E3: PROBABILITY AND STATISTICS
lecture notes

Contents

Chapter 1

PROBABILITY THEORY

1.1 Experiments and random events.

Definition 1.1.1. *In probability theory,* ***random experiment*** *means a repeatable process that yields a result or an observation.*

Tossing a coin, rolling a die, extracting a ball from a box are random experiments.

When tossing a coin, we get one of the following elementary results:

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

When throwing a die, if we denote by (1) the appearance of the face with one dot, with (2) the appearance of the face with two dots, etc., then we get the following elementary results:

$$(1),\ (2),\ (3),\ (4),\ (5),\ (6).$$

Definition 1.1.2. *A* ***random event*** *is an event that either happens or fails to happen as a result of an experiment.*

When tossing a coin, the event (heads) may happen or may fail to happen, so this is a *random event*. On the other hand, if we consider the event (the coin falls down), we observe that this is a *certain event*, due to gravity.

A random event depends on the combined action of several factors which may not have been taken into consideration when setting up the experiment. When tossing a coin, such factors may be: the way we move our hand, the characteristics of the coin, the position of the coin in our hand.

A priori, we cannot say anything about the outcome of a random experiment. We cannot foresee if we obtain heads or tails when tossing a coin. Probability theory deals with such random events, giving us a tool to evaluate the chances of various outcomes of experiments.

1.2 Certain event. Impossible event.

There are two special events for every experiment: the certain event and the impossible event.

Definition 1.2.1. *The **certain event** (denoted by $\mathcal{S}$) is an event which happens with certitude at each repetition of an experiment.*

When tossing a coin, the event (one of the two faces appears) is a certain event of the experiment. When rolling a die, the event (one of the six faces appears) is a certain event of the experiment.

Definition 1.2.2. *The **impossible event** (denoted by $\emptyset$) is an event which never happens in a random experiment.*

When extracting a ball from a box which contains only white balls, the event (a red ball is extracted) is an impossible event.

1.3 Contrary events.

In the case of rolling a die, let's denote by A the event consisting of the appearance of one of the faces 2 or 5, and B the event consisting of the appearance of one of the faces 1, 3, 4 or 6. We observe that if the event A does not take place, then the event B takes place, and the other way round.

Definition 1.3.1. *The **contrary** of an event A is an event B satisfying the property that, at any repetition of the experiment, if the event A occurs then B does not occur, and if the event B occurs then A does not occur. The events A and B are also called **mutually exclusive events**.*

If B is the contrary of A then A is the contrary of B.

We denote the contrary of an event A by $\overline{A}$ or $\complement A$.

1.4 Compatible events. Incompatible events.

Definition 1.4.1. *The events A and B are **compatible** if they can occur simultaneously.*

When throwing a die, the event A=(an even number appears) and the event B=(one of the numbers 2 or 6 appears), are compatible. If the outcome of the experiment is the appearance of the face with the number 2, then both events A and B take place.

Definition 1.4.2. *The events A and C are **incompatible** if they cannot occur simultaneously.*

When rolling a die, the events A=(an even number appears) and C=(an odd number appears) are incompatible. They cannot take place at the same time. One may notice that the events A and C are contrary events.

On the other hand, if we consider the event D=(the number 5 appears), we can see that A and D are incompatible, but they are not contrary events: the non-occurrence of A does not imply the occurrence of D.

Definition 1.4.3. *The events $A_1, A_2, ..., A_n$ are **compatible** if they can occur simultaneously.*

When throwing a die, the events:

A_1=(one of the faces 2 or 4 appears)

A_2=(one of the faces 2 or 6 appears)

A_3=(one of the faces 2, 4 or 6 appears)

are compatible: if the outcome of the experiment is the appearance of the face with the number 2, all three events take place.

1.5 Event implied by another event.

Definition 1.5.1. *We say that* ***the event*** A ***implies the event*** B *(or* ***the event*** B ***is implied by the event*** A*) if the occurrence of the event A means that the event B occurs as well.*

When throwing a die, the event A=(one of the faces 1 or 3 appears) implies the event B=(one of the faces 1, 2, 3 or 5 appears).

Any event implies the certain event.

1.6 Operations with events.

In the framework of an experiment, when we study the occurrence of an event, we actually analyze the occurrence of a part of the set of elementary results of an experiment.

When rolling a die, if we study the occurrence of the event A=(one of the faces 1 or 3 appears), we actually analyze whether we obtain one of the results (1) or (3) from the set of elementary events (1), (2), (3), (4), (5), (6). The event A is completely determined by the set formed by these two elementary results, and we can identify it by $A = \{1, 3\}$.

Consider the experiment in which two dice are rolled. We are interested in the event A=(the sum of the numbers on the two dice is 7). We look at the following elementary events:

$$(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)$$

and we write A as the following set:

$$A = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}.$$

The set which represents the impossible event, which never takes place in our experiment, is the empty set $\emptyset$. The certain event is represented by the set of all elementary events.

If A is the set of results which represent an event, then the set $\complement A$ (the complementary of A) is the set of results which represent the contrary event.

We have seen that the fact that the event A implies the event B means that whenever A takes place, B takes places as well. Therefore, the set of results representing the event A is included in the set of results representing the event B: $A \subset B$.

The sets representing two incompatible events are disjoint.

Definition 1.6.1. *The* ***union*** $A \cup B$ *of two events A and B is the event which takes place when at least one of the events A or B occur.*

Definition 1.6.2. *The **intersection** $A \cap B$ of two events A and B is the event which occurs when both events A and B take place at the same time.*

Example 1.6.1. For the experiment of rolling one die, let's consider the following events:

$$A = \{1, 2, 5\},\ B = \{3, 4, 5\}.$$

The event A occurs if one of the following results is obtained: $\{1\}$, $\{2\}$ or $\{5\}$; the event B occurs if one of the results $\{3\}$, $\{4\}$ or $\{5\}$ is obtained.

To insure that at least one of the events A or B take place, we must obtain one of the results $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$. Therefore:

$$A \cup B = \{1, 2, 3, 4, 5\}.$$

On the other hand, both events take place at the same time only in the case when the face with number 5 is obtained, so we get:

$$A \cap B = \{5\}.$$

1.7 Sample space of an experiment.

Consider the following example:

Example 1.7.1. The experiment consists of tossing two coins.

If we take a look at the repartition of heads and tails on the two coins, all the possible outcomes of this experiment form the following set:

$$\{(H, H), (H, T), (T, H), (T, T)\} = \mathcal{A}_1.$$

The first letter of each couple corresponds to the first coin, while the second letter corresponds to the second coin. In this case, every elementary result of the experiment can be regarded as an element of the set $\mathcal{A}_1$.

From another point of view, if we look at the number of obtained heads and the number of obtained tails, then the set of elementary results of the experiment can be expressed as the set:

$$\{(2, 0), (1, 1), (0, 2)\} = \mathcal{A}_2.$$

In this case, the first number of each couple represents the number of obtained heads, and the second number represents the number of obtained tails. Every elementary result of the experiment can be regarded as an element of the set $\mathcal{A}_2$.

Finally, if we analyze whether the two symbols obtained on the two coins are different or the same, then the set of elementary results of the experiment is:

$$\{\text{same}, \text{different}\} = \mathcal{A}_3.$$

Every elementary result of the experiment can be regarded as an element of the set $\mathcal{A}_3$.

Each of the three sets $\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{A}_3$ represents a set of elementary results of the same experiment. However, in each case, the concept of elementary result of the experiment is specific (it means something different) and the sets $\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{A}_3$ are called sample spaces.

The sample space $\mathcal{A}_1$ provides more information then the spaces $\mathcal{A}_2$ or $\mathcal{A}_3$. If we know which result of the sample space $\mathcal{A}_1$ has occurred, we can say which was the result of the space $\mathcal{A}_2$ or $\mathcal{A}_3$.

Definition 1.7.1. *The **sample space** of an experiment is a set of all elementary results (possible outcomes) of the experiment. The elements of a sample space are called **sample points**.*

1.8 Frequency.

Let's consider an experiment and an event A associated to this experiment. We repeat the experiment n times (in given conditions) and we denote by α the number of occurrences of the event A. The number of occurrences of the event $\bar{A}$ is $n - \alpha$.

Definition 1.8.1. *The number* $f_n(A) = \dfrac{\alpha}{n}$ *is called* ***relative frequency*** *of the event* A.

The number α, called **absolute frequency** of the event A, is between 0 and n; $\alpha = 0$ if during n repetitions of the experiment, the event A did never occur; $\alpha = n$ if the event A occurred at every repetition of the experiment. Therefore

$$0 \leq \alpha \leq n \quad \text{and} \quad 0 \leq f_n(A) \leq 1, \qquad \forall n \in \mathbb{N}^\star$$

Proposition 1.8.1 (Properties of the relative frequency)**.**

1. $f_n(\mathcal{S}) = 1$, *where* $\mathcal{S}$ *is the certain event;*
2. *If* $A \cap B = \emptyset$ *then* $f_n(A \cup B) = f_n(A) + f_n(B)$.

1.9 Equally possible events.

Example 1.9.1. We consider the experiment of tossing a coin. The possible outcomes of this experiment are: obtaining heads or obtaining tails, and we cannot know a priori which will be the result. If there is no reason to believe that the occurrence of one of these two events is favored, we can say that these two events are equally possible.

Example 1.9.2. When rolling a die, we can obtain any of the six faces. If there is no reason to suppose that the occurrence of a face is favored, we can say that the events (1), (2), (3), (4), (5), (6) are equally possible. In the framework of the same experiment, the events $A = \{1, 2\}$ and $B = \{3, 4\}$ are also equally possible.

Definition 1.9.1. *Consider two events* A *and* B *associated to an experiment. If there is no reason to suppose that the occurrence of one of the two events is favored with respect to the other, then we say that the two events are* ***equally possible****.*

1.10 Probability of an event.

Example 1.10.1. Let's consider the experiment of tossing a coin and the sample space $\mathcal{A}$ which includes the two possible elementary results of this experiment:

H=(the outcome is heads)

T=(the outcome is tails)

$\mathcal{A} = \{H, T\}$.

As these two events H and T are equally possible, it is natural to estimate (to measure) the chance of occurrence of each of them by $1/2$ = the inverse of the number of elementary events from $\mathcal{A}$ (the relative frequency $50\% = 1/2$ of each of the two events).

Example 1.10.2. We consider the experiment of rolling a die and the associated sample space $\mathcal{A} = \{(1), (2), (3), (4), (5), (6)\}$. As these six events are equally possible, it is natural to evaluate the chance of occurrence of each of them by $1/6$ = the inverse of the number of events from $\mathcal{A}$.

Example 1.10.3. Consider the experiment of tossing two coins and the sample space $\mathcal{A} = \{(H, H), (H, T), (T, H), (T, T)\}$. As the four events are equally possible, the chance of occurrence of each of them is evaluated by $1/4$ = the inverse of the number of events from $\mathcal{A}$.

Example 1.10.4. When tossing two coins and considering the sample space

$$\mathcal{A} = \{(\text{same symbol}), (\text{different symbols})\}$$

as the events are equally possible, we evaluate the chance of occurrence of each event by $1/2$ = the inverse of the number of events from $\mathcal{A}$.

Definition 1.10.1. *If the events of the sample space $\mathcal{A}$ associated to an experiment are equally possible, we say that they are* ***equally probable*** *and* ***the probability of each event is equal to the inverse of the number of events from the sample space.***

In the followings, we extend the definition of the probability of an event to events which do not belong to the sample space associated to an experiment, but which are parts of $\mathcal{A}$, meaning that they belong to $\mathcal{P}(\mathcal{A})$ (the set of parts of $\mathcal{A}$). Let's start with an example.

Example 1.10.5. Consider the experiment of rolling a die and the sample space $\mathcal{A} = \{(1), (2), (3), (4), (5), (6)\}$. The event A =(an even number appears) is actually $A = \{(2), (4), (6)\}$. The occurrence of any of the events (2), (4), (6) is favorable to the occurrence of A. That's why we evaluate the chance of occurrence of the event A (the probability of A) by 3 times the chance of occurrence of any of the events (2), (4), (6). The ratio $\frac{3}{6} = \frac{1}{2}$ represents the chance (probability) of occurrence of the event A and is obtained by dividing the number of events from $\mathcal{A}$ which are favorable for the occurrence of A to the number of all events of $\mathcal{A}$.

Definition 1.10.2. *If the sample space $\mathcal{A}$ associated to an experiment is made up of n equally probable events and A is an event belonging to $\mathcal{P}(\mathcal{A})$, then the* ***probability of the event*** *A is the ratio between the number of equally probable events of $\mathcal{A}$ which define A and the total number of events of $\mathcal{A}$.*

It results that if $A = \emptyset$ then $P(A) = 0$ and if $A = \mathcal{A}$ then $P(A) = 1$. In general, $P(A) \in [0, 1]$.

Taking into account the definition of a contrary event, it results that if $\mathcal{A}$ has n elements and A is made up of $m \leq n$ elements, then $\bar{A}$ has $n - m$ elements and:

$$P(\bar{A}) = \frac{n - m}{n} = 1 - \frac{m}{n} = 1 - P(A).$$

1.11 Finite sample space. Elementary event.

Definition 1.11.1. *A* ***finite sample space*** *associated to an experiment is a finite set $\mathcal{S} = \{e_1, e_2, ..., e_n\}$ of abstract elements.*

The parts of the set $\mathcal{S}$ are called ***events****. An vent is called* ***elementary*** *if it consists of a single point of the space $\mathcal{S}$. The empty part of $\mathcal{S}$, $\emptyset$, is called* ***impossible event****, and $\mathcal{S}$ is called* ***certain event****.*

Example 1.11.1. The teachers of a school are asked the following questions:

1. Is it necessary to modernize the school?

2. Is it necessary to build a sport facility for the school?

The answers given by one teacher can be one of the followings: $e_1 = (\text{YES}, \text{YES})$, $e_2 = (\text{YES}, \text{NO})$, $e_3 = (\text{NO}, \text{YES})$, $e_4 = (\text{NO}, \text{NO})$. The set

$$\mathcal{S} = \{e_1, e_2, e_3, e_4\}$$

is a possible sample space for this experiment for a given teacher. The subsets of this space are:

$$\begin{aligned}\mathcal{P}(\mathcal{S}) \;=\; & \{\emptyset, \{e_1\}, \{e_2\}, \{e_3\}, \{e_4\}, \{e_1, e_2\}, \{e_1, e_3\}, \{e_1, e_4\}, \{e_2, e_3\}, \{e_2, e_4\}, \{e_3, e_4\}, \\ & \{e_1, e_2, e_3\}, \{e_1, e_2, e_4\}, \{e_2, e_3, e_4\}, \{e_1, e_2, e_3, e_4\}\}.\end{aligned}$$

Each of these subsets is an event. The subsets $E_1 = \{e_1\}$, $E_2 = \{e_2\}$, $E_3 = \{e_3\}$, $E_4 = \{e_4\}$ contain only one point and they are elementary events. Any event (besides the impossible event) is a union of elementary events.

1.12 Axiomatic definition of probability

Definition 1.12.1. *We call* ***probability on the sample space*** $\mathcal{S} = \{e_1, e_2, ..., e_n\}$ *a function* P *which associates to every event* $A \in \mathcal{P}(\mathcal{S})$ *a number* $P(A)$*, called probability of* A*, such that the following conditions (axioms) are fulfilled:*

i) $P(A) \geq 0$, $\forall A \in \mathcal{P}(\mathcal{S})$;

ii) $P(\mathcal{S}) = 1$;

iii) $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$, $\forall A, B \in \mathcal{P}(\mathcal{S})$.

The function $P : \mathcal{P}(\mathcal{S}) \to \mathbb{R}_+^1$ *is called* ***probability measure****.*

The sample space $\mathcal{S}$ *together with the probability measure* P *(the pair* $(\mathcal{S}, P)$*) is called* ***probability space****.*

Proposition 1.12.1. *Let* $A \in \mathcal{P}(\mathcal{S})$.

1. *If* $A = \emptyset$ *then* $P(A) = 0$.

2. *If* $A = \{e_1, e_2, ..., e_k\}$ *then* $P(A) = \sum_{i=1}^{k} P(\{e_i\})$.

Proof. As $P(\emptyset \cup \mathcal{S}) = P(\emptyset) + P(\mathcal{S})$ and $\emptyset \cup \mathcal{S} = \mathcal{S}$, it results that $P(\emptyset) + P(\mathcal{S}) = P(\mathcal{S})$ and therefore $P(\emptyset) = 0$.

If $A = \{e_1, e_2, ..., e_k\}$ it can be written as $A = \{e_1, e_2, ..., e_{k-1}\} \cup \{e_k\}$, hence $P(A) = P(\{e_1, e_2, ..., e_{k-1}\}) + P(\{e_k\})$. Therefore, we have:

$$\begin{aligned} P(\{e_1, e_2, ..., e_k\}) &= P(\{e_1, e_2, ..., e_{k-1}\}) + P(\{e_k\}) \\ P(\{e_1, e_2, ..., e_{k-1}\}) &= P(\{e_1, e_2, ..., e_{k-2}\}) + P(\{e_{k-1}\}) \\ \cdots \quad &\cdots \quad \cdots \\ P(\{e_1, e_2\}) &= P(\{e_1\}) + P(\{e_2\}) \end{aligned}$$

Summing up all these equalities, we obtain:

$$P(\{e_1, e_2, ..., e_k\}) = \sum_{i=1}^{k} P(\{e_i\}).$$

□

Consequence 1.12.1. *If all n elementary events e_1, e_2, ..., e_n of the sample space $\mathcal{S}$ have the same probability (are equally probable), i.e. $P(\{e_i\}) = P(\{e_j\}), \forall i, j = \overline{1,n}$, then $P(\{e_i\}) = \frac{1}{n}, \forall i = \overline{1,n}$.*

Remark 1.12.1. In many applications, the elementary events from a sample space $\mathcal{S}$ have different probabilities. In Example 1.11.1, it is possible that the number of the teachers answering e_i is different of those answering e_j. Suppose that 60% of the teachers answer e_1, 20% answer e_2, 15% answer e_3 and 5% answer e_4. It is obvious that in this case, we associate the following probabilities to the elementary events:

$$P(\{e_1\}) = 0.6 \quad P(\{e_2\}) = 0.2 \quad P(\{e_3\}) = 0.15 \quad P(\{e_4\}) = 0.05.$$

Proposition 1.12.2. *For any $A \in \mathcal{P}(\mathcal{S})$, we have:*

$$P(\complement A) = 1 - P(A).$$

Proof. As $A \cap \complement A = \emptyset$ and $A \cup \complement A = \mathcal{S}$, we have $P(A) + P(\complement A) = P(\mathcal{S}) = 1$, so $P(\complement A) = 1 - P(A)$. □

Proposition 1.12.3. *If $A, B \in \mathcal{P}(\mathcal{S})$ and $A \subset B$ then $P(A) \leq P(B)$.*

Proof. $A \subset B$, so $B = A \cup (B \cap \complement A)$. As $A \cap (B \cap \complement A) = \emptyset$, it results that $P(B) = P(A) + P(B \cap \complement A)$ and as $P(B \cap \complement A) \geq 0$, it results that $P(B) \geq P(A)$. □

Proposition 1.12.4. *If $A_1, A_2, ..., A_n \in \mathcal{P}(\mathcal{S})$ and $A_i \cap A_j = \emptyset, \forall i \neq j$, then*

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

Proof. For $n = 2$ the equality $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ is true, by means of axiom iii) of the definition of the probability measure.

For $n = 3$ we have $(A_1 \cup A_2) \cap A_3 = \emptyset$, so

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

By mathematical induction, suppose that for any $A_1, A_2, .., A_n$ such that $A_i \cap A_j = \emptyset, \forall i \neq j$ we have

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

and consider $A_1, A_2, .., A_n, A_{n+1}$ such that $A_i \cap A_j = \emptyset, \forall i \neq j$, $i, j = \overline{1, n+1}$. We have:

$$\begin{aligned} P\left(\bigcup_{i=1}^{n+1} A_i\right) &= P\left(\bigcup_{i=1}^{n} A_i \cup A_{n+1}\right) = P\left(\bigcup_{i=1}^{n} A_i\right) + P(A_{n+1}) = \\ &= \sum_{i=1}^{n} P(A_i) + P(A_{n+1}) = \sum_{i=1}^{n+1} P(A_i) \end{aligned}$$

□

Proposition 1.12.5. *For any $A, B \in \mathcal{P}(\mathcal{S})$ the following equality holds:*

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

Proof. Consider $C = A \cap \complement B$, $D = B \cap \complement A$ and remark that we have

$$A \cup B = C \cup (A \cap B) \cup D$$

leading us to

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

Taking into account the equalities

$$P(A) = P(A \cap B) + P(A \cap \complement B) = P(A \cap B) + P(C)$$

and

$$P(B) = P(A \cap B) + P(B \cap \complement A) = P(A \cap B) + P(D)$$

we obtain:

$$\begin{aligned} P(A \cup B) &= P(A) - P(A \cap B) + P(A \cap B) + P(B) - P(A \cap B) = \\ &= P(A) + P(B) - P(A \cap B). \end{aligned}$$

□

Proposition 1.12.6. *For any $A_1, A_2, .., A_n \in \mathcal{P}(\mathcal{S})$ we have:*

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)\,, \forall n \in \mathbb{N}.$$

Proof. For $n = 2$, taking into account that $P(A_1 \cap A_2) \geq 0$, we have

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq P(A_1) + P(A_2)$$

By mathematical induction, suppose that for any $A_1, A_2, .., A_n$ we have

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$$

and we want to show that:

$$P\left(\bigcup_{i=1}^{n+1} A_i\right) \leq \sum_{i=1}^{n+1} P(A_i)$$

We have:

$$P\left(\bigcup_{i=1}^{n+1} A_i\right) \leq P\left(\bigcup_{i=1}^{n} A_i\right) + P(A_{n+1}) \leq \sum_{i=1}^{n} P(A_i) + P(A_{n+1}) = \sum_{i=1}^{n+1} P(A_i).$$

□

Proposition 1.12.7. *For any $A_1, A_2, .., A_n \in \mathcal{P}(\mathcal{S})$ we have:*

$$P\left(\bigcap_{i=1}^{n} A_i\right) \geq 1 - \sum_{i=1}^{n} P(\complement A_i)\,, \forall n \in \mathbb{N}.$$

Proof.

$$P\left(\bigcap_{i=1}^{n} A_i\right) = 1 - P\left(\complement \bigcap_{i=1}^{n} A_i\right) = 1 - P\left(\bigcup_{i=1}^{n} \complement A_i\right) \geq 1 - \sum_{i=1}^{n} P(\complement A_i).$$

□

Example 1.12.1. If the probabilities of the elementary events are those from Remark 1.12.1, compute the probability that, randomly choosing a teacher, he/she would answer YES to the question concerning:

i) modernizing the school;

ii) necessity of a sports facility;

iii) modernizing the school or necessity of a sports facility.

Solution: i) Choosing a teacher that would answer YES to the question concerning modernizing the school means the event $\{e_1, e_2\}$ and its probability is $P(\{e_1, e_2\}) = 0.6 + 0.2 = 0.8$.
ii) Choosing a teacher that would answer YES to the question concerning the necessity of a sports facility means the event $\{e_1, e_3\}$ and its probability is $P(\{e_1, e_3\}) = 0.6 + 0.15 = 0.75$.
iii) Choosing a teacher that would answer YES to one of the two questions means the event $\{e_1, e_2, e_3\}$ and its probability is $P(\{e_1, e_2, e_3\}) = 0.6 + 0.2 + 0.15 = 0.95$.

1.13 Independent and dependent events.

Definition 1.13.1. *The events A and B from $\mathcal{P}(\mathcal{S})$ are called **independent** if*

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

Theorem 1.13.1. *If $A, B \in \mathcal{P}(\mathcal{S})$ are independent events having non-zero probabilities, then $A \cap B$ is a set which contains at least one point e_i of the sample space $\mathcal{S}$.*

Proof. We have to show that $A \cap B \neq \emptyset$. Considering that $A \cap B = \emptyset$ it results that $P(A \cap B) = 0$ and as $P(A \cap B) = P(A) \cdot P(B)$ it results that $P(A) \cdot P(B) = 0$. Hence, we obtain that either $P(A) = 0$ or $P(B) = 0$, which is absurd, taking into consideration the hypothesis of the theorem. Therefore, $A \cap B \neq \emptyset$. □

Definition 1.13.2. *We say that the events $A_1, A_2, ..., A_n$ are **totally independent**, or independent, if for any $1 \leq i_1 < i_2 < ... < i_s \leq n$, we have:*

$$P(A_{i_1} \cap A_{i_2} \cap ... \cap A_{i_s}) = P(A_{i_1}) \cdot P(A_{i_2}) \cdot ... \cdot P(A_{i_s}).$$

Definition 1.13.3. *We say that the events $A_1, A_2, ..., A_n \in \mathcal{P}(\mathcal{S})$ are k-**independent**, $k \leq n$, if the events of any family of k events are independent as in Definition 1.13.2.*

Remark 1.13.1. The independence of the events $A_1, A_2, ..., A_n$ means that

$$C_n^2 + C_n^3 + ... + C_n^n = 2^n - n - 1$$

relations have to be satisfied.
The independence of three events A_1, A_2, A_3 means that we must have:

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2)$$
$$P(A_1 \cap A_3) = P(A_1) \cdot P(A_3)$$
$$P(A_2 \cap A_3) = P(A_2) \cdot P(A_3)$$
$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2) \cdot P(A_3)$$

Theorem 1.13.2. *If A and B are independent events, then the events A and $\complement B$; $\complement A$ and B; $\complement A$ and $\complement B$ are also independent.*

Proof. We know that $P(A \cap B) = P(A) \cdot P(B)$. We have to prove the following equalities: $P(A \cap \complement B) = P(A) \cdot P(\complement B)$; $P(\complement A \cap B) = P(\complement A) \cdot P(B)$; $P(\complement A \cap \complement B) = P(\complement A) \cdot P(\complement B)$.

In order to obtain $P(A \cap \complement B) = P(A) \cdot P(\complement B)$ we write $A = (A \cap B) \cup (A \cap \complement B)$. We find that:

$$P(A) = P(A \cap B) + P(A \cap \complement B) = P(A) \cdot P(B) + P(A \cap \complement B)$$

or:

$$P(A) \cdot [1 - P(B)] = P(A \cap \complement B).$$

As $1 - P(B) = P(\complement B)$, we obtain that $P(A) \cdot P(\complement B) = P(A \cap \complement B)$.

The other equalities are proved similarly. □

Definition 1.13.4. *We say that the events $B_1, B_2, ..., B_k \in \mathcal{P}(\mathcal{S})$ form a **partition** of the sample space $\mathcal{S}$ if the following conditions are fulfilled:*

i) $B_i \cap B_j = \emptyset$ *for* $i \neq j$;

ii) $\bigcup_{i=1}^{k} B_i = \mathcal{S}$;

iii) $P(B_i) > 0$, $\forall i = 1, 2, ..., k$.

*The events of a partition of the sample space are called **hypotheses**.*

Definition 1.13.5. *Let $A_1, A_2, ..., A_n$, $B_1, B_2, ..., B_k$ two partitions of the sample space $\mathcal{S}$. We say that these partitions are **independent** if*

$$P(A_i \cap B_j) = P(A_i) \cdot P(B_j)$$

for any i, j, $i = 1, 2, ..., n$, $j = 1, 2, ..., k$.

Example 1.13.1. If A is an event of the sample space $\mathcal{S}$, then A and $\mathcal{S}$ are independent.

Solution: $A = A \cap \mathcal{S}$, therefore $P(A) = P(A \cap \mathcal{S}) = P(A) \cdot P(\mathcal{S})$, because $P(\mathcal{S}) = 1$.

Example 1.13.2. We toss two coins. The events $A =$"obtain heads on the first coin" and $B =$"obtain tails on the second coin" are independent.

Solution: A sample space of this experiment is

$$\mathcal{S} = \{e_1 = (H, H), e_2 = (H, T), e_3 = (T, H), e_4 = (T, T)\}.$$

The events A and B are

$$A = \{e_1, e_2\}, \quad B = \{e_2, e_4\}.$$

The elementary events e_1, e_2, e_3, e_4 of the sample space are equally probable and $P(e_i) = \frac{1}{4}$, $i = 1, 2, 3, 4$. Therefore, $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{2}$. The event $A \cap B$ is actually $A \cap B = \{e_2\}$ and its probability is $P(A \cap B) = \frac{1}{4}$. Hence, $P(A \cap B) = P(A) \cdot P(B)$.

Example 1.13.3. When tossing two coins, consider the following events: A_1 ="obtain heads on the first coin", A_2 ="obtain tails on the second coin", A_3 ="obtain heads and tails". The events A_1, A_2, A_3 are not 3-independent.

Solution: A sample space of this experiment is

$$\mathcal{S}=\{e_1=(H,H), e_2=(H,T), e_3=(T,H), e_4=(T,T)\}.$$

The events A_1, A_2, A_3 are

$$A_1=\{e_1,e_2\}, \quad A_2=\{e_2,e_4\}, \quad A_3=\{e_2,e_3\}.$$

We have:

$$A_1\cap A_2=\{e_2\}\Rightarrow P(A_1\cap A_2)=\frac{1}{4}=P(A_1)\cdot P(A_2)$$
$$A_1\cap A_3=\{e_2\}\Rightarrow P(A_1\cap A_3)=\frac{1}{4}=P(A_1)\cdot P(A_3)$$
$$A_2\cap A_3=\{e_2\}\Rightarrow P(A_2\cap A_3)=\frac{1}{4}=P(A_2)\cdot P(A_3)$$
$$A_1\cap A_2\cap A_3=\{e_2\}\Rightarrow P(A_1\cap A_2\cap A_3)=\frac{1}{4}\neq\frac{1}{8}=P(A_1)\cdot P(A_2)\cdot P(A_3).$$

Example 1.13.4. When tossing two coins, consider the following events:

A_1 ="obtain heads on the first coin";

A_2 ="obtain tails on the first coin";

A_3 ="obtain heads on the second coin";

A_4 ="obtain tails on the second coin".

Show that the events A_1, A_2, A_3, A_4 are not totally independent.

Example 1.13.5. If $\mathcal{S}=\{e_1,e_2,e_3,e_4\}$ is the sample space associated to the experiment from Example 1.13.2, then the events $\{e_1\}$, $\{e_2\}$, $\{e_3\}$, $\{e_4\}$ form a partition of the sample space $\mathcal{S}$.

1.14 Conditional probability

We illustrate the meaning of "conditional probability" using the following example:

Example 1.14.1. Consider rolling two dice. Let a be the number which appears on the first die and b the number appearing on the second die. What is the probability that $b=3$, knowing that $a+b>8$?

Solution: The sample space associated to this experiment is the set $\mathcal{S}$ of the following pairs:

$$\begin{array}{cccccc}
(1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6)\\
(2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6)\\
(3,1) & (3,2) & (3,3) & (3,4) & (3,5) & \underline{(3,6)}\\
(4,1) & (4,2) & (4,3) & (4,4) & \underline{(4,5)} & \underline{(4,6)}\\
(5,1) & (5,2) & (5,3) & \underline{(5,4)} & \underline{(5,5)} & \underline{(5,6)}\\
(6,1) & (6,2) & \underline{(6,3)} & \underline{(6,4)} & \underline{(6,5)} & \underline{(6,6)}
\end{array}$$

All these events are equally probable and $P((i,j)) = \frac{1}{36}$, for any $i = \overline{1,6}$, $j = \overline{1,6}$. Looking at all the elementary events of the sample space $\mathcal{S}$, only in the case of the events $(6,3)$, $(5,4)$, $(4,5)$, $(3,6)$, $(6,4)$, $(5,5)$, $(4,6)$, $(6,5)$, $(5,6)$, $(6,6)$ the condition $a+b>8$ is fulfilled. We consider the set $\mathcal{S}'$ formed by these events:

$$\mathcal{S}' = \{(6,3),(5,4),(4,5),(3,6),(6,4),(5,5),(4,6),(6,5),(5,6),(6,6)\}.$$

The set $\mathcal{S}'$ is another sample space associated to this experiment, built up by taking into account that $a+b>8$. The elementary events of the sample space $\mathcal{S}'$ are equally probable and their probability is $\frac{1}{10}$.

We find only one element of the sample space $\mathcal{S}'$ for which $b=3$: $(6,3)$. Therefore, in the sample space $\mathcal{S}'$, the probability of the event $b=3$ is $\frac{1}{10}$. This result will be called the probability of "$b=3$" conditioned by "$a+b>8$".

We can rethink everything in the following manner: first, we compute the probability in the sample space $\mathcal{S}$ of the event $A =$"$a+b>8$". This is $P(A) = \frac{10}{36}$. Then, we determine, in the same sample space $\mathcal{S}$ the probability of the event $B =$"$b=3$". We get $P(B) = \frac{6}{36}$. The probability in $\mathcal{S}$ that both events A and B take place is $P(A \cap B) = P((6,3)) = \frac{1}{36}$.

If we denote by $P(B|A)$ the probability of the event B, conditioned by the occurrence of the event A, we have:

$$P(B|A) = \frac{1}{10}, \quad P(A) = \frac{10}{36}, \quad P(A \cap B) = \frac{1}{36},$$

and therefore

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

Definition 1.14.1. *The **probability of the event** A **conditioned by the occurrence of the event** B is denoted by $P(A|B)$ or $P_B(A)$ and is defined by*

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad dac'a\ P(B) \neq 0.$$

Instead of "probability of the event A conditioned by the occurrence of the event B" we simply say "probability of A, given B".

The reduced sample space is B (the conditioning event).

Remark 1.14.1. The probability which has been introduced by the axiomatic Definition 1.12.1 can be regarded as a conditional probability, conditioned by the certain event, which is a sample space $\mathcal{S}$, such that $P(\mathcal{S}) = 1$.

Proposition 1.14.1. *For a fixed event $B \in \mathcal{P}(\mathcal{S})$ such that $P(B) \neq 0$, for any two events A_1, A_2 from $\mathcal{P}(\mathcal{S})$, we have:*

A1) $0 \leq P(A_1|B) \leq 1$;

A2) $P(\mathcal{S}|B) = 1$;

A3) A_1, A_2 - *incompatible* $\Rightarrow P((A_1 \cup A_2)|B) = P(A_1|B) + P(A_2|B)$.

Proof. From $P(A_1|B) = \dfrac{P(A_1 \cap B)}{P(B)}$ it results that $P(A_1|B) \geq 0$ and from $P(A_1 \cap B) \leq P(B)$ we get that $P(A_1 \cap B) \leq 1$.

$P(\mathcal{S}|B) = \dfrac{P(\mathcal{S} \cap B)}{P(B)} = \dfrac{P(B)}{P(B)} = 1.$

$P((A_1 \cup A_2)|B) = \dfrac{P((A_1 \cup A_2) \cap B)}{P(B)} = \dfrac{P(A_1 \cap B)}{P(B)} + \dfrac{P(A_2 \cap B)}{P(B)} = P(A_1|B) + P(A_2|B).$ □

Theorem 1.14.1. *If A and B are independent events with non-zero probabilities, then:*

$$P(A|B) = P(A) \quad \textit{and} \quad P(B|A) = P(B).$$

Proof. As A and B are independent and $A \cap B = B \cap A$, we have

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

Hence:

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

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

□

Theorem 1.14.2. *If $A_1, A_2, ..., A_n$ are events such that $P(A_1 \cap A_2 \cap ... \cap A_n) \neq 0$ (they can occur simultaneously, i.e. they are compatible), then*

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

Proof.

$$\begin{aligned} &P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|(A_1 \cap A_2)) \cdot ... \cdot P(A_n|(A_1 \cap ... \cap A_{n-1})) = \\ &\quad = P(A_1) \cdot \frac{P(A_1 \cap A_2)}{P(A_1)} \cdot P(A_3|(A_1 \cap A_2)) \cdot ... \cdot P(A_n|(A_1 \cap ... \cap A_{n-1})) = \\ &\quad = P(A_1 \cap A_2) \cdot \frac{P(A_1 \cap A_2 \cap A_3)}{P(A_1 \cap A_2)} \cdot ... \cdot P(A_n|(A_1 \cap ... \cap A_{n-1})) = \\ &\quad \cdots\cdots\cdots \\ &\quad = P(A_1 \cap ... \cap A_{n-1}) \cdot \frac{P(A_1 \cap ... \cap A_{n-1} \cap A_n)}{P(A_1 \cap ... \cap A_{n-1})} = P(A_1 \cap ... \cap A_n). \end{aligned}$$

□

Consequence 1.14.1. *If $A_1, A_2, ..., A_n$ are independent events then:*

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

Example 1.14.2. A box contains 3 white balls and 5 black balls. We extract 2 balls from the box, one after the other. Give a sample space for this experiment and the probabilities of the elementary events of this sample space.

