha_1, \alpha_2, \dots, \alpha_m$ ,  $m < n$  so that the known  $n$  pairs of the values of  $x$  and  $y$  would suit the formula in the best way.

We can consider as the best values  $\alpha_1, \alpha_2, \dots, \alpha_m$  which turn into minimum the sum

$$\sum_{k=1}^n [f(x_k, \alpha_1, \alpha_2, \dots, \alpha_m) - y_k]^2,$$

that is, the sum of the squares of deviations of the values of  $y$ , calculated by the formula, from preassigned values. This explains the name of the **method of least squares**.

This condition yields a system of  $m$  equations which are used to determine  $\alpha_1, \alpha_2, \dots, \alpha_m$ :

$$\sum_{k=1}^n [f(x_k, \alpha_1, \alpha_2, \dots, \alpha_m) - y_k] \cdot \frac{\partial f(x_k, \alpha_1, \alpha_2, \dots, \alpha_m)}{\partial \alpha_j} = 0, \quad (5.14)$$

where  $j = 1, 2, \dots, m$ .

## 5.6 Special cases - a)

Let us put:

$$y = f(x, \alpha_1, \alpha_2, \dots, \alpha_m) = a_0 x^m + a_1 x^{m-1} + \dots + a_m.$$

It is necessary to define  $m + 1$  parameters  $a_0, a_1, \dots, a_m$ ;  $n > m + 1$ . System (5.14) assumes the form

$$\begin{aligned}
 a_0 \sum_{k=1}^n x_k^m + a_1 \sum_{k=1}^n x_k^{m-1} + \dots + na_m &= \sum_{k=1}^n y_k, \\
 a_0 \sum_{k=1}^n x_k^{m+1} + a_1 \sum_{k=1}^n x_k^m + \dots + a_m \sum_{k=1}^n x_k &= \sum_{k=1}^n x_k y_k, \\
 a_0 \sum_{k=1}^n x_k^{m+2} + a_1 \sum_{k=1}^n x_k^{m+1} + \dots + a_m \sum_{k=1}^n x_k^2 &= \sum_{k=1}^n x_k^2 y_k, \\
 &\dots \\
 a_0 \sum_{k=1}^n x_k^{2m} + a_1 \sum_{k=1}^n x_k^{2m-1} + \dots + a_m \sum_{k=1}^n x_k^m &= \sum_{k=1}^n x_k^m y_k.
 \end{aligned} \tag{5.15}$$

This system of  $m + 1$  equations with  $m + 1$  unknowns always has a unique solution since its determinant is nonzero.

## 5.7 Special cases - b)

Let us put:

$$y = Ae^{cx}.$$

To simplify system (5.14), we first take logarithms of this formula and replace it by

$$\log y = \log A + c \cdot x \cdot \log e.$$

In this case, system (5.14) assumes the form

$$\begin{aligned}
 c \cdot \log e \cdot \sum_{k=1}^n x_k + n \cdot \log A &= \sum_{k=1}^n \log y_k, \\
 c \cdot \log e \cdot \sum_{k=1}^n x_k^2 + \log A \cdot \sum_{k=1}^n x_k &= \sum_{k=1}^n x_k \log y_k.
 \end{aligned}$$

Then we determine  $c$  and  $\log A$ .

## 5.8 Special cases - c)

Let us put:

$$y = Ax^q.$$

To simplify system (5.14), we again take logarithms of this formula and replace it by

$$\log y = \log A + q \cdot \log x.$$

Now system (5.14) assumes the form

$$\begin{aligned} q \cdot \sum_{k=1}^n \log x_k + n \cdot \log A &= \sum_{k=1}^n \log y_k, \\ q \cdot \sum_{k=1}^n \log^2 x_k + \log A \cdot \sum_{k=1}^n \log x_k &= \sum_{k=1}^n \log x_k \cdot \log y_k. \end{aligned}$$

Then we determine  $q$  and  $\log A$ .

## 5.9 Special cases - d)

It is often necessary to choose the best way to replace some given function  $y = f(x)$  on the interval  $[a, b]$  by an  $m$ th-degree polynomial

$$y \approx \varphi(x) = a_0 x^m + a_1 x^{m-1} + \dots + a_m.$$

In that case, the application of the method of least squares helps in finding the coefficients  $a_0, a_1, \dots, a_m$  from the condition of the minimum of the integral

$$\int_a^b [\varphi(x) - f(x)]^2 dx = \int_a^b [a_0 x^m + a_1 x^{m-1} + \dots + a_m - f(x)]^2 dx.$$

The necessary condition for the minimum of that integral lead to a system of  $m + 1$  equations with  $m + 1$  unknowns  $a_0, a_1, \dots, a_m$ , which is used to determine all these coefficients:

$$\begin{aligned} \int_a^b [a_0 x^m + a_1 x^{m-1} + \dots + a_m - f(x)] \cdot x^m dx &= 0, \\ \int_a^b [a_0 x^m + a_1 x^{m-1} + \dots + a_m - f(x)] \cdot x^{m-1} dx &= 0, \\ &\dots \\ \int_a^b [a_0 x^m + a_1 x^{m-1} + \dots + a_m - f(x)] dx &= 0. \end{aligned}$$

# Chapter 6

## Numerical Integration

In practice, we seldom succeed in finding the exact value of a definite integral or in integrating an ordinary differential equation. For instance, the integral

$$\int_1^2 \frac{dx}{\ln x}$$

cannot be expressed in elementary functions, and the equation

$$u' = \exp(-x^2 - u^2)$$

cannot be integrated. In the next we describe some numerical methods for solving definite integrals and differential equations.

### 6.1 Quadrature Formulae

Let us first formulate a theorem from integral calculus.

**Theorem 23.** *Let  $f, g \in C[a, b]$ , and  $g(x) \geq 0$  on  $[a, b]$ . Then there is a point  $\xi \in [a, b]$  such that*

$$\int_a^b f(x)g(x) dx = f(\xi) \int_a^b g(x) dx.$$

Let us introduce the notion of a quadratic formula. Let there be given the definite integral

$$I = \int_a^b f(x) dx$$

of a function  $f$  which is continuous on the interval  $[a, b]$ . The approximate equality

$$\int_a^b f(x) dx \approx \sum_{j=1}^n q_j \cdot f(x_j),$$

where  $q_j$  are certain numbers, and  $x_j$  are certain points of the interval  $[a, b]$ , is called a *quadrature formula* defined by *weights*  $q_j$  and *nodes*  $x_j$ .

## 6.2 The Rectangular Formula

Suppose that  $f \in C^2[-h/2, h/2]$ ,  $h > 0$ . We set approximately

$$\int_{-h/2}^{h/2} f(x) dx \approx h \cdot f_0, \quad (6.1)$$

where  $f_0 = f(0)$ , that is, the area of the curvilinear trapezoid bounded from above by the graph of the function  $f$  is approximated by the area of the hatched rectangle whose altitude is equal to the value of  $f$  at the midpoint of the trapezoid's base. We then find the remainder, that is, the error of formula (6.1). Let

$$F(x) = \int_0^x f(t) dt, \quad F_{\pm 1/2} = F(\pm h/2). \quad (6.2)$$

Since  $F(0) = 0$ ,  $F'(0) = f_0$ ,  $F''(0) = f'_0 = f'(0)$  and  $F'''(x) = f'''(x)$ , according to Taylor's formula with its remainder in Lagrange's form, we have

$$F_{\pm 1/2} = \pm \frac{h}{2} \cdot f_0 + \frac{h^2}{8} \cdot f'_0 \pm \frac{h^3}{48} \cdot f''(\xi_{\pm})$$

where  $\xi_+$  and  $\xi_-$  are some points such that

$$-\frac{h}{2} < \xi_- < \xi_+ < \frac{h}{2}.$$

Then

$$\int_{-h/2}^{h/2} f(x) dx = F_{1/2} - F_{-1/2} = h \cdot f_0 + \frac{h^3}{24} \cdot \frac{f'''(\xi_-) + f'''(\xi_+)}{2}.$$

Since there is a  $\xi$  such that  $f''(\xi) = \frac{1}{2}(f''(\xi_-) + f''(\xi_+))$  we obtain the *rectangular formula with its remainder*:

$$\int_{-h/2}^{h/2} f(x) dx = h \cdot f_0 + \frac{h^3}{24} \cdot f'''(\xi), \quad |\xi| \leq \frac{h}{2}.$$

### 6.3 The Trapezoidal Formula

Let  $f \in C^2[0, h]$ . We set

$$\int_0^h f(x) dx \approx h \cdot \frac{f_0 + f_1}{2},$$

where  $f_0 = f(0)$  and  $f_1 = f(h)$ , i.e. the integral is approximately replaced by the area of the hatched trapezoid. Let us express  $f_1$  and  $F_1 = F(h)$ , where  $F$  is given by (6.2), by Taylor's formula with its remainder in integral form:

$$f_1 = f_0 + hf'_0 + \int_0^h (h-t)f''(t) dt, \quad (6.3)$$

$$\begin{aligned} F_1 = f(0) + hF'(0) + \frac{h^2}{2} \cdot F''(0) + \frac{1}{2} \cdot \int_0^h (h-t)^2 F''(t) dt = \\ h \cdot f_0 + \frac{h^2}{2} \cdot f'_0 + \frac{1}{2} \cdot \int_0^h (h-t)^2 f''(t) dt. \end{aligned} \quad (6.4)$$

According to (6.3), we have

$$h \cdot \frac{f_0}{2} = h \cdot \frac{f_1}{2} - \frac{h^2}{2} f'_0 - \frac{h}{2} \cdot \int_0^h (h-t)f''(t) dt.$$

Isolating the term  $hf_0/2$  on the right-side of (6.4) we obtain

$$\int_0^h f(x) dx = h \cdot \frac{f_0 + f_1}{2} - \frac{1}{2} \cdot \int_0^h (h-t)t f''(t) dt.$$