Solution: If w means extracting a white ball and b means extracting a black ball, then a sample space associated to this experiment is:

$$\mathcal{S} = \{(w,w),(w,b),(b,w),(w,w)\}.$$

The event (b,w) means that the first ball is black and the second ball is white. As the balls are randomly extracted from the box, Based on the conditional probability formula, we have:

$$P(w,w) = \frac{3}{8}\cdot\frac{2}{7} = \frac{6}{56},\ P(w,b) = \frac{3}{8}\cdot\frac{5}{7} = \frac{15}{56},\ P(b,w) = \frac{5}{8}\cdot\frac{3}{7} = \frac{15}{56},\ P(b,b) = \frac{5}{8}\cdot\frac{4}{7} = \frac{20}{56}.$$

Theorem 1.14.3 (total probability formula)**.** *If the events $A_1, A_2, ..., A_n$ form a partition of the sample space $\mathcal{S}$ and $X \in \mathcal{P}(\mathcal{S})$, then:*

$$P(X) = \sum_{i=1}^{n} P(A_i)\cdot P(X|A_i).$$

Proof. We can write X as:

$$X = \bigcup_{i=1}^{n}(X \cap A_i).$$

As $(X \cap A_i) \cap (X \cap A_j) = \emptyset$ for $i \neq j$, we obtain:

$$P(X) = \sum_{i=1}^{n} P(X \cap A_i).$$

But $P(X \cap A_i) = P(A_i)\cdot P(X|A_i)$, and by replacement, we obtain the requested equality. □

Example 1.14.3. Three boxes have the following structure: the box i contains a_i white balls and b_i black balls, $i = 1,2,3$. The event A_i consists of choosing the box i. It is know that $P(A_i) = p_i$ and $p_1 + p_2 + p_3 = 1$. A box is randomly chosen and a ball is extracted. What is the probability that the extracted ball is black?

Solution: Let X be the event "the extracted ball is black". The probability that a black ball is extracted, given that the the box i has been chosen is:

$$P(X|A_i) = \frac{b_i}{a_i + b_i}.$$

Therefore, based on the total probability formula, the probability that a black ball is extracted is:

$$\begin{aligned} P(X) &= P(A_1)\cdot P(X|A_1) + P(A_2)\cdot P(X|A_2) + P(A_3)\cdot P(X|A_3) = \\ &= p_1\frac{b_1}{a_1+b_1} + p_2\frac{b_2}{a_2+b_2} + p_3\frac{b_3}{a_3+b_3}. \end{aligned}$$

Theorem 1.14.4 (Bayes' formula)**.** *If the events $A_1, A_2, ..., A_n$ form a partition of the sample space $\mathcal{S}$ and are the cause of the occurrence of an event X, then:*

$$P(A_k|X) = \frac{P(A_k)\cdot P(X|A_k)}{\sum_{i=1}^{n} P(A_i)\cdot P(X|A_i)}.$$

Proof. takes into account the equalities

$$P(A_i) \cdot P(X|A_i) = P(X) \cdot P(A_i|X)$$

and the total probability formula. □

Definition 1.14.2. *The probabilities $P(A_i)$, $P(X|A_i)$, $i = \overline{1,n}$ are called **prior probabilities** and $P(A_i|X)$ are called **posterior probabilities**. The event X is called **evidence**.*

Before we receive the evidence, we have a set of prior probabilities $P(A_i)$, $i = \overline{1,n}$ for the hypotheses. If we know the correct hypothesis, we know the probability for the evidence. That is, we know $P(X|A_i)$, $i = \overline{1,n}$. If we want to find the probabilities for the hypothesis, given the evidence, that is, we want to find $P(A_i|X)$, we can use the Bayes' formula.

Example 1.14.4. Consider two boxes. The first one contains 2 white balls and 3 black balls, and the second contains 7 white balls and 5 black balls. The event A_1 means choosing the first box, and the event A_2 means choosing the second box. It is known that the probability of the event A_1 is $P(A_1) = 0.4$, and the probability of the event A_2 is $P(A_2) = 0.6$. We randomly choose one box and a black ball. What is the probability that this black ball is chosen from the second box?

Solution: Let X be the event "a black ball has been extracted". By Bayes' formula, we have:

$$P(A_1|X) = \frac{P(A_1) \cdot P(X|A_1)}{P(A_1) \cdot P(X|A_1) + P(A_2) \cdot P(X|A_2)} = \frac{0.4 \cdot \frac{3}{5}}{0.4 \cdot \frac{3}{5} + 0.6 \cdot \frac{5}{12}} \approx 0.49;$$

$$P(A_2|X) = \frac{P(A_2) \cdot P(X|A_2)}{P(A_1) \cdot P(X|A_1) + P(A_2) \cdot P(X|A_2)} = \frac{0.6 \cdot \frac{5}{12}}{0.4 \cdot \frac{3}{5} + 0.6 \cdot \frac{5}{12}} \approx 0.51.$$

Bayes's formula is particularly appropriate for medical diagnosis, as it can be seen from the following example:

Example 1.14.5. A doctor gives a patient a test for a particular cancer. Before the results of the test, the only evidence the doctor has to go on, is that 1 man in 1000 has cancer. Experience has shown that, in 99 percent of the cases in which cancer is present, the test is positive; and in 95 percent of the cases in which it is not present, it is negative. If the test turns out to be positive, what probability should the doctor assign to the event that cancer is present?

Solution: Let's denote by C the event "the patient has cancer". For our experiment, C and $\overline{C}$ are hypotheses. We will also denote by $+$ the event "the test is positive" and by $-$ the event "the test is negative", which are evidences. We are given the prior probabilities $P(C) = 0.001$, $P(+|C) = 0.99$ and $P(-|\overline{C}) = 0.95$. We can easily compute $P(\overline{C}) = 0.999$ and $P(+|\overline{C}) = 0.05$. Using Bayes' formula, we compute:

$$P(C|+) = \frac{P(C) \cdot P(+|C)}{P(C) \cdot P(+|C) + P(\overline{C}) \cdot P(+|\overline{C})} = \frac{0.001 \cdot 0.99}{0.001 \cdot 0.99 + 0.999 \cdot 0.05} \approx 0.019$$

Hence, we find that among positive results, only 1.9 percent are cancers, and 98.1 percent are false positives. When a group of medical students was asked this question, almost all students incorrectly guessed the probability to be larger then 50 percent.

1.15 One-dimensional discrete random variables

In order to introduce the term discrete random variable and its distribution, we consider the following example:

Example 1.15.1. We extract 3 balls from a box containing an equal number of white and black balls, and we put the ball back into the box after each extraction. How many white balls can occur and what are the associated probabilities?

Solution: We will give the answer to this question indicating the possible outcomes and the associated probabilities.

Sample space	Nr. of white balls	Probability
WWW	3	$\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{8}$
WWB	2	$\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{8}$
WBW	2	$\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{8}$
BWW	2	$\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{8}$
WBB	1	$\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{8}$
BWB	1	$\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{8}$
BBW	1	$\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{8}$
BBB	0	$\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{8}$

The information concerning the number of white balls and their probabilities, is given in the following table:

Nr. of white balls	0	1	2	3
Probability	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

If the variable X represents the number of white balls that can occur, then the table shows the values X can take and the probabilities to take these values.

The set of ordered pairs, each of the form (number of white balls, corresponding probability), defines the distribution of the random variable X. Because the values of X are determined by events resulting from a random experiment, we call X a random variable.

The function f defined by $f(x)=P(X=x)$, is called frequency or probability function.

In our case

$$f(0)=f(X=0)=\frac{1}{8};\ f(1)=f(X=1)=\frac{3}{8};\ f(2)=f(X=2)=\frac{3}{8};\ f(3)=f(X=3)=\frac{1}{8}.$$

We observe that

$$f(x)=P(X=x)=C_3^x\cdot\left(\frac{1}{2}\right)^3,\quad x=0,1,2,3;$$

$$f(x)\geq 0\quad\text{and}\quad\sum_{i=0}^{3}f(x)=\sum_{i=0}^{3}C_3^x\cdot\left(\frac{1}{2}\right)^3=2^3\cdot\left(\frac{1}{2}\right)^3=1.$$

Definition 1.15.1. *A variable whose value is a number determined by the elementary event resulting from an experience, is called* ***random variable****.*

Definition 1.15.2. *If X is a random variable which can take the values x_1, x_2, ..., x_n with the probabilities $f(x_1)$, $f(x_2)$, ..., $f(x_n)$ then the set of ordered pairs $(x_i,f(x_i))$, $i=\overline{1,n}$ is called* ***the probability distribution of the random variable*** *X.*

In our previous example, the distribution is: $(0, \frac{1}{8})$, $(1, \frac{3}{8})$, $(2, \frac{3}{8})$, $(3, \frac{1}{8})$ or:

$$X : \begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{pmatrix}.$$

Example 1.15.2. Three balls, a, b, c, are randomly distributed in three boxes. Determine the distribution of the random variable $X =$"the number of non-empty boxes".

Solution:

$$X : \begin{pmatrix} 1 & 2 & 3 \\ \frac{3}{27} & \frac{18}{27} & \frac{6}{27} \end{pmatrix}.$$

Remark 1.15.1. In the Kolmogorov approach, the random variable X is a function defined on a sample space corresponding to the experience, that is a "point function". So, if we consider, in the previous example, the following events:

$$\begin{array}{lll} e_1 = \{abc|0|0\} & e_{10} = \{c|ab|0\} & e_{19} = \{0|b|ac\} \\ e_2 = \{0|abc|0\} & e_{11} = \{0|ab|c\} & e_{20} = \{a|0|bc\} \\ e_3 = \{0|0|abc\} & e_{12} = \{b|ac|0\} & e_{21} = \{0|a|bc\} \\ e_4 = \{ab|c|0\} & e_{13} = \{0|ac|b\} & e_{22} = \{a|b|c\} \\ e_5 = \{ab|0|c\} & e_{14} = \{a|bc|0\} & e_{23} = \{a|c|b\} \\ e_6 = \{ac|b|0\} & e_{15} = \{0|bc|a\} & e_{24} = \{b|c|a\} \\ e_7 = \{ac|0|b\} & e_{16} = \{c|0|ab\} & e_{25} = \{b|a|c\} \\ e_8 = \{bc|a|0\} & e_{17} = \{0|c|ab\} & e_{26} = \{c|a|b\} \\ e_9 = \{bc|0|a\} & e_{18} = \{b|0|ac\} & e_{27} = \{c|b|a\} \end{array}$$

and the sample space $\mathcal{S} = \{e_1, ..., e_{27}\}$, we can consider the function $X : \mathcal{S} \to \{1, 2, 3\}$, defined as: $X(e_k) =$ the number of non-empty boxes if the event e_k takes place. It is clear that $X(e_k) = 1$ if $k = 1, 2, 3$; $X(e_k) = 2$ if $k = 4, 5, ...21$; and $X(e_k) = 3$ if $k = 22, 23, 24, 25, 26, 27$.

If the random variable is given in this manner, then the values of the variable are known for each event $e_k \in \mathcal{S}$ and $P(e_k)$. From here one can determine the possible values of the variable and the probabilities corresponding to these values. In this way one can obtain the distribution of the random variable X.

Example 1.15.3. For the experience from the Example 1.15.2, let us denote Y the random variable whose values are the number of balls from the first box.

The random variable Y can take the values 0, 1, 2 or 3, because we can have 0, 1, 2 or 3 balls in the first box.

Y takes the value 0 if one of the following events from the sample space S takes place: e_2, e_3, e_{11}, e_{13}, e_{15}, e_{17}, e_{19}, e_{21}, that is

$$(Y = 0) = (e_2 \text{ or } e_3 \text{ or } e_{11} \text{ or } e_{13} \text{ or } e_{15} \text{ or } e_{17} \text{ or } e_{19} \text{ or } e_{21}).$$

It follows that

$$\begin{aligned} P(Y = 0) =& P(\{e_2\}) + P(\{e_3\}) + P(\{e_{11}\}) + P(\{e_{13}\}) + P(\{e_{15}\}) + P(\{e_{17}\}) + \\ &+ P(\{e_{19}\}) + P(\{e_{21}\}) = \frac{8}{27}. \end{aligned}$$

Y takes the value 1 if one of the events: e_{10}, e_{12}, e_{14}, e_{16}, e_{18}, e_{20}, $e_{22} - e_{27}$ take place; the value 2 , for one of the events $e_4 - e_9$ and the value 3 if the event e_1 takes place.

So the distribution of the random variable Y is:

$$Y : \begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{8}{27} & \frac{12}{27} & \frac{6}{27} & \frac{1}{27} \end{pmatrix}.$$

Remark 1.15.2. Onicescu considered the random variable as "an event function". A possible value for the random variable can correspond to an elementary event from the sample space (the value 3 of the random variable Y from the Example 1.15.3, corresponds to the elementary event e_1) or to a subset of elementary events which determine an event (foe example, the event $Y = 2$ is determined by a set of elementary events, $e_4 - e_9$).

Remark 1.15.3. A random variable which takes the distinct values x_1, x_2, ..., x_n determines a partition A_1, A_2, ..., A_n of the sample space $\mathcal{S}$. The event A_i is defined by $e_k \in A_i \Leftrightarrow X(e_k) = x_i$.

In the Example 1.15.3 the random variable Y determines the following partition of the sample space $\mathcal{S}$:

A_1: "There is no ball in the first box."

A_2: "There is a ball in the first box."

A_3: "There are two balls in the first box."

A_4: "There are three balls in the first box."

We have:

$$\begin{aligned} &A_1 \cup A_2 \cup A_3 \cup A_4 = \mathcal{S} \\ &A_i \cap A_j = \emptyset \quad \text{for} \quad i \neq j \\ &P(A_1) = P(Y = 0) = \frac{8}{27}, \quad P(A_2) = P(Y = 1) = \frac{12}{27}, \\ &P(A_3) = P(Y = 2) = \frac{6}{27}, \quad P(A_4) = P(Y = 3) = \frac{1}{27}, \end{aligned}$$

which proves the statement we made.

Definition 1.15.3. *A random variable with a countable set of possible values is called* ***discrete random variable****.*

Definition 1.15.4. *We will say that the random variable X is* ***symmetric*** *with respect to the point c if the following conditions are satisfied:*

i) if $c + a$ is a value of the random variable X, then $c - a$ is also a value of the random variable X;

ii) $P(X = c + a) = P(X = c - a)$.

The condition ii) can be rewritten as

$$P(X - c = a) = P(c - X = a)$$

which shows that "X is symmetric with respect to the point c".

$X - c$ and $c - X$ have the same distribution.

In particular, the symmetry with respect to zero shows that X and $-X$ have the same distribution.

Example 1.15.4. If $P(X = i) = \frac{1}{n}$, $i = 1, 2, ..., n$ then X is symmetrically distributed with respect to $\frac{n+1}{2}$, which is the middle point of the two extreme possible values: 1 and n.

Remark 1.15.4. If the random variables X, Y are seen as functions defined on the sample space $\mathcal{S}$, then we can define the sum $X + Y$, the product $X \cdot Y$, and the scalar product $k \cdot X$ of the random variables. The meaning of these operations is that of the corresponding function operations.

Also, if k is a real function defined on the value set of the variable X, $k : X(\mathcal{S}) \to \mathbb{R}^1$, then we can make the composition $k \circ X$ and we still obtain a random variable, whose values determine the set $k(\mathcal{S}(X))$.

1.16 The distribution function of a discrete one-dimensional random variable

Some questions:

- If we throw two dice, what is the probability that the sum of the obtained numbers is less than 7?
- If three balls are randomly distributed in three boxes, what is the probability that two or less than two of these boxes shall be non-empty?
- If three balls are randomly distributed in three boxes, what is the probability that there shall be at most two balls in the first box?

Generally: What is the probability that a random variable X shall take values less than a given number?

The need of answering these kind of questions has led to the following definition:

Definition 1.16.1. *Let X be a random variable and x a real number. The function F defined as: "$F(x)$ is the probability that X shall take values less than x", or*

$$F(x) = P(X < x)$$

*is called **the distribution function** (or **probability density function**) of the random variable X.*

Proposition 1.16.1. *If X is a discrete random variable having the distribution*

$$X : \begin{pmatrix} x_1 & x_2 & ... & x_n \\ f(x_1) & f(x_2) & ... & f(x_n) \end{pmatrix}$$

then

$$F(x) = \sum_{x_i < x} f(x_i)$$

that is, the value of the distribution function in x is given by the sum of the left-hand-side (with respect to x) probability values.

Proof. Immediate. □

Proposition 1.16.2. *The following equalities take place:*

i) $\lim\limits_{\substack{x\to x_i\\ x>x_i}} F(x) = F(x_i+0) = \sum\limits_{j=1}^{i} f(x_j);$

ii) $\lim\limits_{\substack{x\to x_i\\ x<x_i}} F(x) = F(x_i-0) = \sum\limits_{j=1}^{i-1} f(x_j) = F(x_i).$

Proof. i) For $x \in (x_i, x_{i+1})$ we have:

$$F(x) = \sum_{j=1}^{i} f(x_j).$$

It follows that

$$F(x_i+0) = \sum_{j=1}^{i} f(x_j).$$

ii) For $x \in (x_{i-1}, x_i)$ we have:

$$F(x) = \sum_{j=1}^{i-1} f(x_j).$$

It follows that

$$F(x_i-0) = \sum_{j=1}^{i-1} f(x_j) = F(x_i).$$

□

Proposition 1.16.3. *The following inequalities take place:*

i) $0 \le F(x) \le 1$, $\forall x \in \mathbb{R}^1$;

ii) $x < y \Rightarrow F(x) \le F(y)$.

Proof. i) Because $F(x) = P(X < x)$ and $P(X < x) \in [0,1]$, we have that $F(x) \in [0,1]$.

ii) $x < y$. If $x_i < x$, then $x_i < y$, and so

$$\sum_{x_i<y} f(x_i) = \sum_{x_i<x} f(x_i) + \sum_{x\le x_i<y} f(x_i),$$

that is $F(x) \le F(y)$. □

Proposition 1.16.4. *If $x < y$, then $F(y) - F(x) = P(x \le X < y)$.*

Proof. If $x < y$, we have:

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

and from here it follows that $F(y) - F(x) = P(x \le X < y)$. □

Remark 1.16.1. If X is a discrete random variable, then the distribution function of the random variable X is a left-continuous step function. We have a discontinuity (a leap) in every point x which is a value for the random variable $X(x = x_i)$, and the hight of the leap is $f(x_i)$.

Definition 1.16.2. *We call* α***-quantile*** *the value* x_α *with the property*

$$F(x_\alpha) = P(X < x_\alpha) = \alpha.$$

If X is a discrete random variable, it is not sure that for each $\alpha \in [0, 1]$ there is an α-quantile. But if there exists an α-quantile, then we have an infinity (the interval that separates two possible values).

The 1/2-quantile is called **median** and we denote it Me; so, $F(Me) = 1/2$.

The 1/4- and 3/4-quantiles are called the **lower quantile**, Q_1, and the **upper quantile**, Q_2; so $F(Q_1) = 1/4$ and $F(Q_2) = 3/4$.

Definition 1.16.3. *We call* ***modulus*** *the value* x_i *such that* $f(x_i)$ *is maximal.*

A random variable may have more then one modulus. When throwing a die, the six faces occur with equal probabilities; in this case, all values of the random variable X="the obtained number" are modula.

Example 1.16.1. Let's consider the experiment consisting of distributing three balls a, b, c in three boxes, taking into consideration the random variable X from Example 1.15.2) and Y from Example 1.15.3.

$$X : \begin{pmatrix} 1 & 2 & 3 \\ \frac{3}{27} & \frac{18}{27} & \frac{6}{27} \end{pmatrix} \quad \text{and} \quad Y : \begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{8}{27} & \frac{12}{27} & \frac{6}{27} & \frac{1}{27} \end{pmatrix}.$$

We obtain the distribution functions:

$$F(x) = \begin{cases} 0 & , x \leq 1 \\ \frac{3}{27} & , 1 < x \leq 2 \\ \frac{21}{27} & , 2 < x \leq 3 \\ \frac{27}{27} = 1 & , 3 < x \end{cases} \quad \text{and} \quad F(y) = \begin{cases} 0 & , y \leq 0 \\ \frac{8}{27} & , 0 < y \leq 1 \\ \frac{20}{27} & , 1 < y \leq 2 \\ \frac{26}{27} & , 2 < y \leq 3 \\ \frac{27}{27} = 1 & , 3 < y \end{cases}.$$

The random variables X and Y do not have medians and lower/upper quantiles. The modulus of X is 2, and the modulus of Y is 1.

1.17 Two-dimensional discrete random variables (random vectors)

Sometimes, it is necessary to consider two or more random variables defined on the same sample space, at the same time. In the followings, we will present the case of two random variables; the generalization to three or more variables can be achieved with no difficulty.

Example 1.17.1. We consider the experiment consisting of distributing three balls a, b, c in three boxes.

The following sample space is considered: $\mathcal{S} = \{e_1, e_2, ..., e_{27}\}$, where e_i are given by:

$$\begin{array}{lll} e_1 = \{abc|0|0\} & e_{10} = \{c|ab|0\} & e_{19} = \{0|b|ac\} \\ e_2 = \{0|abc|0\} & e_{11} = \{0|ab|c\} & e_{20} = \{a|0|bc\} \\ e_3 = \{0|0|abc\} & e_{12} = \{b|ac|0\} & e_{21} = \{0|a|bc\} \\ e_4 = \{ab|c|0\} & e_{13} = \{0|ac|b\} & e_{22} = \{a|b|c\} \\ e_5 = \{ab|0|c\} & e_{14} = \{a|bc|0\} & e_{23} = \{a|c|b\} \\ e_6 = \{ac|b|0\} & e_{15} = \{0|bc|a\} & e_{24} = \{b|c|a\} \\ e_7 = \{ac|0|b\} & e_{16} = \{c|0|ab\} & e_{25} = \{b|a|c\} \\ e_8 = \{bc|a|0\} & e_{17} = \{0|c|ab\} & e_{26} = \{c|a|b\} \\ e_9 = \{bc|0|a\} & e_{18} = \{b|0|ac\} & e_{27} = \{c|b|a\}. \end{array}$$

The 27 events are equally probable and their probabilities are $\frac{1}{27}$.

Let X be the random variable which associates to the elementary event $e_i \in \mathcal{S}$ the number of non-empty boxes. We have $X(e_i) = 1$ for $i = 1, 2, 3$, $X(e_i) = 2$ for $i = \overline{4, 21}$, $X(e_i) = 3$ for $i = \overline{22, 27}$. Hence: $P(X = 1) = \frac{3}{27}$, $P(X = 2) = \frac{18}{27}$, $P(X = 3) = \frac{6}{27}$ and the probability distribution of the random variable X is:

$$X : \begin{pmatrix} 1 & 2 & 3 \\ \frac{3}{27} & \frac{18}{27} & \frac{6}{27} \end{pmatrix}.$$

Let Y be the random variable associating to the elementary event $e_i \in \mathcal{S}$ the number of balls from the first box. We have: $Y(e_1) = 3$, $Y(e_i) = 2$, for $i = 4 - 9$, $Y(e_i) = 1$ for $i = 10, 12, 14, 16, 18, 20, 22 - 27$, $Y(e_i) = 0$ for$i = 2, 3, 11, 13, 15, 17, 19, 21$. Hence $P(Y = 0) = \frac{8}{27}$, $P(Y = 1) = \frac{12}{27}$, $P(Y = 2) = \frac{6}{27}$, $P(Y = 3) = \frac{1}{27}$ and the probability distribution of the random variable Y is:

$$Y : \begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{8}{27} & \frac{12}{27} & \frac{6}{27} & \frac{1}{27} \end{pmatrix}.$$

We now consider the random variable Z which associates to the elementary event $e_i \in \mathcal{S}$ the pair of numbers (number of non-empty boxes, number of balls from the first box). As the values of Z are two-dimensional vectors, the random variable Z is called two-dimensional random variable. We have: $Z(e_1) = (1, 3)$; $Z(e_2) = (1, 0)$; $Z(e_3) = (1, 0)$; $Z(e_i) = (2, 2)$, $i = \overline{4, 9}$; $Z(e_i) = (2, 1)$, $i = 10, 12, 14, 16, 18, 20$; $Z(e_i) = (2, 0)$, $i = 11, 13, 15, 17, 19, 21$; $Z(e_i) = (3, 1)$, $i = \overline{22, 27}$.

Therefore, the values of this random variable are the vectors $(1, 3)$; $(1, 0)$; $(2, 2)$; $(2, 1)$; $(2, 0)$; $(3, 1)$. The corresponding probabilities are:

$$P(X = 1, Y = 3) = \frac{1}{27}; \qquad P(X = 1, Y = 0) = \frac{2}{27}; \qquad P(X = 2, Y = 2) = \frac{6}{27};$$

$$P(X = 2, Y = 1) = \frac{6}{27}; \qquad P(X = 2, Y = 0) = \frac{6}{27}; \qquad P(X = 3, Y = 1) = \frac{6}{27}.$$

The probability distribution of Z is

$$Z : \begin{pmatrix} (1,3) & (1,0) & (2,2) & (2,1) & (2,0) & (3,1) \\ \frac{1}{27} & \frac{2}{27} & \frac{6}{27} & \frac{6}{27} & \frac{6}{27} & \frac{6}{27} \end{pmatrix}.$$

In general, let's consider two random variables X, Y defined on the same sample space $\mathcal{S} = \{e_1, e_2, ..., e_n\}$. Let x_1, x_2,..., x_k be the values of the random variable X and y_1, y_2, ..., y_l the values of the random variable Y.

Definition 1.17.1. *Using the random variables X, Y we can build up the **two-dimensional random vector** $Z = (X, Y)$, whose values are the ordered pairs (x_i, y_j) (two dimensional vectors), and the corresponding probabilities are*

$$r_{ij} = P(X = x_i \text{ and } Y = y_j) , \quad 1 \leq i \leq k, \ 1 \leq j \leq l.$$

The probability distribution of Z is given by the following table:

$X \diagdown Y$	y_1	y_2	y_3	...	y_j	...	y_l	$P(X = x_i)$
x_1	r_{11}	r_{12}	r_{13}	...	r_{1j}	...	r_{1l}	p_1
x_2	r_{21}	r_{22}	r_{23}	...	r_{2j}	...	r_{2l}	p_2
x_3	r_{31}	r_{32}	r_{33}	...	r_{3j}	...	r_{3l}	p_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	r_{i1}	r_{i2}	r_{i3}	...	r_{ij}	...	r_{il}	p_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_k	r_{k1}	r_{k2}	r_{k3}	...	r_{kj}	...	r_{kl}	p_k
$P(Y = y_j)$	q_1	q_2	q_3	...	q_j	...	q_k	1

As the events $(X = x_i; Y = y_j)$ form a partition of the sample space , the sum of the probabilities from this table must be equal to 1:

$$\sum_{i=1}^{k} \sum_{j=1}^{l} r_{ij} = 1.$$

If we know the probability distribution of the random vector $Z = (X, Y)$, we can easily find the probability distribution of each component X and Y. Sice the events $(X = x_i, Y = y_1)$, $(X = x_i, Y = y_2)$, ..., $(X = x_i, Y = y_l)$, $1 \leq i \leq k$ are incompatible two by two, and

$$(X = x_i) = (X = x_i, Y = y_1) \cup (X = x_i, Y = y_2) \cup ... \cup (X = x_i, Y = y_l),$$

we have:

$$p_i = P(X = x_i) = r_{i1} + r_{i2} + ... + r_{ik} = \sum_{j=1}^{l} r_{ij} , \quad 1 \leq i \leq k.$$

Similarly, we obtain:

$$q_j = P(Y = y_j) = r_{1j} + r_{2j} + ... + r_{kj} = \sum_{i=1}^{k} r_{ij} , \quad 1 \leq j \leq l.$$

It follows that in order to obtain the probability that X (Y) takes the value x_i (y_j), we will sum up the probabilities from the line (column) of x_i (y_j).

Hence, the random variable X (Y) is associated with the probability distribution given by the marginal column (line) of the previous table. For this reason, the first column (line) together with the last column (line) of the table form the **marginal probability distribution of the random variable X (Y)**.

Definition 1.17.2. ***The variable X, given $Y = y_j$** has the probability distribution:*

$$\begin{pmatrix} x_1 & x_2 & ... & x_i & ... & x_k \\ P(x_1|y_j) & P(x_2|y_j) & ... & P(x_i|y_j) & ... & P(x_k|y_j) \end{pmatrix}$$

where

$$P(x_i|y_j) = P(X = x_i|Y = y_j) = \frac{P(X = x_i,\, Y = y_j)}{P(Y = y_j)} = \frac{r_{ij}}{q_j}\,, \quad 1 \leq j \leq l.$$

Similarly, ***the variable*** Y***, given*** $X = x_i$ *has the probability ditribution:*

$$\begin{pmatrix} y_1 & y_2 & ... & y_j & ... & y_l \\ P(y_1|x_i) & P(y_2|x_i) & ... & P(y_j|x_i) & ... & P(y_l|x_i) \end{pmatrix}$$

where

$$P(y_j|x_i) = P(Y = y_j|X = x_i) = \frac{P(Y = y_j,\, X = x_i)}{P(X = x_i)} = \frac{r_{ij}}{p_i}\,, \quad 1 \leq i \leq k.$$

We have:

$$\sum_{i=1}^{k} P(x_i|y_j) = \frac{1}{q_j} \sum_{i=1}^{k} r_{ij} = 1$$

and

$$\sum_{j=1}^{l} P(y_j|x_i) = \frac{1}{p_i} \sum_{j=1}^{l} r_{ij} = 1.$$

1.18 The distribution function of a random vector

Definition 1.18.1. *We call* ***distribution function*** *of the random vector* (X, Y) *the function defined by*

$$F(x, y) = P(X < x \text{ and } Y < y) = \sum_{x_i<x} \sum_{y_j<y} P(X = x_i \text{ and } Y = y_j) = \sum_{x_i<x} \sum_{y_j<y} r_{ij}$$

where $r_{ij} = P(X = x_i \text{ and } Y = y_j)$.

Proposition 1.18.1. *The distribution function of the random vector* (X, Y) *satisfies the following properties:*

i) $F(x_i + 0, y_j + 0) = \sum_{m=1}^{i} \sum_{s=1}^{j} p_{ms}\,, \qquad F(x_i + 0, y_j - 0) = \sum_{m=1}^{i} \sum_{s=1}^{j-1} p_{ms}$

$F(x_i - 0, y_j + 0) = \sum_{m=1}^{i-1} \sum_{s=1}^{j} p_{ms}\,, \qquad F(x_i - 0, y_j - 0) = \sum_{m=1}^{i-1} \sum_{s=1}^{j-1} p_{ms}.$

ii) $F(x_2, y) \geq F(x_1, y)$ *if* $x_2 > x_1$,
$F(x, y_2) \geq F(x, y_1)$ *if* $y_2 > y_1$.

iii) $F(x, -\infty) = F(-\infty, y) = 0$ *and* $F(\infty, \infty) = 1$.

iv) $F(x, \infty)$ *is the distribution function of the random variable* X,
$F(\infty, y)$ *is the distribution function of the random variable* Y.

v) As

$$P(Y < y|X = x_i) = \frac{P(Y < y,\ X = x_i)}{P(X = x_i)} = \frac{F(x_i + 0, y) - F(x_i - 0, y)}{F(x_i + 0, \infty) - F(x_i - 0, \infty)},$$

the distribution function of the random variable $Y|X = x_i$ *is:*

$$\frac{F(x_i + 0, y) - F(x_i - 0, y)}{F(x_i + 0, \infty) - F(x_i - 0, \infty)}.$$

Proof. using Definition 1.18.1. □

Definition 1.18.2. *We say that the random variables X and Y are **independent** if for every pair (i, j) we have*

$$r_{ij} = p_i \cdot q_j.$$

Proposition 1.18.2. *If the random variables X, Y are independent, then:*

1. *the conditional distributions are the same as the marginal distributions:*

$$P(x_i|y_j) = \frac{r_{ij}}{q_j} = \frac{p_i \cdot q_j}{q_j} = p_i,$$

$$P(y_j|x_i) = \frac{r_{ij}}{p_i} = \frac{p_i \cdot q_j}{p_i} = q_j.$$

2. $F(x,y) = \sum\limits_{x_i<x} \sum\limits_{y_j<y} r_{ij} = \sum\limits_{x_i<x} \sum\limits_{y_j<y} p_i \cdot q_j = \left(\sum\limits_{x_i<x} p_i\right) \cdot \left(\sum\limits_{y_i<y} q_j\right) = F(x,\infty) \cdot F(y,\infty).$

1.19 Expected value. Variance. Moments. (for discrete one-dimensional random variables)

Definition 1.19.1. *The **expected value** (or **mean**) of a random variable X with the distribution:*

$$X : \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix},$$

is the number

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

Proposition 1.19.1. *The expected value of a random variable has the following properties:*

1. *The expected value of a constant random variable is equal to the constant itself:*

$$E(a) = a.$$

2. *The expected value of the product of a constant a and a random variable X, is equal to the product of a and the expected value of X:*

$$E(a \cdot X) = a \cdot E(X).$$

3. *The expected value of the sum of two random variables X and Y is equal to the sum of their expected values:*

$$E(X + Y) = E(X) + E(Y).$$

4. *The expected value of the product of two **independent** random variables is equal to the product of the expected values of the two random variables:*

$$E(X \cdot Y) = E(X) \cdot E(Y).$$

5. *The expected value of the random variable X satisfies:*

$$\inf X \leq E(X) \leq \sup X.$$

6. *The expected value of the deviation from the mean of the random variable, is equal to zero:*

$$E(X - E(X)) = 0.$$

Proof. The proof of these properties is based on the definition and it is left to the reader. □

Example 1.19.1. We roll a die. The random variable X is "the number of dots that appear". The expected value of this random variable is

$$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2} = 3.5.$$

We roll k dies. Let Y be the total number of dots that appear. The expected value of this random variable is $E(Y) = \dfrac{7k}{2}$.

Example 1.19.2. We toss a coin three times. Let X (Y) be the random variable which gives the number of heads from the first (last) two tosses. Show that:

1. $E(X) = E(Y) = 1$;
2. $E(X \cdot Y) = \dfrac{5}{4}$;
3. $E(\frac{1}{1+Y}) = \dfrac{7}{12}$;
4. $\dfrac{X}{1+Y}$ has the distribution

$$\frac{X}{1+Y} : \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{2} & \frac{2}{3} & 1 \\ \\ \frac{2}{8} & \frac{1}{8} & \frac{2}{8} & \frac{1}{8} & \frac{2}{8} \end{pmatrix}.$$

Example 1.19.3. We roll a die two times. Let X (Y) be the random variable which gives the number from the first (second) rolling. Show that:

1. $E(X \cdot Y) = E(X) \cdot E(Y)$;
2. $E(\frac{X}{Y}) = E(X) \cdot E(\frac{1}{Y})$.

Definition 1.19.2. *By definition, the **variance** of the random variable X is the expected value of the square of the deviation $X - E(X)$:*

$$V(X) = E\left[(X - E(X))^2\right].$$

The variance is usually denoted $V(X)$ or $\sigma^2(X)$.

Proposition 1.19.2. *The variance of a random variable has the following properties:*

1. *The variance of a constant random variable is zero:*

$$V(a) = 0.$$

2. *The variance of the product of a constant and a random variable X is equal to the product of the square of the constant and the variance of the variable X:*

$$V(a \cdot X) = a^2 \cdot V(X).$$

3. *The variance of the sum of two independent random variables is equal to the sum of the variances of the two random variables:*

$$V(X + Y) = V(X) + V(Y).$$

4. *For all $L > 0$ we have the following inequality:*

$$P\left[|X - E(X)| < L\right] \geq 1 - \frac{V(X)}{L^2} \quad \textit{(Chebyshev's inequality).}$$

Proof. The proof of the properties 1-3 is left to the reader.

We only prove 4.

We consider

$$X : \begin{pmatrix} x_1 & x_2 & ... & x_n \\ p_1 & p_2 & ... & p_n \end{pmatrix}$$

and

$$\sigma^2 = \sum_{i=1}^{n} p_i \cdot (x_i - E(X))^2 = \sum_{i=1}^{n} p_i \, \alpha_i^2 \quad \text{where } \alpha_i = x_i - E(X).$$

We can admit that α_i^2 are ordered:

$$\alpha_1^2 \leq \alpha_2^2 \leq ... \leq \alpha_n^2.$$

We interpolate L between two values α_i^2, α_{i+1}^2, that is:

$$\alpha_1^2 \leq \alpha_2^2 \leq ... \leq \alpha_i^2 \leq L \leq \alpha_{i+1}^2 \leq ... \leq \alpha_n^2.$$

If in the expression of σ^2 we put zero for all $\alpha_j^2 \leq L$ and replace with L all $\alpha_j^2 \geq L$, we obtain:

$$\sigma^2 \geq L\left[p_{i+1} + ... + p_n\right]$$

or

$$p_{i+1} + ... + p_n \leq \frac{\sigma^2}{L}.$$

The sum $p_{i+1} + ... + p_n$ represents the probability that the deviation is greater or equal to L and so

$$P[|X - E(X)| \geq L] \leq \frac{\sigma^2}{L}.$$

From

$$P[|X - E(X)| \geq L] + P[|X - E(X)| < L] = 1$$

we have the inequality in 4. □

Remark 1.19.1. If $L = k \cdot \sigma$ then:

$$P\left[|X - E(X)| < k \cdot \sigma\right] \geq 1 - \frac{1}{k^2}.$$

If $k = 3$ we have

$$P\left[|X - E(X)| < 3\sigma\right] \geq \frac{8}{9}.$$

This means that 89% ($\frac{8}{9}$) of the absolute deviations of the variable X do not exceed 3σ (the "three sigma" rule).

So we observe that the variance works as a concentration indicator of the deviations around the expected value.

Problem 1.19.1. Determine the variances for the random variables in the Examples 1.19.1, 1.19.2, 1.19.3.

Definition 1.19.3. *By definition, the* ***standard deviation*** *of a random variable X is the square root of the variance of this variable; we will denote it $D(X)$ or μ or σ:*

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

The standard deviation has the same measurement units as the random variable we considered.

Definition 1.19.4. *We call the k**th moment** of a random variable X, the expected value of the variable X^k. If we denote such a moment by μ'_k, we can write:*

$$\mu'_k = E(X^k) = \sum_{i=1}^{n} p_i \cdot x_i^k.$$

Definition 1.19.5. *The k**th central moment** of a random variable X is the espected value of the random variable $[X - M(X)]^k$. If we denote these moments by μ_k, we can write:*

$$\mu_k = E[X - E(X)]^k.$$

In particular we have:

$$\mu_1 = E[X - E(X)] = 0\ , \qquad \mu_2 = E[X - E(X)]^2 = V(X).$$

Proposition 1.19.3. *The moments and the central moments satisfy the relationship:*

$$\mu_k = \mu'_k - C_k^1\, \mu'_1\, \mu'_{k-1} + C_k^2\, (\mu'_1)^2\, \mu'_{k-2} + ... + (-1)^k (\mu'_1)^k.$$

Proof. By definition $\mu_k = M(X - \mu'_1)^k$. But

$$(X - \mu'_1)^k = X^k - C_k^1 \mu'_1 X^{k-1} + C_k^2 (\mu'_1)^2 X^{k-2} - ... + (-1)^k (\mu'_1)^k.$$

So:

$$\mu_k = E(X - \mu'_1)^k = E(X^k) - C_k^1\, \mu'_1\, E(X^{k-1}) + C_k^2\, (\mu'_1)^2\, E(X^{k-2}) - ... + (-1)^k (\mu'_1)^k.$$

□

Based on this formula, for $k = 2$, we have

$$\mu_2 = \mu'_2 - (\mu'_1)^2,$$

that is, the variance is equal to the difference between the second moment and the square of the first moment.

If we put successively $k = 3, 4$, we have

$$\mu_3 = \mu'_3 - 3\mu'_1\mu'_2 + 2(\mu'_1)^3,$$

$$\mu_4 = \mu'_4 - 4\mu'_1\mu'_3 + 6(\mu'_1)^2\mu'_2 - 3(\mu'_1)^4.$$

1.20 Covariance. Correlation coefficient

Definition 1.20.1. *If X and Y are random variables defined on the same sample space $\mathcal{S}$, we call* ***covariance*** *of the variables X and Y, the number*

$$Cov(X,Y) = E\left([X-E(X)][Y-E(Y)]\right) = \sum_{x\in V_X}\sum_{y\in V_Y}(x-E(X))\cdot(y-E(Y))\cdot P(X=x,Y=y)$$

where V_X and V_Y are the sets of values of the random variables X and Y.

Proposition 1.20.1. *The following equality takes place:*

$$Cov(X,Y) = E(X\cdot Y) - E(X)\cdot E(Y).$$

Proof. We can write :

$$\begin{aligned}
Cov(X,Y) =& \sum_{x\in V_X}\sum_{y\in V_Y}(x-E(X))\cdot(y-E(Y))\cdot P(X=x,Y=y) = \\
=& \sum_{x\in V_X}\sum_{y\in V_Y} x\cdot y\cdot P(X=x,Y=y) - E(Y)\sum_{x\in V_X} x\sum_{y\in V_Y} P(X=x,Y=y) - \\
&- E(X)\sum_{y\in V_Y} y\sum_{x\in V_X} P(X=x,Y=y) + \sum_{x\in V_X}\sum_{y\in V_Y} E(X)\,E(Y)\,P(X=x,Y=y) = \\
=& E(X\cdot Y) - E(Y)\cdot\sum_{x\in V_X} x\cdot P(X=x) - E(X)\cdot\sum_{y\in V_Y} y\cdot P(Y=y) + \\
&+ E(X)\,E(Y)\sum_{x\in V_X}\sum_{y\in V_Y} P(X=x,Y=y) = \\
=& E(X\cdot Y) - E(Y)\cdot E(X) - E(X)\cdot E(Y) + E(X)\cdot E(Y) = \\
=& E(X\cdot Y) - E(X)\cdot E(Y).
\end{aligned}$$

□

Proposition 1.20.2. *If X and Y are independent random variables, then:*

$$Cov(X,Y) = 0.$$

Proof. Immediate, based on Proposition 1.20.1. □

Proposition 1.20.3. *If X_1, X_2, ..., X_n are n random variables defined on the same sample space, then:*

$$V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i) + 2\sum_{i<j} Cov(X_i,X_j).$$

Proof. We will prove the proposition for $n = 3$. We have:

$$\begin{aligned}
V(X_1+X_2+X_3) =& E((X_1-E(X_1)+X_2-E(X_2)+X_3-E(X_3))^2) = \\
=& E[(X_1-E(X_1))^2+(X_2-E(X_2))^2+(X_3-E(X_3))^2+ \\
&+2(X_1-E(X_1))(X_2-E(X_2))+2(X_1-E(X_1))(X_3-E(X_3))+ \\
&+2(X_2-E(X_2))(X_3-E(X_3))] = \\
=& E(X_1-E(X_1))^2+E(X_2-E(X_2))^2+E(X_3-E(X_3))^2+ \\
&+2E[(X_1-E(X_1))(X_2-E(X_2))]+ \\
&+2E[(X_1-E(X_1))(X_3-E(X_3))]+ \\
&+2E[(X_2-E(X_2))(X_3-E(X_3))] = \\
=& \sum_{i=1}^{3} V(X_i) + 2\sum_{1\le i<j\le 3} Cov(X_i,X_j).
\end{aligned}$$

□

Definition 1.20.2. *If X and Y are two variables defined on the same sample space $\mathcal{S}$, we call* ***correlation coefficient*** *of the variables X and Y, the number*

$$\rho(X,Y) = \frac{E[(X-E(X))\cdot(Y-E(Y))]}{\sqrt{V(X)\cdot V(Y)}} = \frac{Cov(X,Y)}{\sigma(X)\cdot\sigma(Y)}.$$

Remark 1.20.1. If X_1, X_2, ..., X_n are n random variables defined on the sample space $\mathcal{S}$, then:

$$V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i) + 2\sum_{i<j} \sigma(X_i)\cdot\sigma(X_j)\cdot\rho(X_i,X_j).$$

where $\sigma(X_i) = \sqrt{V(X_i)}$.

Example 1.20.1. In a library, there are books numbered from 1 to n. We randomly take out the books. We say we have a meeting, if the number from the book is the same as the extraction number. Compute the expected value and the variance for the total number of meetings.

Solution: We associate to each book a random variable X_i, $i = 1, 2, ..., n$, defined as follows: if the book has the number i at the ith extraction, then $X_i = 1$, in the rest of the situations $X_i = 0$. The probability to obtain the book number i at the ith extraction is $P(X_i) = \frac{1}{n}$, because there is only one book with this number among the other n in the library. As each variable X_i can only take the values 1 or 0, we have that:

$$P(X_i = 0) = 1 - P(X_i = 1) = 1 - \frac{1}{n}.$$

From here we have that the distribution of the random variable X_i is:

$$X_i : \begin{pmatrix} 1 & 0 \\ \frac{1}{n} & 1-\frac{1}{n} \end{pmatrix}.$$

We have $E(X_i) = \frac{1}{n}$, from where $V(X_i) = E(X_i^2) - E^2(X_i) = \frac{1}{n} - \frac{1}{n^2} = \frac{n-1}{n^2}$.

The total number of meetings is given by the random variable

$$Y = X_1 + X_2 + ... + X_n.$$

We have:

$$E(Y) = \frac{1}{n} + \frac{1}{n} + ... + \frac{1}{n} = 1$$

and

$$V(Y) = \sum_{i=1}^{n} V(X_i) + 2 \sum_{i<j} Cov(X_i, X_j).$$

To compute the covariance, we have:

$$Cov(X_i, X_j) = E(X_i \cdot X_j) - E(X_i) \cdot E(X_j)$$

and

$$E(X_i \cdot X_j) = 1 \cdot P(X_i X_j = 1) + 0 \cdot P(X_i X_j = 0) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)},$$

because $X_i \cdot X_j = 1$ if and only if the book with the numbers i and j have been extracted at their turn, and there are $(n-2)!$ arrangements in which this event can take place.

So, if $i \neq j$, we have:

$$Cov(X_i, X_j) = \frac{1}{n(n-1)} - \frac{1}{n} \cdot \frac{1}{n} = \frac{1}{n^2(n-1)}.$$

Taking into consideration the results obtained so far, we have:

$$V(Y) = n \cdot V(X_i) + n(n-1)\, Cov(X_i, X_j) = n \cdot \frac{n-1}{n^2} + n(n-1) \cdot \frac{1}{n^2(n-1)} = 1.$$

1.21 Convergence of sequences of random variables.

We consider a sequence of random variables X_1, X_2, ..., X_n, ... defined on the same sample space $\mathcal{S}$.

In the probability theory we can find different concepts of convergence for the sequences of random variables $(X_n)_n$.

Definition 1.21.1. *We say that the sequence of random variables (X_n)* ***converges surely*** *or* ***everywhere*** *towards X if*

$$\lim_{n\to\infty} X_n(e) = X(e) \quad \forall e \in \mathcal{S}.$$

Definition 1.21.2. *We say that the sequence of random variables (X_n) converges towards X* ***in probability****, if*

$$\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1$$

for all $\varepsilon > 0$.

Definition 1.21.3. *We say that the sequence of random variables (X_n) converges* ***almost surely*** *towards the random variable X, if*

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

Definition 1.21.4. *Let $F_n(x)$ be the distribution function of the variable X_n, $(n = 1, 2, ...)$ and $F(x)$ the distribution function of the variable X. The sequence X_n converges towards X* ***in distribution*** *if*

$$\lim_{n\to\infty} F_n(x) = F(x).$$

Definition 1.21.5. *If*

$$\lim_{n\to\infty} D^2(X_n - X) = 0,$$

we say that the sequence X_n ***converges in mean square*** *to* X*.*

Proposition 1.21.1. *If the sequence* X_n *converges in mean square to* X*, then* X_n *converges to* X *in probability.*

Proof. From Chebyshev's inequality we have

$$1 - \frac{V(X_n - X)}{\varepsilon} \leq P(|X_n - X| < \varepsilon) \leq 1,$$

from where, passing to the limit when $n \to \infty$, we obtain that if $V(X_n - X) \overrightarrow{_{n\to\infty}} 0$, then $P(|X_n - X| < \varepsilon) \overrightarrow{_{n\to\infty}} 1$ □

Proposition 1.21.2. *If the sequence* X_n *converges almost surely to* X*, then* $X_n \overrightarrow{_{n\to\infty}} X$ *in probability.*

Proof. If $X_n \overrightarrow{_{n\to\infty}} X$ almost surely, then for all $\varepsilon > 0$ we have

$$\lim_{N\to\infty} P\left(\sup_{n\geq N} |X_n - X| > \varepsilon\right) = 0.$$

It follows from here that

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0.$$

□

Proposition 1.21.3. *If* $V(X_n - X) \overrightarrow{_{n\to\infty}} 0$ *and* $\sum_{i=1}^{\infty} E(X_n - X)^2 < +\infty$*, then* $X_n \overrightarrow{_{n\to\infty}} X$ *almost surely.*

1.22 Law of large numbers

Theorem 1.22.1 (Chebyshev)**.** *Let* (X_n) *be a sequence of random variables defined on a sample space* $\mathcal{S}$*. If the random variables are independent and* $V(X_n) \leq c$*,* $\forall n$*, then for all* $\varepsilon > 0$ *we have*

$$\lim_{n\to\infty} P(|\bar{X}_n - E(\bar{X}_n)| < \varepsilon) = 1,$$

where $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + ... + X_n)$*.*

Proof. From Chebyshev's inequality we have:

$$1 - \frac{V(\bar{X}_n)}{\varepsilon} \leq P(|\bar{X}_n - E(\bar{X}_n)| < \varepsilon) \leq 1.$$

Because

$$V(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_j) \leq \frac{nc}{n^2} = \frac{c}{n},$$

we obtain

$$1 - \frac{c}{n} \leq P(|\bar{X}_n - E(\bar{X}_n)| < \varepsilon) \leq 1$$

and from here

$$\lim_{n\to\infty} P(|\bar{X}_n - E(\bar{X}_n)| < \varepsilon) = 1.$$

□

Remark 1.22.1. Chebyshev's Theorem shows that even if the independent random variables can take values far away from their expected values, the arithmetic mean of a sufficiently large number of such random variables takes, with a large probability, values in the neighborhood of the constant $\frac{1}{n}\sum_{j=1}^{n} E(X_j)$. So, there is a big difference between the behavior of the random variables and their arithmetic mean. In the case of the random variables we cannot predict their value with a large probability, while, in the case of their arithmetic mean we can give its value with a probability close to 1.

The arithmetic mean of a sufficiently large number of random variables loses the property of being a random variable

Theorem 1.22.2 (Bernoulli)**.** *We suppose we make n independent experiences, in each experience the probability of the event A being p, and let ν be the number of times the event A is accomplished during the n experiences. For each ε we have*

$$\lim_{n\to\infty} P\left(\left|\frac{\nu}{n} - p\right| < \varepsilon\right) = 1.$$

Proof. We associate to each experience a random variable X_j which takes the value 1 if the event has been accomplished in the jth experience, and 0 if it hasn't. So, the number of the accomplishments of the event A during the n experiences is given by

$$\nu = X_1 + X_2 + ... + X_n$$

where each of the variables X_1, X_2, ..., X_n has the distribution

$$X_i : \begin{pmatrix} 1 & 0 \\ p & 1-p \end{pmatrix}.$$

It follows from here that: $E(X_i) = p$, $V(X_i) = p\,(1-p)$, and $E(\nu) = np$, $V(X) = n\,p\,(1-p)$.

Considering Chebyshev's inequality for the variable $\frac{1}{n}\nu$ we have:

$$P\left(\left|\frac{\nu}{n} - E\left(\frac{\nu}{n}\right)\right| < \varepsilon\right) \geq 1 - \frac{V(\frac{\nu}{n})}{\varepsilon^2}$$

or

$$P\left(\left|\frac{\nu}{n} - p\right| < \varepsilon\right) \geq 1 - \frac{p\,(1-p)}{n\varepsilon^2}.$$

Taking into account that $p\,(1-p) \leq \frac{1}{4}$, we get

$$P\left(\left|\frac{\nu}{n} - p\right| < \varepsilon\right) \geq 1 - \frac{1}{4n\varepsilon^2}.$$

Passing to the limit, we come to

$$\lim_{n\to\infty} P\left(\left|\frac{\nu}{n} - p\right| < \varepsilon\right) = 1,$$

which proves the theorem. □

Remark 1.22.2. If we have a large population, if we make a selection of volume n and get ν favorable results, we can state, with a probability near to 1, that the probability of the studied event is given by the relative frequency.

Thus, in the study of large populations for which we cannot determine the probability of accomplishing an event a priori, we can express this probability by means of relative frequency of the considered event, $\frac{\nu}{n}$, which theoretically justifies the use of the relative frequency instead of the probability.

Example 1.22.1. We toss a coin n times. Hoe large should n be, so that the probability of the inequality

$$\left|\frac{\alpha}{n}-\frac{1}{2}\right|<\frac{1}{100}$$

is larger than 0.99, knowing that α is the number of times a face (before-chosen) appears.

Solution: Having no reason to suppose one face has a greater chance to occur than another one, we have that $\alpha \sim \frac{n}{2}$.

thus $E(\alpha)=\frac{n}{2}$, $V(\alpha)=\frac{n}{4}$, $E(\frac{\alpha}{n})=\frac{1}{2}$, $V(\frac{\alpha}{n})=\frac{1}{4n}$.

it follows, from Chebyshev's inequality, that:

$$P\left(\left|\frac{\alpha}{n}-\frac{1}{2}\right|<\frac{1}{100}\right)\geq 1-\frac{D^2(\frac{\alpha}{n})}{\frac{1}{10^2}}.$$

We infer that, as $P\left(\left|\frac{\alpha}{n}-\frac{1}{2}\right|<\frac{1}{100}\right)>0.99$, it is sufficient to enforce

$$1-\frac{V(\frac{\alpha}{n})}{\frac{1}{10^2}}>0.99 \;\Rightarrow\; 1-\frac{(100)^2}{4n}>\frac{99}{100} \;\Rightarrow\; 4n>10^6 \;\Rightarrow\; n>250.000.$$

1.23 Binomial distribution

We consider an experiment whose sample space is made up of two elementary events $\mathcal{S}_1=\{e_1,e_2\}$, with probabilities

$$P(\{e_1\})=p \;\text{ and }\; P(\{e_2\})=1-p,$$

where p is a number in the interval $[0,1]$.

We assume that we repeat the experiment for n times and we consider the cartesian product:

$$\mathcal{S}=\underbrace{\mathcal{S}_1\times\mathcal{S}_1\times ...\times\mathcal{S}_1}_{n \text{ times}}$$

that is the set of elements of the form $(e_{i_1},e_{i_2},...,e_{i_n})$ where i_j is 0 or 1.

The sample space $\mathcal{S}$ has 2^n elements. We define the probability P on $\mathcal{S}$ as follows:

$$P(\{(e_{i_1},e_{i_2},...,e_{i_n})\})=P_1(\{e_{i_1}\})\cdot P_1(\{e_{i_2}\})\cdot ...\cdot P_1(\{e_{i_n}\}).$$

Let A_k, $k=0,1,...,n$, be the event of those $(e_{i_1},e_{i_2},...,e_{i_n})$ which have k elements of 1. (A_k is the event which consists of k successes).

Proposition 1.23.1.
$$P(A_k) = C_n^k \cdot p^k \cdot (1-p)^{1-k}.$$

Proof. The number of systems of the form $(e_{i_1}, e_{i_2}, ..., e_{i_n})$ containing k numbers of 1 is C_n^k. □

Remark 1.23.1. The events A_1, A_2, ..., A_n are mutually exclusive and their union is the sure event, thus:
$$A_0 \cup A_1 \cup ... \cup A_n = \mathcal{S}$$
and
$$P(A_0) + P(A_1) + ... + P(A_n) = 1.$$
the probabilities $P(A_0)$, $P(A_1)$, ..., $P(A_n)$ can be useful in defining a distribution in a sample space of $n+1$ points, in which the kth event is A_k.

Definition 1.23.1. *The random variable X having the distribution*
$$X : \begin{pmatrix} 0 & 1 & ... & k & ... & n \\ C_n^0 p^0 (1-p)^n & C_n^1 p^1 (1-p)^{n-1} & ... & C_n^k p^k (1-p)^{n-k} & ... & C_n^n p^n (1-p)^0 \end{pmatrix}$$
*is called **binomial variable** it is usually denoted by $B(n,p)$ (or $X \sim B(n,p)$).*

The term "binomial random variable" comes from the fact that the probabilities
$$b(k; n, p) = C_n^k p^k (1-p)^{n-k}$$
are consecutive terms from the expansion $[p + (1-p)]^n$.

Remark 1.23.2. The choices of p and n uniquely determine the binomial distribution; different choices lead to different distributions.

Definition 1.23.2. *The set of all binomial distributions is called the **family of binomial distributions**.*

Remark 1.23.3. Binomial distributions have been analyzed by James Bernoulli. The diagram he gave and which models a large number of real phenomena, is: consecutive independent extractions from an urn containing a white and b black balls, a and b remain the same during the extractions (the balls are put back in the urn). The probability to extract a white ball is $p = \dfrac{a}{a+b}$ and that of extracting a black one: $1 - p = q = \dfrac{b}{a+b}$.

Proposition 1.23.2. *The random variable $X \sim B(n,p)$ has the following properties:*

1. $E(X) = n \cdot p$;
2. $V(X) = n \cdot p \cdot (1-p)$;
3. *If we denote by Mo the modulus of the variable X (the most probable value), then:*
$$np - (1-p) \leq Mo \leq np + p.$$

Proof. By computation. □

Proposition 1.23.3. *If $X_1 \sim B(n_1, p)$, $X_2 \sim B(n_2, p)$ and X_1, X_2 are independent, then*

1. $X_1 + X_2 \sim B(n_1 + n_2, p)$;

2. $P(X_1 = k | X_1 + X_2 = n) = \dfrac{C_{n_1}^k \cdot C_{n_2}^{n-k}}{C_{n_1+n_2}^n}$, *for* $\max(0, n_1 - n_2) \leq k \leq \min(n_1, n)$.

Proposition 1.23.4. *The distribution function for the variable* $X \sim B(n,p)$ *is:*

$$F(x) = \begin{cases} 0 & , x \leq 0 \\ (1-p)^n & , 0 < x \leq 1 \\ (1-p)^n + C_n^1 p\,(1-p)^{n-1} & , 1 < x \leq 2 \\ \dots\ \dots\ \dots & \\ (1-p)^n + C_n^1 p\,(1-p)^{n-1} + \dots + C_n^k p^k\,(1-p)^{n-k} & , k < x \leq k+1 \\ \dots\ \dots\ \dots & \\ 1 & , x > n. \end{cases}$$

1.24 The Poisson distribution as an approximation of the binomial distribution

Binomial distributions $b(k; n, p) = C_n^k p^k (1-p)^{n-k}$ are often hard to compute. The tables built for such probabilities depend on two parameters, n and p. In some conditions we can find simple expressions, which approximate the probabilities $b(k; n, p)$ for $n \to \infty$, and the approximation is acceptable even for small values of n.

In the following we will present such approximations.

In the beginning we will prove a theorem concerning the binomial distribution for a large n and a reasonably large $n \cdot p$.

Theorem 1.24.1 (Poisson)**.** *If* $n \cdot p_n \to \lambda$, *then*

$$b(k; n, p) \to \frac{\lambda^k}{k!}\, e^{-\lambda}.$$

Proof. We have:

$$\begin{aligned} b(k; n, p_n) =& \frac{1}{k!}\, n\,(n-1)\dots(n-k+1) \cdot p_n^k\,(1-p_n)^{n-k} = \\ =& \frac{1}{k!} \cdot \frac{n}{n} \cdot \frac{n-1}{n} \cdot \dots \cdot \frac{n-k+1}{n} \cdot (n\,p_n)^k\,(1-p_n)^{n-k}. \end{aligned}$$

Because

$$\frac{n}{n} \cdot \frac{n-1}{n} \cdot \dots \cdot \frac{n-k+1}{n} \xrightarrow[n\to\infty]{} 1 \quad \text{and} \quad (n\,p_n)^k \xrightarrow[n\to\infty]{} \lambda^k$$

proving the theorem comes to proving that

$$(1-p_n)^{n-k} \xrightarrow[n\to\infty]{} e^{-\lambda}.$$

Because

$$(1-p_n)^{n-k} = (1-p_n)^n\,(1-p_n)^{-k}$$

and $(1-p_n)^{-k} \xrightarrow[n\to\infty]{} 1$ and $p_n \xrightarrow[n\to\infty]{} 0$, it is sufficient to prove that

$$(1-p_n)^n \xrightarrow[n\to\infty]{} e^{-\lambda}.$$

It is known that $\left(1 - \dfrac{a}{n}\right)^n \xrightarrow[n\to\infty]{} e^{-a}$ and the convergence is uniform on each finite interval $a_0 < a < a_1$.

All the numbers $n\,p_n$ can be restrained to such an interval around λ and thus $\forall \varepsilon > 0$, $\exists n_1(\varepsilon)$ so that $\forall n > n_1(\varepsilon)$ we have

$$\left|\left(1 - \frac{n\,p_n}{n}\right)^n - e^{-n\,p_n}\right| < \varepsilon.$$

From the continuity of the function e^{-x} for $n > n_2(\varepsilon)$ we have:

$$|e^{-n\,p_n} - e^{-\lambda}| < \varepsilon.$$

So, for $n > n(\varepsilon) = \max(n_1(\varepsilon), n_2(\varepsilon))$, we have:

$$\begin{aligned}|(1-p_n)^n - e^{-\lambda}| &= \left|\left(1-\frac{n\,p_n}{n}\right)^n - e^{-\lambda}\right| = \left|\left(1-\frac{n\,p_n}{n}\right)^n - e^{-n\,p_n} + e^{-n\,p_n} - e^{-n\,p_n}\right| \leq \\ &\leq \left|\left(1-\frac{n\,p_n}{n}\right)^n - e^{-n\,p_n}\right| + |e^{-n\,p_n} - e^{-\lambda}| < 2\varepsilon,\end{aligned}$$

which proves the theorem. □

Definition 1.24.1. *The random variable X with the probability mass function (probability function)*

$$p(k,\lambda) = P(X = k) = \frac{\lambda^k}{k!}\,e^{-\lambda}\,, \quad k = 0, 1, 2, ...$$

*is called **Poisson variable** with parameter λ, and we denote it by $X \sim Pois(\lambda)$.*

Example 1.24.1. 3% of the screws produced by a machine are out of order, the technical hitches occurring randomly during the production process.If the screws are packed in boxes containing each 100 pieces, which is the probability that a box shall have x screws that are out of order?

Solution: The probability is given by the binomial probability mass function

$$b\left(x; 100, \frac{3}{100}\right) = C_{100}^x \cdot \left(\frac{3}{100}\right)^x \cdot \left(1 - \frac{3}{100}\right)^{100-x}\,, \quad x = 0, 1, 2, ..., 100.$$

in this case $n = 100$, $p = 0.03$ and $n\,p = 3$, and the Poisson approximation for $b(x; 100, \frac{3}{100})$ is

$$b\left(x; 100, \frac{3}{100}\right) = \frac{3^x \cdot e^{-3}}{x!}.$$

Remark 1.24.1. The Poisson distribution occurs in different situations, for example:

- it gives the probabilities of a given number of phone calls in a certain time interval;
- it gives the probabilities of a given number of flaws on a length unit of a wire;
- it gives the probabilities of a specific number of faults on an area unit of a fabric;
- it gives the probabilities of a specific number of bacteria in a volume unit of a solution;
- it gives the probabilities of a specific number of accidents on time unit.

Let us see how the Poisson distribution occurs in one of the before mentioned situations.

Example 1.24.2. We consider a wire of length L and suppose that the probability of a flaw on a ΔL long part is $\lambda \cdot \Delta L$. We admit that this probability is independent of the position of the ΔL part.

We divide the wire of length L in $\Delta L = \dfrac{L}{n}$ long parts and we have:

1. the probability of a flaw along the ΔL part is $\lambda \cdot \Delta L$.
2. the probability of having no flaw along the ΔL part is $1 - \lambda \cdot \Delta L$.
3. the probability of having two or more flaws along the ΔL part is $\Theta(\Delta L)$, so that $\frac{\Theta(\Delta L)}{\Delta L} \to 0$ when $\Delta L \to 0$.

The event of having x flaws along a $L + \Delta L$ part, is the reunion of the following events:

a) there are x flaws on the L-long part and no flaws on the ΔL-long part;

b) there are $x - 1$ flaws on the L-long part and one flaw on the ΔL-long part;

c) there are $x - i$ flaws on the L-long part and i flaws on the ΔL-long part.

The probabilities of these events are: $P_x(L) \cdot [1 - \lambda \cdot \Delta L]$, $P_{x-1}(L) \cdot \lambda \cdot \Delta L$ and $\Theta(\Delta L)$. So

$$P(L + \Delta L) = P_x(L) \cdot [1 - \lambda \cdot \Delta L] + P_{x-1}(L) \cdot \lambda \cdot \Delta L + \Theta(\Delta L) \ , \ x = 1, 2, ...$$

and

$$P_0(L + \Delta L) = P_0(L) \cdot [1 - \lambda \cdot \Delta L] + \Theta(\Delta L).$$

These equalities can be written as:

$$\frac{P_x(L + \Delta L) - P_x(L)}{\Delta L} = \lambda \, [P_{x-1}(L) - P_x(L)] + \frac{\Theta(\Delta L)}{\Delta L} \ , \ x = 1, 2, ...$$

$$\frac{P_0(L + \Delta L) - P_0(L)}{\Delta L} = -\lambda \, P_x(L) + \frac{\Theta(\Delta L)}{\Delta L}.$$

For $\Delta L \to 0$ we get:

$$\frac{dP_x}{dL} = \lambda \, [P_{x-1}(L) - P_x(L)]$$

$$\frac{dP_0}{dL} = -\lambda \, P_0(L).$$

If we impose $P_0(0) = 1$ and $P_x(0) = 1$, $x > 1$, we have:

$$P_x(L) = \frac{(\lambda \, L)^x \, e^{-\lambda L}}{x!} \ , \ x = 0, 1, 2, ...$$

which is the probability mass function with $\lambda' = \lambda \, L$.

Theorem 1.24.2. *If X and Y are Poisson distributed independent random variables of parameters μ and ν, then $X + Y \sim Pois(\mu + \nu)$.*

Proof. We have

$$P(X = x) = \frac{\mu^x \cdot e^{-\mu}}{x!} \ , \quad x = 0, 1, 2, ...$$

$$P(Y = y) = \frac{\nu^y \cdot e^{-\nu}}{y!} \ , \quad y = 0, 1, 2, ...$$

and as X and Y are independent:

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y) = \frac{\mu^x \nu^y \cdot e^{-(\mu+\nu)}}{x! \, y!}.$$

Let there be $Z = X + Y$ and $g(z) = P(Z = z)$. The value $g(z)$ is obtained by summing up $P(X = x, Y = y)$ after all the pairs (x, y) with $x + y = z$, that is, after all the pairs $(x, z - x)$. So

$$g(z) = \sum_{x=0}^{z} P(X = x, Y = z - x) = e^{-(\mu+\nu)} \sum_{x=0}^{z} \frac{\mu^x \, \nu^{z-x}}{x! \, (z-x)!} =$$

$$= \frac{e^{-(\mu+\nu)} \, \nu^z \sum_{x=0}^{z} C_z^x \left(\frac{\mu}{\nu}\right)^x}{z!} = \frac{e^{-(\mu+\nu)} \, \nu^z \left(1 + \frac{\mu}{\nu}\right)^z}{z!} =$$

$$= e^{-(\mu+\nu)} \, \frac{(\mu+\nu)^z}{z!} \ , \ z = 0, 1, 2, ...$$

□

Proposition 1.24.1. *The distribution function of the variable $X \sim Pois(\lambda)$ is:*

$$F(x) = \begin{cases} 0 & , x \leq 0 \\ \cdots\cdots\cdots & \\ \sum_{j=1}^{k} \frac{\lambda^j}{j!} \, e^{-\lambda} & , k < x \leq k+1 \\ \cdots\cdots\cdots & \\ 1 & , n < x. \end{cases}$$

Proposition 1.24.2. *If $X \sim Pois(\lambda)$, then*

1. $\lambda - 1 < Mo < \lambda$;
2. $E(X) = V(X) = \lambda$.

1.25 The multinomial distribution

A box contains $N = N_1 + N_2 + ... + N_k$ balls, out of which: N_1 have the color 1, N_2 have the color 2, ..., N_k have the color k. The probability to extract a ball from the box is $\frac{1}{N}$. The probability to extract a ball of color i from the box, is $p_i = \frac{N_i}{N}$. This kind of box is called **Bernoulli's ...urna lui Bernoulli cu mai multe st'ari**.

Let A_i be the event consisting of extracting a ball of color i from the box, $1 \leq i \leq k$.

Proposition 1.25.1. *The probability that in n independent extractions (each of them being able to give rise to one of the k events, A_1, A_2, ..., A_k) the A_i occurs n_i times, is*

$$P_n(n_1, n_2, ..., n_k) = \frac{n!}{n_1! \, n_2! \, ... \, n_k!} \cdot p_1^{n_1} \, p_2^{n_2} \, ... \, p_k^{n_k}.$$

Proof. We suppose that the n balls extracted are in the following order

$$\underbrace{A_1 ... A_1}_{n_1 \text{ times}} \underbrace{A_2 ... A_2}_{n_2 \text{ times}} ... \underbrace{A_k ... A_k}_{n_k \text{ times}}.$$

Because A_i are independent, the probability of this event is $p_1^{n_1} \, p_2^{n_2} \, ... \, p_k^{n_k}$. Because the succession of the events A_1, ..., A_k is arbitrary, we have to see in how many ways we can write n symbols,

n_1 of them equal to A_1, n_2 equal to A_2, ..., n_k equal to A_k. This number is given by the number of the permutations with repetitions, and this is

$$\frac{n!}{n_1!\, n_2! \ldots n_k!}.$$

So, the sought probability is

$$P_n(n_1, n_2, ..., n_k) = \frac{n!}{n_1!\, n_2! \ldots n_k!} \cdot p_1^{n_1}\, p_2^{n_2} \ldots p_k^{n_k}.$$

Because

$$1 = (p_1 + p_2 + ...p_k)^n = \sum_{n_1,...,n_k} \frac{n!}{n_1!\, n_2! \ldots n_k!} \cdot p_1^{n_1}\, p_2^{n_2} \ldots p_k^{n_k}$$

we have that

$$\sum_{n_1,...,n_k} P_n(n_1, n_2, ..., n_k) = 1$$

where n_j can take all the possible values, such that $n_j \geq 0$ and $\sum_{j=1}^{k} n_j = n$. □

Definition 1.25.1. *A vector $(n_1, n_2, ..., n_k)$ having the probability mass function*

$$P_n(n_1, n_2, ..., n_k) = \frac{n!}{n_1!\, n_2! \ldots n_k!} \cdot p_1^{n_1}\, p_2^{n_2} \ldots p_k^{n_k}.$$

is called multinomial. Its corresponding distribution is called ***multinomial distribution*** *and is denoted $M(n; p_1, p_2, ..., p_n)$.*

1.26 Geometric distribution. Negative binomial distribution

When we defined the binomial distribution, we considered the representation of an experience with two possible results: success (S) and failure (F). We fixed a number of n experiences and determined the distribution of the numbers of successes during n repeats. We called this distribution the binomial distribution.

In what comes, we consider the situation in which the total number of events is not given afore.

We repeat the experience until we obtain r successes, r being fixed afore; the number of failures (x) and the total number of events ($x + r$) are variable.

Let X be the variable of the "total number of failures before r successes are accomplished". We want to find the probability mass function $f(x)$ of the variable X.