Since  $(h-t)t \geq 0$ ,  $t \in [0, h]$  we (by previous theorem) have

$$\int_0^h (h-t)t f''(t) dt = f''(\xi) \cdot \int_0^h (h-t)t dt = \frac{h^3}{6} \cdot f''(\xi),$$

where  $\xi \in [0, h]$  is a point. We have arrived at the *trapezoidal formula with its remainder*:

$$\int_0^h f(x) dx = h \cdot \frac{f_0 + f_1}{2} - \frac{h^3}{12} \cdot f''(\xi), \quad \xi \in [0, h].$$

## 6.4 Simpson Formula

Suppose that  $f \in C^4[-h, h]$ . We replace approximately the integral

$$\int_{-h}^h f(x) dx$$

by the area of the hatched curvilinear trapezoid bounded from above by a parabola passing through the points  $(-h, f_{-1})$ ,  $(0, f_0)$ ,  $(h, f_1)$ , where  $f_i = f(ih)$ . This parabola is given by the equation

$$y = f_0 + \frac{f_1 - f_{-1}}{2h} \cdot x + \frac{f_{-1} - 2f_0 + f_1}{2h^2} \cdot x^2,$$

which is readily verified by setting, in turn,  $x$  equal to  $-h$ ,  $0$  and  $h$ . Hence we easily find

$$\int_{-h}^h y(x) dx = \frac{h}{3} \cdot (f_{-1} + 4f_0 + f_1).$$

Thus *Simpson's formula*, which is also called the *parabolic formula*, has the form

$$\int_{-h}^h f(x) dx = \frac{h}{3} \cdot (f_{-1} + 4f_0 + f_1).$$

Let us set  $F_{\pm} = F(\pm h)$ , where  $F$  is the function (6.2). Since  $F(0) = 0$  and  $F^{(k)}(x) = f^{(k-1)}(x)$ ,  $1 \leq k \leq 5$ , by Taylor's formula with its remainder in integral form we have

$$F_{\pm 1} = \pm h f_0 + \frac{h^2}{2} \cdot f'_0 \pm \frac{h^3}{6} \cdot f''_0 + \frac{h^4}{24} \cdot f'''_0 \pm \frac{1}{24} \cdot \int_0^h (h-t)^4 f^{(4)}(\pm t) dt,$$

$$f_{\pm 1} = f_0 \pm h \cdot f'_0 + \frac{h^2}{2} \cdot f''_0 \pm \frac{h^3}{6} \cdot f'''_0 + \frac{1}{6} \cdot \int_0^h (h-t)^3 f^{(4)}(\pm t) dt.$$

Hence we obtain

$$F_1 - F_{-1} - \frac{h}{3} \cdot (f_{-1} + 4f_0 + f_1) = \frac{-1}{24} \cdot \int_0^h (h-t)^3 \left( \frac{h}{3} + t \right) (f^{(4)}(t) + f^{(4)}(-t)) dt.$$

Since  $(h-t)^3(h/3+t) \geq 0$  for  $t \in [0, h]$ , we find

$$\begin{aligned} & -\frac{1}{24} \int_0^h (h-t)^3 \left( \frac{h}{3} + t \right) (f^{(4)}(t) + f^{(4)}(-t)) dt = \\ & -\frac{1}{12} \frac{f^{(4)}(\eta) + f^{(4)}(-\eta)}{2} \cdot \int_0^h (h-t)^3 \left( \frac{h}{3} + t \right) dt = \frac{-h^5}{90} \cdot f^{(4)}(\xi), \end{aligned}$$

where  $\eta \in [0, h]$  and  $\xi \in [-h, h]$  are some points. Taking into account that

$$F_1 - F_{-1} = \int_{-h}^h f(x) dx$$

we arrive at *Simpson's formula with its remainder*:

$$\int_{-h}^h f(x) dx = \frac{h}{3} \cdot (f_{-1} + 4f_0 + f_1) - \frac{h^5}{90} \cdot f^{(4)}(\xi).$$

The quadrature formulas considered above are called *canonical*.

## 6.5 Composite Quadrature Formulas

In practice, if it is required to evaluate approximately the integral, the given interval  $[a, b]$  is divided into  $N$  equal subintervals, one of the canonical quadrature formulas is applied on each of the subintervals, and the results obtained are summed. The quadrature formula thus constructed on the interval  $[a, b]$  is said to be *composite*. When applying the rectangular and trapezoidal formulas, it is convenient to take subintervals of length  $h$ , and when using Simpson's formula, of length  $2h$ .

Let us dwell on the use of the rectangular formula in more detail. Let  $f \in C^2$ . We denote the subintervals by  $[x_i, x_{i+1}]$ , where  $x_i = a + ih$ ,  $i = 0, 1, \dots, N-1$ ,  $x_N = b$ ,  $h = (b-a)/N$ . In accordance with rectangular formula

$$\int_{x_i}^{x_{i+1}} f(x) dx \approx hf_{i+1/2}, \tag{6.5}$$

where  $f_{i+1/2} = f(a + (i + 1/2)h)$  is the value of  $f$  at the midpoint of the subinterval  $[x_i, x_{i+1}]$ . Moreover,

$$\int_{x_i}^{x_{i+1}} f(x) dx = hf_{i+1/2} + \frac{h^3}{24} \cdot f''(\xi_i),$$

where  $\xi_i \in [x_i, x_{i+1}]$  is some point. Summing over all the approximation (6.5) leads to the *composite rectangular formula*:

$$\int_a^b f(x) dx \approx h (f_{1/2} + f_{3/2} + \dots + f_{N-1/2}).$$

Since it can be easily proved:

$$\sum_{i=0}^{N-1} f''(\xi_i) = N \cdot \left( \frac{1}{N} \cdot \sum_{i=0}^{N-1} f''(\xi_i) \right) = Nf''(\xi) = \frac{b-a}{h} \cdot f''(\xi),$$

where  $\xi \in [a, b]$ , we have obtained the *composite rectangular formula with its remainder*:

$$\int_a^b f(x) dx = h (f_{1/2} + f_{3/2} + \cdots + f_{N-1/2}) + h^2 \cdot \frac{b-a}{24} \cdot f''(\xi).$$

Just in the same way, under the condition that  $f \in C^2[a, b]$ , we obtain the *composite trapezoidal formula*:

$$\int_a^b f(x) dx \approx h \left( \frac{f_0}{2} + f_1 + \cdots + f_{N-1} + \frac{f_N}{2} \right)$$

and the corresponding *composite trapezoidal formula with its remainder*:

$$\int_a^b f(x) dx = h \left( \frac{f_0}{2} + f_1 + \cdots + f_{N-1} + \frac{f_N}{2} \right) - h^2 \cdot \frac{b-a}{12} \cdot f''(\xi),$$

where  $f_i = f(a + ih)$ ,  $h = (b-a)/N$ , and  $\xi \in [a, b]$  is some point.

Let now  $h = (b-a)/2N$  and  $x_j = a + jh$ ,  $f_j = f(x_j)$ . We rewrite Simpson's canonical formula in connection with subinterval  $[x_{2i}, x_{2i+2}]$  of length  $2h$ :

$$\int_{x_{2i}}^{x_{2i+2}} f(x) dx \approx \frac{h}{3} (f_{2i} + 4f_{2i+1} + f_{2i+2}).$$

Summing both sides of this relationship over  $i$  from 0 to  $N-1$ , we obtain *Simpson's composite formula*:

$$\int_a^b f(x) dx \approx \frac{h}{3} (f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 4f_{2N-1} + f_{2N}).$$

The corresponding *Simpson's composite formula with its remainder*, which is obtained by summing equalities over the subintervals  $[x_{2i}, x_{2i+2}]$  provided that  $f \in C^4$ , is:

$$\int_a^b f(x) dx = \frac{h}{3} \left( f_0 + f_{2N} + 4 \sum_{i=1}^N f_{2i-1} + 2 \sum_{i=1}^{N-1} f_{2i} \right) - h^4 \cdot \frac{b-a}{180} \cdot f^{(4)}(\xi),$$

where  $f_i = f(a + ih)$ ,  $h = (b-a)/(2N)$ , and  $\xi \in [a, b]$ .

For the sake of brevity, let us introduce the notation

$$I_h^{\text{rect}} = h \cdot \sum_{i=0}^{N-1} f_{i+1/2},$$

$$I_h^{\text{trap}} = h \cdot \left( \frac{f_0 + f_N}{2} + \sum_{i=1}^{N-1} f_i \right),$$

where  $h = (b - a)/N$  and  $f_\mu = f(a + \mu h)$ , and

$$I_h^{\text{Simp}} = \frac{h}{3} \cdot \left( f_0 + f_{2N} + 4 \sum_{i=1}^N f_{2i-1} + 2 \sum_{i=1}^{N-1} f_{2i} \right),$$

where  $h = (b - a)/(2N)$  and  $f_i = f(a + ih)$ . From the expressions for the remainders it is seen that the rectangular and trapezoidal formulas are exact for polynomials of first degree, whereas Simpson's formula is exact for polynomials of third degree. The first two formulas have the second order of accuracy with respect to  $h$ , while Simpson's formula is of the fourth order of accuracy if  $f$  is smooth. Therefore for the functions of class  $C^4$  for a small  $h$  Simpson's formula usually yields a higher accuracy as compared with previous method.

The error of the rectangular formula and Simpson's formula satisfies the inequalities

$$|I - I_h^{\text{rect}}| \leq h^2 \cdot \frac{b - a}{24} \cdot \max_{[a,b]} |f''(x)|,$$

$$|I - I_h^{\text{Simp}}| \leq h^4 \cdot \frac{b - a}{180} \cdot \max_{[a,b]} |f^{(4)}(x)|.$$

There is a similar inequality for the error of the trapezoidal formula. Estimates from below are also useful. In particular, for the error of the rectangular formula the lower estimate is:

$$|I - I_h^{\text{rect}}| \geq h^2 \cdot \frac{b - a}{24} \cdot \min_{[a,b]} |f''(x)|.$$

**Example 55.** As an example, let us analyse the errors of the quadrature formulas for the integral

$$I = \int_0^{1/2} e^{-x^2} dx$$

which is not expressible in terms of elementary functions and is frequently used in applications.