First of all, we consider the case $r = 1$. Then $f(x)$ is the probability to have x failures before the first success. This event can be realized in a single way: "we must obtain failures in the first x replays of the experience and then a success for the $x + 1$th experience". Because the replays of the experience are independent, we have

$$P(\underbrace{I, I, ..., I}_{x \text{ times}}, S) = \underbrace{(1-p)\,(1-p) \ldots (1-p)}_{x \text{ times}}\, p\ ,$$

where p is the probability to have a success. So,

$$f(x) = (1-p)^x \cdot p\ , \quad x = 0, 1, 2, ...$$

Definition 1.26.1. *The random variable X having the probability mass function*

$$f(x) = (1-p)^x \cdot p$$

is called ***geometric variable****, and $(x, f(x))$, $x = 0, 1, 2, ...$ is called* ***geometric distribution****.*

As $0 < p < 1$, we have

$$\sum_{x=0}^{\infty} f(x) = p\left[1 + (1-p) + (1-p)^2 + ...\right] = \frac{p}{1-(1-p)} = 1.$$

In the general case, $f(x)$ is the probability to obtain exactly x failures before the rth success. To carry out this event, we must obtain a success in the experience $r + x$, and $r - 1$ successes and x failures in the previous $r + x - 1$ experiences. the probability to have $r - 1$ successes in the first $r + x - 1$ experiences is

$$C_{r+x-1}^{r-1} \cdot p^{r-1} (1-p)^x ,$$

where the probability to have a success in the experience $r + x$ is p.

Because the experiences are independent, the sought probability is

$$f(x) = C_{r+x-1}^{r-1} \cdot p^{r-1} (1-p)^x , \quad x = 0, 1, 2, ...$$

Definition 1.26.2. *The variable X having the probability mass function*

$$f(x) = C_{r+x-1}^{r-1} \cdot p^{r-1} (1-p)^x$$

is called the ***negative binomial variable****, and $(x, f(x))$ the* ***negative binomial distribution****.*

Proposition 1.26.1. *If X, Y are independent random variables, with negative binomial distributions of parameters (r_1, p), respectivly (r_2, p), then the variable $X + Y$ has a negative binomial distribution with parameters $(r_1 + r_2, p)$.*

1.27 Continuous random variables

The lengths and the theoretic time can take any real value in an interval. If the values of a random variable are such entities, then the probability function is defined on an interval, and associates "probabilities" to all the events of this interval, not only to a finite number of points (point-probabilities).

Example 1.27.1. If a clock stops randomly, what is the probability that the hand indicating the hours should stop between 7 and 10?

Solution: In this case, we deal with a random variable T, whose possible values are in a continuous interval. Any value between 0 and 12 is possible and there are an infinite number of possibilities.

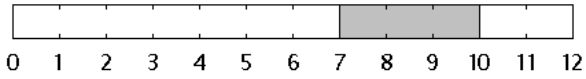
We cannot count the equiprobable cases and we cannot associate point-probabilities. Therefore, to each subinterval of [0, 12] we associate a probability proportional to the length of the subinterval. Because to the interval [0,12] of length 12, we must associate the probability 1,

we associate to an interval of unit length, the probability $\frac{1}{12}$, and therefore we will associate the probability $\frac{3}{12}$ to an interval of length $3 = 10 - 7$.

So, the probability that T takes values from 7 to 10, is $\frac{3}{12}$.

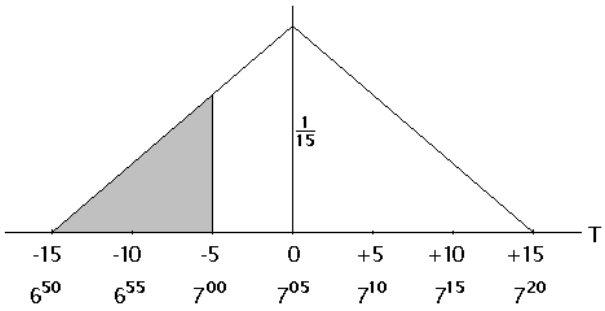
$$P(7 < T < 12) = \frac{10-7}{12-0} = \frac{3}{12}.$$

In this case, we can reason by measuring the probabilities through rectangular areas as follows: we represent the time interval $[0, 12]$ horizontally as in the figure and we consider the rectangle of length 12 and width $\frac{1}{12}$.



We considered the width $\frac{1}{12}$, so that the rectangle's area shall be equal to the probability of the sure event ("the hand stops somewhere between 7 and 12 "), that is 1. The shaded rectangle corresponds to the interval $[7, 10]$, and has the area $3 \cdot \frac{1}{12} = P(7 < T < 12)$.

Example 1.27.2. A bus must arrive in the station A at 7 o'clock. For unknown reasons, the arrival hour varies between 6^{50} and 7^{20}. The relative frequencies of the different arrivals show that these are distributed as in the figure below:



What is the probability that for the bus to arrive in time?

Solution: Let T be the random variable that gives the arrival hour. Suppose that $T = 0$ for 7^{05}, which is the middle point of the interval for the arrival hour. Because the triangle's area is 1, and the base varies from $T = -15$ to $T = 15$, its height will be $h = \frac{1}{15}$.

The area of the shaded triangle represents the probability that the bus arrives in the station at 7^{00} o'clock. The height of the triangle is $\frac{2}{45}$, and so its area will be $\frac{1}{2} \cdot 10 \cdot \frac{2}{45} = \frac{2}{9}$.

Therefore, the sought probability is $\frac{2}{9}$.

Example 1.27.3. On the line segment AB we randomly choose a point U. What is the probability that the distance from point U to A is at least twice as large as the distance from U to B?

Solution: To randomly choose the point U comes to saying that no region on the line segment AB is privileged. This means that the probability that U belongs to a subinterval I of AB is

proportional to I's length. We have from here that

$$P(U \in I) = \frac{\text{length of } I}{\text{length of } AB}$$

and therefore

$$P(|AU| > 2|BU|) = \frac{1}{3}.$$

1.28 The distribution function for the continuous random variables. Probability distribution

If X is a random variable whose values are the points of an interval, we expect to have small changes in $P(X < x)$ for small changes in x, in other words, the distribution function $F(x)$ is continuous.

Definition 1.28.1. *A variable X with real values is called **a continuous variable** if its distribution function $F(x)$ is continuous.*

In fact, we will suppose more. We will admit that $F(x)$ is derivable and that $\frac{dF}{dx}$ is continuous.

Definition 1.28.2. *The function $f(x) = \frac{dF}{dx}$ is called **the probability distribution** of the variable X.*

Since F is continuous, we have that

$$f(x) \geq 0\,, \quad \forall x.$$

We also have that:

$$F(x) = \int_{-\infty}^{x} f(t)dt \quad \text{and} \quad F(\infty) = \int_{-\infty}^{\infty} f(t)dt.$$

The distribution function $F(x)$ is also called **the probabilistic integral** of X.

Remark 1.28.1. The probability distribution f is defined up to a multiplicative constant, which is determined from the condition

$$\int_{-\infty}^{\infty} f(t)dt = 1.$$

Proposition 1.28.1. *If a and b are two real numbers, the following equalities take place:*

1. $P(a \leq X < b) = P(X < b) - P(X < a) = F(b) - F(a)$;
2. $P(a \leq X < b) = F(b) - F(a) = \int_a^b f(t)dt$.

Proof. Immediate. □

Remark 1.28.2. If $a \to b$, then $F(a) \to F(b)$, that is

$$P(a \leq X < b) \xrightarrow[a \to b]{} 0,$$

so the probability that X takes the value b is zero. The expression $f(b)$ does not give the probability that $X = b$;we only use the probability distribution as an integrand.

Remark 1.28.3. If $b-a$ is small and $f(x)$ is continuous, then

$$P(a \leq X < b) = \int_a^b f(t)dt \approx (b-a) \cdot f(t)$$

where $t = \dfrac{a+b}{2}$.

Remark 1.28.4.

$$f(x) = \lim_{\Delta x \to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{P(x \leq X < x+\Delta x)}{\Delta x}.$$

Aceast'a rela'tie explic'a 'si originea termenului de "densitate de reparti'tie": $P(x \leq X < x+\Delta x)$ reprezint'a probabilitatea ca variabila aleatoare X s'a ia valori 'in intervalul $[x, x+\Delta x)$ (adic'a "masa"), iar Δx este lungimea acestui interval (adic'a "volumul").

Definition 1.28.3. *We call **mode**, and denote it with Mo, the value of the variable X for which the probability distribution is maximum.*

The mode is among the roots of the equation

$$\frac{df}{dx} = 0.$$

If the equation has only one solution, the distribution is called **unimodal**.

Because $f(x) = \dfrac{dF}{dx}$, we have that the mode is among the roots of the equation

$$\frac{d^2F}{dx^2} = 0.$$

This shows that in $x = Mo$ the graph of the distribution function has an inflexion point.

Definition 1.28.4. ***The α-quantile** of the continuous random variable X is the value x_α of the variable X for which $P(x_\alpha) = \alpha$.*

If the function $F(x)$ is strictly increasing, then the variable X admits quantiles of any order, and these are unique.

The $\frac{1}{2}$-quantile is called **median** and we denote it Me. For the median we have

$$P(X < Me) = P(X > Me) = \frac{1}{2}.$$

Example 1.28.1. Determine the distribution function, the median, the mode for the random variable T from the example 1.27.2.

1.29 The expected values and the variance of a continuous random variable

Definition 1.29.1. *If X is a continuous random variable with the probability distribution f, then its **expected value** is defined as*

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

if the integral converges absolutely, that is

$$\int_{-\infty}^{+\infty} |x\, f(x)|dx < +\infty.$$

In the contrary we say that X ***does not have a finite expected value (mean).***

Proposition 1.29.1. *The expected value of a continuous random variable has the following properties:*

1. $E(aX) = a \cdot E(X)$;
2. $E(X + Y) = E(X) + E(Y)$;
3. $E(X - E(X)) = 0$.

Proof. We use the properties of the integral. □

Example 1.29.1. In the conditions of the example 1.27.2, determine the expected value of the arrival time.

Definition 1.29.2. ***The variance*** *of the random variable X is*

$$var(X) = \int_{-\infty}^{+\infty} (x - E(X))^2\, f(x)dx.$$

Example 1.29.2. Determine the variance of the random variable from the example 1.27.2.

1.30 The normal distribution

Definition 1.30.1. *We say that a variable X follows a* ***normal distribution of parameters*** *μ* ***and*** *σ^2 if it has the probability distribution*

$$n(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x - \mu)^2\right], \quad -\infty < x < \infty,$$

and we will write $X \sim N(\mu, \sigma^2)$.

Proposition 1.30.1. *If $X \sim N(\mu, \sigma^2)$, then*

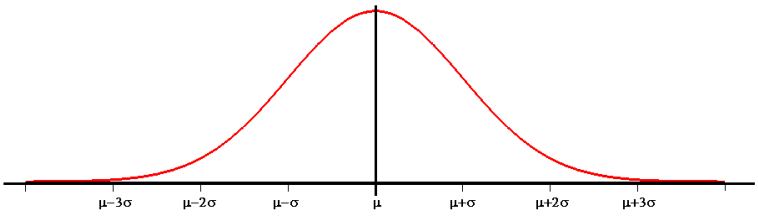
$$E(X) = \mu \quad \text{and} \quad var(X) = \sigma^2.$$

Proof. By direct calculations. □

Remark 1.30.1. The function $n(x; \mu, \sigma^2)$ is symmetric with respect to $x = \mu$, it has a maximum in $x = \mu$ and has inflection points in $x = \mu \pm \sigma$. The graph of the function is represented in the figure:

Definition 1.30.2. *We say that the variable X has* ***the standard normal probability distribution*** *if it has the probability distribution*

$$n(x; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \quad -\infty < x < +\infty.$$



Definition 1.30.3. *The distribution function of the standard normal variable,*

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

is called ***Laplace function****.*

From the symmetry of $\Phi(x)$ we have that:

$$\Phi(-x) = P(X < -x) = P(X > x) = 1 - P(X < x) = 1 - \Phi(x)$$

from where:

$$\Phi(x) + \Phi(-x) = 1.$$

The normal distribution function has tables for different values of x. With these tables we can find the probabilities associated with the events regarding any normal variable.

Proposition 1.30.2. *If $X \sim N(\mu, \sigma^2)$, then:*

i) $P(X < b) = \Phi\left(\dfrac{b-\mu}{\sigma}\right)$;

ii) $P(a < X < b) = \Phi\left(\dfrac{b-\mu}{\sigma}\right) - \Phi\left(\dfrac{a-\mu}{\sigma}\right)$.

Proof. We use the variable change $u = \dfrac{x-\mu}{\sigma}$. □

Proposition 1.30.3. *If $X_i \sim N_i(\mu_i, \sigma_i^2)$, $i = \overline{1,n}$, then*

$$Y = \sum_{i=1}^{n} c_i X_i$$

has the property

$$Y \sim N_i\left(\sum_{i=1}^{n} c_i \mu_i, \sum_{i=1}^{n} c_i^2 \sigma_i^2\right).$$

Chapter 2

STATISTICS

2.1 What is Statistics?

Problem 2.1.1. *The West University of Timi'soara wants to make a development plan of the accommodating facilities. To proceed to actions, the council decides it is necessary to answer the following question: How many students will have to be lodged in the next ten years?*
To answer this question we have to know at least the answer to the following two questions: How many high school graduates will there be? How many want to attend an university? (And maybe other questions as well).
To answer these two questions we need information concerning the number of high school graduates in the next ten years and information indicating the percentage of those graduates who wish to attend the U.V.T.
One way to obtain the data concerning the number of high school graduates in the next ten years is to see this number for the past ten years and extrapolate it. We have to observe that this idea assumes that there is a connection between the past and the future. This is not always true. In this context, an additional question would be whether we have to count all the graduates from all the high schools in the past ten years or we can limit our counting to those of certain schools? In other words, can we consider only sample?
One way to obtain the data concerning the percent of high school graduates who wish to attend the U.V.T. in the next ten years is to see this number for the past ten years and extrapolate it. Other questions that arising, are: how do we interpret these data? How do we formulate a conclusion based on these data? How do we make a decision based on these data?
We haven't finished the enumeration of the questions that can be relevant. At this moment the important thing is to start thinking about such problems and about the questions that need a clarification in order to obtain an answer.

Remark 2.1.1. The relationship between statistics and probabilities
Statistics and probabilities are two strongly connected, but still distinct fields of mathematics. It is said that "probabilities are the vehicle of statistics". This is true, meaning that if it weren't for the probabilistic laws, statistics wouldn't be possible. To illustrate the difference between probabilities and statistics, let us consider two boxes: a probabilistic and a statistical one. For the probabilistic box we know that it contains 5 white, 5 black and 5 red balls; the probabilistic problem is that if we take a ball, what is the chance that it were white? For a statistical box we do not know the combination of balls in the box. We consider a sample and from this sample we conjecture what we think the box contains. We have to remember the difference: the probability sets the question of the chance that something (an event) happens when we know the probabilities (we know the population). Statistics asks us to make a sample, to analyze it and then make a

prediction concerning the population, based on the information provided by the sample.

Remark 2.1.2. The correct an the incorrect use of statistics
The use of statistics is unlimited. It is hard to find a domain where statistics is not used. We give here some examples of where and how we use statistics:

- *in education; descriptive statistics is often used to present results;*
- *in science; experimental results have to be collected and analyzed;*
- *government; it collects statistical data all the time.*

many people are indifferent to statistical descriptions, others think they are lies. Most of the statistical lies are innocent ones and come from an inadequate usage or from data coming from an improper sample. All these lead to a wrong understanding of information from the consumer's side. The wrong use of statistics leads sometimes to confusions.

Remark 2.1.3. Statistics and the computer
In the last decade the computer has had an important role in nearly every aspect of life. Statistics does not make exception, as it uses a lot of techniques that are repetitive; formulas to calculate descriptive statistics, procedures to be made in order to formulate predictions. The computer is very good to make such repetitive computations. The analysis of statistical data is much easier if the computer has a standard statistical software. The most popular statistical softwares are: Minitab, Biomed (biomedical program), SAS (Statistical analysis system), IBM Scientific Subroutine Packages and SPSS (statistics package for social sciences).

2.2 Basics

Definition 2.2.1. The population is a collection (a set) of individuals, objects or numerical data obtained by measurements, whose properties need to be analyzed.

Remark 2.2.1. *The population is the **complete** collection of individuals, objects or numerical data obtained by measurements and which are of interest (to the one collecting the sample). In statistics the population concept is fundamental. The population has to be carefully defined and is considered completely defined only if the member list is specified. The set of the Mathematics and Informatics' students is a well defined population.*
Usually if we hear the word population, we think of a set of people. In statistics, the population can be a set of animals, of manufactured objects or of numerical data obtained through measurements. For example, the set of the "heights" of the students of the Faculty of Mathematics and Informatics is a population.

Definition 2.2.2. The sample is a subset of a population.

Remark 2.2.2. *A sample is made of individuals, objects or measured data selected from a population (by the sample collector).*

Definition 2.2.3. A response variable (or simply **variable**) is a characteristic (usually a numerical one) which is of interest for each element (individual) of a population.

Remark 2.2.3. *the age of the student, his grade point average, his hair color, his height, his weight a.s.o. are answer variables for the population: the students from the Faculty of Mathematics and Informatics.*

Definition 2.2.4. A data is the value of a response variable for an element of the population or of the sample.

Example 2.2.1. *Popescu Nicolae is "19 years" old, his GPA is 8.50, his hair color is "brown", his height is "1 m and 75 cm", and his weight is "65 kg". These five "values" of the five variables (Remark 2.2.3) for Popescu Nicolae are "five" data.*

Definition 2.2.5. The "values" of a variable for a population or a sample make up **a data set** . In a data set, a data occurs as many times as the variable has this "value".

Example 2.2.2. *The 25 heights in a 25 students' sample is a set of 25 not necessary different data.*

Definition 2.2.6. A planned activity whose result is a data set is called **experiment** or **survey**.

Definition 2.2.7. A parameter is a numeric characteristic of a population.

Example 2.2.3. *The percentage of students which have passed all the exams in the winter term, is an example of parameter for the population made up of the students from the Faculty of Mathematics and Informatics.*

Remark 2.2.4. *The parameter ia a numerical value concerning the whole population. In statistics a parameter is usually denoted by a Greek letter.*

Definition 2.2.8. A statistic is a numerical characteristic of a sample.

Example 2.2.4. *The mean height of 25 students in a sample is an example of a statistic (sample statistic).*

Remark 2.2.5. *A statistic is a numerical characteristic referring to a sample. Sample statistics are denoted by Latin letters.*

2.3 Data collection

The first problem of a statistician is the collection of a data set. This requires a preliminary definition of the survey's (event's) objectives, of the population and of the variable.
Examples of objectives:

a) The comparison of the efficiency of a standard medicine;

b) The estimation of the average income of a family from the county.

Examples of populations and their corresponding variables:

a) the patients suffering from a disease undergoing a treatment with the considered medicine represent the population, and the recovery time represents the variable;

b) the families of the county represent the population, and the total income of a family is the variable.

Also before collecting the data set, we have to decide whether the data set is made up for the whole population or just for a sample. If the data set is made up for the whole population, then we consider a census.

Definition 2.3.1. A census is an enumeration of each element of the population together with the data (the value of the variable) corresponding to the element.

In the case of a large population, making up a data set is difficult and expensive. Therefore, if organizing a census is not possible, the data set is considered for just a part of the population, for a sample. The selection for the sample elements is made from a sampling frame.

Definition 2.3.2. A sampling frame is a list of elements belonging to a population, from which we extract the sample.

Remark 2.3.1. *Because only the elements from the sampling frame have a chance to be selected in a sample, the sampling frame has to be representative for the population (from the variable's perspective).*

Remark 2.3.2. *For a population of individuals the elector lists or the telephone books are often used as a sampling frame. Depending on the variable, these can be proper or improper sampling frames.*

Remark 2.3.3. *After defining the sampling frame we pass to establishing the way of choosing the elements of a sample. This process is called sample design.*

Definition 2.3.3. The sample design means the establishing of the procedure in order to choose the elements from the sampling frame.

There are many choice procedures for the elements of the sample. Altogether, the choice procedures together with their corresponding samples, can be divided in two categories: representability procedures and probabilistic procedures.

Definition 2.3.4. Judgement samples are samples for which the elements are chosen so that, from the variable's perspective, the chosen element is representative for the population.

Example 2.3.1. From the variable's perspective: "course A is useful or not for your professional development?", the students from a sample who have not attended the course are not representative. So they are not chosen in the sample.

Definition 2.3.5. A sample for which the elements are chosen on a probabilistic basis (every element from the sample has a certain nonzero chance to be selected) is called **probability sample**.

Remark 2.3.4. *Statistic inferences require that the sample should be probabilistic. The random probabilistic samples are the most familiar probabilistic samples.*

Definition 2.3.6. A sample of size n is a **random probability sample** if any sample of size n chosen from the same frame has the same probability to be chosen.

Remark 2.3.5. *The most popular method to collect data uses a simple random sample.*

Definition 2.3.7. A random probabilistic sample for which the elements are selected from a frame where the elements are equiprobable is called **simple random sample**.

Remark 2.3.6. *When we build a simple random sample we have to take care that each element from the sampling frame has the same probability to be selected. We often make mistakes because the "random" term is mistaken for a "randomly chosen" one. A correct manner to select a simple probabilistic sample is that which uses a random number generator or a random number table. First we number the elements from the sampling frame. After that we choose as many numbers as necessary for the sample from the random number table. We will choose for the sample each element from the sampling frame, whose number coincides with a number selected from the random number table.*

Example 2.3.2. *If the sampling frame is a list of 4265 students, then they are numbered* $0001; 0002; ...; 4265$. *For a sample of 50 students we choose 50 four-digits random numbers and and we identify the corresponding students from the sampling frame.*

Definition 2.3.8. The systematic sample is build by choosing every k-th element from the sampling frame.

Remark 2.3.7. *In this selection we only use the random number table once, to determine the start point.*

Example 2.3.3. *If we consider a sampling frame of 245 students from the Faculty of Mathematics and Informatics and we want a systematic sample of 15 students then:*

1) we associate to each student a number between 1 and 245;

2) we compute k (the counting step) using the following relationship:

$$k = \left[\frac{\textit{number of elements from the sampling frame}}{\textit{number of elements from the sample}}\right] = \left[\frac{245}{15}\right] = 16$$

3) we choose the starting point between 1 *and k by using the random number table.*

If this number is 10*, we get the sample:*

$$10, 16, 32, 48, 64, 80, 96, 112, 128, 144, 160, 176, 192, 208, 234.$$

Because $k = \dfrac{245}{15} = 16,33$*, is not an integer, the counting step can also be chosen 17. In this case the systematic sample obtained has only 14 elements.*

Remark 2.3.8. *There is a good procedure to sample a percentage for large populations. In order to select a systematic sample of x% from a population, we will select an element from* $100/x$ *(if* $100/x$ *is not an integer we take its integral part).*

Remark 2.3.9. *Using a systematic sample is improper if the given population is repetitive or cyclic.(from the variable's perspective)*

Example 2.3.4. If we want to estimate the number of students admitted at the Faculty of Mathematics and Informatics who are more than 20 years old and we use systematic sampling, choosing from the nominee's list those from the positions which are a multiple of 5, there is the possibility that all the nominees in those positions are less than 20 years old. Such a sample indicates there are no nominees over 20 years admitted, a fact which cannot be asserted.

When possible, when we sample very large populations we part the population into two subpopulations based on certain characteristics. These subpopulations are called **strata**; the strata are separate samples.

Definition 2.3.9. A sample obtained by stratifying the sampling frame and by selecting a number of elements from each of the strata, is called **stratified sample**.

Remark 2.3.10. *When we design a stratified sample, we divide the frame in two or more strata and we design a subsample on each of the strata. these underlayers can be random, systematic or of another kind. After that the subsamples are joined to one sample to collect a set of data.*

Example 2.3.5. *To study one characteristic of the population of the students from the Faculty of Mathematics and Informatics, we can divide this population based on:*
- areas: informatics, mathematics
- academic years.

Definition 2.3.10. A quota sample (or proportional sample) is a stratified sample built by selecting a number of elements from each of the strata by a certain quota or proportional to the size of the stratum.

Example 2.3.6. *If we want to build a sample of 150 students from the population of students from the Faculty of Mathematics and Informatics we can make the stratification based on academic years. In this case the number of selected students in each academic year will be proportional to the total number of students in that year:*

Academic year	*Number of students*	*Quota*	*Nr. of students selected in the sample:*
First Year	*431*	*36.49%*	*54*
Second Year	*303*	*25.65%*	*40*
Third Year	*206*	*17.44%*	*26*
Forth Year	*240*	*20.40%*	*30*

the sample will be made of 54 *freshmen,* 40 *second-year students,* 26 *third year students and* 30 *senior students.*

Another sampling method starting with a stratification of the population is the cluster sampling.

Definition 2.3.11. The cluster sample is a stratified sample built by selecting samples from certain strata (not from all of them).

Example 2.3.7. *If we wish to build a cluster sample of students from the West University of Timi'soara, this population can be stratified based on their area of study, selecting samples only from some areas (not from all of them).*

Remark 2.3.11. *The cluster sample is obtained using random numbers or a systematic method to identify the strata to be sampled, and then sampling each of these strata. The assembled subsamples build a cluster sample.*

In given circumstances the sampling method used depends on the population, on the variable, on the sampling difficulty and on the expenses. After establishing the sample we can go on to collecting the data set.

2.4 Determining the frequency and grouping the data

After collecting a data set comes the primary processing of the data. Establishing the frequency and grouping the data is a primary processing of data and it is used when we have a large amount of information.
to illustrate the concept of frequency let use consider the following data set:

$$\begin{array}{ccccc} 3 & 2 & 2 & 3 & 2 \\ 4 & 4 & 1 & 2 & 2 \\ 4 & 3 & 2 & 0 & 2 \\ 2 & 1 & 3 & 3 & 1 \end{array}$$

0 appears only once, so the frequency for 0 is one.
1 appears three times in this set, so the frequency for 1 is three.
2 appears eight times in this set, so the frequency of 2 is eight.
3 appears five times in this set, so the frequency of 3 is five.
4 appears twice in this set, so the frequency of 2 is two.
the frequency of 0,1,2,3,4 occurring in the data set is given in the following table:

x	f
0	1
1	3
2	8
3	5
4	3

Definition 2.4.1. *The frequency f (from the second column) shows the number of times the value of the variable x appears in the data set.*

When we have a lot of distinct data in a set (instead of just a few, like in the previous example) we group the data in classes and then we build frequencies for these classes.

To illustrate this procedure, let us consider the following data set:

$$\begin{array}{lllll} 82 & 74 & 88 & 66 & 58 \\ 62 & 68 & 72 & 92 & 86 \\ 74 & 78 & 84 & 96 & 76 \\ 76 & 52 & 76 & 82 & 78 \end{array}$$

We will put in the same class all the data which have the same first digit and we obtain the following five classes:

$$50-59; \quad 60-69; \quad 70-79; \quad 80-89; \quad 90-99$$

($50-59$ is the class made of all the data having the first digit 5, a.s.o.).
These classes do not intersect (there aren't data belonging to two classes) and every data belongs to a class.
The lower class limits are $50, 60, 70, 80, 90$, and the upper class limits are $59, 69, 79, 89, 99$.
Data belonging to a class are greater than the lower class limit and smaller than its upper limit.

Definition 2.4.2. *The class width is the difference between the lower class limit and the next lower class limit (in the previous example it is equal to* 10 *and it is the same for all classes); the class width is not the difference between the lower and the upper limit of the same class.*

Definition 2.4.3. *The class boundary is the arithmetic mean of the upper class limit and the next lower class limit inferior (in our example they are:* $49,5$; $59,5$; $69,5$; $79,5$; $89,5$; $99,5$.*)*

Definition 2.4.4. *The class mark is the arithmetic mean of the upper and lower class limit.*

In our example they are:

$$54.5 = \frac{50+59}{2} \quad \text{for the class} \quad 50-59$$

$$64.5 = \frac{60+69}{2} \quad \text{for the class} \quad 60-69$$

$$74.5 = \frac{70+79}{2} \quad \text{for the class} \quad 70-79$$

$$84.5 = \frac{80+89}{2} \quad \text{for the class} \quad 80-89$$

$$94.5 = \frac{90+99}{2} \quad \text{for the class} \quad 90-99$$

In this case the frequency is the number of data in class. The grouped frequency of data is:

$$\text{for the cluster} \quad 50-59 \quad 2 \text{ data}$$

$$\text{for the cluster} \quad 60-69 \quad 3 \text{ data}$$

$$\text{for the cluster} \quad 70-79 \quad 8 \text{ data}$$

$$\text{for the cluster} \quad 80-89 \quad 5 \text{ data}$$

$$\text{for the cluster} \quad 90-99 \quad 2 \text{ data}$$

Generally, in order to group the data and establish the frequency, we have to keep the following rules:

1) The classes must not intersect and each data from the set must belong to a class;

2) All the classes must have the same width.

The grouping procedure implies:

i) An identification of the greatest data H and the smallest data L and the determination of $R = H - L$.

ii) A choice of the number of classes m and of the class width c (if possible, an odd number) so that the product $m \cdot c$ shall be a little bit bigger than R.

iii) The choice of a starting point I which is a little bit smaller than the smallest data L. We add to I the multiples of c (c is the class width) and we obtain the numbers:

$$I, I+c, I+2c, I+3c, ..., I+(m-1)c$$

These numbers are the lower class limits.

iv) The establishment of the upper class limits so that the conditions 1) and 2) hold.

v) Determining the frequency of each class by counting its elements.

2.5 Data presentation

The presentation of a data set can be made in different ways and is part of the primary processing of data.

Presenting the data as series

Definition 2.5.1. The frequency distribution is an ensemble of two finite sequences, the first representing the distinct elements from the data set or the classes obtained by grouping the elements from the statistical data, and the second, the corresponding frequencies.

Example 2.5.1. *For the statistical data set we have:*

$$\begin{array}{ccccc} 3 & 2 & 2 & 3 & 2 \\ 4 & 4 & 1 & 2 & 2 \\ 4 & 3 & 2 & 0 & 2 \\ 2 & 1 & 3 & 3 & 1 \end{array}$$

the frequency distribution is:

$$X\begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 3 & 8 & 5 & 3 \end{pmatrix}$$

Example 2.5.2. *For the classes* $50-59;\ 60-69;\ 70-79;\ 80-89;\ 90-99$ *obtained by grouping the data from the data set:*

$$\begin{array}{cccccccccc} 82 & 74 & 88 & 66 & 58 & 74 & 78 & 84 & 96 & 76 \\ 62 & 68 & 72 & 92 & 86 & 76 & 52 & 76 & 82 & 78 \end{array}$$

the frequency distribution is:

$$X\begin{pmatrix} 50-59 & 60-69 & 70-79 & 80-89 & 90-99 \\ 2 & 3 & 8 & 5 & 2 \end{pmatrix}$$

Generally, a frequency distribution looks like:

$$X\begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ f_1 & f_2 & f_3 & \cdots & f_n \end{pmatrix}$$

and whatever the level of data grouping, x_i with the frequency f_i, is called the frequency distribution term.

Remark 2.5.1. *When presenting the frequency distributions, instead of the frequency* f_i*, we often use the relative frequency:*

$$f_i' = \frac{f_i}{\sum_{j=1}^{n} f_j}$$

or, in percentaged form:

$$f_i'' = f_i' \cdot 100$$

Definition 2.5.2. The value of the piece of data that occurs with the greatest frequency in the frequency distribution, is called **mode**.

Definition 2.5.3. The class with the highest frequency in a frequency distribution, is called **modal class**.

Definition 2.5.4. A bimodal frequency distribution is a frequency distribution which has two highest frequency classes separated by classes with lower frequencies.

Definition 2.5.5. The cumulative frequency of a class is the sum of the frequencies of all classes with smaller marks.

Definition 2.5.6. The dynamical (temporal, chronological) distribution is a double sequence of which the first is the sequence of the values of the response variable, and the second is the sequence of time moments where the variable takes these values. Generally we denote a dynamical (temporal) distribution:

$$X\begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ t_1 & t_2 & t_3 & \cdots & t_n \end{pmatrix}$$

Presentation of data as statistical tables

Statistical tables are extremely variate and are used to arrange the statistical data from a data set in order to apply the statistical computation and interpretation methods.

According to the number of characteristics presented, we have simple tables, double entrance tables, group tables, etc.

The presentation of data in graphical form

There are several methods of graphical presentation of a statistical data set. The graphical presentation method is given by the data type and by the presentation idea. We have to make it clear from the beginning that there are several ways to represent statistical data in graphical form. The analyst's judgement and the circumstances of the problem have a major role in choosing the graphic display method.

Definition 2.5.7. The graphs representing the ungrouped frequency distributions, are called **diagrams**.

Definition 2.5.8. The pie chart of the (ungrouped) frequency distribution

$$X\begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ f_1 & f_2 & f_3 & \cdots & f_n \end{pmatrix}$$

is a circle dived in n circular sectors $S_1, S_2, ..., S_n$ so that the area of the sector S_i is equal to

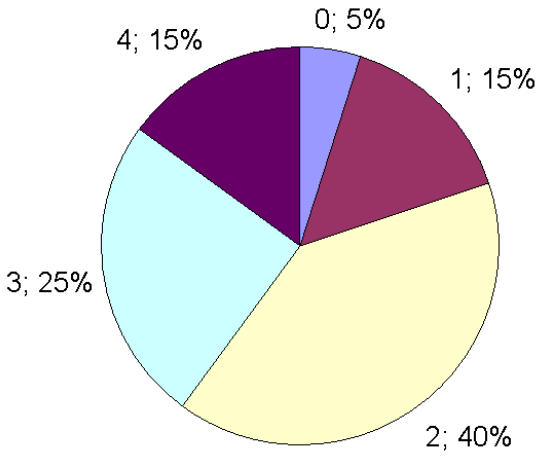
$$f_i'' = \frac{f_i}{\sum_{j=1}^{n} f_j} \cdot 100$$

percents of the circle area.

Example 2.5.3. *For the frequency distribution in example 2.5.1*

$$X\begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 3 & 8 & 5 & 3 \end{pmatrix}$$

the circle is divided in five sectors with areas equal to 5%, , 15%, 40%, 25%, 15% *of the circle area*



Definition 2.5.9. The bar chart of the (ungrouped) frequency distribution

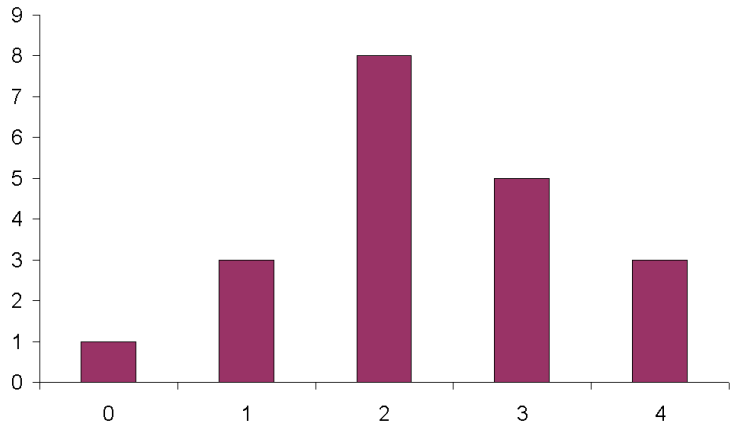
$$X\begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ f_1 & f_2 & f_3 & \cdots & f_n \end{pmatrix}$$

is a set of n rectangles. $f_1, f_2, ..., f_n$

Example 2.5.4. *For the frequency distribution from example 2.5.1:*

$$X\begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 3 & 8 & 5 & 3 \end{pmatrix}$$

the bar chart is:



Definition 2.5.10. The stem and leaf diagram of the (ungrouped) frequency distribution

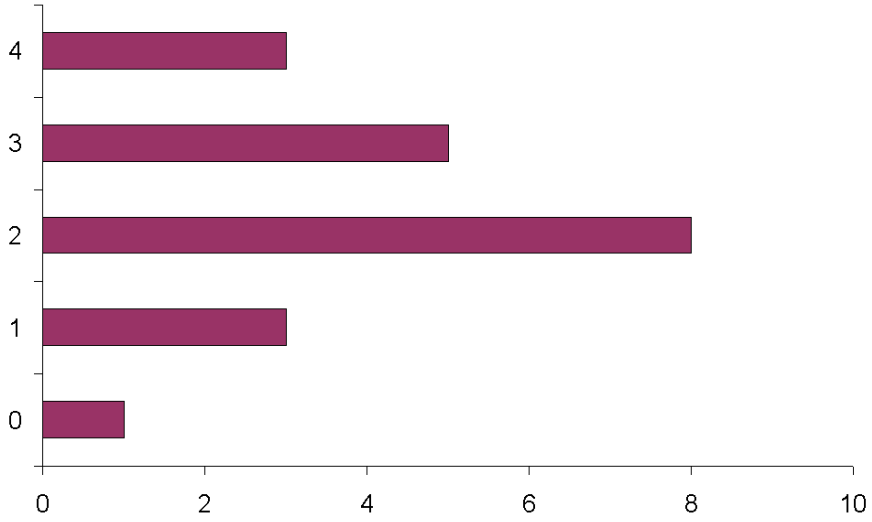
$$X\begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ f_1 & f_2 & f_3 & \cdots & f_n \end{pmatrix}$$

is a set of n rectangles. The basis of these rectangles are equal and they are situated on the Oy axis, their heights being $f_1, f_2, ..., f_n$.

Example 2.5.5. *For the frequency distribution in example 2.5.1:*

$$X\begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 3 & 8 & 5 & 3 \end{pmatrix}$$

the stem and leaf diagram is:



Definition 2.5.11. The histogram of the grouped frequency distribution

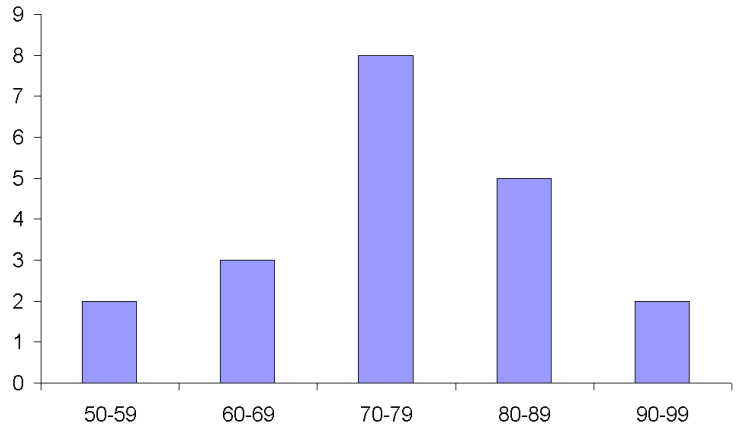
$$X\begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ f_1 & f_2 & f_3 & \cdots & f_n \end{pmatrix}$$

is a set of n rectangles representing the classes. The basis of these rectangles are equal (the classes have the same width) and are situated on the Ox axis, their heights being $f_1, f_2, ..., f_n$.

Example 2.5.6. *For the frequency distribution in example 2.5.2:*

$$X\begin{pmatrix} 50-59 & 60-69 & 70-79 & 80-89 & 90-99 \\ 2 & 3 & 8 & 5 & 2 \end{pmatrix}$$

the histogram is:



Remark 2.5.2. *Unlike a bar chart, for a histogram a column is a number of distinct data.*

Remark 2.5.3. *A histogram has the following components:*

i) A title, which identifies the population;

ii) A horizontal scale, which identifies the variable X, the values of the class limits, the class boundaries, the class marks.

iii) A vertical scale which identifies the frequencies for each class.

Definition 2.5.12. A relative frequency histogram is a histogram in which the frequencies are substituted by the relative frequencies.

The relative frequency (is a measure proportional to the given frequency) **is obtained by dividing the class frequency to the total number of elements from the data set**.

Definition 2.5.13. The ogive of a class frequency distribution with cumulative relative frequencies is a set of rectangles. The basis of the rectangles are equal and situated on the Ox axis, their heights being the cumulative relative frequencies.

The ogive has the following components:

1. A title, which identifies the population.
2. A horizontal scale which identifies the upper class boundaries.
3. A vertical scale which identifies the cumulative relative frequencies of each class.

2.6 Parameters and statistics of the central tendency

A category of numerical characteristics associated to a statistical data set are: **the central tendency parameters**, for the populations, and **the central tendency statistics**, for the samples. As they have analogous definitions, we will only present the central tendency statistics.

Definition 2.6.1. The statistics of central tendency are numerical values associated to a statistical data set, that locate in some sense the middle of this set of data.

Definition 2.6.2. The arithmetic mean of the set of statistical data $\{x_1, x_2, ..., x_n\}$ is the sum of these data, divided by the number of data

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

Remark 2.6.1. *When the data are given as a frequency distribution (without class grouping), the arithmetic mean is given by the formula:*

$$\overline{x} = \frac{\sum_{j=1}^{m} x_j \cdot f_j}{\sum_{j=1}^{m} f_j}$$

Remark 2.6.2. *For a frequency distribution (with class grouping) with compute the mean with the following formula:*

$$\overline{x} = \frac{\sum x \cdot f_x}{\sum f_x}$$

where x represents the class mark and f_x the corresponding frequency, and the sum is over the whole set of classes.

Definition 2.6.3. The root mean square of the statistical data set $\{x_1, x_2, ..., x_n\}$ is the number:

$$\overline{x}_p = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$$

Remark 2.6.3. *If the data are presented as a frequency distribution (without class grouping), the root mean square is computed with the formula:*

$$\overline{x}_p = \sqrt{\frac{\sum_{j=1}^{m} x_j^2 \cdot f_j}{\sum_{j=1}^{m} f_j}}$$

Remark 2.6.4. *For a frequency distribution with class grouping, the root mean square is given by the following relationship:*

$$\overline{x}_p = \sqrt{\frac{\sum x^2 \cdot f_x}{\sum f_x}}$$

where x represents the class mark and f_x the corresponding frequency, and the sum is over the whole set of classes.

Definition 2.6.4. The harmonic mean of the statistical data set $\{x_1, x_2, ..., x_n\}$ is the number:

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

Remark 2.6.5. *If the data are presented as a frequency distribution (without class grouping), the harmonic mean is computed with the formula:*

$$\overline{x}_h = \frac{\sum_{j=1}^{m} f_j}{\sum_{j=1}^{m} \frac{1}{x_j} \cdot f_j}$$

Remark 2.6.6. *For a frequency distribution with class grouping, the harmonic mean is given by the following relationship:*

$$\overline{x}_h = \frac{\sum_{i=1}^{n} f_x}{\sum_{i=1}^{n} \frac{1}{x} \cdot f_x}$$

where x represents the class mark and f_x the corresponding frequency, and the sum is over the whole set of classes.

Definition 2.6.5. The geometric mean of the statistical data set $\{x_1, x_2, ..., x_n\}$ is the number:

$$\overline{x}_p = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Remark 2.6.7. *If the data are presented as a frequency distribution (without class grouping), the geometric mean is computed with the formula:*

$$\overline{x}_g = \sum_{j=1}^{m} f_j \sqrt{\prod_{j=1}^{n} x_j^{f_j}}$$

Remark 2.6.8. *For a frequency distribution with class grouping, the geometric mean is given by the following relationship:*

$$\overline{x}_g = \sqrt[\Sigma f_x]{\prod x^{f_x}}$$

where x represents the class mark and f_x the corresponding frequency, and the sum is over the whole set of classes.

Definition 2.6.6. ***Median*** *m_e of a distinct statistical data set where the data are ranked in order according to size $x_1 < x_2 < ... < x_n$ is the number that divides the data set in two sets of the same size:*

- *if $n = 2 \cdot k + 1$, then m_e is the value with the rank $k + 1$: $m_e = x_{k+1}$;*
- *if $n = 2 \cdot k$, then any number between the values x_k and x_{k+1} satisfies the condition from the definition of m_e. In this case we agree that m_e is the arithmetic mean of x_k and x_{k+1}: $m_e = \frac{x_k + x_{k+1}}{2}$.*

Example 2.6.1. *For the statistical data set:*

$$4 \quad 7 \quad 12 \quad 26 \quad 32 \quad 38 \quad 59$$

the median is $m_e = 26$.
For the statistical data set:

$$4 \quad 7 \quad 12 \quad 26 \quad 32 \quad 38$$

the median is $m_e = \frac{12 + 26}{2} = 19$.

Remark 2.6.9. In this case, the median m_e has the property that the frequencies sum of the values greater than m_e is equal to the frequencies sum of the values smaller than m_e.

Remark 2.6.10. *If the data are not necessarily distinct, the property from the Observation 2.6.9 may not be. For the set of data:*

$$1 \quad 1 \quad 1 \quad 2 \quad 3 \quad 3 \quad 4$$

The corresponding frequency distribution is:

1	2	3	4
3	1	2	1

According to the definition of m_e, we have $m_e = 2,5$. This value of m_e does not satisfy the request that the values greater or smaller than it have equal cumulative frequencies; The frequency of the smaller ones is 4, and the frequency of the others is 3.

Remark 2.6.11. *When the data are presented as a (grouped or ungrouped) frequency distribution, m_e is computed by linear interpolation, based on the uniform distribution of frequencies in the median interval hypothesis.*

Definition 2.6.7. The midrange of the sample is, by definition, the number:

$$M_r = \frac{L + H}{2}$$

where L is the lowest value and H is the highest value of the variable X

2.7 Parameters and statistics of dispersion

After establishing the midrange of a set of data, the next natural question is: what are the parameters and the statistics that characterize the dispersion of the data.

The parameters and the statistics of dispersion are: the range, the mean absolute deviation, the variance, the standard deviation and the variation coefficient. These numerical values describe the size of the spread or of the variability of the data. Closely grouped data will have a smaller spread, while more widely spread-out data will have a larger dispersion.

Definition 2.7.1. The range P is the difference between the highest-valued (H) and the lowest-valued (L) pieces of data:

$$P = H - L$$

The mean absolute deviation, the variance and the standard deviation measure the dispersion about the arithmetic mean.

Definition 2.7.2. The deviation from the mean $\overline{x}$ of the value x_i of the variable X is $d_i = x_i - \overline{x}$.

The deviation is zero if and only if $x_i = \overline{x}$.
The deviation is positive if and only if $x_i > \overline{x}$.
The deviation is negative if and only if $x_i < \overline{x}$.

One might believe that the sum of the deviations $\sum_{i=1}^{n}(x_i - \overline{x})$ might serve as a measurement of dispersion about the mean. But this sum is always zero:

$$\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - n \cdot \overline{x} = n \cdot \overline{x} - n \cdot \overline{x} = 0$$

To reduce the cancelling effect we will use the the absolute value of these deviations: $x_i - \overline{x}$.

Definition 2.7.3. The mean absolute deviation of the set of distinct statistical data $\{x_1, x_2, ..., x_n\}$ is, by definition:

$$\overline{d} = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$$

Remark 2.7.1. *We calculate the mean absolute deviation, for data presented as a frequency distribution, with the formula:*

$$\overline{d} = \frac{\sum_{j=1}^{m} |x_j - \overline{x}| \cdot f_j}{\sum_{j=1}^{m} f_j}$$

Remark 2.7.2. *We calculate the mean absolute deviation, for data presented as a grouped frequency distribution, with the formula:*

$$\overline{d} = \frac{\sum |x - \overline{x}| \cdot f_x}{\sum f_x}$$

where x represents the class mark and f_x the corresponding frequency, and the sum is over the whole set of classes.

Even if we do not frequently use this dispersion parameter, it is a dispersion measurement and it shows the mean distance from value of the variable X to the arithmetic mean.
There is another way of reducing the cancelling effect of the deviations. Squaring the deviations will cause all these values to be positive (or zero). When we add these squares, the result will be a positive number. The sum of the deviation squares about the arithmetic mean $\sum_{i=1}^{n}(x_i-\overline{x})^2$ is used in defining the variance.

Definition 2.7.4. The variance s^2 of the set of distinct statistical data $\{x_1, x_2, ..., x_n\}$ is, by definition:

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

Remark 2.7.3. *If the data set is presented as an ungrouped frequency distribution, we compute the variance s^2 with the formula:*

$$s^2 = \frac{\sum_{j=1}^{m}(x_j - \overline{x})^2 \cdot f_j}{\sum_{j=1}^{m} f_j}$$

Remark 2.7.4. *If the data set is presented as an grouped frequency distribution, we compute the variance s^2 with the formula:*

$$s^2 = \frac{\sum(x - \overline{x})^2 \cdot f_x}{\sum f_x}$$

where x represents the class mark and f_x the corresponding frequency, and the sum is over the whole set of classes.

Definition 2.7.5. The standard deviation s of the set of distinct statistical data $\{x_1, x_2, ..., x_n\}$ is, by definition:

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

Remark 2.7.5. *If the data set is presented as an ungrouped frequency distribution, we compute the variance s with the formula:*

$$s = \left[\frac{\sum_{j=1}^{m}(x_j - \overline{x})^2 \cdot f_j}{\sum_{j=1}^{m} f_j}\right]^{\frac{1}{2}}$$

Remark 2.7.6. *If the data set is presented as an grouped frequency distribution, we compute the variance s with the formula:*

$$s = \left[\frac{\sum(x - \overline{x})^2 \cdot f_x}{\sum f_x}\right]^{\frac{1}{2}}$$

where x represents the class mark and f_x the corresponding frequency, and the sum is over the whole set of classes.

Remark 2.7.7. We have defined the standard deviation by a formula. So the question arises what it actually represents. An answer to this question can be given with Chebyshev inequality from which we have that for every frequency distribution, the fraction of data situated at at most k units from the standard deviation is at least $1 - \frac{1}{k^2}$, where k is a arbitrary positive number, greater than 1. In particular we have that, for every frequency distribution the fraction situated at at most $k = 2$ units from the standard deviation represents at least 75% of the entire data. If $k = 3$ then it represents 89% of the data.
According to the empirical rule that if a frequency distribution is normal, then the fraction of data situated at at most one unit from the standard deviation σ is about 68% about the mean, and the fraction of data situated at at most two units from the standard deviation σ is about 95%about the mean.

Definition 2.7.6. The variation coefficient V is by definition:

$$V = \frac{s}{\overline{x}} \cdot 100$$

Remark 2.7.8. *The variation coefficient is a relative statistic of the dispersion and is used when comparing the dispersion of different variables (characteristics).*

Remark 2.7.9. *V can take values between 0 and 100%. If V is close to zero ($V < 35\%$), then we have a homogenous population and the mean $\overline{x}$ is representative for this population. If V is close to 100% ($V > 75\%$), then we have a heterogenous population and the mean $\overline{x}$ is not representative. Most of the times, in these cases, we need to separate the population in several homogenous groups which are separately studied.*

2.8 Factorial parameters and statistics of the variance

When analyzing the variance of a set of data we use the following factorial parameters of the variance:

- group variance (partial) s_j^2
- mean group variance $\overline{s}^2$
- the variance of the group means about the general mean δ^2
- total variance s^2.

Definition 2.8.1. For a group of m data $x_1, x_2, ..., x_m$, **the group variance** is defined by the formula:

$$s_j^2 = \frac{\sum_{i=1}^{m} (x_i - \overline{x}_j)^2 \cdot n_{ij}}{\sum_{i=1}^{m} n_{ij}}$$

where j is the group index, $\overline{x}_j$ is the group mean, x_i are the data from the j-th group having the frequencies n_{ij}

Remark 2.8.1. *The group variances are smaller than the variance and have lower or higher values depending on the group's heterogeneity.*

Definition 2.8.2. By definition **the mean group variance** is:

$$\overline{s}^2 = \frac{\sum_{j=1}^{k} s_j^2 \cdot n_j}{\sum_{j=1}^{k} n_j}$$

where k is the number of groups, $n_j = \sum_{i=1}^{m} n_{ij}$ is the number of data from a group.

Definition 2.8.3. The variance of the group means about the general mean is, by definition:

$$\delta^2 = \frac{\sum_{j=1}^{k} (\overline{x}_j - \overline{x})^2 \cdot n_j}{\sum_{j=1}^{k} n_j}$$

Comment 2.8.1. For a frequency distribution with 2-dimensional data we define **the covariance**:

$$cov(X,Y) = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{k} (x_i - \overline{x}) \cdot (y_j - \overline{y}) \cdot n_{ij}$$

where: x_i are the values of the variable X, y_i are the values of the variable Y, n_{ij} is the frequency of the pair (x_i, y_j) and $n = \sum^{i=1} m \sum_{j=1}^{k} n_{ij}$, $\overline{x}$ is the mean of the variable X and $\overline{y}$ is the mean of the variable Y. The covariance tells us upon what degree the variables X, Y tend to vary concomitantly and it is imprudent to interpret it in terms of causality. High values of the covariance may come from:

- an immediate link between phenomena: work productivity - qualification of workers; work productivity - technical equipment; request of wares - monetary incomes; national income - accumulation; production - request; productivity - remuneration.
- covariation of phenomena due to certain causes: the covariance of the request of wares and of the monetary savings is determined by the monetary income of the population.

2.9 Parameters and statistics of position

The parameters and statistics of position are used to describe the location of a specific piece of data in relation to the rest of the sample.

Definition 2.9.1. The quantiles are numerical values which divide the set of data in q equal groups. q is called the order of the quantile.

The median is the second order quantile.
The forth order quantiles divide the set of data in four equal groups; they are called **quartiles**. There are three quartiles, usually denoted Q_1, Q_2, Q_3.
The Q_1 quartile is a number with the property that one quarter of the data have values less than Q_1 and three quarters have values greater than Q_1.

The Q_2 quartile is a number with the property that a half of the data have values less than Q_2 and half of them have values greater than Q_2. The Q_2 quartile is actually the median.
The Q_3 quartile is a number with the property that three quarters of the data have values less than Q_3 and the other quarter have values greater than Q_3.
Other categories of common quantiles are:

- the deciles which divide the set of data in 10 equal groups.
- the percentiles which divide the set of data in 100 equal groups.
- the permillages which divide the set of data in 1000 equal groups.

Every set of data has 99 percentiles $P_k, k = 1..99$. P_k is a numerical value having the property that $k\%$ of the data have values smaller than P_k, and $(100-k)\%$ of the data have values greater than P_k.

Remark 2.9.1. $Q_1 = P_{25}$; $Q_3 = P_{75}$; $m_e = Q_2 = P_{50}$

Remark 2.9.2. *The procedure to determine the percentile P_k is as follows:*

1) we order the data increasingly;

2) we find the position i of the percentile k. First, we determine the number $\frac{n \cdot k}{100}$, where n is the number of data. If $\frac{n \cdot k}{100}$ is not an integer, then i is the next integer ($\frac{n \cdot k}{100} = 17.2 \rightarrow i = 18$). If $\frac{n \cdot k}{100}$ is an inter, then i is $\frac{n \cdot k}{100} + 0.5$ ($\frac{n \cdot k}{100} = 23 \rightarrow i = 23.5$).

3) we locate the value P_k: we count from L (the lowest value) i values, if i is an integer. If i is not an integer, then it is and integer and a half. In this case P_k is the half sum of the $\frac{n \cdot k}{100}$-th and the $\frac{n \cdot k}{100} + 1$-th data.

Another measure of position is the z or standard score.

Definition 2.9.2. The z or **standard score** is the position of the value x in terms of the number of standard deviations it is located from the mean $\overline{x}$:

$$z = \frac{x - \overline{x}}{s}$$

2.10 The sampling distribution of the sample statistics

In order to make predictions upon the population parameters we need to analyze the sampling statistics. The mean $\overline{x}$ for a sample is not necessarily equal to the mean μ of the population. But we are satisfied if $\overline{x}$ is close to μ. If we consider $\overline{x}\prime$ for a second sample, it may differ from $\overline{x}$ and μ. What we can hope is that it is close to μ and $\overline{x}$. We are interested in the validity of this type of behavior for any kind of population and statistic.

The question that naturally arises is about the meaning of "close". How do we measure and determine this closeness? What is the sampling distribution of the sample statistics?

Definition 2.10.1. The sampling distribution of the sample statistics is the frequency distribution of the statistics of a certain type obtained for all the samples of the same size. The type can be any of those presented in sections 6 and 7.

Example 2.10.1. *We consider a population of N elements from which we can obtain the following distinct data: $\{0, 2, 4, 6, 8\}$. For this population we form samples of size 2:*

$$\begin{array}{ccccc} (0,0) & (2,0) & (4,0) & (6,0) & (8,0) \\ (0,2) & (2,2) & (4,2) & (6,2) & (8,2) \\ (0,4) & (2,4) & (4,4) & (6,4) & (8,4) \\ (0,6) & (2,6) & (4,6) & (6,6) & (8,6) \\ (0,8) & (2,8) & (4,8) & (6,8) & (8,8) \end{array}$$

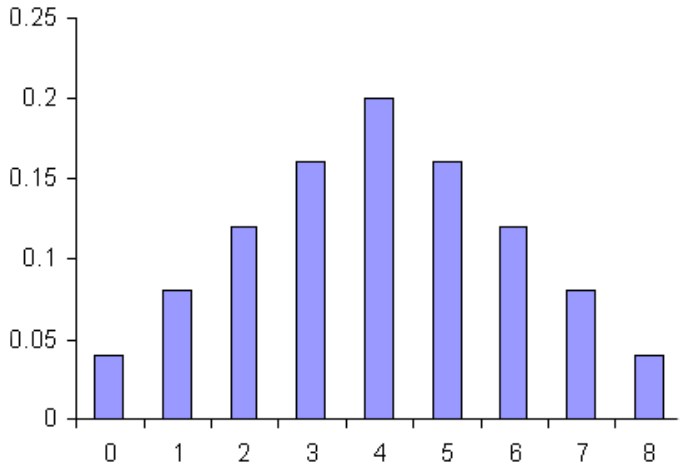
For these samples we have the means:

$$\begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 4 & 5 & 6 \\ 3 & 4 & 5 & 6 & 7 \\ 4 & 5 & 6 & 7 & 8 \end{array}$$

Because the samples are random, each one is chosen with the probability and the sampling distribution of their means is:

$\overline{x}$	$f'(\overline{x})$
0	*0.04*
1	*0.08*
2	*0.12*
3	*0.16*
4	*0.20*
5	*0.16*
6	*0.12*
7	*0.08*
8	*0.04*

where $f'(\overline{x})$ is the relative frequency of the mean $\overline{x}$. The bar chart of the sample means is:



For the same set of 25 samples we can determine the sampling distribution of the ranges R.

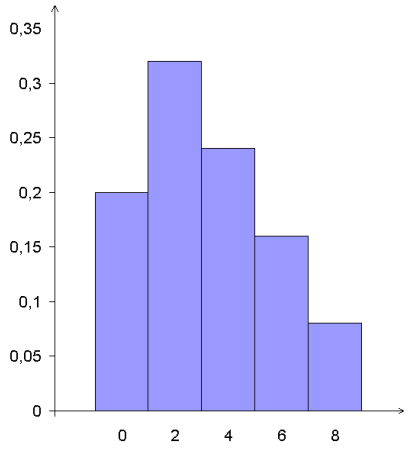
The ranges R of these samples are given in the following table:

0	2	4	6	8
2	0	2	4	6
4	2	0	2	4
6	4	2	0	2
8	6	4	2	0

The sampling distribution of the sample ranges is:

R	$f'(R)$
0	0.20
2	0.32
4	0.24
6	0.16
8	0.08

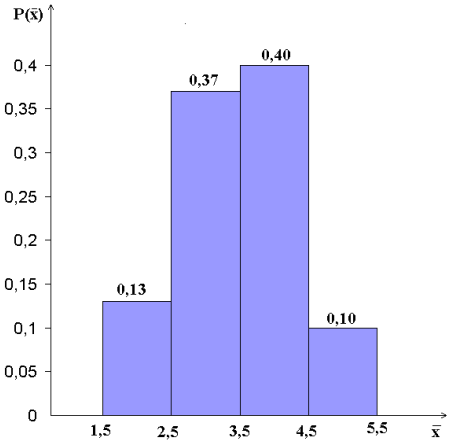
and the bar chart of the sample range is:



Example 2.10.2. *When rolling a die N times, the set of data giving the number on the appearing side is* $1, 2, 3, 4, 5, 6$.
We form samples consisting of 5 throws. Each of these samples has the mean $\overline{x}$. *We consider* 30 *such samples (that means* $30 \times 5 = 150$ *throws) and we represent the results and the corresponding means:*

Throw	*Sample*	$\overline{x}$	*Throw*	*Sample*	$\overline{x}$
1	*1 2 3 2 2*	*2.0*	*16*	*5 2 1 3 5*	*3.2*
2	*4 5 5 4 5*	*4.6*	*17*	*6 1 3 3 5*	*3.6*
3	*3 1 5 2 4*	*3.0*	*18*	*6 5 5 2 6*	*4.8*
4	*5 6 6 4 2*	*4.6*	*19*	*1 3 5 5 6*	*4.0*
5	*5 4 1 6 4*	*4.0*	*20*	*3 1 5 3 1*	*2.6*
6	*3 5 6 1 5*	*4.0*	*21*	*5 1 1 4 3*	*2.8*
7	*2 3 6 3 2*	*3.2*	*22*	*4 6 3 1 2*	*3.2*
8	*5 3 4 6 2*	*4.0*	*23*	*1 5 3 4 5*	*3.6*
9	*1 5 5 3 4*	*3.6*	*24*	*3 4 1 3 3*	*2.8*
10	*4 1 5 2 6*	*3.6*	*25*	*1 2 4 1 4*	*2.4*
11	*5 1 3 3 2*	*2.8*	*26*	*5 2 1 6 3*	*3.4*
12	*1 5 2 3 1*	*2.4*	*27*	*4 2 5 6 3*	*4.0*
13	*2 1 1 5 3*	*2.4*	*28*	*4 3 1 3 4*	*3.0*
14	*5 1 4 4 6*	*4.0*	*29*	*2 6 5 3 3*	*3.8*
15	*5 5 6 3 3*	*4.4*	*30*	*6 3 5 1 1*	*3.2*

The histogram of the sampling distribution of the means of the 30 samples is presented in the following figure:



This distribution law seems to have the characteristics of a normal distribution; there is a maximum and a symmetry with respect to its own mean 3.5.

2.11 The central limit theorem

In the preceding section we discussed the sample distribution of sample means and of sample ranges. The mean is the most commonly used sample statistic and thus it is very important. The central limit theorem is about the sampling distribution of sample means of random samples of size n.

Let us establish what we are interested in when studying this distribution:

1) Where is the center?
2) How wide is the dispersion?
3) What are the characteristics of the distribution?

The central limit theorem gives us an answer to all these questions.

Theorem 2.11.1. The central limit theorem
Let μ be the mean and σ the standard deviation of a population variable. If we consider all possible random sample of size n taken from this population, the sampling distribution of sample means will have the following properties:

a) the mean $\mu_{\overline{x}}$ of this sampling distribution is μ;
b) the standard deviation $\sigma_{\overline{x}}$ of this sampling distribution is $\dfrac{\sigma}{\sqrt{n}}$;
c) if the parent population is normally distributed the sampling distribution of the sample means is normal;if the parent population is not normally distributed, the sampling distribution of the sample means is approximately normal for samples of size 30 or more. The approximation to the normal distribution improves with samples of larger size.

In short, the central limit theorem states the following:

1) $\mu_{\overline{x}} = \mu$, where $\overline{x}$ is the mean of the sample x;

2) $\sigma_{\overline{x}} = \sigma/\sqrt{n}$, the standard deviation of the mean is equal to the standard deviation of the population divided by the square root of the sample size.

3) the sample distribution of the sample means is approximatively normal regardless of the shape of the parent population.

Remark 2.11.1. *The standard deviation $\sigma_{\overline{x}}$ of the sampling distribution of the sample means is the standard deviation of the sample means to the mean of the sampling distribution.*
We won't prove the central limit theorem, but we will illustrate its validity by examining an illustration.
Let us consider a population whose relative frequency distribution for the variable X is:

$$X : \begin{pmatrix} 2 & 4 & 6 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}$$

The mean μ and the standard deviation σ for this variable, are:

$$\mu = \sum_{j=1}^{3} x_j \cdot f'_{x_j} \qquad \sigma = \sqrt{\sum_{j=1}^{3} x_j^2 \cdot f'_{x_j} - \left(\sum_{j=1}^{3} x_j \cdot f'_{x_j}\right)^2}$$

$$\mu = \frac{12}{3} = 4 \qquad \sigma = 1,63$$

For this population, each sample of size 2 has the following possible data:

$$\begin{array}{ccc} (2,2) & (2,4) & (2,6) \\ (4,2) & (4,4) & (4,6) \\ (6,2) & (6,4) & (6,6) \end{array}$$

The samples have the following means:

$$\begin{array}{ccc} 2 & 3 & 4 \\ 3 & 4 & 5 \\ 4 & 5 & 6 \end{array}$$

Sample	*Mean*
(2,2)	*2*
(2,4)	*3*
(2,6)	*4*
(4,2)	*3*
(4,4)	*4*
(4,6)	*5*
(6,2)	*4*
(6,4)	*5*
(6,6)	*6*

Because they are random, each sample is chosen with the probability $\frac{1}{9}$ and the sampling distribution of the sample means is:

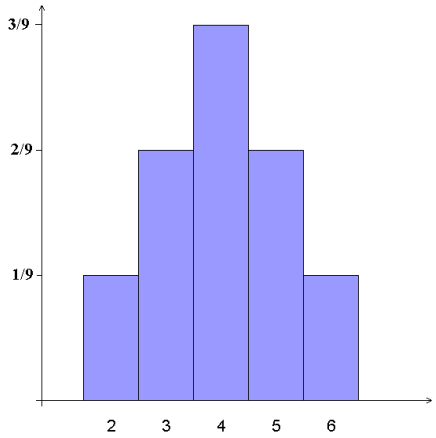
$$\overline{X} \begin{pmatrix} 2 & 3 & 4 & 5 & 6 \\ 1/9 & 2/9 & 3/9 & 2/9 & 1/9 \end{pmatrix}$$

The mean of the sampling distribution of the sample means $\mu_{\overline{x}}$ is $\mu_{\overline{x}} = 36/9 = 4,0$. So, $\mu = \mu_{\overline{x}}$, and the standard deviation of the sampling means distributions is:

$$\sigma_{\overline{x}} = \sqrt{\sum_{j=1}^{5} x_j^2 \cdot f'_{x_j} - \left(\sum_{j=1}^{5} x_j \cdot f'_{x_j}\right)^2} = \sqrt{\frac{156}{9} - \left(\frac{36}{9}\right)^2} = 1,15$$

$$\frac{\sigma}{\sqrt{n}} = \frac{1,63}{\sqrt{2}} = \frac{1,63}{1,44} = 1,15 = \sigma_{\overline{x}}$$

Representing the sapling distribution of the sample means we get:



This diagram shows us that the sampling distribution of the sample means is normal.

2.12 An application of the central limit theorem

The central limit theorem tells us about the sampling distribution of the sample means by describing the shape of the distribution of all sample means (almost normal). It establishes the relationship between the mean μ of the population and the mean $\mu_{\overline{x}}$ of the sampling distribution, and the relationship between the standard deviation of the population, σ, and the standard deviation $\sigma_{\overline{x}}$ of the sampling distribution. Since sampling means are approximatively normally distributed, we can establish probabilistic links between the population mean and a sample mean.

Example 2.12.1. *Consider a normal population with $\mu = 100$ and $\sigma = 20$. If we choose a random sample of size $n = 16$, what is the probability that the mean value of this sample is between 90 and 110? In other words, what is $P(90 < \overline{x} < 110)$?*

Solution: *According to the central limit theorem, the sampling distribution of the sample means is normal. So we will have to transform the condition $P(90 < \overline{x} < 110)$ in a condition that allows us to use the standard normal distribution table. We do this by writing:*

$$P(90 < \overline{x} < 110) = \Phi\left(\frac{110 - \mu_{\overline{x}}}{\sigma_{\overline{x}}}\right) - \Phi\left(\frac{90 - \mu_{\overline{x}}}{\sigma_{\overline{x}}}\right) =$$

$$= \Phi\left(\frac{110 - 100}{\sigma_{\overline{x}}}\right) - \Phi\left(\frac{-10}{\sigma_{\overline{x}}}\right) = 2 \cdot \Phi\left(\frac{10}{\sigma_{\overline{x}}}\right) - 1 = F\left(\frac{10}{\sigma_{\overline{x}}}\right)$$

where $\Phi(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X} e^{-\frac{1}{2}t^2}\, dt$ *and* $F(X) = \Phi(X) - \frac{1}{2}$.

Because $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$, *we have* $\sigma_{\overline{x}} = \frac{20}{\sqrt{16}} = 5$ *and we obtain:*

$$P(90 < \overline{x} < 110) = 2 \cdot \Phi(2) - 1 = 2F(2) = 0.9544$$

Increasing the size of the sample, n, does not affect $\mu_{\overline{x}} = \mu$ *but decreases* $\sigma_{\overline{x}}$. *So* $P(90 < \overline{x} < 110)$ *increases if n increases.*

Example 2.12.2. *Kindergarten children have heights that are normally distributed about a mean of* $\mu = 100$ *cm and a standard deviation of* $12,5$ *cm. We determine the mean* $\overline{x}$ *for a random sample of 25 children. What is the probability that this mean value is between 90 and 110 cm?*

Solution:

$$P(90 < \overline{x} < 110) = 2 \cdot \Phi\left(\frac{10}{\sigma_{\overline{x}}}\right) - 1 = 2 \cdot \Phi(4) - 1 = 2 \cdot F(4) = 2 \cdot 0.499968$$

2.13 Point estimation for a parameter

Let us consider a population whose mean μ is unknown and the problem of finding this mean. For this purpose, we consider a random sample of size n for which we determine the mean $\overline{x}$. The sample mean $\overline{x}$ is a point estimation of the population mean μ.

Definition 2.13.1. A point estimation of the parameter γ of a population is a value g of the corresponding statistic.

Remark 2.13.1. *If* $\overline{x}$ *is the sample mean used to estimate the unknown mean* μ *of the population, it doesn't imply that* $\overline{x} = \mu$. *In general,* $\overline{x} \neq \mu$ *and we expect that* $\overline{x}$ *is close to* μ. *This closeness can be fixed by the specification of an interval (centered in* μ*) called interval estimate.*

Definition 2.13.2. A bounded interval (a, b) used to estimate the value of a certain parameter γ of a population, is called **interval estimate**. The values a, b (the boundary of the interval) are calculated from the sample that is used as a basis for the estimation.

The way of determining an unknown interval centered in μ using only data provided by a sample is going to be described in the following.

Example 2.13.1. *Consider a population with the known standard deviation* σ, *an unknown mean* μ *and a simple random sample of size n and mean* $\overline{x}$, *both known. The condition* $\overline{x} \in (\mu - 1, \mu + 1)$ *means that the standard score z given by:*

$$z = \frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

satisfies

$$z \in (-\frac{1}{\frac{\sigma}{\sqrt{n}}}, \frac{1}{\frac{\sigma}{\sqrt{n}}}) = (-\frac{\sqrt{n}}{\sigma}, \frac{\sqrt{n}}{\sigma})$$

So, in terms of standard score, the interval estimate is (a, b) *with* $a = -\frac{\sqrt{n}}{\sigma}$ *and* $b = \frac{\sqrt{n}}{\sigma}$.
More generally, the condition $\overline{x} \in (\mu - \delta, \mu + \delta)$, *means that the standard score z given by:*

$$z = \frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

satisfies:

$$z \in (-\frac{\delta \cdot \sqrt{n}}{\sigma}, \frac{\delta \cdot \sqrt{n}}{\sigma})$$

The interval estimate *is* $(-\frac{\delta \cdot \sqrt{n}}{\sigma}, \frac{\delta \cdot \sqrt{n}}{\sigma})$.

Definition 2.13.3. The level of confidence α is the probability that the sample statistic has the value outside the interval estimate.

Example 2.13.2. *In the example 2.13.1,* $\overline{x}$ *is normally or approximatively normally distributed and we have:*

$$P(\mu - 1 < \overline{x} < \mu + 1) = P\left(-\frac{\sqrt{n}}{\sigma} < z < \frac{\sqrt{n}}{\sigma}\right) =$$

$$2 \cdot P\left(0 < z < \frac{\sqrt{n}}{\sigma}\right) = 2 \cdot F\left(\frac{\sqrt{n}}{\sigma}\right)$$

where $F(z) = \frac{1}{\sqrt{2 \cdot \pi}} \int_0^z e^{-\frac{1}{2}t^2}\, dt$.