We have

$$f(x) = e^{-x^2}, \quad f'(x) = -2xe^{-x^2}, \quad f''(x) = (4x^2 - 2)e^{-x^2},$$

$$f'''(x) = (-8x^3 + 12x)e^{-x^2}, \quad f^{(4)}(x) = 4(4x^4 - 12x^2 + 3)e^{-x^2},$$

and

$$e^{-1/4} \leq |f''(\mathbf{x})| \leq 2, \quad |f^{(4)}(\mathbf{x})| \leq 12$$

if  $x \in [0, 1/2]$ .

Hence for  $h = 0.05$  we get

$$0.4 \cdot 10^{-4} \leq |I - I_h^{\text{rect}}| \leq 0.11 \cdot 10^{-3}$$

and

$$|I - I_h^{\text{Simp}}| \leq 0.21 \cdot 10^{-6}.$$

The upper estimate of the error of Simpson's formula is considerably less than the lower estimate of the error of the rectangular formula.

## 6.6 Newton-Cotes Quadrature Formulas

Suppose, for a given function  $y = f(x)$ , it is required to compute the integral

$$\int_a^b f(x) dx.$$

Choosing a spacing  $h = (b - a)/n$  divide the integral  $[a, b]$  by means of equally spaced points  $x_0 = a$ ,  $x_i = x_0 + ih$ ,  $i = 0, 1, \dots, n - 1$ ,  $x_n = b$  into  $n$  equal parts, and let  $y_i = f(x_i)$ ,  $i = 0, 1, \dots, n$ .

Replacing  $f(x)$  by Lagrange interpolation polynomial  $L_n(x)$ , we obtain the approximate quadrature formula

$$\int_{x_0}^{x_n} f(x) dx \approx \sum_{i=0}^n A_i f(x_i) \quad (6.6)$$

where  $A_i$  are certain constant coefficients. We derive explicit expressions for the coefficients  $A_i$  of formula (6.6). Let us recall:

$$L_n(x) = \sum_{i=0}^n p_i(x) y_i$$

where

$$p_i(x) = \frac{(x - x_0) \dots (x - x_{i-1})(x - x_{i+1}) \dots (x - x_n)}{(x_i - x_0) \dots (x_i - x_{i-1})(x_i - x_{i+1}) \dots (x_i - x_n)}.$$

Introducing the notation  $q = (x - x_0)/h$  and

$$q^{[n+1]} = q(q - 1) \dots (q - n)$$

we have

$$L_n(x) = \sum_{i=0}^n \frac{(-1)^{n-i}}{i!(n-i)!} \cdot \frac{q^{[n+1]}}{q-i} \cdot y_i.$$

Replacing in (6.6)  $y$  by  $L_n(x)$  we have

$$A_i = \int_{x_0}^n \frac{(-1)^{n-i}}{i!(n-i)!} \cdot \frac{q^{[n+1]}}{q-i} dx$$

or, since  $q = (x - x_0)/h$ ,  $dq = dx/h$ ,

$$A_i = \int_0^n \frac{(-1)^{n-i}}{i!(n-i)!} \cdot h \cdot \frac{q^{[n+1]}}{q-i} dq = h \cdot \frac{(-1)^{n-i}}{i!(n-i)!} \cdot \int_0^n \frac{q^{[n+1]}}{q-i} dq, \quad i = 0, 1, \dots, n.$$

Since  $h = (b - a)/n$  we ordinarily put  $A_i = (b - a)H_i$  where

$$H_i = \frac{1}{n} \cdot \frac{(-1)^{n-i}}{i!(n-i)!} \cdot \int_0^n \frac{q^{[n+1]}}{q-i} dq, \quad i = 0, 1, \dots, n \quad (6.7)$$

are constants called *Cotes coefficients*.

Then the quadrature formula (6.6) assumes the form

$$\int_a^b f(x) dx \approx (b - a) \sum_{i=0}^n H_i y_i$$

where  $h = (b - a)/n$  and  $y_i = f_i(a + ih)$ ,  $i = 0, 1, \dots, n$ .

It is easy to see that the following relations are valid:

$$\sum_{i=0}^n H_i = 1, \quad H_i = H_{n-i}.$$

Applying (6.7) for  $n = 1$ , we have

$$H_0 = -1 \cdot \int_0^1 \frac{q(q-1)}{q} dq = \frac{1}{2}, \quad H_1 = \int_0^1 q dq = \frac{1}{2}$$

whence

$$\int_{x_0}^{x_1} f(x) dx \approx \frac{h}{2} \cdot (y_0 + y_1).$$

We thus obtain *trapezoidal formula* for approximate computation of a definite integral.

Applying (6.7) for  $n = 2$ , we get

$$H_0 = \frac{1}{2} \cdot \frac{1}{2} \cdot \int_0^2 (q-1)(q-2) dq = \frac{1}{4} \left( \frac{8}{3} - 6 + 4 \right) = \frac{1}{6},$$

$$H_1 = -\frac{1}{2} \cdot \frac{1}{1} \cdot \int_0^2 q(q-2) dq = \frac{2}{3},$$

$$H_2 = \frac{1}{2} \cdot \frac{1}{2} \cdot \int_0^2 q(q-1) dq = \frac{1}{6}.$$

Hence, since  $x_2 - x_0 = 2h$ , we have

$$\int_{x_0}^{x_2} \approx \frac{h}{3}(y_0 + 4y_1 + y_2)$$

which is the formula of *Simpson's rule*.

Carrying out the appropriate computations for  $n = 3$  we obtain from (6.7) *Newton's quadrature formula*:

$$\int_{x_0}^{x_3} \approx \frac{3h}{8}(y_0 + 3y_1 + 3y_2 + y_3)$$

which is sometimes called the *three-eighths rule*. The remainder of this formula is

$$R = -\frac{3h^5}{80} \cdot y^{(4)}(\xi)$$

where  $\xi \in (x_0, x_3)$ .

# Chapter 7

## Methods of Solving Nonlinear Equations and Systems

### 7.1 The Halving Method (Method of Dividing a Line Segment into Two Equal Parts)

Suppose we have an equation

$$f(x) = 0$$

where the function  $f(x)$  is continuous on  $[a, b]$  and

$$f(a) \cdot f(b) < 0.$$

In order to find a root lying in the interval  $[a, b]$ , divide the interval in half. If  $f((a + b)/2) = 0$ , then  $\xi = (a + b)/2$  is a root of the equation. If

$$f\left(\frac{a + b}{2}\right) \neq 0$$

than we choose that half,  $[a, (a + b)/2]$  or  $[(a + b)/2, b]$ , at the endpoints of which the function  $f(x)$  has opposite signs. The newly reduced interval  $[a_1, b_1]$  is again halved and the same investigation is made, etc. Finally, at some stage in the process we get either the exact root or an infinite sequence of nested intervals

$$[a_1, b_1], [a_2, b_2], \dots, [a_n, b_n], \dots$$

such that

$$f(a_n) \cdot f(b_n) < 0, \quad n = 1, 2, \dots \quad (7.1)$$

and

$$b_n - a_n = \frac{1}{2^n}(b - a).$$

Since the left endpoints

$$a_1, a_2, a_3, \dots, a_n, \dots$$

form a monotonic nondecreasing bounded sequence, and the right endpoints

$$b_1, b_2, b_3, \dots, b_n, \dots$$

form a monotonic nonincreasing bounded sequence, then there exists a common limit

$$\xi = \lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} b_n.$$

Passing to the limit in (7.1) as  $n \rightarrow \infty$  we get  $[f(\xi)]^2 \leq 0$ , whence  $f(\xi) = 0$ , which means that  $\xi$  is a root of equation. It is obvious that

$$0 \leq \xi - a_n \leq \frac{1}{2^n}(b - a).$$

## 7.2 The Method of Chords (Method of Proportional Parts)

Let us suppose that  $f(a) < 0$ ,  $f(b) > 0$ . Then instead of halving the interval  $[a, b]$  it is more natural to divide it in the ratio

$$f(a) : f(b) .$$

This yields an appropriate value of the root

$$x_1 = a + h_1$$

where

$$h_1 = \frac{-f(a)}{-f(a) + f(b)} \cdot (b - a) = \frac{-f(a)}{f(b) - f(a)} \cdot (b - a).$$

Then, applying this device to the interval  $[a, x_1]$  or  $[x_1, b]$  at the endpoints of which the function  $f(x)$  has opposite signs, we get a second approximation to the root  $x_2$ , etc. Geometrically, the method of proportional parts is equivalent to replacing the curve

$$y = f(x)$$

by a chord that passes through the points  $A[a, f(a)]$ ,  $B[b, f(b)]$ . Indeed, the equation of the chord  $AB$  is

$$\frac{x - a}{b - a} = \frac{y - f(a)}{f(b) - f(a)}.$$

Whence, assuming  $x = x_1$  and  $y = 0$ , we get

$$x_1 = a - \frac{f(a)}{f(b) - f(a)} \cdot (b - a).$$

Suppose, for definiteness, that  $f''(x) > 0$  for  $a \leq x \leq b$  (the case  $f''(x) < 0$  reduces to our case if we write the equation as:  $-f(x) = 0$ ). Then the curve  $y = f(x)$  will be *convex down* and, hence, will be located below its chord  $AB$ . Two cases are possible:  $f(a) > 0$  and  $f(a) < 0$ .