So the level of confidence α *is* $1 - 2 \cdot F\left(\frac{\sqrt{n}}{\sigma}\right)$.

Definition 2.13.4. The level of confidence (confidence coefficient) $1 - \alpha$ is the probability that the sample statistic is in the chosen interval estimate.

Definition 2.13.5. The confidence interval is an interval estimate with a specified level of confidence $1 - \alpha$.

Example 2.13.3. *For the example 2.13.1, the interval estimate* $\left(-\frac{\sqrt{n}}{\sigma}, \frac{\sqrt{n}}{\sigma}\right)$ *is a confidence interval with the confidence coefficient* $1 - \alpha = 2 \cdot F\left(\frac{\sqrt{n}}{\sigma}\right)$.

Definition 2.13.6. The maximum error of estimate is one-half the width of the confidence interval with the confidence coefficient $1 - \alpha$.

In terms of standard score, this error is:

$$E = \overline{z}\left(\frac{\alpha}{2}\right) \cdot \frac{\sigma}{\sqrt{n}}$$

where $\overline{z}\left(\frac{\alpha}{2}\right)$ is the solution of the equation $F(\overline{z}) = \frac{\alpha}{2}$, and the confidence interval $1 - \alpha$ for μ is:

$$\left(\overline{x} - \overline{z}\left(\frac{\alpha}{2}\right) \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + \overline{z}\left(\frac{\alpha}{2}\right) \cdot \frac{\sigma}{\sqrt{n}}\right)$$

$\overline{x} - \overline{z}\left(\frac{\alpha}{2}\right) \cdot \frac{\sigma}{\sqrt{n}}$ is the lower confidence limit and $\overline{x} + \overline{z}\left(\frac{\alpha}{2}\right) \cdot \frac{\sigma}{\sqrt{n}}$ is the upper confidence limit.

2.14 Generalities regarding the problem of hypothesis testing

To illustrate the analysis that precedes the decision making regarding the credibility of an assertion (called hypothesis test) let us consider the following example:
The admission candidate Popescu Nicolae has to fill in a form test with ten questions. Each

question has five answers of which only one is correct. Popescu Nicolae has completed the form and answered correctly to seven of those questions. He claims to have filled in the form without reading the questions and their answers and that he randomly marked the answers.
The question is to what extent can we believe that he randomly marked the answers?
Such a question determines us to analyze and decide: is it or is it not reasonable that Popescu Nicolae obtains seven correct by answers choosing randomly the answers? In the following we describe an analysis, called the hypothesis test, which leads to the formulation of a conclusion.
The hypothesis testing is, in general, a procedure which consists of 5 steps. Each of these steps will be presented and illustrates for the considered example.

Step 1. **Formulation of the null hypothesis H_0**
A hypothesis is a statement that something is true. In general, the null hypothesis is a statement relative to a population parameter, and it states that the parameter has a given value. Often the phrase "there is no difference" is used in its interpretation, thus the name "null hypothesis" (the difference is zero).

Step 2. **Formulation of the alternative hypothesis H_a**
The alternative hypothesis H_a is a statement about the same parameter that is used in the null hypothesis H_0. In H_a it is specified that the parameter has a value different from the value given in H_0.

The hypothesis H_0 and the hypothesis H_a are formulated after an analysis of the assertion that is to be investigated.
For the considered example:
The population is a set of 5^{10} (distinct) elements. An element is an ordered system of 10 answers $(R'_{i_1}, R'_{i_2}, \ldots, R'_{i_{10}})$, $i_1, i_1, \ldots, i_{10} \in \{1,2,3,4,5\}$; R'_{i_1} is one of the five possible answers to the first question, $\ldots$, $R'_{i_{10}}$ is one of the five possible answers to the tenth question.
For a person that marks the answers randomly (without reading them), all the answers are equally possible. In other words all of the five possible answers for a question have the same chance to be correct. From the statement of Popescu Nicolae it follows that he randomly marked the answers, so he admitted that the probability (the parameter p) for each element is $\frac{1}{5^{10}}$.
The analysis of the statement of Popescu Nicolae leads to the following formulation of the null hypothesis:

$$H_0 : p(X) = \frac{1}{5^{10}} = p \text{ for every element } X \text{ of the population} \quad \Leftrightarrow \quad \text{Popescu Nicolae has completed the form randomly.}$$

The alternative hypothesis is:

$$H_a : \text{there are two elements } X_1,\ X_2 \text{ in the population for which } p(X_1) \neq p(X_2) \quad \Leftrightarrow \quad \text{Popescu Nicolae hasn't completed the form randomly}$$

Starting from this point we admit that the null hypothesis is true. This situation can be compared with a trial, where the accused is supposed to be innocent until the contrary is proved.
We will make one of the two possible decisions only in the 5-th step of the hypothesis testing: we decide agreeing to the null hypothesis and we say that we accept H_0 or we decide agreeing to the alternative hypothesis and we say that we reject H_0.
Depending on the truth value of H_0 and on its rejection or its failure to reject, we present the decisions that are made in the following table:

Decision	The null hypothesis H_0 is	
	True	False
Fail to reject H_0 (we accept it)	correct decision	error
	Type A	Type II
Reject H_0	error	correct decision
	Type I	Type B

A type A correct decision: occurs when H_0 is true and we decide in its favor
A type B correct decision: occurs when H_0 is false and our decision is in opposition to H_0
A type I error: occurs when H_0 is true and H_0 is rejected
A type II error: occurs when H_0 is false but we fail to reject it

It would be nice that each time we make a decision, it were a correct one, but this is statistically impossible because they are based on sample information. The best we can hope for is to control the risk and the probability with which an error occurs.
The probability of the type I error is called α and the one of the type II error, β:

Error	Error type	Probability
Rejection of a true hypothesis	I	α
Acceptance of a false hypothesis	II	β

Step 3 **Determining the test criteria**: it consists of (1) the identification of a statistic test; (2) the specification of α; (3) the determination of the critical region.
(1) A test statistic is a random variable whose value is used in rejecting or accepting H_0. The test statistic is a sample statistic or some other value obtained from the sample results. The probabilities that appear in this test statistic are determined supposing that H_0 is true.

In the considered example, the random variable "X= the number of correct answers" is used as a test statistic. The probabilities for each value x of the variable X supposing that H_0 is true, are given in the following table:

X	0	1	2	3	4	5
P(X)	0.1074	0.2684	0.302	0.20133	0.0881	0.0264
X	6	7	8	9	10	
P(X)	0.0055	$7.92 \cdot 10^{-4}$	$7.38 \cdot 10^{-5}$	$4.098 \cdot 10^{-6}$	$1.02 \cdot 10^{-7}$	

This distribution shows that the probability to guess the correct answers for 5 or more questions, is 0.0327, and for 4 or less than 4 questions, is 0.9673. We can say that the occurrence of the values $5, 6, 7, 8, 9, 10$ does not support H_0. If anyone said he guessed the correct answer to $0, 1, 2, 3, 4$ questions, we say it is very likely. If anyone said he guessed the correct answer to $5, 6, 7, 8, 9, 10$ questions, we say this is quite unlikely.
The level of significance is the probability α to commit a type I error, that is to reject H_0 when it is true. Currently α is given from the beginning, and it determines the critical region. For the example, if $\alpha = 0.033$, then from $P(x \geq 5) = 0.0327$ we have that the critical region is $x = 5, 6, 7, 8, 9, 10$.

The critical region is the set of values (W) for which $P(X \in W) \geq \alpha$ and it determines us o reject the null hypothesis H_0.
The critical value is the first value from the critical region.
If, for a sample, the value of the test statistic X surpasses the critical value, we reject the hypothesis H_0.

Step 4. **Determining the value of the test statistic**
After finishing steps 1,2,3 we observe or compute the value x of the test statistic.

For our example, $x = 7$ (the number of correct answers) is the value of the test statistic and it is given. We usually compute the value of the test statistic based on the sample information.

Step 5. **Making a decision and interpreting it**
The decision is made comparing the value of the test statistic determined in **Step 4** with the critical region found in **Step 3**.
Decision rule: If the value of the test statistic is in the critical region, we reject H_0, if not, we will fail to reject H_0.
The set of values of the test statistic which are not in the critical region is called the acceptance region. The test is completed by taking and motivating the taken decision.

For our example: $x = 7$ is in the critical region, and we reject H_0.

Remark 2.14.1. *Thus we did not prove that Popescu Nicolae has not guessed the 7 answers. We only showed that if he guessed them, he is very lucky for this is a rare event and has the probability equal to or less than* 0.033*.*

2.15 Hypothesis test: A classical approach

In the previous section we presented generalities regarding hypothesis testing. In this section we present the hypothesis testing for assertions regarding the mean μ of a population. To simplify this presentation, we first suppose that the standard deviation σ of the population is known.

The following three examples refer to different formulations of the null hypothesis H_0 and of the alternative hypothesis H_a.

Example 2.15.1. *An ecologist claims that Timi'soara has an air pollution problem. Specifically, he claims that the mean level of carbon monoxide in downtown air is higher than* $4,9/10^6 =$*, the normal mean value.*

To formulate the hypotheses H_0 and H_a, we have to identify: the population, the population parameter in question and the value to which it is being compared.
In this case, the population can be the set of the downtown Timisoara inhabitants. The variable X is the carbon monoxide concentration, whose values x vary according to the location, and the population parameter is the mean value μ of this variable. The value to which this mean has to be compared is $4,9/10^6$, the normal mean value. The ecologist makes an assertion concerning the value of μ. This value can be: $\mu < 4,9/10^6$ or $\mu = 4,9/10^6$ or $\mu > 4,9/10^6$. The three situations can be arranged to form two statements, one that states what the ecologist is trying to show, and the other, the opposite.

The inequality $\mu > 4,9/10^6$ represents the statement: "the mean value is higher than $4,9/10^6$".

The inequality $\mu \leq 4,9/10^6$ is equivalent to "$\mu < 4,9/10^6$ or $\mu = 4,9/10^6$" and it represents the opposite statement: "the mean value is not higher than $4,9/10^6$".

The ecologist claims that $\mu > 4,9/10^6$. To formulate H_0 and H_a, we remind that:

1) generally, H_0 states that the mean μ (parameter in question) has a specified value.
2) The inference regarding the mean μ of the population is based on the mean of a sample, and the sample means are approximatively normally distributed. (according to the central mean theorem).
3) A normal distribution is completely determined if the mean value and the standard deviation of the distribution are known.

All these suggest that $\mu = 4,9/10^6$ should be the null hypothesis and $\mu > 4,9/10^6$ the alternative hypothesis:

$$\begin{aligned} H_0 &: \quad \mu = 4,9/10^6 \\ H_a &: \quad \mu > 4,9/10^6 \end{aligned}$$

Recall that once the null hypothesis is stated, we proceed with the hypothesis test that H_0 is true. This means that $\mu = 4,9/10^6$ is equal to the distribution mean of the sample means $\mu_{\overline{x}}$ and thus:

$$H_0 : \mu = 4,9/10^6.$$

If we admit that the statement "$\mu = 4,9/10^6$ or $\mu < 4,9/10^6$" is the null hypothesis H_0, then:

$$\begin{aligned} H_0 &: \quad \mu \leq 4,9/10^6 \\ H_a &: \quad \mu > 4,9/10^6. \end{aligned}$$

Remark 2.15.1. *The equal sign must always be present in the null hypothesis. In our example the ecologist's statement is actually expressed in H_a and we must analyze it.*

Example 2.15.2. *Let us consider now a second assertion; for instance, that of the Chamber of Commerce, which claims that the mean level of carbon monoxide in downtown Timisoara is less than $4,9/10^6$ (the normal value). This is a good commercial for tourism.*

We have in this case as well, that the parameter is the mean μ of the carbon monoxide distribution. The specific value is $4,9/10^6$ which is the normal value.

$$\begin{aligned} \text{"}\mu < 4,9/10^6\text{"} \Leftrightarrow &\quad \textit{"the mean value is less than the normal value"} \\ \text{"}\mu \geq 4,9/10^6\text{"} \Leftrightarrow &\quad \textit{"the mean value is higher or equal to} \\ &\quad \textit{the normal mean value"} \end{aligned}$$

H_0, H_a can be formulated as follows:

$$\begin{aligned} H_0 &: \quad \mu \geq 4,9/10^6 \\ H_a &: \quad \mu < 4,9/10^6 \end{aligned}$$

We have again that the assertion of the Chamber of Commerce is expressed in H_a and it must be analyzed.

Example 2.15.3. *A third assertion (a more neutral one) claims that the mean level μ of the carbon monoxide in downtown air is different from $4,9/10^6$ (the normal value is different from μ).*
In this case:

$$H_0 : \quad \mu = 4.9/10^6 \quad \text{and} \quad H_a : \quad \mu \neq 4,9/10^6$$

The three examples show that the assertion that has to be analyzed determine in some way the formulation of the hypotheses H_0, H_a. More exactly: in these cases the assertion claims that the value of the parameter μ is different from the normal one, while the null hypothesis claims it is the same.

In these examples, those who formulate the assertion expect the rejection of the null hypothesis, H_0, and the acceptance of the alternative one, H_a, which is a statement in accordance with their assertion.

The situations from the judicial trials have some similarity with the facts stated before. If the public prosecutor does not believe in the guilt of the accused, he does not sue him (the null hypothesis H_0: the innocence presumption is supposed as true). The trial takes place only if the public prosecutor has enough proofs to start it.

It is somehow the same with statistics; if the experimenter believes the null hypothesis is true, he will not challenge its truth and will not be testing it. He proceeds to test the null hypothesis only if he wishes to show that H_a is correct.

The following example illustrates all the five steps of the hypothesis test procedure for a statement regarding the population mean.

Example 2.15.4. *For several years, a teacher has recorded his students' grades, and the mean μ for all these students' grades is 72 and the standard deviation is $\sigma = 12$. The current class of 36 students has an average $\overline{x} = 75,2$ (higher than $\mu = 72$) and the teacher claims that this class is superior to his previous ones. The question is whether the class mean $\overline{x} = 75,2$ is a sufficient argument to sustain the statement of the teacher for the level of significance $\alpha = 0,05$.*
We mention that, in order for this class to be superior, it has to have a mean grade that is higher than the mean of all the previous classes. If its mean is equal or less than the mean of a previous class, it is not superior.
if we consider random samples of size $n = 36$ from a population with the mean $\mu = 72$, many sample will have the mean $\overline{x}$ close to 72, for instance 71; 71,8; 72; 72,5; 73. Only the means $\overline{x}$ considerably higher than 72 will sustain the teacher's statement.
Therefore:

Step 1. $H_0 : \mu_{\overline{x}} = \mu = 72 \Leftrightarrow$ the class is not superior

Step 2. $H_a : \mu_{\overline{x}} = \mu > 72 \Leftrightarrow$ the class is superior

Step 3.

- When in the null hypothesis H_0 the population mean and the standard deviation are known, we use the standard score z as a test statistic.
- the level of significance $\alpha = 0,05$ is given;
- We recall that based on the central limit theorem, the sample means are approximatively normally distributed. So, the normal distribution will be used to determine the critical region. The critical region is equal to the set of the standard score values z which determine the rejection of H_0 and it is located at the extreme right of the distribution. The critical region is on the right, because large values of the sample mean support the null hypothesis H_0 while values close to, or beneath 72 support the null hypothesis.

The critical value, the cutoff between "not superior" and "superior", is determined by α, the probability of the type I error. $\alpha = 0,05$ has been given. So the critical region, the shaded region on Figure 2., has the area $0,05$ and the critical value $1,65$ is the solution of the equation:

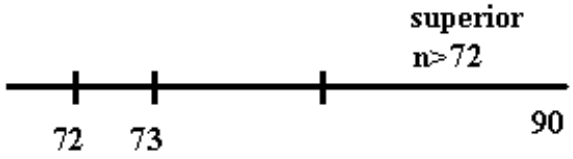


Figure 2.1:

$$\frac{1}{\sqrt{2\cdot\pi}}\int_{z}^{\infty} e^{-\frac{t^2}{2}}\,dt = 0,05.$$

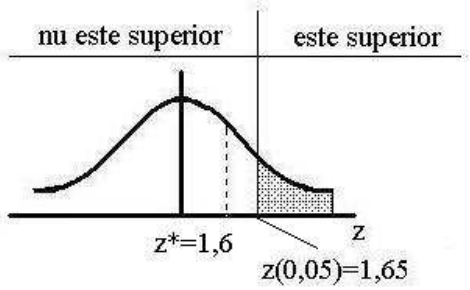


Figure 2.2:

Step 4. The value of the test statistic is given by:

$$z^* = \frac{\overline{x}-\mu}{\sigma\sqrt{n}} = \frac{75,2-72}{12/6} = 1,6$$

Step 5. We compare the value we found $1,6$ with the critical value $1,65$ and we have $1,6 < 1,65$. The decision is that we cannot reject the null hypothesis H_0. The test ends with the formulation of the conclusion.
Conclusion: there is not sufficient evidence to show that the current class is superior.

Does this conclusion seem realistic, regarding the fact that $75,2$ is higher than 72? We have to keep in mind that $\overline{x} = 75,2$ is the mean of a sample of size 36 taken from a population of mean $\mu = 72$ and standard deviation $\sigma = 12$ and the study shows that the probability that the sample mean is higher than all sample means, is larger than the risk, α, with which we are willing to make a type I error.

Example 2.15.5. *It has been claimed that the mean weight of women students at a college is* $\mu = 54,4$ *kg, and the standard deviation* $\sigma = 5,4$ *kg. The sports professor does not believe this statement. To test the claim, he makes a random sample of size* 100 *among the women students and finds the mean* $\overline{x} = 53,75$ *kg. Is this sufficient evidence to reject the statement at the significance level* $\alpha = 0,05$*?*

Step 1. $H_0 : \mu = 54,4$ kg

Step 2. $H_a : \mu \neq 54,4$ kg

Step 3. - because we use a sample mean distribution, the test statistic will be the standard score.

- the level $\alpha = 0,05$ is given;

- the sample mean is an estimation of the population mean. The alternative hypothesis, "not equal to", is supported by sample means considerably larger or considerably smaller than $54,4$. The null hypothesis is supported by sample means close to $54,4$. The critical region is made up of two equal parts, one at each extreme of the normal distribution. The area of each region will be $\frac{\alpha}{2}$ and its probability is $0,025$. We have that $z\left(\frac{\alpha}{2}\right) = 1,96$ $\left(z\left(\frac{\alpha}{2}\right)$ is the solution of the equation: $\frac{1}{\sqrt{2\cdot\pi}}\int_z^{\infty} e^{-\frac{t^2}{2}}\,dt = \frac{\alpha}{2}\right)$.

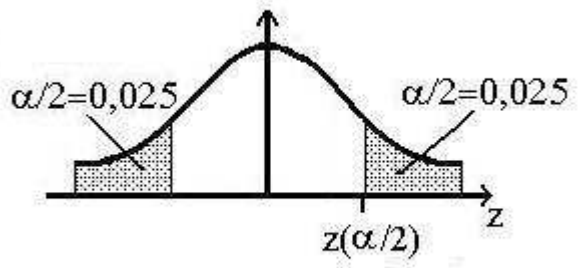


Figure 2.3:

Step 4. We determine the value of the test statistic:

$$z^* = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = -1,204$$

whose location is given in the next figure:

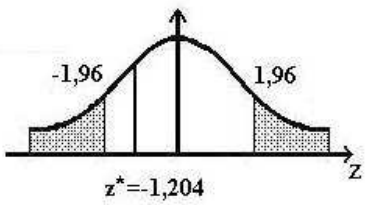


Figure 2.4:

We recall that: If the value of the test statistic is in the critical region, we reject H_0, if not, we fail to reject H_0.

Step 5. The value of the test statistic is not in the critical region.
Decision: We do not reject H_0.
Decision justification: The value of the test does not disagree with H_0 at the risk level $\alpha = 0,05$. This does not mean that H_0 is true.

Conclusion: The mean $\overline{x}$ found by the professor does not run counter to the hypothesis that the mean μ is 54,4 kg, when the dispersion, σ, is $5,4$ kg.
A decision to reject H_0, means that the test value implies that H_0 is false and indicates the truth of H_a.

A summary regarding the classical hypothesis test upon the mean:

1. The null hypothesis, H_0, specifies a particular value of the population mean.

2. The alternative hypothesis, H_a, can take three forms. Each of these determine a specific location of the critical region, as shown in the table below:

Sign in the alternative hypothesis'a	$<$	$\neq$	$>$
Critical region	One region left side	Two regions one on each side	One region right side
	Left one-sided test	**Two-sided test**	**Right one-sided test**

3. In many cases, the sign in the alternative hypothesis H_a indicates the direction in which the critical region is located.

The value of α is called level of significance, and it represents **the risk (probability) of rejecting H_0 when it is actually true. We cannot determine whether H_0 is true or false. We can only decide whether to reject it or to accept it.**
The probability with which we reject the true hypothesis is α, but we do not know the probability with which we make a wrong decision. A type I error and a decision error are two different things.

2.16 Hypothesis test: a probability-value approach

We have described the classical approach of the hypothesis test upon the mean μ of a population in the previous section. A probabilistic approach implies the calculation of a probability value called **p-value** (prob-value) related to the observed sample statistic, which is compared to the level of significance α.

Definition 2.16.1. ***The p-value*** *of a hypothesis test is the smallest level of significance, α, for which the observed sample information becomes significant (H_0 true, is rejected).*

We consider again the example 2.15.4 from the previous section and analyze it from this point of view.

Example 2.16.1. *For several years, a teacher has recorded his students' grades, and the mean μ for all these students' grades is 72 and the standard deviation is $\sigma = 12$. The current class of 36 students has an average $\overline{x} = 75,2$ (higher than $\mu = 72$) and the teacher claims that this class is superior to his previous ones. The question is whether the class mean $\overline{x} = 75,2$ is a sufficient argument to sustain the statement of the teacher for the level of significance $\alpha = 0,05$.*
We mention that, in order for this class to be superior, it has to have a mean grade that is higher than the mean of all the previous classes. If its mean is equal or less than the mean of a previous class, it is not superior.

Step 1. The formulation of the null hypothesis H_0: $\mu_{\overline{x}} = \mu = 72$.
This hypothesis corresponds to the assertion the the current class is not superior to the other classes.

Step 2. The formulation of the alternative hypothesis H_a : $\mu_{\overline{x}} = \mu > 72$.
This hypothesis corresponds to the assertion the the current class is superior to the other classes.

We notice that steps 1 and 2 are exactly the same when using either the classical or the probabilistic approach to hypothesis testing.

Step 3. The specification of the level of significance α, of the probability of the type I error: $\alpha = 0,005$.

Step 4. using the formula for the standard score (z-score) and the sample mean $\overline{x} = 75,2$, where the sample size is $n = 36$, we determine the value of the test statistic:

$$z^* = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = 1,60$$

We observe that **Step 4** in the probabilistic approach is the same as in the classical approach of hypothesis testing.

Step 5. We draw a sketch of the probability distribution for the means (test statistic) in this case, and we locate the value z^* that we found at **Step 4** (which parts the distribution in two parts) and we determine which part of the distribution represents the $p-$value.
Afterwards, we calculate the $p-$value. The alternative hypothesis H_a shows that in our case:

$$p = P(z > z^*) = P(z > 1,6) = 0,0548$$

Step 6. In our case, the p-value is $0,0548$. Therefore we fail to reject the null hypothesis for every level of significance $\alpha \leq 0,0548$ and the conclusion is that we haven't got enough evidence to demonstrate the superiority of the current class. But if the level of significance α is fixed from the beginning and is greater than $0,0548$ (for ex. $\alpha = 0,1$) then we will reject the null hypothesis and our conclusion will be the superiority of the current class.

Before looking at another example, let us summarize some details regarding the prob-value approach of the hypothesis testing:

1. The hypotheses H_0 and H_a are formulated in the same manner as before.
2. We determine the level of significance, α, to be used.

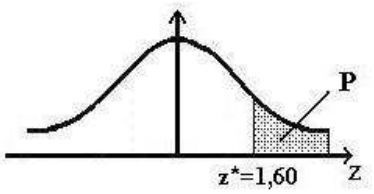


Figure 2.5:

3. The value of the test statistic is calculated in **Step 4** in the same manner as before.

4. The p-value is the area between the probability distribution curve, the Oz axis and $z = z^*$. There are three possible cases: two are one-sided and one is two-sided. the direction (or sign) in H_a is the key:

Case 1. If H_a is one-sided to the right (" $>$ ") then $p = P(z > z^*)$ and the area is to the right of z^*.

case 2. If H_a is one-sided to the left (" $<$ "), then $p = P(z < z^*)$ the area to the left of z^*.

Case 3. If H_a is two-sided (" $\neq$ "), then $p = P(z < -|z^*|) + P(z > |z^*|) = 2 \cdot P(z > |z^*|)$

5. We make the decision comparing the p-value with the level of significance α:

a) If $P \leq \alpha$ then H_0 is rejected;

b) If $P > \alpha$ we fail to reject H_0.

6. We formulate the conclusion in the same manner as in the classical approach.

Let us now consider an example where H_a is two-sided.

Example 2.16.2. *Large companies use specialized agencies to test the prospective employees. The A Agency uses a selection test which has resulted, in time, in scores distributed about a mean of 82 and a standard deviation of 8. The B Agency has developed a new test that is quicker, easier to administer and less expensive. The B Agency claims that their test results are the same as those obtained by the A Agency test.*
To reduce the cost, many of the companies are considering a change from the A Agency to the B Agency, but they do not wish to make this change unless the B test results have the same mean as the A test results. An independent agency, C, has tested 36 *employees and obtained a mean of* 80.
Determine the p−value associated with this hypothesis test.
The results of the B Agency test are the same if $\mu = 82$ *and differ if* $\mu \neq 82$. *Therefore:*

Step 1. $H_0 : \mu = 82$ (the tests have the same mean)

Step 2. $H_a : \mu \neq 82$ (the tests have different means)

Step 3. Step 3 is omitted when a question asks for the p-value and not for a decision.

Step 4. The sample information $n = 36$ and $\overline{x} = 80$:

$$z^* = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Step 5. We locate z^* on a normal distribution; because H_a is two-sided, we will consider $P(z < -|z^*|)$ and $P(z > |z^*|)$ and we have:

$$\begin{aligned} p &= P(z < -1,50) + P(z > 1,50) \\ &= 0,5 - 0,4332 + 0,5 + 0,4332 = 0,1336 \end{aligned}$$

so the $p-$value is $0,1336$.

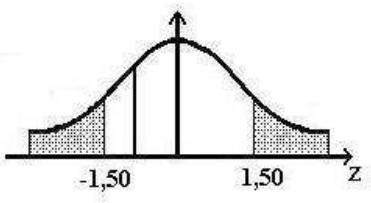


Figure 2.6:

Each company will make a decision whether to: a) continue with A or b) make a change and go to B. Each will need to establish the level of significance that best fits their own situation and then make a decision using the rule described above.

2.17 Statistical inference about the population mean when the standard deviation is not known

Until now we discussed two types of statistical inference about the population mean: the confidence interval estimation and the hypothesis test. In these two types of inference the standard deviation σ was known. However, the population standard deviation σ is not generally known. This section deals with **inferences about the mean μ when the standard deviation σ is unknown**.

If the sample size is sufficiently large (generally talking, samples of size greater than $n = 30$ are considered sufficiently large), the sample standard deviation s is a good estimate of the standard deviation of the population and we can substitute σ with s in the already discussed procedure. If the population investigated is approximately normal and $n \leq 30$, we will base our procedure on the Student's t distribution.

The Student's t distribution (or simple, the t distribution) is the distribution of the t statistic, which is defined as:

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$$

In 1908 W.S. Gosset, an Irish brewery clerk, published a paper about the t distribution under the pseudonym "Student". In Gosset's paper the population is assumed to be normal. This restriction has later proved to be too restrictive, because there were satisfactory results for nonnormal populations. We do not give here the equations that defines the t-distribution, we only mention some of the distribution's properties:

1) t is distributed with a mean of 0;

2) t is distributed symmetrically about its mean;

3) dt is distributed with a variance greater than 1, but as the sample size increases, the variance approaches 1;

4) t is distributed so as to be less peaked at the mean and thicker at the tails than the normal distribution;

5) t is distributed so as to form a family of distributions, a separate distribution for each sample size. The t distribution approaches the normal distribution as the sample size increases.

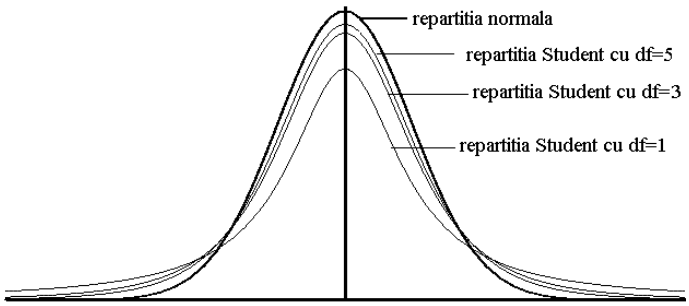


Figure 2.7:

Although there is a separated t distribution for each sample size (n=2,3,4,...), only certain key critical values are practically used. These critical values, to the right of the mean, are given in the following table:

α df	0,40	0,30	0.25	0,20	0,10	0,05	0,025	0,010	0,005	0,001	0,0005
1	0,325	0,727	1,000	1,376	3,078	6,314	12,71	31,82	63,66	318,3	636,6
2	0,289	0,617	0,816	1,061	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	0,277	0,584	0,765	0,978	1,638	2,353	3,182	4,541	5,841	10,22	12,94
4	0,271	0,569	0,741	0,941	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,267	0,559	0,727	0,920	1,476	2,015	2,571	3,365	4,032	5,893	6,859
6	0,265	0,553	0,718	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,263	0,549	0,711	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,405
8	0,262	0,546	0,706	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,261	0,543	0,703	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,260	0,542	0,700	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,260	0,540	0,697	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,259	0,539	0,695	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,259	0,538	0,694	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,258	0,537	0,692	0,868	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,258	0,536	0,691	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,258	0,535	0,690	0,865	l,337	1,746	2,120	2,583	2,921	3,686	4,015

α df	0,40	0,30	0,25	0,20	0,10	0,05	0,025	0,010	0,005	0,001	0,0005
17	0,257	0,534	0,689	0,863	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,257	0,534	0,688	0,862	1,330	1,734	2,101	2,552	2,878	3,611	3,922
19	0,257	0,533	0,688	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,257	0,533	0,687	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	0,257	0,532	0,686	0,859	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,256	0,532	0,686	0,858	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,256	0,532	0,685	0,858	1,319	1,714	2,069	2,500	2,807	3,485	3,767
24	0,256	0,531	0,685	0,857	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	0,256	0,531	0,684	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,256	0,531	0,684	0,856	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,256	0,531	0,684	0,855	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	0,256	0,530	0,683	0,855	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,256	0,530	0,683	0,854	1,311	1,699	2,045	2,462	2,756	3,396	3,659
z	0,256	0,530	0,674	0,854	1,310	1,697	2,042	2,457	2,750	3,385	3,646

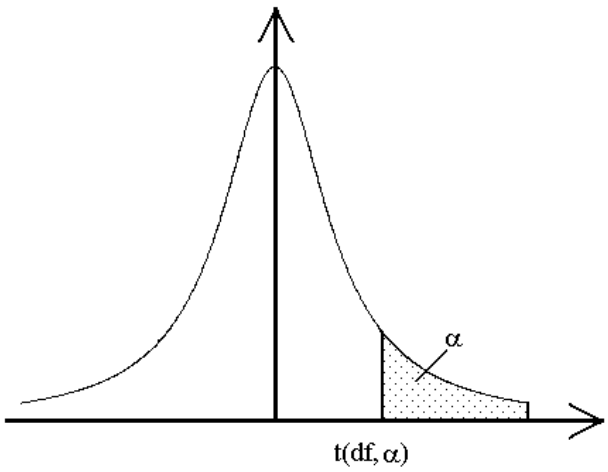


Figure 2.8:

In this table *df* takes values from 1 to 29 and it represents the **degrees of freedom**. The proximity of the values belonging to the lines that correspond to $df = 29$ and z, is due to the fact that if $n \geq 30$ the t distribution is the normal one (the central limit theorem).

The degrees of freedom, *df*, is a parameter that is difficult to define. It is an index used to identify the correct distribution to be used. In our considerations $df = n - 1$, where n is the sample size. The critical value of the test t that we should use either in the estimation of the confidence interval or in the hypothesis test is obtained from the above given table. In order to obtain this value we need to know:

1) *df* - the degrees of freedom;

2) the α area determined by the distribution curved situated to the right of the critical value. We will denote this value $t(df, \alpha)$.

Example 2.17.1. *Determine* $t(10, 0.05)$ *from the table. We have* $df = 10$ *and* $\alpha = 0.05$*, so* $t(10, 0.05) = 1.81$.

The critical values of the test statistic t situated to the left of the mean are obtained with the formula: $-t(df, \alpha)$*, considering the symmetry of the t distribution.*

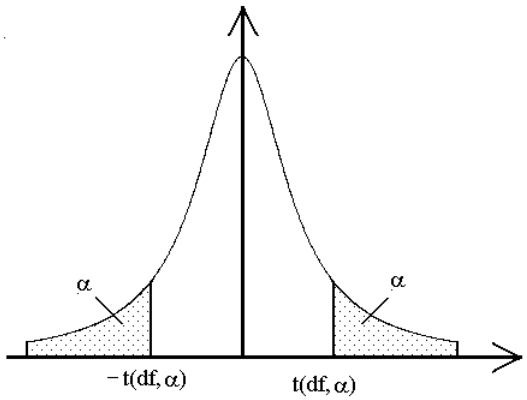


Figure 2.9:

We easily observe that $-t(df, \alpha) = t(df, 1-\alpha)$. therefore: $-t(df; 0,05) = t(df; 0,95)$.

Example 2.17.2. *Determine* $t(15; 0,95)$*. We have:* $t(15; 0,95) = -t(15; 0,05) = -1,75$*.*

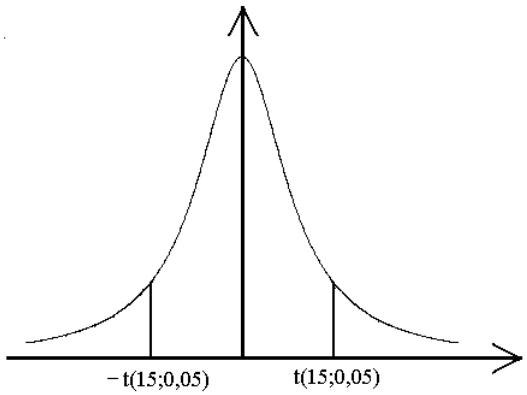


Figure 2.10:

The statistic t is used in the hypothesis tests about the mean μ in the same manner as the statistic z.

Example 2.17.3. *Let us return to the example concerning the air pollution; the ecologist's point of view: "the level of carbon monoxide in the air, is higher than* $4,9/10^6$*". Does a sample of* 25 *readings with the mean* $\overline{x} = 5,1/10^6$ *and* $s = 2,1/10^6$ *present sufficient evidence to sustain the statement? We will use the level of significance* $\alpha = 0,05$*.*

Step 1. $H_0 : \mu = 4,9/10^6$

Step 2. $H_a : \mu > 4,9/10^6$

Step 3. $\alpha = 0,05$; $df = 25 - 1 = 24$ and $t(24; 0,05) = 1,71$ from the table.

Step 4.

$$t^* = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{5,1 - 4,9}{2,1/\sqrt{25}} = \frac{0,20}{0,42} = 0,476 \simeq 0,48$$

Step 5. **Decision:** We cannot reject H_0 (t^* is not in the critical region).
Conclusion: We do not have sufficient evidence to reject the hypothesis that the level of carbon monoxide in the air is $4,96/10^6$.