In the former case, the endpoint  $a$  is fixed and the successive approximations:

$$\begin{aligned} x_0 &= b, \\ x_{n+1} &= x_n - \frac{f(x_n)}{f(x_n) - f(a)} \cdot (x_n - a), \quad n = 0, 1, 2, \dots \end{aligned}$$

form a bounded decreasing sequence and

$$a < \xi < \dots < x_{n+1} < x_n < \dots < x_1 < x_0.$$

In the latter case, the endpoint  $b$  is fixed and the successive approximations:

$$\begin{aligned} x_0 &= a, \\ x_{n+1} &= x_n - \frac{f(x_n)}{f(b) - f(x_n)} \cdot (b - x_n), \quad n = 0, 1, 2, \dots \end{aligned}$$

form a bounded increasing sequence and

$$x_0 < x_1 < \dots < x_n < x_{n+1} < \dots < \xi < b.$$

It can be proved that

$$\lim_{n \rightarrow \infty} x_n = \xi, \quad \text{and} \quad f(\xi) = 0.$$

### 7.3 Newton's Method (Method of Tangens)

Let there is a root of the equation  $f(x) = 0$ . Newton's method is equivalent to replacing a small arc of the curve  $y = f(x)$  by a tangent line drawn to a

point of the curve. Suppose, for definiteness, that  $f''(x) > 0$  for  $a \leq x \leq b$  and  $f(b) > 0$ . Choose, say,  $x_0 = b$  for which  $f(x_0) \cdot f''(x_0) > 0$ . Drawn the tangent line to the curve  $y = f(x)$  at the point  $B_0(x_0, f(x_0))$ . For the first approximation  $x_1$  of the root  $\xi$  let us take the abscissa of the point of intersection of this tangent with  $x$ -axis. Through the point  $B_1(x_1, f(x_1))$  again draw a tangent line whose abscissa of the intersection point with the  $x$ -axis yields a second approximation  $x_2$  of the root  $\xi$ , and so on. It is plain that the equation of the tangent at the point  $B_n(x_n, f(x_n))$ ,  $n = 0, 1, 2, \dots$  is

$$y - f(x_n) = f'(x_n)(x - x_n).$$

Putting  $y = 0$ ,  $x = x_{n+1}$ , we get formula

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \quad (7.2)$$

Note that if in our case we put  $x_0 = a$  and hence  $f(x_0) \cdot f''(x_0) < 0$ , then drawing the tangent to the curve  $y = f(x)$  at the point  $A(a, f(a))$  we would get point  $x'_1$  lying outside the interval  $[a, b]$  and method does not work.

**Theorem 24.** *If  $f(a) \cdot f(b) < 0$ ,  $f'(x)$ ,  $f''(x)$  are nonzero and preserve signs over  $a \leq x \leq b$ , then, proceeding from the initial approximation  $x_0 \in [a, b]$  for which  $f(x_0) \cdot f''(x_0) > 0$ , it is possible, by using Newton's method (7.2), to compute the sole root  $\xi$  of equation  $f(x) = 0$  to any degree of accuracy.*

For the accuracy we have the formula

$$|\xi - x_n| \leq |x_n - x_{n-1}|.$$

## 7.4 The Method of Iteration

Suppose we have an equation

$$f(x) = 0 \quad (7.3)$$

where  $f(x)$  is a continuous function and it is required to determine its real roots. Replace (7.3) with an equivalent equation

$$x = \varphi(x). \quad (7.4)$$

In some way choose a roughly approximate value of the root,  $x_0$ , and substitute it into the right member of (7.4) to get a number

$$x_1 = \varphi(x_0). \quad (7.5)$$

Now inserting  $x_1$  in the right member of (7.5) in place of  $x_0$ , we get a new number

$$x_2 = \varphi(x_1).$$

Repeating this process, we get a sequence of numbers

$$x_n = \varphi(x_{n-1}), \quad n = 1, 2, \dots$$

If this sequence is convergent then the limit

$$\xi = \lim_{n \rightarrow \infty} x_n$$

is a root of (7.3).

Geometrically, the method of iteration can be explained as follows (see figure):

**Theorem 25.** *Let a function  $\varphi$  be defined and differentiable on an interval  $[a, b]$  with all values  $\varphi(x) \in [a, b]$ . Then if there exist a proper fraction  $q$  such that*

$$|\varphi'(x)| \leq q < 1$$

*for  $a < x < b$  then the process of iteration*

$$x_n = \varphi(x_{n-1}), \quad n = 1, 2, \dots$$

*converges irrespective of the initial value  $x_0 \in [a, b]$ ; the limiting value  $\xi = \lim_{n \rightarrow \infty} x_n$  is the only root of the equation*

$$x = \varphi(x)$$

*on the interval  $[a, b]$ .*

**Remark 2.** The process of iteration may be divergent (see figure):

## 7.5 The Method of Iteration for a System of Two Equations

Let there be given two equations in two unknowns:

$$\begin{aligned} F_1(x, y) &= 0, \\ F_2(x, y) &= 0 \end{aligned} \tag{7.6}$$

whose real roots it is required to find to within a specified degree of accuracy. We offer an iteration process which, under certain circumstance, permits improving the given approximate values of the roots. To do this, represent (7.6) as

$$\begin{aligned} x &= \varphi_1(x, y), \\ y &= \varphi_2(x, y), \end{aligned} \tag{7.7}$$

and construct the successive approximations according to the following formulas

$$\begin{aligned} x_{n+1} &= \varphi_1(x_n, y_n), \\ y_{n+1} &= \varphi_2(x_n, y_n), \\ n &= 1, 2, \dots \end{aligned} \tag{7.8}$$

If there exist the limits

$$\xi = \lim_{n \rightarrow \infty} x_n, \quad \eta = \lim_{n \rightarrow \infty} y_n$$

then the point  $(\xi, \eta)$  is a root of (7.6).

**Theorem 26.** *In a closed neighbourhood  $R = \{a \leq x \leq A, b \leq y \leq B\}$  let there be one and only one root  $x = \xi, y = \eta$  of (7.7). If  $\varphi_1(x, y), \varphi_2(x, y)$  are continuously differentiable in  $R$ ; the initial approximations  $(x_0, y_0)$  and all the succeeding approximations  $(x_n, y_n), n = 1, 2, \dots$  belong to  $R$ ;*

$$\begin{aligned} |\varphi'_{1x}(x, y)| + |\varphi'_{2x}(x, y)| &\leq q_1 < 1, \\ |\varphi'_{1y}(x, y)| + |\varphi'_{2y}(x, y)| &\leq q_2 < 1 \end{aligned}$$

*then the process of successive approximations (7.8) converges to the root  $(\xi, \eta)$  of the system (7.7).*

## 7.6 Estimate of an Approximation

For the method of iteration we have

$$|\xi - x_n| \leq \frac{q^n}{1 - q} \cdot |x_1 - x_0|.$$

It is possible to prove the inequality:

$$|\xi - x_n| \leq \frac{q}{1 - q} \cdot |x_n - x_{n-1}|.$$

**Example 56.** Find the real roots of the equation

$$x - \sin x = 0,25$$

to three significant digits.

**Solution.** Write the equation as

$$x = \sin x + 0,25.$$

We establish graphically that the equation has one real root  $\xi$  approximately equal to  $x_0 = 1,2$  in the interval  $[1,1; 1,3]$ . Let us put

$$\varphi(x) = \sin x + 0,25.$$

Since  $\varphi'(x) = \cos x$  and  $|\varphi'(x)| \leq \approx 0,62 = q$ ,  $x \in (0,9; 1,5)$  then

$$x_n = \sin x_{n-1} + 0,25, \quad n = 1, 2, \dots$$

These approximations lie in the interval  $(0,9; 1,5)$  and  $x_n \rightarrow \xi$  as  $n \rightarrow \infty$ . We construct the successive approximations  $x_n$ ,  $n = 1, 2, \dots$  until two adjacent approximations  $x_{n-1}$ ,  $x_n$  coincide to within the limits of error equal to

$$\frac{1 - q}{q} \cdot \varepsilon = 0,51 \cdot \frac{1}{2} \cdot 10^{-2} \approx 0,0025.$$

We have

$$\begin{aligned} x_1 &= \sin 1,2 + 0,25 = 1,182, \\ x_2 &= 1,175, \\ x_3 &= 1,173, \\ x_4 &= 1,172, \\ x_5 &= 1,172. \end{aligned}$$

Therefore  $\xi = 1,17 \pm 0,005$ .

## 7.7 The Method of Iteration in a Common Case

Given a system of nonlinear equations of a form

$$\begin{aligned}x_1 &= \varphi_1(x_1, x_2, \dots, x_n), \\x_2 &= \varphi_2(x_1, x_2, \dots, x_n), \\&\dots \\x_n &= \varphi_n(x_1, x_2, \dots, x_n)\end{aligned}$$

where the functions  $\varphi_1, \varphi_2, \dots, \varphi_n$  are real, defined and continuous in some neighbourhood  $\omega$  of an isolated solution  $x_1^*, x_2^*, \dots, x_n^*$ . Introducing the vectors

$$\begin{aligned}x &= (x_1, x_2, \dots, x_n), \\ \varphi &= (\varphi_1, \varphi_2, \dots, \varphi_n)\end{aligned}$$

we can write this system as

$$x = \varphi(x). \tag{7.9}$$

In finding the vector root

$$x^* = (x_1^*, x_2^*, \dots, x_n^*)$$

of equation (7.9) it is often convenient to use the *method of iteration*

$$x^{(p+1)} = \varphi(x^{(p)}), \quad p = 0, 1, 2, \dots \tag{7.10}$$

where the initial approximation  $x^{(0)} \approx x^*$ . If all the approximations  $x^{(p)}$ ,  $p = 0, 1, 2, \dots$  belong to the domain  $\omega$  and  $x^*$  is a *unique root* of the system (7.9) in  $\omega$ , then

$$x^* = \lim_{p \rightarrow \infty} x^{(p)}.$$

The method of iteration can also be applied to the general system

$$f(x) = 0$$

where  $f(x)$  is a vector function defined and continuous in the neighbourhood  $\omega$  of an isolated root  $\omega^*$ . For example, rewrite this system as

$$x = x + \Lambda f(x)$$

where  $\Lambda$  is a nonsingular matrix. Introducing the notation

$$x + \Lambda f(x) = \varphi(x)$$

we will have

$$x = \varphi(x).$$

## 7.8 Contracting Mapping

Let there be given a nonlinear system

$$y = \varphi(x) \tag{7.11}$$

where the functions  $\varphi_1, \varphi_2, \dots, \varphi_n$  are defined and continuous in a known domain  $G$  of a real  $n$ -dimensional space  $E_n$ , their values  $(y_1, y_2, \dots, y_n)$  for  $(x_1, x_2, \dots, x_n) \in G$  filling some domain  $G' \subset E_n$ . System (7.11) establishes a *mapping* of domain  $G$  onto  $G'$ . The mapping  $\varphi$  is termed a *contraction* mapping in the domain  $G$  if there exists a proper fraction  $q$  such that for any two points  $x^*, x^{**} \in G$  their images

$$y^* = \varphi(x^*), \quad y^{**} = \varphi(x^{**})$$

satisfy the condition

$$\|y^* - y^{**}\| \leq q \|x^* - x^{**}\|.$$

That is

$$\|\varphi(x^*) - \varphi(x^{**})\| \leq q \|x^* - x^{**}\|.$$

**Theorem 27.** *Let domain  $G$  be closed and let mapping (7.11) be a contracting mapping in  $G$ . Then if for the iteration process (7.10) all successive approximations  $x^{(p)} \in G$ ,  $p = 0, 1, 2, \dots$ , it follows that*

1. *irrespective of the choice  $x^{(0)} \in G$  the process (7.10) converges, i.e. there is the limit*

$$x^* = \lim_{p \rightarrow \infty} x^{(p)};$$

2. *the vector  $x^*$  is the sole solution of (7.11) in the domain  $G$ ;*

3. *the estimate*

$$\|x^* - x^{(p)}\| \leq \frac{q^p}{1 - q} \cdot \|x^{(1)} - x^{(0)}\|$$

*holds true.*

**Theorem 28.** *Let  $\varphi(x)$ ,  $\varphi'(x)$  be continuous in the domain  $G$ , and, in  $G$ , let the inequality*

$$\|\varphi'(x)\|_I = \max_{x \in G} \|\varphi'(x)\|_m = \max_{x \in G} \max_i \sum_{j=1}^n \left| \frac{\partial \varphi_i(x)}{\partial x_j} \right| \leq q < 1,$$

*where  $q$  is a constant, hold true. If the successive approximations lie in  $G$ , then the iteration process converges to the sole solution in  $G$ .*

**Remark 3.** The following inequality can be proved:

$$\|x^* - x^{(p)}\|_m \leq \frac{q^p}{1 - q} \cdot \|x^{(1)} - x^{(0)}\|_m, \quad p = 0, 1, 2, \dots$$

where  $x^{(1)} = \varphi(x^{(0)})$  and  $\|x\|_m = \max_i |x_i|$ .

# Chapter 8

## Numerical Methods for Ordinary Differential Equations

### 8.1 Euler's Method

Consider a differential equation

$$y' = f(x, y) \tag{8.1}$$

with the initial condition

$$y(x_0) = y_0.$$

Let us construct a system of equally spaced points  $x_i = x_0 + ih$ ,  $i = 0, 1, 2, \dots$  where  $h > 0$ .

In Euler's method approximate values of

$$y(x_i) \approx y_i$$

are computed successively by the formula

$$y_{i+1} = y_i + hf(x_i, y_i), \quad i = 0, 1, 2, \dots$$

Here the required integral curve  $y = y(x)$ , passing through the point  $M_0(x_0, y_0)$ , is replaced by a polygonal line  $M_0M_1M_2 \dots$  with the vertices  $M_i(x_i, y_i)$ ,  $i = 0, 1, 2, \dots$ . Each segment  $M_iM_{i+1}$  of this line, called *Euler's polygon*, has the direction coinciding with that of the integral curve of equation (8.1) which passes through the point  $M_i$ .

If the right-hand member of (8.1) in some rectangle

$$R\{|x - x_0| \leq a, |y - y_0| \leq b\}$$

satisfies the conditions

$$|f(x, y_1) - f(x, y_2)| \leq N|y_1 - y_2|, \quad N = \text{const},$$

$$\left| \frac{df}{dx} \right| = |f'_x + f \cdot f'_y| \leq M, \quad M = \text{const}$$

then we have the following error estimate:

$$|y(x_n) - y_n| \leq \frac{hM}{2N} \cdot [(1 + hN)^n - 1],$$

where  $y(x_n)$  is the value of the exact solution of the equation for  $x = x_n$ , and  $y_n$  is the approximate value obtained for the  $n$ th pitch. This formula has only theoretical application. For practical purposes sometimes it turns out to be more convenient to use *double computation*: the calculation is repeated with the spacing  $h/2$ , and the error of the more accurate value  $y_n^*$  (obtained at  $h/2$ ) is estimated in the following way:

$$|y_n^* - y(x_n)| \approx |y_n^* - y_n|.$$

Euler's method is readily applied to systems of differential equations, as well as to differential equations of higher orders. The latter should be first reduced to a system of differential equations of the first order.

Consider the system of two equations of the first order

$$\begin{aligned} y' &= f_1(x, y, z), \\ z' &= f_2(x, y, z) \end{aligned}$$

for the initial conditions

$$y(x_0) = y_0, \quad z(x_0) = z_0.$$

The approximative values

$$y(x_i) \approx y_i \quad \text{and} \quad z(x_i) \approx z_i$$

are computed successively by the formulas:

$$\begin{aligned} y_{i+1} &= y_i + hf_1(x_i, y_i, z_i), \\ z_{i+1} &= z_i + hf_2(x_i, y_i, z_i), \\ i &= 0, 1, 2, \dots \end{aligned}$$

## 8.2 Modifications of Euler's Method

*Euler's first improved method* for solving the problem (8.1) consists in that first the intermediate values are computed

$$x_{i+1/2} = x_i + \frac{h}{2},$$

$$y_{i+1/2} = y_i + \frac{h}{2} \cdot f_i,$$

$$f_{i+1/2} = f(x_{i+1/2}, y_{i+1/2}),$$

and then put

$$y_{i+1} = y_i + h f_{i+1/2}.$$

Using the *second improved method* (the *Euler-Cauchy method*), first determine the “rough approximation”

$$\tilde{y}_{i+1} = y_i + h f_i,$$

then compute

$$\tilde{f}_{i+1} = f(x_{i+1}, \tilde{y}_{i+1})$$

and put approximately

$$y_{i+1} = y_i + h \cdot \frac{f_i + \tilde{f}_{i+1}}{2}.$$

The remainder terms of Euler's first and second improved methods have the order  $O(h^3)$  for each spacing.

The error at the point  $x_n$  can be estimated with the aid of double computation: the calculation is repeated with the spacing  $h/2$  and the error of the more accuracy value  $y_n^*$  (for  $h/2$ ) is estimated approximately in the following way

$$|y_n^* - y(x_n)| \approx \frac{1}{3} |y_n^* - y_n|,$$

where  $y(x)$  is the exact solution of the differential equation.

### 8.3 Euler's Method Complete with an Iterative Process

The Euler-Cauchy method of solving the problem (8.1) can be made still more accurate by applying an iterative process to each value  $y_i$ . Namely, proceeding from the rough approximation

$$y_{i+1}^{(0)} = y_i + h \cdot f(x_i, y_i),$$

let us form an iterative process

$$y_{i+1}^{(k)} = y_i + \frac{h}{2} \left[ f(x_i, y_i) + f(x_{i+1}, y_{i+1}^{(k-1)}) \right].$$

Iterations are continued until the corresponding decimal digits of two subsequent approximations  $y_{i+1}^{(k)}, y_{i+1}^{(k+1)}$  coincide. Then we put

$$y_{i+1} \approx y_{i+1}^{(k+1)}.$$

As a rule, for a sufficiently small  $h$  iterations converge rapidly. If after three-four iterations the necessary number of decimal digits do not coincide, the spacing  $h$  must be decreased.

### 8.4 The Runge-Kutta Method

Consider the Cauchy problem for a differential equation

$$y' = f(x, y)$$

with the initial condition

$$y(x_0) = y_0.$$

Denote the approximate value of the sought-for solution at point  $x_i$  by  $y_i$ . According to the Runge-Kutta method, the approximate value  $y_{i+1}$  at the next point  $x_{i+1} = x_i + h$  is computed by the formulas

$$y_{i+1} = y_i + \Delta y_i,$$

$$\Delta y_i = \frac{1}{6} \left( K_1^{(i)} + 2K_2^{(i)} + 2K_3^{(i)} + K_4^{(i)} \right)$$

where

$$\begin{aligned} K_1^{(i)} &= hf(x_i, y_i), \\ K_2^{(i)} &= hf\left(x_i + \frac{h}{2}, y_i + \frac{K_1^{(i)}}{2}\right), \\ K_3^{(i)} &= hf\left(x_i + \frac{h}{2}, y_i + \frac{K_2^{(i)}}{2}\right), \\ K_4^{(i)} &= hf(x_i + h, y_i + K_3^{(i)}). \end{aligned}$$

It is advisable to arrange all the computations according to the computational scheme shown in a table below.