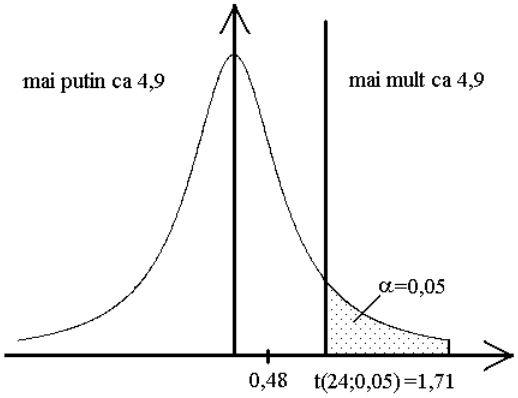


Figure 2.11:

Remark 2.17.1. *If df* ($df = n - 1$) *is higher than* 29, *then the critical value of* $t(df, \alpha)$ *is very close to* $z(\alpha)$ *(the z-score is listed at the end of the table) and therefore instead of* $t(df, \alpha)$ *we use* $z(\alpha)$. *Because the considered table contains only critical values of the t distribution, we cannot determine the p-value from the table for the hypothesis test, because it needs the complete t distribution. Still, we can estimate the p-value using the table.*

Example 2.17.4. *Let us return to the example 2.17.3. We have* $t^* = 0,48$, $df = 24$ *and* $H_a : \mu > 49$. *Therefore, to solve the problem using a probabilistic approach for* ***Step 5*** *with the p-value, we have:*

$$p = P(t > 0,48, \text{provided } df = 24)$$

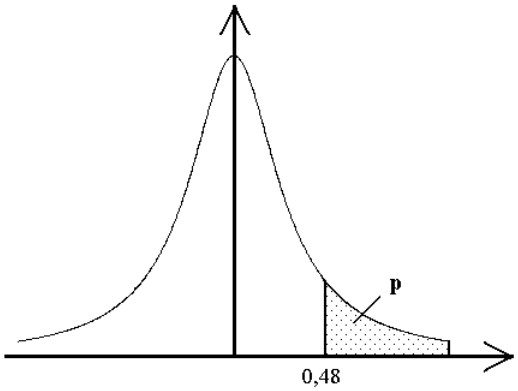


Figure 2.12:

The line $df = 24$ from the table shows that the p-value is higher than $0,25$. The value $0,685$ from the table, shows that $P(t > 0,685) = 0,25$ as in the following figure:
Comparing $t^* = 0,48$, we see that the p-value is higher than $0,25$.

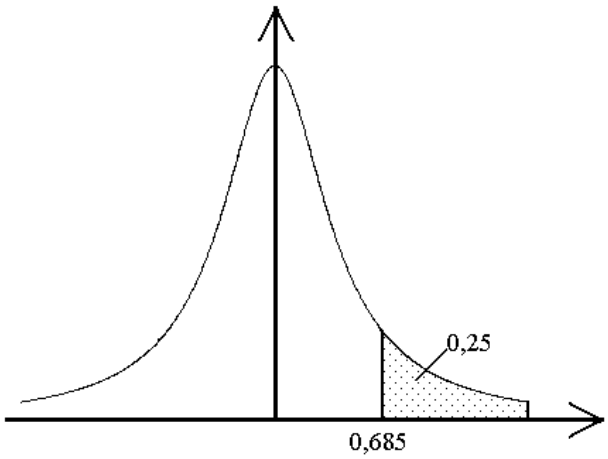


Figure 2.13:

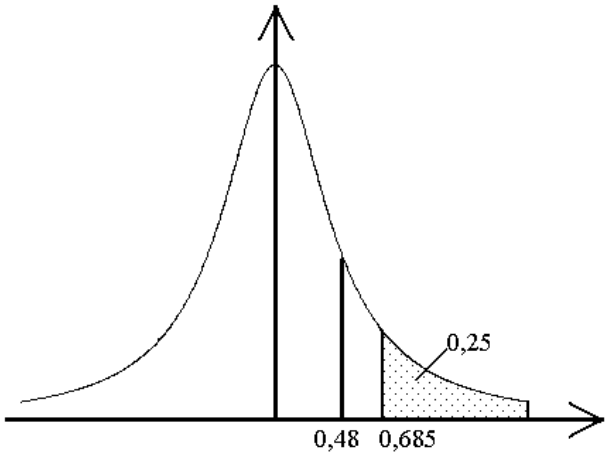


Figure 2.14:

Example 2.17.5. *Determine the p-value for the following hypothesis test:*

$$H_0 : \mu = 55$$

$$H_a : \mu \neq 55$$

provided $df = 15$ *and* $t^* = -1,84$.

Solution: $p = P(t < -1,84) + P(t > 1,84) = 2 \cdot P(t > 1,84)$. The line $df = 15$ from the table shows that $P(t > 1,84)$ is between $0,025$ and $0,05$. therefore, we have: $0,05 < p < 0,10$.

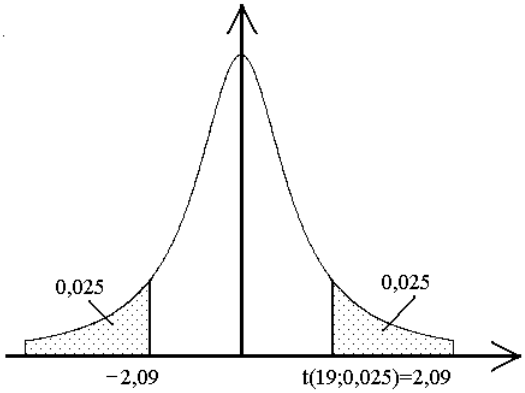
The population mean can be estimated when σ is unknown in a manner similar to that used when σ is known. The difference is the use of Student's T in place of z and the use of s, the sample standard deviation, as an estimate of σ. The formula for the $1 - \alpha$ confidence interval is:

$$\left(\overline{x} - t(df, \frac{\alpha}{2}) \cdot \frac{s}{\sqrt{n}}, \overline{x} + t(df, \frac{\alpha}{2}) \cdot \frac{s}{\sqrt{n}}\right)$$

where $df = n - 1$.

Example 2.17.6. *For a random sample of size 20 taken from the weights of new-born babies, we have a weight mean of* $3,4$ *kg and a standard deviation of* $0,9$ *kg. Estimate, with* 95% *confidence, the mean weight of all new-born babies.*

Solution: $\overline{x} = 3,4$ kg, $s = 0,9$ kg and $n = 20$, and $1 - \alpha = 0,95$, implies: $\alpha = 0,05$; $df = 19$, and we have from the table that $t(19; 0,025) = 2,09$. The interval boundaries are:



$$\overline{x} \pm t(19; 0,025) \cdot \frac{s}{\sqrt{n}} = 3,4 \pm 2,09 \cdot \frac{0,9}{\sqrt{20}}$$

$$3,4 \pm 2,09 \cdot \frac{0,9}{4,472} = 3,4 \pm 0,46$$

The 95% confidence interval is $(2,94; 3,86)$.

2.18 Inferences about the variance and the estimation of the variance

Often problems arise that require us to make inferences about the variance. For example, a soft drink bottling company has a machine that fills $0,32$ l$=$ 32 cl bottles. The mean amount placed in each bottle is important, but the correct mean amount does not ensure that the machine is working correctly. If the variance is too large, there could be many bottles that are overfilled and many that are underfilled. Thus this bottling company will want to maintain as small a variance as possible.

We will study two kinds of inferences in this section. The first is the inference about the variance (or the standard deviation) and the second is the estimation of the variance (the standard deviation) of a population. Often in these two inferences it is customary to talk about the standard deviation instead of the variance. We must underline that the standard deviation is the square root of the variance; therefore talking about the variance is comparable to talking about the standard deviation.
Let us return to the example of the soft drink bottling company. Let us imagine that this company wishes to detect when the variability in the amount of soft drink placed in each bottle gets out of control. A variance of $0,0004$ is considered acceptable and the company will want to adjust the bottle-filling machine when the variance becomes larger than this value.The decision will be made by using the hypothesis test procedure. The null hypothesis, H_0, is that the variance has the value $0,0004$, while the alternative hypothesis, H_a, states that the variance is

larger than $0,0004$:

$$\begin{array}{lll} H_0: & \sigma^2 = 0,0004 & \text{(the variance is controlled)} \\ H_a: & \sigma^2 > 0,0004 & \text{(the variance is out of control)} \end{array}$$

The test statistic that will be used in making a decision about the null hypothesis is the chi-square, χ^2. The calculated value of χ^2 will be obtained by using the formula:

$$\chi^2 = \frac{n \cdot s^2}{\sigma^2}$$

where s^2 is the sample variance, n is the sample size, and σ^2 is the value specified in the null hypothesis.

If we take a sample from a normal population, having the variance σ^2, then the quantity $n \cdot s^2/\sigma^2$ has a distribution called χ^2. We shall not give here the formula that defines the χ^2 distribution, but in order to use the χ^2 distribution we will mention the following properties:

1. χ^2 is nonnegative in value; it is zero or positively valued;

2. χ^2 is not symmetrical; it is skewed to the right;

3. There are many χ^2 distributions. Like the t distribution, there is a different χ^2 distribution for each degree-of-freedom value. The inference studied here is corresponding to $df = n-1$.

The critical values of χ^2 are given in the following table:

df	0.995	0.990	0.975	0.950	0.900	0.10	0.05	0.025	0.01	0.005
2	0.01	0.020	0.050	0.103	0.211	4.61	6.0	7.38	9.21	10.6
3	0.071	0.115	0.216	0.352	0.584	6.25	7.82	9.35	11.4	12.9
4	0.207	0.297	0.484	0.711	1.06	7.78	9.50	11.1	13.3	14.9
5	0.412	0.554	0.831	1.15	1.61	9.24	11.1	12.8	15.1	16.8
6	0.676	0.872	1.24	1.64	2.20	10.6	12.6	14.5	16.8	18.6
7	0.990	1.24	1.69	2.17	2.83	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	14.7	17.0	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	16.0	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.58	5.58	17.2	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	18.6	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.90	7.04	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.79	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.55	22.3	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	9.31	23.5	26.3	28.9	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	10.9	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	28.4	31.41	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	13.2	29.6	32.7	35.5	39.0	41.4
22	8.64	9.54	11.0	12.3	14.0	30.8	33.9	36.8	40.3	42.8
23	9.26	10.2	11.0	13.1	14.9	32.0	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.9	15.7	33.2	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	16.5	34.4	37.7	40.7	44.3	46.9
26	11.2	12.2	13.8	15.4	17.3	35.6	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	18.1	36.7	40.1	43.2	47.0	49.7
28	12.5	13.6	15.3	16.9	18.9	37.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.1	17.7	19.8	39.1	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	20.6	40.3	43.8	47.0	50.9	53.7
40	20.7	22.2	24.4	26.5	29.1	51.8	55.8	59.3	63.7	66.8
50	28.0	29.7	32.4	34.8	37.7	63.2	67.5	71.4	76.2	79.5
60	5.5	37.5	40.5	43.2	46.5	74.4	79.1	83.3	88.4	92.0
70	43.3	45.4	48.8	51.8	55.3	85.5	90.5	95.0	100.0	104.0
80	51.2	53.5	57.2	60.4	64.3	96.6	102.0	107.0	112.0	116.0
90	59.2	61.8	65.7	69.1	73.3	108.0	113.0	118.0	124.0	128.0
100	67.3	70.1	74.2	77.9	82.4	114.0	124.0	130.0	136.0	140.0

The critical values will be identified by two values: degrees of freedom and the area under the curve to the right of the critical value to be sought. Thus $\chi^2(df, \alpha)$ is the symbol used to identify the critical value χ^2 with df degrees of freedom and with the area α to the right as shown in the following figure:

Example 2.18.1. *Find* $\chi^2(20; 0,05)$ *and* $\chi^2(14; 0,90)$ *using the table.*
We have from the table: $\chi^2(20; 0,05) = 31,4$ *and* $\chi^2(14; 0,90) = 7,79$.

Remark 2.18.1. *If* $df > 2$ *the mean value of* χ^2 *is* df. *The mean value is located to the right of the mode (the value where the curve reaches its high point).*

Example 2.18.2. *Recall that the soft drink bottling company wanted to control the variance by not allowing it to exceed* $0,0004$. *Does a sample of size* 28 *with a variance of* $0,0010$ *indicate that the bottling process is out of control (with regard to variance)at the* $0,05$ *level of significance?*

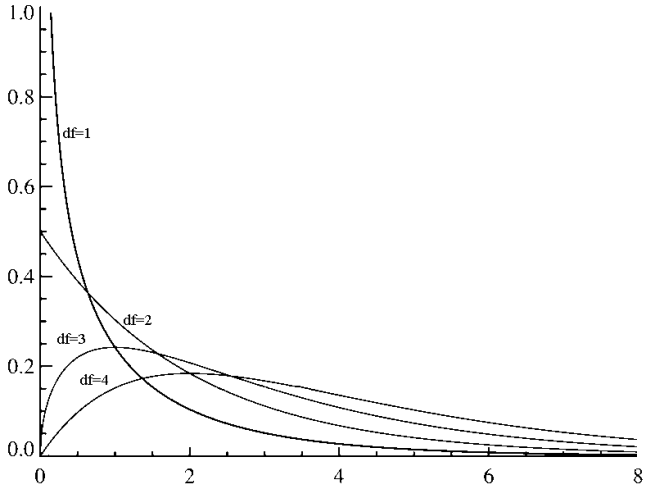


Figure 2.15:

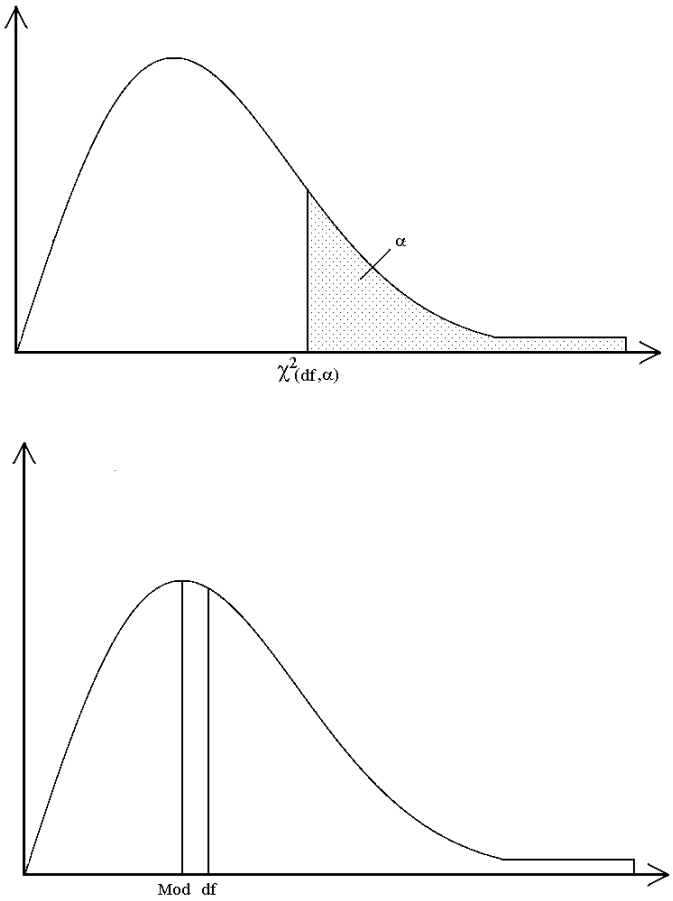


Figure 2.17:

Solution:

Step 1. $H_0 : \sigma^2 = 0,0004$ (the process is under control)

Step 2. $H_a : \sigma^2 > 0,0004$ (the process is out of control)

Step 3. $\alpha = 0,05$, $n = 28$, $df = 27$ and we have from the table:

$$\chi^2(27; 0,005) = 40,1.$$

Step 4.

$$\chi^2_* = \frac{n \cdot s^2}{\sigma^2} = \frac{28 \cdot 0,0010}{0,0004} = 70$$

Step 5. Taking the decision.

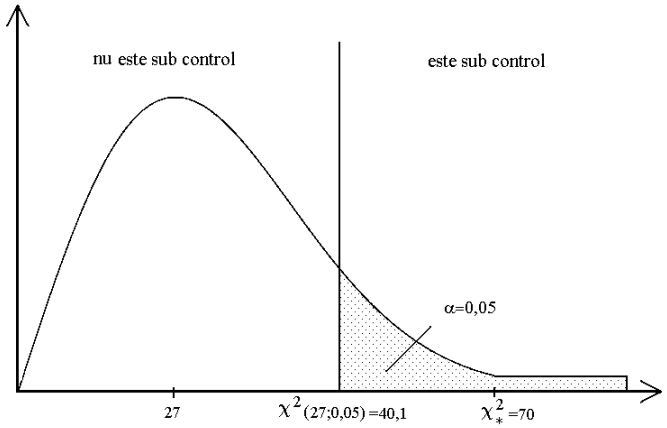


Figure 2.18:

Conclusion: The bottling process is out of control with regard to the variance.

Example 2.18.3. *The specifications of a certain medicine indicate that each pill has to contain 2,5 g active substance. We analyze 100 pills, randomly chosen from the production. They contain a mean of 2,6 g of active substance with a standard deviation of* $s = 0,4g$.
Can we say that the medicine follows the specifications ($\alpha = 0,05$)*?*

Step 1. The H_0 hypothesis is that the medicine is according to the specifications:

$$H_0 : \mu = 2,5$$

Step 2. The H_a hypothesis is that the medicine is not according to the specifications:

$$H_0 : \mu \neq 2,5$$

Step 3. The statistic used here is the mean $\overline{x}$, and the level of significance is $\alpha = 0,05$. The critical region is:

Step 4. The test statistic is:

$$z = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{2,6 - 2,5}{\frac{0,4}{10}} = \frac{0,1}{0,04} = 2,5$$

The value of z in the table is: $z_{0,975} = 1,96 < 2,5$.

Step 5. H_0 is rejected, therefore we cannot say that the medicine is according to the specifications.

In the probabilistic approach of the statistical inference about the variance, the p-value can be estimated for the hypothesis test using the statistic table χ^2 in the same manner as for the Student test.

Example 2.18.4. *Find the p-value for the following statistical hypotheses:*

$$\begin{aligned} H_0: &\quad \sigma^2 = 150 \\ H_a: &\quad \sigma^2 > 150 \end{aligned}$$

We know: $df = 18$ *and* $\chi^2_* = 32,7$.

Solution: $p = P(\chi^2 > 32,7) \in (0,010; 0,025)$ (critical data from the table).

Example 2.18.5. *One of the factors used in determining the usefulness of a particular exam as a measure of students' abilities is the amount of "spread" that occurs in the grades. A set of test results has little value if the range of the grades is very small. However, if the range of grades is quite large, there is a definite difference in the scores achieved by the better students and the scores achieved by the "poorer" students. On an exam with a total of* 100 *points, it has been claimed that a standard deviation of 12 points is desirable. To determine whether or not the last one-hour exam he gave his test was a good test, a professor tested the hypothesis above at* $\alpha = 0,05$ *by using the exam scores of the class. There were 28 scores and the standard deviation found was* $10,5$. *Does this constitute an evidence on the level of significance* $\alpha = 0,05$ *that the exam does not have a specified standard deviation?*

Solution: $n = 28$, $s = 10,5$ and $\alpha = 0,05$

Step 1. $H_0 : \sigma = 12$

Step 2. $H_0 : \sigma \neq 12$

Step 3. $\alpha = 0,05$, $df = 27$ and we obtain the critical values from the table:

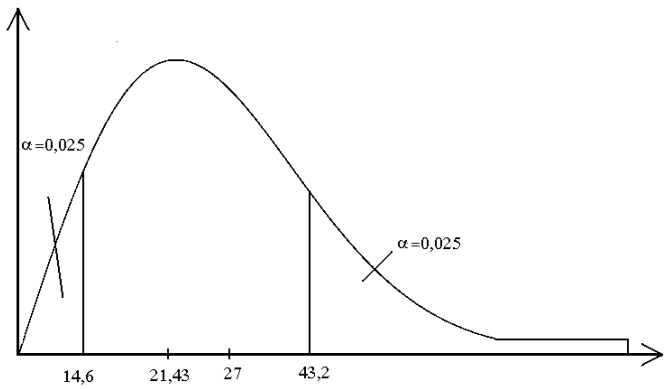
$$\chi^2_1(27; 0,975) = 14,6 \quad \text{and} \quad \chi^2_2(27; 0,025) = 43,2.$$

Step 4.

$$\chi^2_* = \frac{n \cdot s^2}{\sigma^2} = \frac{28 \cdot (10,5)^2}{(12)^2} = \frac{3087}{144} = 21,43$$

Step 5. We cannot reject H_0.

Conclusion: We don't have enough evidence to reject the null hypothesis H_0.



2.19 Generalities about correlation. Linear correlation

In statistics, there occur problems of the following type: for the same population we have two sets of data corresponding to two distinct variables and the question arises whether there is a relationship between those two variables? If the answer is yes, what is that relationship? How are these variables correlated? The relationships discussed here are not necessary of the type cause-and-effect. They are mathematic relationships which predict the behavior of one variable from knowledge about the second variable. Here we have some examples:

Example 2.19.1.

- *Generally, a person who grows taller will also gain in weight. The question arises whether there is a relationship between height and weight.*
- *The students spend their time at the university, learning or taking exams. The question arises whether the more they study, the higher grades they will have.*
- *Research doctors test a new drug by prescribing different amounts and observing the responses of their patients; we could ask, "Does the amount of drug prescribed determine the amount of recovery time needed by the patient?"*

The problems from the previous example require the analysis of the correlation between two variables.

When for a population we have two sets of data corresponding to two distinct variables, we form the pairs (x, y), where x is the value of the first variable and y is the value of the second one. For example, x is the height and y is the weight.
An ordered pair (x, y) is called **bivariate data**.

Traditionally, the variable X (having the values x) is called **input variable (independent variable)**, and the variable Y (having the values y) is called **output variable (dependent variable)**.

The input variable X is the one measured or controlled to predict the variable Y.

'In cazul test'arii medicamentului doctorii (m'asoar'a) controleaz'a cantitatea de medicament prescris'a 'si deci aceast'a cantitate x este valoarea variabilei de intrare (independent'a) X. Timpul de recuperare y este valoarea variabilei de ie'sire (dependente) Y.

In the example with height and weight, any of the two variables can be both input and out variable. The results of the analysis will be depending on the choice made.

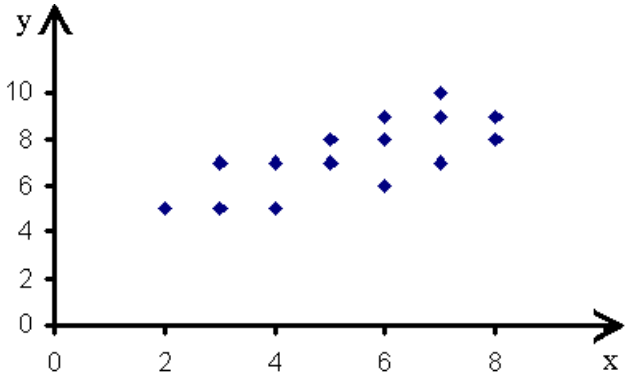
In problems that deal with the analysis of the correlation between two variables, the sample data are presented as a scatter diagram.

Definition 2.19.1. A scatter diagram is the graphical representation of the pairs of data in an orthogonal coordinate system. The values x of the input variable X are represented on the Ox axis, and the values y of the output variable Y are represented on the Oy axis.

Example 2.19.2. *For a sample of 15 students, the following table represents the number of study hours for an exam, x, and the grade obtained to that exam, y:*

x	2	3	3	4	4	5	5	6	6	6	7	7	7	8	8
y	5	5	7	5	7	7	8	6	9	8	7	9	10	8	9

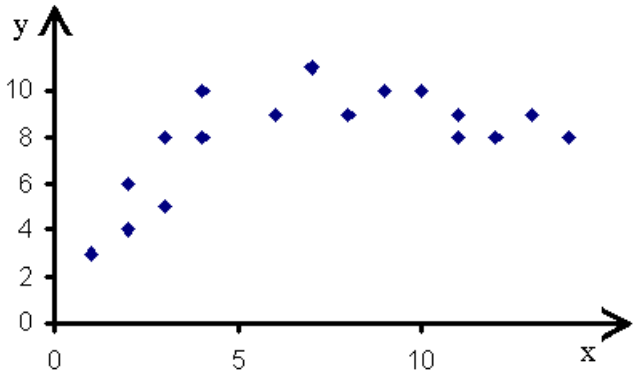
The scatter diagram for this example is:



Example 2.19.3. *The scatter diagram for these data:*

x	2	12	4	6	9	4	11	3	10	11	3	1	13	12	14	7	2	8
y	4	8	10	9	10	8	8	5	10	9	8	3	9	8	8	11	6	9

is:



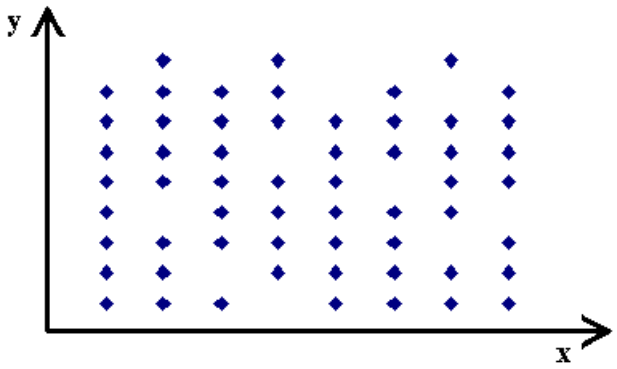
The primary purpose of the correlation analysis is to establish a relationship between the two variables.

We will present some scatter diagrams to illustrate the possible correlations between the input variable X and the output variable Y.

Definition 2.19.2. If for the increasing values x of the input variable X there is no definite displacement of the values y of the variable Y, we then say that there is **no correlation** or **no relationship between X and Y**.

The scatter diagram in the case of no correlation is the following:

Definition 2.19.3. If for the increasing values x of the input variable X there is a definite displacement of the values y of the variable Y, we then say that there is a **correlation**. We have a **positive correlation** if y tends to increase, and we have a **negative correlation** if y tends to decrease while x increases.



The precision of the shift in y as x increases determines the strength of the correlation. The following scatter diagrams demonstrate these ideas:

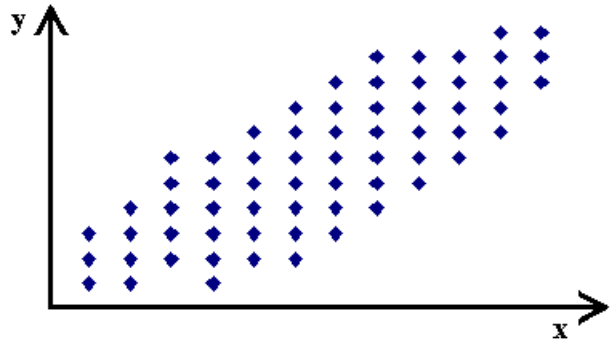


Figure 2.19: The scatter diagram for a positive correlation

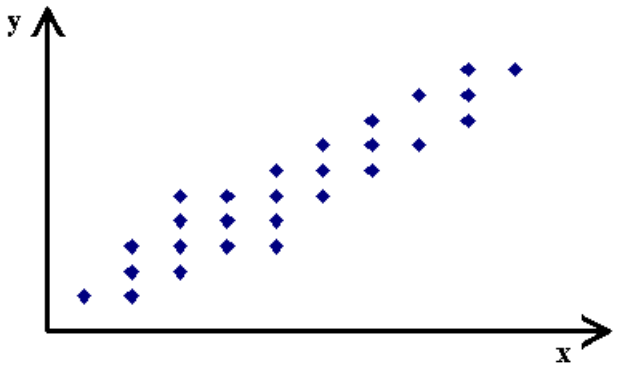


Figure 2.20: The scatter diagram for a high positive correlation

Definition 2.19.4. If the pairs (x, y) tend to follow a line, we say that we have a **linear correlation**.

Definition 2.19.5. If all the pairs (x, y) are on a line (that is not horizontal nor vertical) we say that we have a **perfect linear correlation**.

Remark 2.19.1. *If all the pairs (x, y) are on a horizontal or on a vertical line, there is no correlation between the two variables, that is because the change of one variable does not affect the value of the other one.*

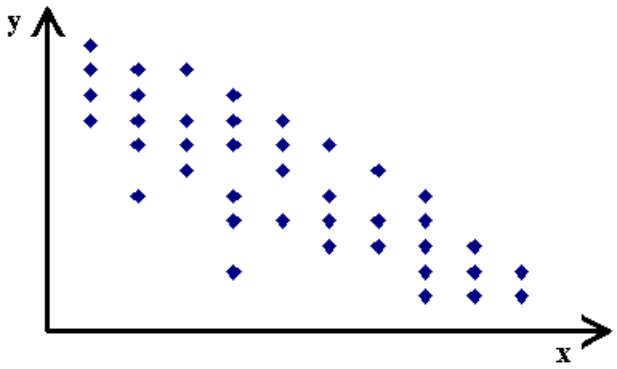


Figure 2.21: The scatter diagram for a negative correlation

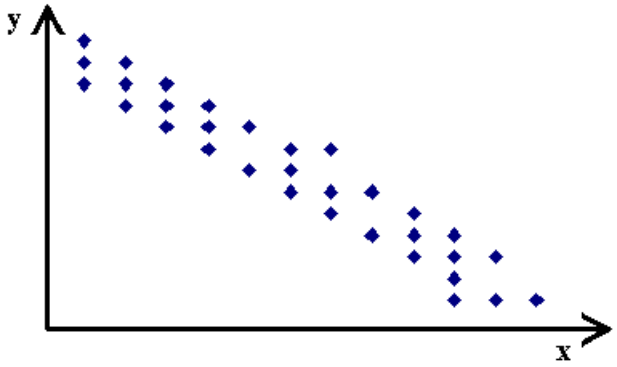


Figure 2.22: The scatter diagram for a high negative correlation

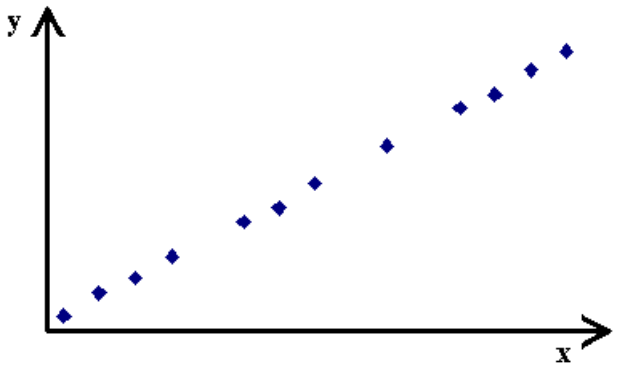


Figure 2.23: The scatter diagram for a perfect linear correlation

Remark 2.19.2. *Scatter diagrams do not always appear in one of the forms shown above and they might suggest correlations of other kinds.*

Definition 2.19.6. The coefficient of linear correlation r measures the strength of the linear correlation between the two variables. It reflects the consistency of the effect that a change in one variable has on the other.

Remark 2.19.3. *The value of the coefficient of linear correlation r allows us to formulate an answer to the question: is there a linear correlation between the two considered variables? The*

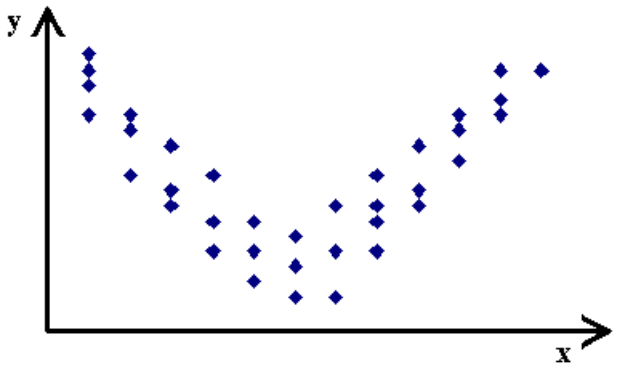


Figure 2.24: The scatter diagram for a nonlinear correlation

coefficient of linear correlation r has a value between -1 and $+1$. The value $r = +1$ signifies a perfect positive correlation, and the value $r = -1$ signifies a perfect negative correlation.

If as x increases there is a general increase in the value of y, then r will indicate a positive linear correlation.

For children, for example, if x is the age and y is the height, we then expect r to be positive, because naturally, he child's height increases as he grows older. For automobiles, if x is the age, and y is its value, we then expect r to be negative, because usually the value of the automobile decreases as the years pass.

Definition 2.19.7. The coefficient of linear correlation r for a sample, is by definition:

$$r = \frac{\sum (x - \overline{x}) \cdot (y - \overline{y})}{n \cdot s_x \cdot s_y}$$

where s_x, s_y are the standard deviations of the variables x, y, and n is the number of pairs (x, y).

Remark 2.19.4. *To calculate r we usually use an equivalent alternative formula:*

$$r = \frac{SS(x,y)}{\sqrt{SS(x) \cdot SS(Y)}}$$

where: $SS(x) = \sum x^2 - \frac{1}{n} \cdot \left(\sum x\right)^2$, $SS(y) = \sum y^2 - \frac{1}{n} \cdot \left(\sum y\right)^2$, $SS(x,y) = \sum x \cdot y - \frac{1}{n} \cdot \left(\sum x \cdot \sum y\right)$.

Example 2.19.4. *Determine the linear correlation coefficient r for a random sample of size 10, if the data table is:*

x	27	22	15	35	30	52	35	55	40	40
y	30	26	25	42	38	40	32	54	50	43

Using these data we have:

$$SS(x) = 1396,9 \quad SS(y) = 858,0 \quad SS(x,y) = 919,0$$

wherefrom we find:

$$r = \frac{919,0}{\sqrt{(1396,9) \cdot (858,0)}} = 0,8394 \approx 0,84.$$

Remark 2.19.5. *If the calculated value r is close to 0, there is **no** linear correlation.*

If the calculated value r is close to $+1$ or -1, we then suppose that between the two variables there is a linear correlation.

Between 0 and 1 there is a value called decision point which indicates whether or not there is a linear correlation. There is also a symmetric point between -1 and 0. The value of the decision point depends on the sample size.
In the next table there are positive decision points for different sample sizes between 5 and 100.

n	decision point	n	decision point	n	decision point	n	decision point
5	0,878	12	0,576	19	0,456	30	0,301
6	0,811	13	0,553	20	0,444	40	0,312
7	0,754	14	0,532	22	0,423	50	0,279
8	0,707	15	0,514	24	0,404	60	0,254
9	0,666	16	0,497	26	0,388	80	0,220
10	0,632	17	0,482	28	0,374	100	0,196
11	0,602	18	0,468				

Table 1: Positive decision points for the linear correlation

The values of the decision points decrease as n increases.

If r is between the negative and the positive decision point, we have no arguments that there is a linear correlation between the two variables.

If r is higher than the positive decision point or lower than the negative one, we have a linear correlation between the two variables.

The existence of a correlation between the two variables, does not imply the existence of a cause-and-effect relationship. Therefore, for example, if X represents the children allocation in the past 10 years and Y is the alcoholic drinks consumption in the last 10 years, a sample will indicate a high positive correlation. Certainly the increase in the allocations has not caused the increase in the sales of alcoholic beverages or viceversa.