The table is filling in the following order:

1. Write the numerical values of  $x_0, y_0$  in the first row of the table.
2. Compute  $f(x_0, y_0)$ , multiply it by  $h$  and enter the result in the table as  $K_1^{(0)}$ .
3. Write the numerical values of  $x_0 + h/2, y_0 + K_1^{(0)}/2$  in the second row.
4. Compute  $f(x_0 + h/2, y_0 + K_1^{(0)}/2)$ , multiply it by  $h$ , and enter the result in the table as  $K_2^{(0)}$ .
5. Write the numerical values of  $x_0 + h/2, y_0 + K_2^{(0)}/2$  in the third row.
6. Compute  $f(x_0 + h/2, y_0 + K_2^{(0)}/2)$ , multiply it by  $h$ , and enter the result in the table as  $K_3^{(0)}$ .
7. Write the numerical values of  $x_0 + h, y_0 + K_3^{(0)}$  in the fourth row of the table.
8. Compute  $f(x_0 + h, y_0 + K_3^{(0)})$ , multiply it by  $h$ , and enter the result in the table as  $K_4^{(0)}$ .
9. Enter the numbers  $K_1^{(0)}, 2K_2^{(0)}, 2K_3^{(0)}, K_4^{(0)}$  in the column  $\Delta y$ .
10. Summate the numbers forming the column  $\Delta y$ , divide the sum by 6, and enter the result as  $\Delta y_0$ .
11. Compute  $y_1 = y_0 + \Delta y_0$ .

Then continue computing following the same order, taking  $(x_1, y_1)$  for the initial point.

$i$	$x$	$y$	$K = hf(x, y)$	$\Delta y$
0	$x_0$	$y_0$	$K_1^{(0)}$	$K_1^{(0)}$
	$x_0 + h/2$	$y_0 + K_1^{(0)}/2$	$K_2^{(0)}$	$2K_2^{(0)}$
	$x_0 + h/2$	$y_0 + K_2^{(0)}/2$	$K_3^{(0)}$	$2K_3^{(0)}$
	$x_0 + h$	$y_0 + K_3^{(0)}$	$K_4^{(0)}$	$K_4^{(0)}$
				$\Delta y_0$
1	$x_1$	$y_1$		

The results of the computation of the right-hand member of  $f(x, y)$  may be included in the table. But if the computations are too cumbersome, it is good practice to enter them in a separate table.

Note that the interval of computation (the spacing) may be changed when passing from one point to another. To check  $h$  for the proper choice it is recommended to compute the fraction

$$\Theta = \left| \frac{K_2^{(i)} - K_3^{(i)}}{K_1^{(i)} - K_2^{(i)}} \right|.$$

The quantity  $\Theta$  should not exceed several hundredths, otherwise  $h$  should be reduced. The order of accuracy of the Runge-Kutta method is  $h^4$  over the entire interval  $[x_0, x_n]$ . The error estimate with this method is rather difficult. The error can be roughly estimated with the aid of double computation by the formula

$$|y_n^* - y(x_n)| \approx \frac{1}{15} |y_n^* - y_n|,$$

where  $y(x_n)$  is the value of the exact solution of equation (8.1) for point  $x_n$ , and  $y_n^*, y_n$  are the approximate values obtained for  $h/2$  and  $h$ .

## 8.5 Adam's Method

Let for the equation

$$y' = f(x, y)$$

with the initial condition

$$y(x_0) = y_0$$

three consecutive values of the required function (the “initial interval” ) be found by one of the above considered methods

$$\begin{aligned} y_1 &= y(x_1) = y(x_0 + h), \\ y_2 &= y(x_2) = y(x_0 + 2h), \\ y_3 &= y(x_3) = y(x_0 + 3h). \end{aligned}$$

With the aid of these methods we compute the quantities

$$\begin{aligned} q_0 &= hy'_0 = hf(x_0, y_0), \\ q_1 &= hy'_1 = hf(x_1, y_1), \\ q_2 &= hy'_2 = hf(x_2, y_2), \\ q_3 &= hy'_3 = hf(x_3, y_3). \end{aligned}$$

Write down the numbers  $x_k, y_k, y'_k, q_k, k = 0, 1, 2, 3$  in the table and compute the finite differences of the quantity  $q$  (the number above the stepped line in table).

$k$	$x_k$	$y_k$	$\Delta y_k$	$y'_k = f(x_k, y_k)$	$q_k = hy'_k$	$\Delta q_k$	$\Delta^2 q_k$	$\Delta^3 q_k$
0	$x_0$	$y_0$	$\Delta y_0$	$f(x_0, y_0)$	$q_0$	$\Delta q_0$	$\Delta^2 q_0$	$\Delta^3 q_0$
1	$x_1$	$y_1$	$\Delta y_1$	$f(x_1, y_1)$	$q_1$	$\Delta q_1$	$\Delta^2 q_1$	$\Delta^3 q_1$
2	$x_2$	$y_2$	$\Delta y_2$	$f(x_2, y_2)$	$q_2$	$\Delta q_2$	$\Delta^2 q_2$	$\Delta^3 q_2$
3	$x_3$	$y_3$	$\Delta y_3$	$f(x_3, y_3)$	$q_3$	$\Delta q_3$		
4	$x_4$	$y_4$	$\Delta y_4$	$f(x_4, y_4)$	$q_4$	$\Delta q_4$		
5	$x_5$	$y_5$	$\Delta y_5$	$f(x_5, y_5)$	$q_5$			
6	$x_6$	$y_6$						

The Adams method consists in expanding the difference table with the aid of the formula

$$\Delta y_k = q_k + \frac{1}{2}\Delta q_{k-1} + \frac{5}{12}\Delta^2 q_{k-2} + \frac{3}{8}\Delta^3 q_{k-3}, \quad k = 3, 4, \dots \tag{8.2}$$

which is called *Adam’s extrapolation formula* and is used for “predicting” the value of

$$y_{k+1} = y_k + \Delta y_k.$$

Let us denote the “*predicted*” value by  $y_{k+1}^{\text{pred}}$ . The value of  $\Delta y_k$  obtained by formula (8.2) has to be specified. To this end we have to enter the values

$$x_{k+1}, y_{k+1}, y'_{k+1}, q_{k+1}$$

in the table, supplement the difference table, and then check the computation by the “*correction*” formula

$$\Delta y_k = q_k + \frac{1}{2}\Delta q_k - \frac{1}{12}\Delta^2 q_{k-1} - \frac{1}{24}\Delta^3 q_{k-2}, \quad (8.3)$$

which is called *Adam’s interpolation formula*. We denote the value specified by formula (8.3) by  $y_{k+1}^{\text{corr}}$ . The formulas (8.2), (8.3) are of a very high accuracy, yielding an error of order  $O(h^4)$ , but the formulas for estimating are rather complicated. For practical purposes the following is usually taken into consideration. The error of the more accurate correction formula (8.3) constitutes about 1/14 the difference between the values of  $\Delta y_k$  calculated by formulas (8.2), (8.3). Therefore, if this difference does not considerably exceed the permissible computational error, then the spacing  $h$  is considered to be chosen adequately, and the computation is continued with the chosen spacing. But if at a certain stage of computation the mentioned difference increases materially (and the calculations themselves are void of errors!) then the spacing  $h$  should be decreased (it is usually reduced by half).

Filling in the previous table:

1. Write down the numbers  $x_k, y_k, y'_k, q_k$  ( $k = 0, 1, 2, 3$ ), and compute the differences  $\Delta q_k$  ( $k = 0, 1, 2$ ),  $\Delta^2 q_k$  ( $k = 0, 1$ ),  $\Delta^3 q_0$ .
2. Using the numbers  $q_3, \Delta q_2, \Delta^2 q_1, \Delta^3 q_0$  placed diagonally in the table, determine by formula (8.2) for  $k = 3$

$$\Delta y_3 = q_3 + \frac{1}{2}\Delta q_2 + \frac{5}{12}\Delta^2 q_1 + \frac{3}{8}\Delta^3 q_0.$$

3. Compute  $x_4 = x_3 + h$ ,  $y_4 = y_3 + \Delta y_3$ .
4. Enter the values  $x_4, y_4$ , find  $y'_4 = f(x_4, y_4)$ ,  $q_4 = hy'_4$ , and supplement the difference table with the values of  $\Delta q_3, \Delta^2 q_2, \Delta^3 q_1$ .

5. Using the obtained values of the differences  $q$ , specify the quantity  $\Delta y_3$  by formula (8.3) for  $k = 3$ :

$$\Delta y_3 = q_3 + \frac{1}{2}\Delta q_3 - \frac{1}{12}\Delta^2 q_2 - \frac{1}{24}\Delta^3 q_1.$$

6. If the corrected value of  $\Delta y_3$  differs from its predicted value by several units of the lower retained order, then introduce the corresponding corrections in the values of  $\Delta y_3$  and  $y_4$ , check to see that the corrections do not affect considerably the value of  $q_4$ , and continue the computation with the chosen spacing. If otherwise, choose a smaller spacing.

7. The computations for  $k = 4, 5, \dots$  are accomplished in a similar way.

When carrying out computational work, it is more convenient to use transformed Adam's formulas, which express  $y_{k+1}$  not in terms of the difference  $\Delta q$ , but directly in terms of the quantity  $q$ . Thus, we obtain the *extrapolation formula of Adams* in the form

$$y_{k+1} = y_k + \frac{h}{24}(55y'_k - 59y'_{k-1} + 37y'_{k-2} - 9y'_{k-3})$$

and the *interpolation formula of Adams* in the form

$$y_{k+1} = y_k + \frac{h}{24}(9y'_{k+1} + 19y'_k - 5y'_{k-1} + y'_{k-2}).$$

The Adams method is readily applicable to systems of differential equations, as also to differential equations of the  $n$ th order.

Suppose we have a system of two equations

$$\begin{aligned} y' &= f_1(x, y, z), \\ z' &= f_2(x, y, z). \end{aligned}$$

Adam's extrapolation formulas for this system are written in the following way:

$$\begin{aligned} \Delta y_k &= p_k + \frac{1}{2}\Delta p_{k-1} + \frac{5}{12}\Delta^2 p_{k-2} + \frac{3}{8}\Delta^3 p_{k-3}, \\ \Delta z_k &= q_k + \frac{1}{2}\Delta q_{k-1} + \frac{5}{12}\Delta^2 q_{k-2} + \frac{3}{8}\Delta^3 q_{k-3}, \end{aligned}$$

where

$$p_k = hy'_k = hf_1(x_k, y_k, z_k), \quad q_k = hz'_k = hf_2(x_k, y_k, z_k).$$

Adams' interpolation formulas for the considered system are written analogously.

## 8.6 The Finite-Difference Method in boundary value problem for second order differential equation

### 1. Formulation of the problem

Consider the boundary-value problem

$$y'' + p(x)y' + q(x)y = f(x),$$

$$ay(0) + by'(0) = c, \quad dy(1) + ey'(1) = f,$$

where  $0 \leq x \leq 1$  and  $a, b, c, d, e, f$  are fixed constant. There are a lot of results stating when this problem has a unique solution. Due to time limitation we not consider theoretical aspects of this problem. On an example we show numerical approach to solution of this problem.

### 2. Example

Consider the following boundary-value problem for second-order differential equation

$$y'' - xy' + 3y = 2x, \tag{8.4}$$

$$y(0) = 0, \quad 3y'(1) + y(1) = 2 \tag{8.5}$$

where  $y = y(x)$ . Solve problem (8.4), (8.5) numerically with the step  $h = 0,25$ .

**Solution.** Denote  $x_0 = 0, x_1 = 0,25, x_2 = 0,5, x_3 = 0,75, x_4 = 1$  and put  $y_0 = 0$ . Then for every  $i = 1, 2, 3, 4$

$$y''(x_i) - x_i y'(x_i) + 3y(x_i) = 2x_i. \tag{8.6}$$

Expressing derivatives in (8.6) numerically we get ( $i = 0, 1, 2, 3, 4$ )

$$\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} - x_i \frac{y_{i+1} - y_{i-1}}{2h} + 3y_i = 2x_i$$

or

$$\left(1 + x_i \frac{h}{2}\right) y_{i-1} + (3h^2 - 2)y_i + \left(1 - x_i \frac{h}{2}\right) y_{i+1} = 2h^2 x_i,$$

and, putting  $h = 0,25$

$$(1 + 0,125x_i) y_{i-1} - 1,8125y_i + (1 - 0,125x_i) y_{i+1} = 0,125x_i.$$

Considering values  $i = 1, 2, 3, 4$  we get

$$\begin{aligned} 1,03125y_0 - 1,8125y_1 + 0,96875y_2 &= 0,03125, \\ 1,0625y_1 - 1,8125y_2 + 0,9375y_3 &= 0,0625, \\ 1,09375y_2 - 1,8125y_3 + 0,90625y_4 &= 0,09375. \\ 1,125y_3 - 1,8125y_4 + 0,875y_5 &= 0,125. \end{aligned}$$

Second condition in (8.5) leads to

$$3y'(x_4) + y(x_4) = 2$$

i.e.

$$\begin{aligned} 3 \frac{y(x_5) - y(x_3)}{2 \cdot 0,25} + y(x_4) &= 2, \\ 3 \frac{y_5 - y_3}{0,5} + y_4 &= 2, \end{aligned}$$

or

$$y_5 = \frac{0,5}{3}(2 - y_4) + y_3 = 0, \bar{3} - 0,1\bar{6}y_4 + y_3.$$

Taking into account this relation and value  $y_0$  which follows from initial problem we get resulting tridiagonal system:

$$\begin{aligned} -1,8125y_1 + 0,96875y_2 &= 0,03125, \\ 1,0625y_1 - 1,8125y_2 + 0,9375y_3 &= 0,0625, \\ 1,09375y_2 - 1,8125y_3 + 0,90625y_4 &= 0,09375, \\ 2y_3 - 1,958\bar{3}y_4 &= -0,1\bar{6}. \end{aligned}$$

Expressing step by step from the first equation  $y_1$ , then from the second one  $y_2$ , from the third one  $y_3$  we get from the fourth equation the value of  $y_4$ . Backward procedure gives values  $y_3, y_2$  and  $y_1$ . Finally, note that the exact solution is  $y(x) = x^3 - 2x$ .

## 8.7 The Finite-Difference Method for partial differential equations. Dirichlet problem

Let  $\Omega \subset \mathbb{R} \times \mathbb{R}$  and  $\bar{\Omega}$  is its boundary. The following problem is called Dirichlet problem: to find the function  $u = u(x, y)$  continuous on  $\bar{\Omega} := \Omega \cup \partial\Omega$  such that

$$\begin{aligned} u''_{xx} + u''_{yy} &= f(x, y), & (x, y) \in \Omega, \\ u(x, y) &= \varphi(x, y), & (x, y) \in \partial\Omega, \end{aligned}$$

where  $f(x, y)$  is a given continuous function on  $\Omega$  and  $\varphi(x, y)$  is a given continuous function on  $\partial\Omega$ . Concerning the terminology - the operator

$$\nabla^2 := \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

is called Laplacian (Laplace's operator), equation

$$\nabla^2 u = 0 \tag{8.7}$$

is called the Laplace's equations and equation

$$\nabla^2 u = f(x, y)$$

is called Poisson's equation. The last equation arise in the study of various time-independent problems, e.g. in the steady-state distribution of heat in a plane region.

For numerical solution of the Dirichlet problem is necessary to express the Laplacian in a discrete form suitable for numerical computations. Introduce rectangular grid with uniformly spaced grid points  $(x_i, y_j) \in \Omega$ ,  $x_{i+1} = x_i + h$ ,  $y_{j+1} = y_j + h$ . Then we calculate the approximation  $u_{ij}$  to the exact value  $u(x_i, y_j)$  of the solution  $u(x, y)$ . From the Taylor expansion of the function  $u(x, y)$  (supposing  $y_j$  is fixed) we have:

$$\begin{aligned} u(x_i + h, y_j) &\approx u(x_i, y_j) + \frac{\partial u(x_i, y_j)}{\partial x} h + \frac{1}{2!} \frac{\partial^2 u(x_i, y_j)}{\partial x^2} h^2, \\ u(x_i - h, y_j) &\approx u(x_i, y_j) - \frac{\partial u(x_i, y_j)}{\partial x} h + \frac{1}{2!} \frac{\partial^2 u(x_i, y_j)}{\partial x^2} h^2, \end{aligned}$$

or, numerically,

$$\frac{\partial^2 u_{ij}}{\partial x^2} = \frac{1}{h^2} (-2u_{ij} + u_{i+1,j} + u_{i-1,j}).$$

Proceed similarly for variable  $y$  (now the variable  $x$  is fixed) we get

$$\frac{\partial^2 u_{ij}}{\partial y^2} = \frac{1}{h^2}(-2u_{ij} + u_{i,j+1} + u_{i,j-1}).$$

Substitution of these 2 numerical relations into the Poisson's equation leads to the numerical scheme

$$-4u_{ij} + u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1} = h^2 f(x_i, y_j). \quad (8.8)$$

In particular, if  $f(x, y) \equiv 0$ , then

$$-4u_{ij} + u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1} = 0. \quad (8.9)$$

Formula (8.9) relates the function value in its four neighboring values ("Laplace stencil")

$$u_{i+1,j}, u_{i-1,j}, u_{i,j+1}, u_{i,j-1}.$$

For any interior grid point  $(x_i, y_j)$  we will write the equation (8.8). When a point  $(x_k, y_l) \in \partial\Omega$  then we substitute corresponding value  $u(x_k, y_l)$  by value  $\varphi(x_k, y_l)$ . In  $n$  is the number of interior grid points then this procedure leads to the system of  $n$  linear equations for  $n$  unknown functions. Summing up we conclude: each interior grid point introduces an equation to be solved. Better approximation require a more dense grid and very many equations might be needed. For computational processing some techniques that reduce the amount of the storage were developed.

**Example 57.** Find an approximate solution to Laplace's equation (8.7) in the region

$$\Omega = \{(x, y) \in [0, 1] \times [0, 1], y < x\}$$

with the stepsize  $h = 0,2$ , if the boundary values are

$$\begin{aligned} u(x, 0) &= -x^4, & x \in [0, 1], \\ u(1, y) &= -y^4 + 6y^2 - 1, & y \in [0, 1], \\ u(x, x) &= 4x^4, & x \in [0, 1]. \end{aligned}$$

**Solution.** Corresponding grid consists of points  $(x_i, y_j)$  with  $i = 0, 1, 2, 3, 4, 5$ ,  $j = 0, 1, 2, 3, 4, 5$  and  $i \leq j$ . For simplicity renumber  $u_{21}$  as  $u_1$ ,  $u_{31}$  as  $u_2$ ,

$u_{41}$  as  $u_3$ ,  $u_{32}$  as  $u_4$ ,  $u_{42}$  as  $u_5$  and  $u_{43}$  as  $u_6$ . Boundary values are

$$\begin{aligned} u_{00} &= 0, \\ u_{10} &= -0,2^4 = -0,0016, \\ u_{20} &= -0,4^4 = -0,0256, \\ u_{40} &= -0,8^4 = -0,4096, \\ u_{50} &= -1, \end{aligned}$$

$$\begin{aligned} u_{50} &= -1, \\ u_{51} &= -0,2^4 + 6 \cdot 0,2^2 - 1 = -0,7616, \\ u_{52} &= -0,4^4 + 6 \cdot 0,4^2 - 1 = -0,0656, \\ u_{53} &= -0,6^4 + 6 \cdot 0,6^2 - 1 = 1,0304, \\ u_{54} &= -0,8^4 + 6 \cdot 0,8^2 - 1 = 2,4304, \\ u_{55} &= -1 + 6 - 1 = 4, \end{aligned}$$

and

$$\begin{aligned} u_{00} &= 0, \\ u_{11} &= 4 \cdot 0,2^4 = 0,0064, \\ u_{22} &= 4 \cdot 0,4^4 = 0,1024, \\ u_{33} &= 4 \cdot 0,6^4 = 0,5184, \\ u_{44} &= 4 \cdot 0,8^4 = 1,6384, \\ u_{55} &= 4 \cdot 1 = 4. \end{aligned}$$