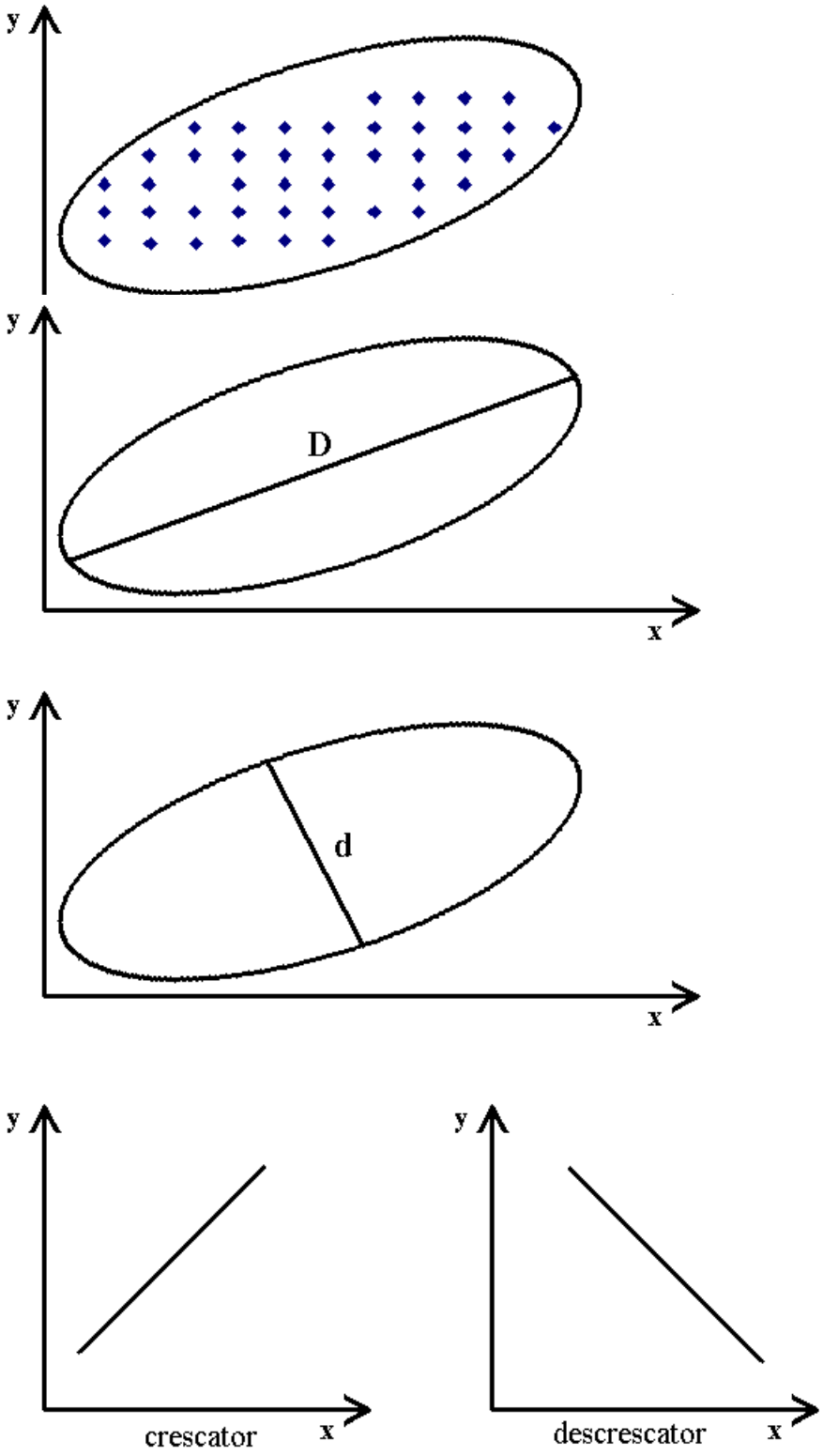
A quick estimation method of the linear correlation coefficient r for a sample, is as follows:

a) Draw a closed curve around the set of pairs (x, y):

c) Determine the length of the minimum diameter d:
b) Determine the length D of the maximum diameter:

d) The value r is estimated with $\pm\left(1-\frac{d}{D}\right)$, where the sign is chosen according to the orientation of the diameter D:

It should be noted that this estimating technique is very crude. It is very sensitive to the "spread" of the diagram. however, if the range of the values of X and the range of the values of Y are approximately equal, the approximation will be helpful.



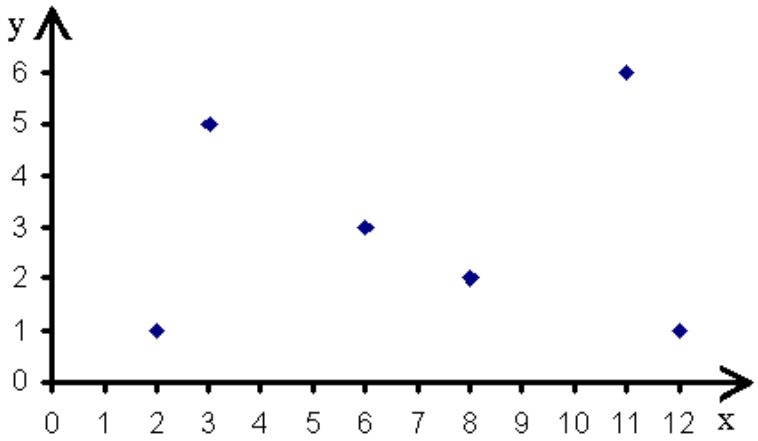
2.20 Linear correlation analysis

In section 20 we have seen the formula of the coefficient of linear correlation r between two variables X, Y that measures the strength of the linear relationship of the two variables.

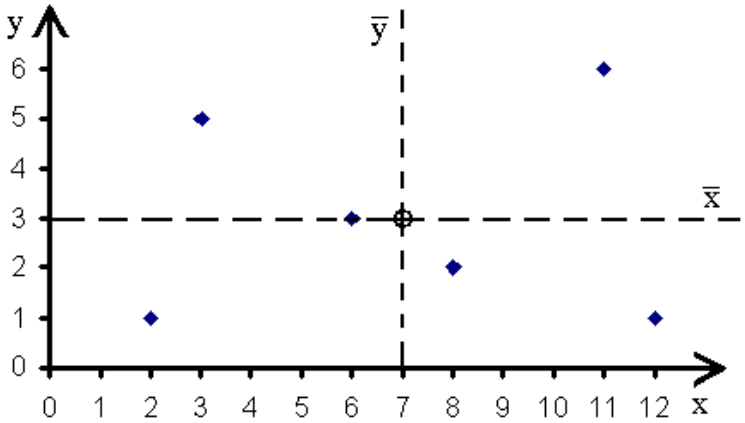
We shall now present a more complex analysis of this formula. We consider the following set of bivariate data:

x	2	3	6	8	11	12
y	1	5	3	2	6	1

The scatter diagram is:

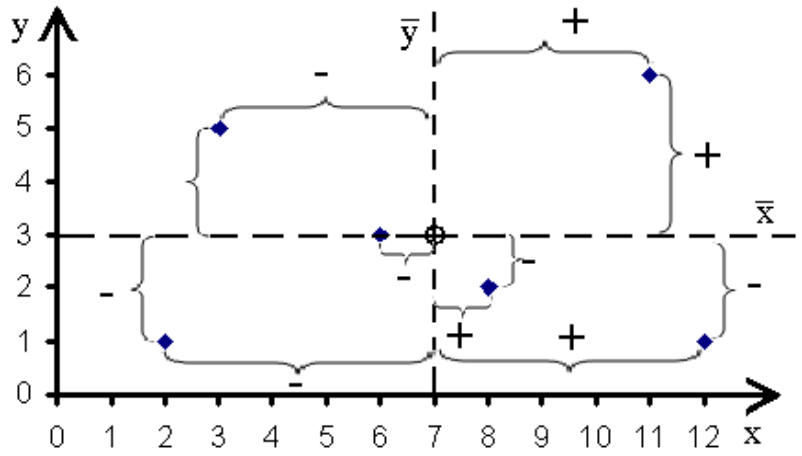


The mean $\overline{x}$ of the variable x is 7: $\overline{x} = 7$, and the mean of the variable y is 3: $\overline{y} = 3$. The point $(\overline{x}, \overline{y})$ is the point $(7, 3)$ and it is called the **centroid** of the data:



If a vertical and a horizontal line are are drawn through $(\overline{x}, \overline{y})$, the graph is divided into four sections. Each data (x, y) lies at a certain distance from each of these lines; $x - \overline{x}$ is the horizontal distance from (x, y) to the vertical line passing through the centroid and $y - \overline{y}$ is the vertical distance from (x, y) to the horizontal line passing through the centroid. The distances may be positive, negative or zero depending on the position of the point (x, y) in reference to $(\overline{x}, \overline{y})$.
One measure of linear dependency is the covariance. The covariance of X and Y is defined as the sum of the products of the distances of all values of X and Y from the centroid, divided by n:

$$cov(x, y) = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y})}{n}$$



The covariance of the data given above is $0,6$.

Positive covariance means that the scatter diagram is dominated by points to the upper right and to the lower left of the centroid. This is because the products $(x-\overline{x})\cdot(y-\overline{y})$ in the points of these regions are positive.

If the scatter diagram is dominated by points to the upper left and lower right of the centroid, the covariance is then negative because the products $(x-\overline{x})\cdot(y-\overline{y})$ for points of these regions, are negative.

The biggest disadvantage of covariance as a measure of linear dependency is that it does not have a standardized unit of measure. One reason for this is that the spread of the data is a strong factor in the size of the covariance.

For example, if we were to multiply each data point in the previous table by 10 we will have the following table:

x	20	30	60	80	110	120
y	10	50	30	20	60	10

The covariance for these data is 60, but this does not mean that the linear dependency between X, Y is stronger. In fact the linear dependency is the same, only the data are more spread out. This is the trouble with covariance as a measure.

We have to find a way to eliminate the effect of the spread of the data when we measure dependency.

If we standardize X and Y by dividing each one's deviation from the mean by its standard deviation:

$$x' = \frac{x-\overline{x}}{s_x} \quad \text{and} \quad y' = \frac{y-\overline{y}}{s_y}$$

and we calculate the covariance of X' and Y', we will have a covariance that is not influenced by the spread of the data. This fact is accomplished when introducing the coefficient of linear correlation, r. Therefore, the coefficient of linear correlation is:

$$r = cov(X', Y') = \frac{covX, Y}{s_x \cdot s_y}$$

The coefficient of linear correlation standardizes the measure of dependency and allows us to compare the relative strengths of dependency of different sets of data. The formula of the coefficient of linear correlation is also commonly referred to as Pearson's product moment.

The value of r for the set of data considered in the beginning is:

$$r = \frac{0,6}{(4,099) \cdot (2,098)} = 0,07$$

Because determining the coefficient of linear correlation with the formula:

$$r = \frac{covar X, Y}{s_x \cdot s_y}$$

is quite tedious, we may use its following workable form:

$$r = \frac{SS(X,Y)}{\sqrt{SS(X) \cdot SS(Y)}}$$

This last one avoids the separate calculations of $\overline{x}$, $\overline{y}$, s_x, s_y and the calculations of the deviations from the mean.

2.21 Inferences about the linear correlation coefficient

After the linear correlation coefficient r has been calculated for the sample data, it seems naturally to ask this question: does the value of r indicate that there is a dependency between the two variables in the population from which the sample was drawn?

To answer this question we shall perform a hypothesis test.

Step 1. The formulation of the null hypothesis H_0:

"The two variables are linearly unrelated."

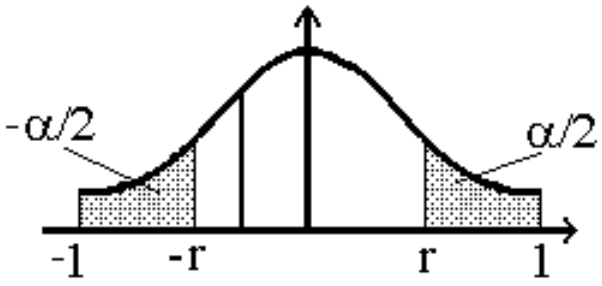
This means that $\rho = 0$, ρ being the linear correlation coefficient for the population.

Step 2. The formulation of the alternative hypothesis.
This can be one- or two-tailed. Most frequently is the two-tailed one $\rho \neq 0$. However, if we suspect that there is only a positive or only a negative correlation, we should use a one-tailed test The alternative hypothesis for a one-tailed test is: $\rho > 0$ or $\rho < 0$.

Step 3. The critical region for the test is on the right is we expect a positive correlation, and it is on the left a negative correlation is expected.

The test statistic used to test the null hypothesis is the value of r from the sample. The critical values for r are found in the following table at the intersectionof the column identified by the appropriate value of α and the row identified by the degrees of freedom $df = n - 2$:

The critical values for r if $\rho = 0$



df|α	0,10	0,05	0,02	0,01
1	0,988	0,997	1,000	1,000
2	0,900	0,950	0,980	0,980
3	0,805	0,878	0,934	0,959
4	0,729	0,811	0,882	0,917
5	0,669	0,754	0,833	0,874
6	0,662	0,707	0,789	0,834
7	0,582	0,666	0,750	0,798
8	0,549	0,632	0,716	0,765
9	0,521	0,602	0,685	0,735
10	0,497	0,576	0,658	0,708
11	0,476	0,553	0,634	0,684
12	0,458	0,532	0,612	0,661
13	0,441	0,514	0,592	0,641
14	0,426	0,497	0,574	0,623
15	0,412	0,482	0,558	0,606
16	0,400	0,468	0,542	0,590
17	0,389	0,456	0,528	0,575
18	0,378	0,444	0,516	0,561
19	0,369	0,433	0,503	0,549
20	0,360	0,423	0,492	0,537
25	0,323	0,381	0,445	0,487
30	0,296	0,349	0,409	0,449
35	0,275	0,325	0,381	0,418
40	0,257	0,304	0,358	0,393
45	0,243	0,288	0,338	0,372
50	0,231	0,273	0,322	0,354
60	0,211	0,250	0,295	0,325
70	0,195	0,232	0,274	0,302
80	0,183	0,217	0,256	0,283
90	0,173	0,205	0,242	0,267
100	0,164	0,195	0,230	0,254

The values from this table are critical values for r for a two-tailed test.
For a one-tailed test the value of α is twice the value of α used in hypothesis testing.

Step 4. Determine r from the sample.

Step 5. Establish whether or not r is in the critical region.

The failure to reject the null hypothesis is interpreted as meaning that linear dependency between the two variables in the population has not been shown.

Caution: This does not mean we have established a cause-and-effect relationship but only a mathematical relationship which allows the prediction of the behavior of the output variable Y from the behavior of the input variable X.

Example 2.21.1. *For our table of data:*

x	2	3	6	8	11	12
y	1	5	3	2	6	1

we have $n = 6$, and $r = 0,07$. The question is whether this value of r is significantly different from zero is the level of significance is $\alpha = 0,02$?

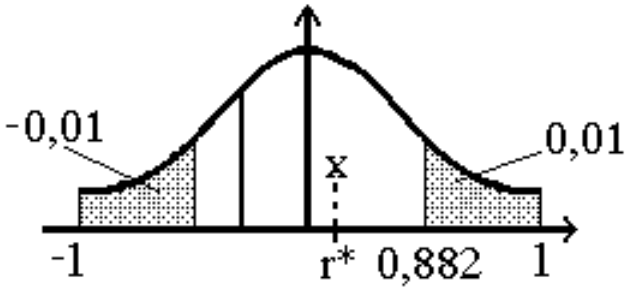
Step 1. $H_0 : \rho = 0$

Step 2. $H_0 : \rho \neq 0$

Step 3. We have $\alpha = 0,02$ and $df = n - 2 = 6 - 2 = 4$. The critical values from the table are: $-0,882$ and $0,882$.

Step 4. The calculated value of r is $r^* = 0,07$

Step 5. We accept H_0.

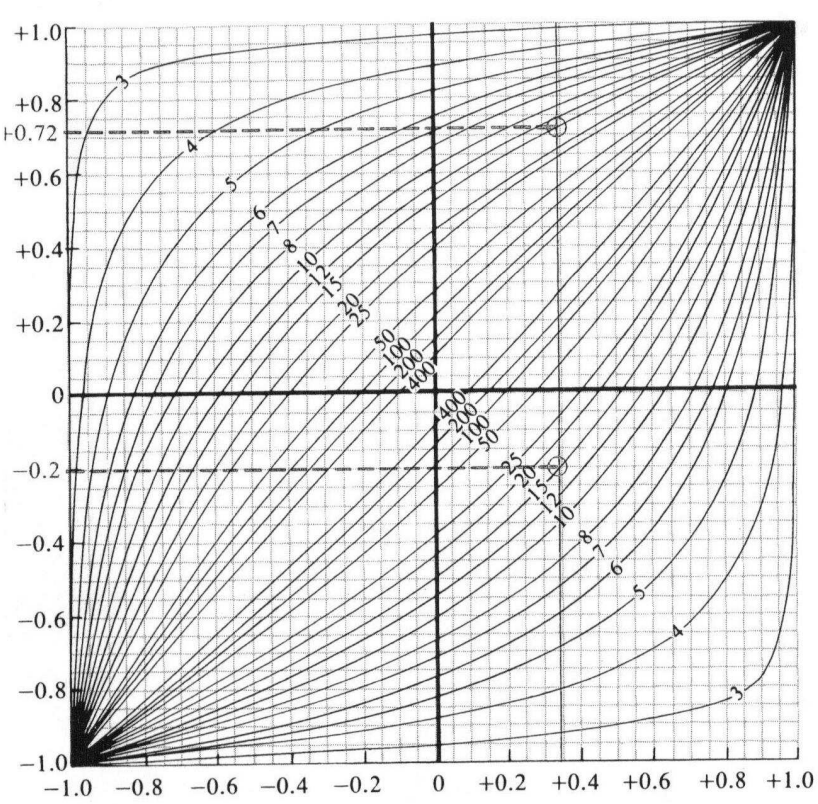


Conclusion: We could not show that X, Y are correlated. If we fail to reject the null hypothesis it means that we have proved the linear independency of the two variables.

As in other problems, sometimes a confidence interval estimate of the population correlation coefficient ρ is required. It is possible to estimate the value of ρ using a table that shows the confidence belts. The next table represents such confidence belts for 95% confidence interval estimates: The following example shows how such a table is to be read.

Example 2.21.2. *For a sample of* 15 *pairs of data, a calculated value of r is $r = 0,35$. Determine the* 95% *confidence interval estimate for the linear correlation coefficient ρ of the population?*

1) We locate $0,35$ on the horizontal axis (the axis of the linear correlation coefficient) and we draw a vertical line.



2) We determine the intersection of the vertical line with the belts corresponding to the sample size (15) and we have two points on the vertical line.

3) The confidence interval is the interval determined by the ordinates of these points $(-0,20, -0,72)$ (the ordinate axis is the axis of the linear correlation coefficient of the population).

2.22 Linear regression

If the value of the linear correlation coefficient r indicates a high linear correlation, there arises the problem of the establishment of an exact numerical relationship. This exact relationship is obtained by linear regression.

Generally, the statistician looks for an equation to describe the relationship between the two variables. The chosen equation is the best fitting of the scatter diagram. The equations found are called prediction equations. Here are some examples of such equations:

$$y = b_0 + b_1 \cdot x \text{ - linear}$$

$$y = a + b \cdot x + c \cdot x^2 \text{ - quadratic}$$

$$y = a \cdot b^x \text{ - exponential}$$

$$y = a \cdot \log_b x \text{ - logarithmic.}$$

The final purpose is to make predictions using these equations. Generally an exact value of the variable Y is not predicted. We are satisfied if the prediction is reasonably close.

Definition 2.22.1. The linear regression establishes the mean linear dependency of y in terms of x.

Next, we shall describe how to establish the best linear dependency for a set of data (x, y).
If a straight-line relationship seems appropriate, the best-fitting straight line is found by using the method of least squares.
Suppose that $\hat{y} = b_0 + b_1 \cdot x$ is the best linear relationship. The least squares method requires that b_0 and b_1 are such that $\sum(y - \hat{y})^2$ is minimum.

From Fermat's theorem we have that the minimum values of the function:

$$F(b_0, b_1) = \sum(y - b_0 - b_1 \cdot x)^2$$

are obtained for

$$b_1 = \frac{\sum(x - \overline{x}) \cdot (y - \overline{y})}{\sum(x - \overline{x})^2}, \qquad b_0 = \frac{1}{n} \cdot \Big(\sum y - b_1 \cdot \sum x\Big)$$

b_1 is the slope, and b_0 is the y-intercept.
To determine the slope b_1 we usually use the equivalent formula:

$$b_1 = \frac{SS(x, y)}{SS(x)}$$

where: $SS(x) = \sum x^2 - \frac{1}{n} \cdot \Big(\sum x\Big)^2$ and $SS(x, y) = \sum x \cdot y - \frac{1}{n} \cdot \Big(\sum x \cdot \sum y\Big)$.
We mention here that the expressions $SS(x, y)$ and $SS(x)$ also appear in the formula of the linear correlation coefficient. Therefore, when computing r we can also compute b_1.

Example 2.22.1. *For a sample of 10 individuals let us consider the following set of data.*

x	27	22	15	35	30	52	35	55	40	40
y	30	26	25	42	38	40	32	54	50	43

To determine the line of best fit $\hat{y} = b_0 + b_1 \cdot x$ we calculate $SS(x, y)$ and $SS(x)$ and we have:

$$SS(x, y) = 919, 0 \quad \text{and} \quad SS(x) = 1396, 9$$

from where we find out that b_1 is:

$$b_1 = \frac{919, 0}{1396, 9} = 0, 6599 \approx 0, 66.$$

To determine the y-intercept b_0 we will eventually have:

$$b_0 = \frac{1}{10}[380 - 0, 65 \cdot 351] = 14, 9077 \approx 14, 9$$

So the best linear relationship is:

$$\hat{y} = 14, 9 + 0, 66 \cdot x$$

Remark 2.22.1.

a) The slope b_1 the predicted change of the variable y according to an unit increase of the variable x.

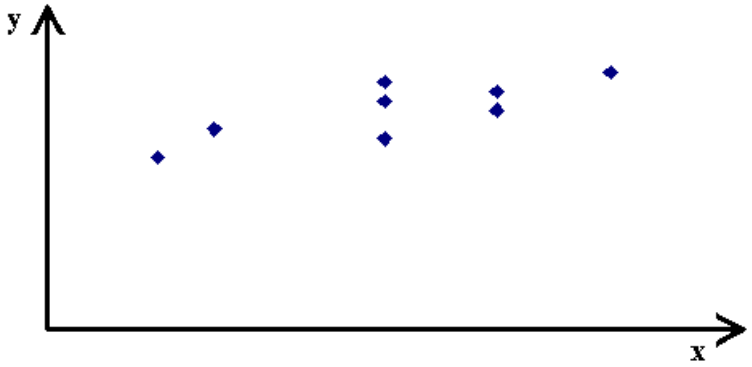
b) The ordinate b_0 represents the value of y in $x = 0$. We can say that b_0 is the predicted value of y for $x = 0$, only if $x = 0$.

c) The best linear relationship is a line passing through the point having the coordinates $(\overline{x}, \overline{y})$. This fact can be used as a verification when we draw the line of best fit.

Example 2.22.2. *For a sample of 8 individuals let us consider the following set of data*

x	65	65	62	67	69	65	61	67
y	105	125	11	120	140	135	95	130

The scatter diagram of this set of data suggests a linear correlation.



To find the line of best fit, we calculate $SS(x, y)$ and $SS(x)$ and we have:

$$SS(x, y) = 230, 0 \quad \text{and} \quad SS(x) = 48, 875$$

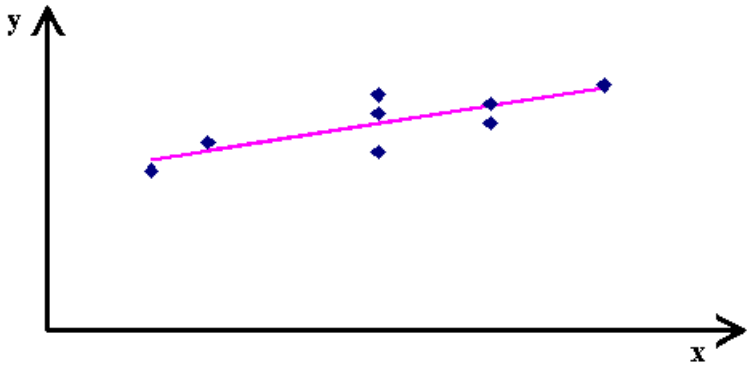
From here we have:

$$b_1 = \frac{230, 0}{48, 875} = 4, 706 \approx 4, 71.$$

$$b_0 = \frac{1}{n}\left[\sum y - b_1 \cdot \sum x\right] = -186, 478 \approx 186, 5$$

wherefrom:

$$\hat{y} = -186, 5 + 4, 71 \cdot x$$



Remark 2.22.2. *A crude estimation of the best-fitting straight line can be made as follows:*

- *as for the approximation of the linear correlation coefficient r we consider a closed curve around the set of pairs (x, y);*
- *the maximum diameter of the set is an approximation of the linear dependency graph;*
- *we right the linear dependency equation as the equation of a line passing through two points of this diameter;*
- *as for the estimation of r this estimation is a crude one and has to be used as such.*

2.23 Linear regression analysis

The linear model used to explain the linear dependence of two variables referring to the same population, is defined by the equation:

$$y = \beta_0 + \beta_1 \cdot x + \varepsilon$$

This equation represents the linear relationship between two variables x and y from a population. In this relationship:
- β_0 is the y-intercept;
- β_1 is the slope;
- y is the value observed at a given value of x;
- $\beta_0 + \beta \cdot x$ is the mean of y for the given value of x

We observe that the value ε depends on x. For the values $x_1, x_2, \ldots, x_n$ of x the linear model is:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i, i = 1, 2, \ldots, n$$

- ε is the random error of the observed value y for a given value of x that represents the deviation from the mean of the observed value y.

The regression line $\hat{y} = b_0 + b_1 \cdot x$ obtained from the sample data $(x_i, y_i), i = 1, 2, \ldots, n$ gives us b_0, which is an estimate for β_0, and b_1 which is an estimate for β_1. We will then be able to write $y_i = b_0 + b \cdot x_i + e_i$. The error is approximated by $y_i - \hat{y}_i$ which is the difference between the observed value y_i and the predicted value $\hat{y}_i$ of y for a given value of x. Because $\hat{y}_i = b_0 + b_1 \cdot x_i$ we have that:

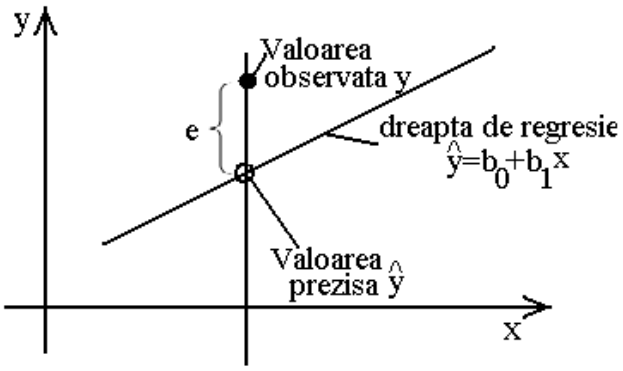
$$e_i = y_i - \hat{y}_i$$

The errors e_i are known as **residues**.
The random variable e has the following properties:

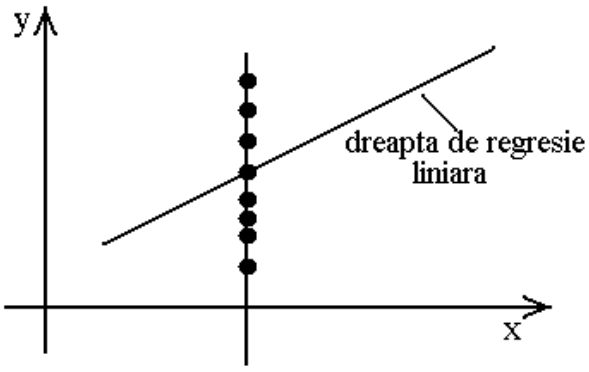
- $e > 0 \Longleftrightarrow y > \hat{y}$;
- $e < 0 \Longleftrightarrow y < \hat{y}$;
- for a given x the sum of the errors for different values of x is zero; this is a consequence of the method of least squares; therefore the mean of the experimental errors is zero: $\sum_{i=1}^{n} e_i = 0$.

We denote with σ_ε^2 the variance of the random errors of the data observed; we want to estimate this variance.

But before we estimate the variance σ_ε^2 let us analyze what the error εrepresents. ε represents the difference between the observed value y and the mean value of y for a given value of x. As we do not know the mean value of y, we will use the regression equation and estimate it with $\hat{y}$ the predicted value of y at this same value of x. therefore the estimation if ε is $e = y - \hat{y}$.



If for a given value x we have more observed values y these can be represented on the vertical axis Ox in x.



There is a similar distribution for each value of x. The mean value of the observed data y depends on x and is estimated by $\hat{y}$.

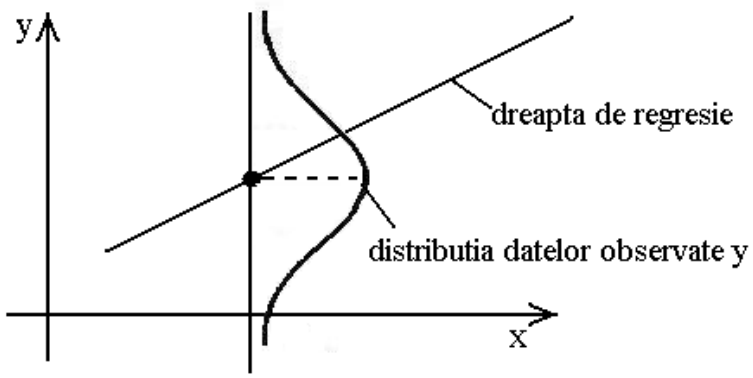
In other words, the standard deviation from the mean of the data distribution y is the same for each x:

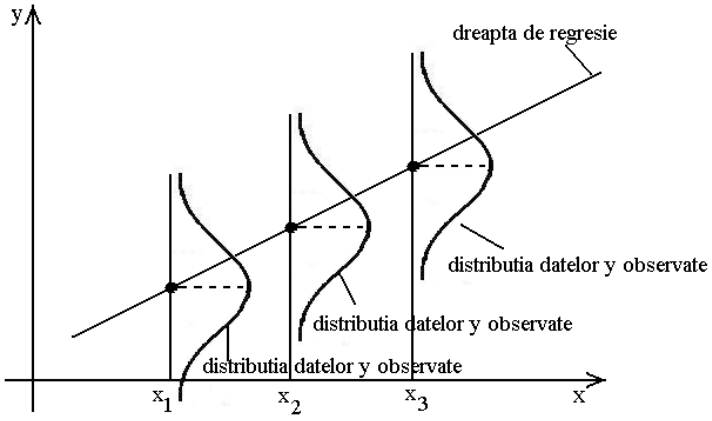
We recall that the variance s^2 of a set of statistical data $x_1, x_2, \ldots, x_n$ has been defined with the formula:

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

The determination introduces a complication because the mean of the data y differs from an x to another. For each x the mean is estimated by the predicted value $\hat{y}$ that corresponds to x by the regression line. Thus, the variance of the error ε is estimated with the formula:

$$s_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$





that shows that the variance of the error ε is the variance of the variable y around the regression line. The variance of the error s_ε^2 can be written as:

$$s_\varepsilon^2 = \frac{1}{n}\sum (y - b_0 - b_1 \cdot x_i)^2 = \frac{1}{n}\left[\sum y_i^2 - b_0 \cdot \sum y - b_1 \cdot x_i \cdot y_i\right]$$

and it is an estimate for σ_ε^2

Example 2.23.1. *Suppose a person moves to Timi'soara and takes a job. He wants to know the time needed to commute to and from work by car. To find an answer to this question he asks 15 colleagues about their one-way travel time and the distance to work. The resulting data are shown in the table below:*

coleg	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x - distance (in km)	3	5	7	8	10	11	12	12	13	15	15	16	18	19	20
y - time (in min)	7	20	20	15	25	17	20	35	26	25	35	32	44	37	45

To find an answer to his problem he has to determine the regression line and the variance s_ε^2.

Using the formulas, he finds:

$$SS(x) = 2,616 - \frac{(184)^2}{15} = 358,9333$$

$$SS(x,y) = 5,623 - \frac{(184)\cdot(403)}{15} = 679,53333$$

$$b_1 = \frac{358,9333}{679,53333} = 1,893202 \approx 1,89$$

$$b_0 = \frac{1}{15}\left[403 - (1,893202)\cdot(184)\right] = 3,643387 \approx 3,64$$

$$\hat{y} = 3,64 + 1,89 \cdot x.$$

This is the formula to estimate the mean time he needs to get to work depending on the distance x between his home and his workplace.

To find the standard deviation from the estimated value he will also have to calculate the variance s_ε^2. He finds: $s_\varepsilon^2 = 29,17$.

2.24 Inferences concerning the slope of the regression line

Now that the equation of the regression line has been determined we wonder when we can use this equation to predict the values of the variable y depending on x?

We will answer this question by hypothesis testing. Before making an inference concerning the regression line we consider the following hypotheses:

- for each x the distribution of the observed data y is approximatively normal;
- for each x the variance of the distribution of the observed data is the same.

Before we pass to the five steps let let us analyze the distribution of the slopes obtained for random samples of size n. These slopes b_1 are almost normally distributed with a mean of β_1, the population value of the slope, and with a variance of $\sigma_{b_1}^2$ where:

$$\sigma_{b_1}^2 = \frac{\sigma_\varepsilon^2}{\sum (x-\overline{x})^2}$$

An appropriate estimator for $\sigma_{b_1}^2$ is obtained by replacing σ_ε^2 with s_e^2:

$$s_{b_1}^2 = \frac{s_e^2}{\sum (x-\overline{x})^2}$$

This formula may be rewritten in the following form:

$$s_{b_1}^2 = \frac{s_e^2}{SS(x)} = \frac{s_e^2}{\sum x - \left[(\sum x)^2/n\right]}$$

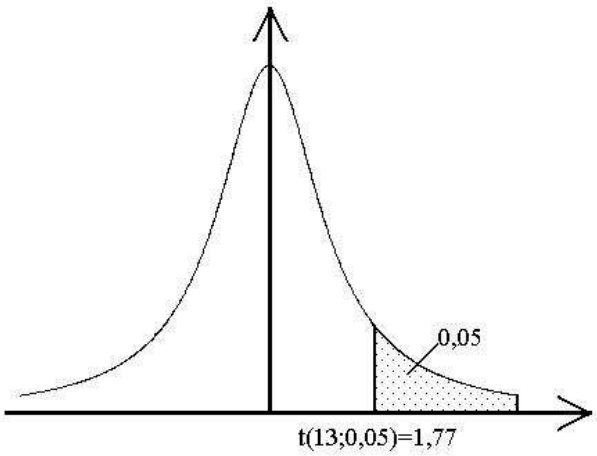
The standard error of the regression (slope) is σ_{b_1} and it is estimated bys_{b_1}.

We are now ready to go to the hypothesis test:

Step 1. **The formulation of** H_0. The null hypothesis will be $\beta_1 = 0$. If $\beta_1 = 0$ then we cannot use the linear equation to predict the value of y this means that: $\hat{y} = \overline{y}$.

Step 2. The alternative hypothesis can be either one-tailed or two-tailed. If we suspect that the slope is positive, a one-tailed test is appropriate: $H_a : \beta_1 > 0$.

Step 3. As a test statistic we use t. The number of degrees of freedom for the test is $df = n - 2$. For the example 2.23.1 of travel times and distances, we have $df = 15 - 2 = 13$. For a level of significance $\alpha = 0,05$, the critical value of t is $t(13; 0,05) = 1,77$.



The formula used to calculate the value of the test statistic t for inferences about the slope is:

$$t^* = \frac{b_1 - \beta_1}{s_{b_1}}$$

Step 4. Considering the equality $s_{b_1}^2 = \dfrac{s_e^2}{SS(X)}$ for the considered example we find that the value of the test statistic is:

$$t^* = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{1,89 - 0}{\sqrt{0,0813}} = 6,629 \approx 6,63$$

Step 5. Decision: reject the null hypothesis H_0 because t^* is in the critical region.
Conclusion: The slope of the line of best fit in the population is greater than zero. The evidence indicates that there is a linear relationship and the travel time to work can be predicted based on the one-way distance.

The slope β_1 of the regression line of the population can be estimated by means of a confidence interval. The confidence interval is given by:

$$b_1 \pm t(n-2; \frac{\alpha}{2}) \cdot s_{b_1}$$

The 95% confidence interval for the estimate of the population slope is:

$$1,89 \pm 2,16 \cdot \sqrt{0,0813} = 1,89 \pm 0,62$$

Thus $1,27$ to $2,51$ is the 95% confidence interval for β_1.

Bibliography

[1] V. Craiu, *Teoria probabilitatilor cu exemple si probleme*, Editura Fundatiei Romania de maine, 1997.

[2] R. Mittelhammer, *Mathematical Statistics for Economics and Business*, Springer, 1996.

[3] R. Johnson, *Elementary Statistics*, Duxbury Press, 1984, Boston

[4] T. Andrei, A. Stancu, *Statistica - teorie si aplicatii*, Editura All, 1995, Bucuresti

[5] T.H. Wonacott, R.J. Wonacott: *Statistique*, Economica, 4me dition, 1991, Paris

[6] Gh. Constantin, N. Surulescu, D. Zaharie, *Lectii de statistica descriptiva*, Universitatea de Vest, 1998, Timisoara

[7] Gh. Bocsan, *Estimarea parametrilor modelelor statistice*, Universitatea de Vest, 1995, Timisoara

[8] Y.G. Udny, M.G. Kendall, *Introducere in teoria statisticii*, Editura Stiintifica, 1969, Bucuresti