Corresponding system of equation is

$$\begin{aligned} -4u_1 + 0,0064 + u_2 - 0,0256 + 0,1024 &= 0, \\ -4u_2 + u_1 + u_3 - 0,1296 + u_4 &= 0, \\ -4u_3 + u_2 - 0,7616 - 0,4096 + u_5 &= 0, \\ -4u_4 + 0,1024 + u_5 + u_2 + 0,5184 &= 0, \\ -4u_5 + u_4 - 0,0656 + u_3 + u_6 &= 0, \\ -4u_6 + 0,5186 + 1,0304 + u_5 + 1,6384 &= 0. \end{aligned}$$

After simplifying we get

$$\begin{aligned} -4u_1 + u_2 &= -0,0832, \\ u_1 - 4u_2 + u_3 + u_4 &= 0,1296, \\ u_2 - 4u_3 + u_5 &= 1,1712, \\ u_2 - 4u_4 + u_5 &= -0,6208, \\ u_3 + u_4 - 4u_5 + u_6 &= 0,0656, \\ u_5 - 4u_6 &= -3,1872. \end{aligned}$$

By Gauss eliminations method we get easily the numerical solution:

$$\begin{aligned}u(0.4, 0.2) &\sim u_1 = 0,0086, \\u(0.6, 0.2) &\sim u_2 = -0,0489, \\u(0.8, 0.2) &\sim u_3 = -0,2612, \\u(0.6, 0.4) &\sim u_4 = 0,1868, \\u(0.8, 0.4) &\sim u_5 = 0,1751, \\u(0.8, 0.6) &\sim u_6 = 0,8406.\end{aligned}$$



# Bibliography

- [1] BUCHANAN, J.L., TURNE, P.R.: *Numerical Methods and Analysis*, McGraw-Hill, Inc., 1992.
- [2] DANKO, P.E., POPOV, R.G., KOZHEVNIKOVA, T. YA.: *Higher Mathematics in Problems and Exercises, Part 2*, Mir, Moscow, 1983.
- [3] MCCLAVE, J.T., DIETRICH II, FRANK, H.: *Statistics, Sixth Edition* Macmillian College Publishing Company, New York, 1994.
- [4] MONTGOMERY, D.C., RUNGER, G.C.: *Applied Statistics and Probability for Engineers, Third Edition*, John Wiley & Sons, Inc., 2003.
- [5] RILEY, K.F., HOBSON, M.P., BENICE, S.J.: *Mathematical Methods for Physics and Engineering, Second Edition*, Cambridge, 2003.
- [6] VOLKOV, E.A.: *Numerical Methods*, Mir Publisher, Moscow, 1986.
- [7] WALPOLE RONALD, E., MYERS, RAYMOND, H.: *Probability and Statistics for Engineers and Scientists, Fourth Edition*, Maxmilian Publishing Company, New York, 1990.
- [8] WOLFE, M.A.: *Numerical Analysis*, VNR, 1972.

# Index

- Additive rule, 9
- Bayes's Formula, 24
- Bernoulli's Formula, 37
- Binomial Distribution, 38
- Collectively Independed Events, 18
- Combinations rule, 21
- Complementary Events, 11
- Compound events, 8
- Conditional probability, 12
- Counting rules, 23
- Event, 3
  - Additive rule, 9
  - Complementary events, 11
  - Compound events, 8
  - Mutually exclusive events, 9
  - Probability, 8
- Events
  - Collectively independent, 18
- Gaussian Distribution, 40
- Independent events, 15
- Laplace Distribution, 40
- Multiplicative rule, 15
- Mutually exclusive events, 9
- Normal Distribution, 40
- Partitions rule, 20
- Permutations rule, 20
- Poisson's Distribution, 38
- Probability, 3, 6
  - Conditional, 12
- Random variable, 26
- Rule
  - Combinations, 21
  - Multiplicative, 15, 19
  - Partitions, 20
  - Permutations, 20
- Total Probability Formula, 24
- Variable
  - Random, 26

# Contents

<b>1</b>	<b>ELEMENTS OF PROBABILITY THEORY AND STATISTICS</b>	<b>3</b>
1.1	Sample Space . . . . .	3
1.2	Probability . . . . .	6
1.3	Unions and Intersections . . . . .	8
1.4	The Additive Rule and Mutually Exclusive Events . . . . .	9
1.5	Complementary Events . . . . .	11
1.6	Conditional Probability . . . . .	12
1.7	The Multiplicative Rule and Independent Events . . . . .	15
1.8	Collectively Independent Events . . . . .	18
1.9	Some Counting Rules . . . . .	19
1.	The Multiplicative Rule . . . . .	19
2.	The Permutations Rule . . . . .	20
3.	The Partitions Rule . . . . .	20
4.	The Combinations Rule . . . . .	21
1.10	Summary of Counting Rules . . . . .	23
1.	Multiplicative rule . . . . .	23
2.	Permutations rule . . . . .	23
3.	Partitions rule . . . . .	23
4.	Combinations rule . . . . .	24
1.11	Total Probability Formula, Bayes's Formula . . . . .	24
1.	Total Probability Formula . . . . .	24
2.	Bayes's Formula . . . . .	25
1.12	Random Variable and the Law of Its Distribution . . . . .	26
1.	Definition of a random variable . . . . .	26
2.	Discrete random variable . . . . .	27

3.	Continuous random variable . . . . .	28
4.	The Mean Value, the Variance and Standard Deviation of the Random variable . . . . .	30
5.	The Mode and the Median . . . . .	34
1.13	Some Special Distributions . . . . .	35
1.	Discrete Uniform Distribution . . . . .	35
2.	Continuous Uniform Distribution . . . . .	36
3.	Binomial Distribution. Poisson's Distribution . . . . .	37
4.	Normal Distribution (Gaussian normal distribution) and Laplace Function . . . . .	40
1.14	Chebyshev's Theorem . . . . .	42
1.15	Normal Approximation to the Binomial . . . . .	44
1.16	The Central Limit Theorem . . . . .	45
1.17	Transformed Random Variables . . . . .	46
1.18	Statistics . . . . .	48
1.	Sample Statistics . . . . .	49
2.	Point Estimator . . . . .	52
3.	Interval Estimation . . . . .	55
<b>2</b>	<b>Polynomial Approximation to Functions</b>	<b>61</b>
2.1	Lagrange's Interpolation Polynomial . . . . .	61
2.2	Interpolation Error . . . . .	63
2.3	Linear Interpolation . . . . .	65
2.4	Finite and Divided Differences . . . . .	66
2.5	Newton's Interpolation Formula . . . . .	68
2.6	The case of equally spaced interpolation points . . . . .	69
<b>3</b>	<b>Numerical Differentiation</b>	<b>73</b>
3.1	Simplest Formulas of Numerical Differentiation . . . . .	73
3.2	Applying Lagrange's Interpolation Polynomial . . . . .	75
3.3	Applications of Newton's Interpolation Polynomial . . . . .	77
3.4	General error estimate . . . . .	77
<b>4</b>	<b>Splines</b>	<b>79</b>
4.1	Methods of Specifying the Inclinations of an Interpolation Cubic Spline . . . . .	80

1.	Method I. (Simplified) . . . . .	80
2.	Method II . . . . .	80
4.2	The Error of Approximation by a Spline . . . . .	82
<b>5</b>	<b>The Method of Least Squares</b>	<b>83</b>
5.1	Introduction . . . . .	83
5.2	A Polynomial of Best Mean-square Approximation . . . . .	87
5.3	Mean-square Approximations by Algebraic Polynomials . . . . .	89
5.4	Application of Orthogonal Polynomials . . . . .	91
5.5	Method of least squares - continuation (a practical approach)	92
5.6	Special cases - a) . . . . .	92
5.7	Special cases - b) . . . . .	93
5.8	Special cases - c) . . . . .	93
5.9	Special cases - d) . . . . .	94
<b>6</b>	<b>Numerical Integration</b>	<b>95</b>
6.1	Quadrature Formulae . . . . .	95
6.2	The Rectangular Formula . . . . .	96
6.3	The Trapezoidal Formula . . . . .	97
6.4	Simpson Formula . . . . .	98
6.5	Composite Quadrature Formulas . . . . .	99
6.6	Newton-Cotes Quadrature Formulas . . . . .	102
<b>7</b>	<b>Methods of Solving Nonlinear Equations and Systems</b>	<b>105</b>
7.1	The Halving Method (Method of Dividing a Line Segment into Two Equal Parts) . . . . .	105
7.2	The Method of Chords (Method of Proportional Parts) . . . . .	106
7.3	Newton's Method (Method of Tangents) . . . . .	107
7.4	The Method of Iteration . . . . .	108
7.5	The Method of Iteration for a System of Two Equations . . . . .	110
7.6	Estimate of an Approximation . . . . .	111
7.7	The Method of Iteration in a Common Case . . . . .	112
7.8	Contracting Mapping . . . . .	113
<b>8</b>	<b>Numerical Methods for Ordinary Differential Equations</b>	<b>115</b>
8.1	Euler's Method . . . . .	115

8.2	Modifications of Euler's Method . . . . .	117
8.3	Euler's Method Complete with an Iterative Process . . . . .	118
8.4	The Runge-Kutta Method . . . . .	118
8.5	Adam's Method . . . . .	120
8.6	The Finite-Difference Method in boundary value problem for second order differential equation . . . . .	124
	1. Formulation of the problem . . . . .	124
	2. Example . . . . .	124
8.7	The Finite-Difference Method for partial differential equa- tions. Dirichlet problem . . . . .	126