9 STATISTICS AND PROBABILITY

ABSTRACT:

In the first four chapters, we start with a review of classical probabilistic tools such as counting principles, probability concepts, definitions and theorems like conditional probability, Bayes' rule, and independence of events.

In chapters 5-8 we introduce the notion of discrete and continuous random variables and their distributions. We present the most important examples like binomial, Poisson's and normal distributions. By presenting numerous examples we show their applications.

The last chapters are devoted mainly to statistics. We give basic definitions and present the main ideas. Finally, we perform statistical tests and present the regression analysis method.

In the end, we show the application of Poisson distribution in reliability theory. A set of exercises for individual activities and learning is available.

AIM: **To acquire skills in solving real-time problems, especially those related to the sea and ships, by using probabilistic and statistical methods. It turns out that many problems in maritime domain are not deterministic and probabilistic tools are necessary.**

Learning Outcomes:

1. Compute with classical definitions probability using basic counting principals
2. Compute total and conditional probability
3. Check the independence of events, and learn their applications
4. Learn and apply discrete random variable distributions
5. Learn the importance of normal distributions
6. Perform statistical tests
7. Calculate the linear regression
8. Learn applications of probability and statistics in the maritime domain

Prior Knowledge:

Elementary mathematics, arithmetic, set theory, elements of Calculus, in particular definite integrals.

Relationship to real maritime problems:

The reason why we apply the probabilistic and statistical methods in maritime problems is that these problems are often unpredictable. The movement of ships in the sea is partially random. Partially means that, of course, we can steer the ship and respond to the actual situation, but

Co-funded by the Erasmus+ Programme of the European Union

in many extremely difficult situations it is not enough. We cannot predict many potentially risky conditions early enough. It is since we cannot describe the sea in the deterministic way. Sea waves are not deterministic. There is no “sea equations” that can describe how waves will behave at a place and time, even if we know the present situation there.

The second reason, why probability and statistics are so important in the maritime domain, is the reliability theory, which we apply to ship equipment. Ships spend usually long days in the sea, so we must be sure that all mechanical, electrical and electronic devices will be working properly.

Co-funded by the
Erasmus+ Programme
of the European Union

Content

9 STATISTICS AND PROBABILITY 1

9.1 COUNTING PRINCIPLES 4

9.2 EVENTS. PROBABILITY 10

9.3 GEOMETRIC PROBABILITY 13

9.4 COMPLETE SYSTEM OF EVENTS. BAYES' RULE 15

9.5 DISCRETE PROBABILITY DISTRIBUTIONS 20

9.6 BINOMIAL DISTRIBUTION AND POISSON DISTRIBUTION 26

9.7 CONTINUOUS PROBABILITY DISTRIBUTIONS 28

9.8 NORMAL DISTRIBUTION 32

9.9 STATISTICS, BASIC DEFINITIONS AND NOTATIONS 36

9.10 TESTING OF STATISTICAL HYPOTHESES. STATISTICAL CONCLUSION ERRORS 39

9.11 REGRESSION, CORRELATION 44

9.12 CONNECTIONS AND APPLICATIONS 48

9.13 EXERCISES 49

Co-funded by the
Erasmus+ Programme
of the European Union

9.1 COUNTING PRINCIPLES

9.1.1 Multiplication principle (the rule of product)

If a set A contains n elements, a set B contains m elements then there are $n \cdot m$ different ordered pairs (a, b) where a belongs to A and b belongs to B.

Example 9.1

How many two-digit even natural numbers are there?

Solution:

Two-digit even natural numbers can be represented as ordered pairs (a, b), where $a \in \{1,2,3,4,5,6,7,8,9\} = A$ and $b \in \{0,2,4,6,8\} = B$. The number of elements in the set A is $n = 9$ and the number of elements in the set B is $m = 5$. Thus, the total number of two-digit even natural numbers is

$$9 \cdot 5 = 45.$$

Note that, since ordered triples (a, b, c) can be represented as ordered pairs $((a, b), c)$, we can extend the multiplication principle to the case of three sets.

9.1.2 Multiplication principle (three and more sets)

If a set A contains n elements, a set B contains m elements and a set C contains p elements, then there are $n \cdot m \cdot p$ different ordered triples (a, b, c), where a belongs to A, b belongs to B and c belongs to C. The similar rule holds for four, five and more sets.

Example 9.2

How many three-digit odd natural numbers are there?

Solution:

Three-digit odd natural numbers can be represented as ordered triples (a, b, c), where $a \in \{1,2,3,4,5,6,7,8,9\} = A$, $b \in \{0,1,2,3,4,5,6,7,8,9\} = B$ and $c \in \{1,3,5,7,9\} = C$. The number of elements in the set A is $n = 9$, the number of elements in the set B is $m = 10$ and the number of elements in the set C is $p = 5$. Thus, the total number of three-digit odd natural numbers is

$$9 \cdot 10 \cdot 5 = 450.$$

Co-funded by the
Erasmus+ Programme
of the European Union

9.1.3 Addition principle (the rule of sum)

If a set A contains n elements, a set B contains m elements and $A \cap B$ contains l elements, then $A \cup B$ contains $n + m - l$ elements.

Example 9.3

Alice bought at the grocery store: apples, plums, bananas, blueberries and oranges. Her husband John bought, in a different store, blackberries, gooseberries, grapes, plums, pears, grapefruits and bananas. How many different types of fruits will they eat today at home?

Solution:

We have

$n = 5,\ m = 7,\ l = 2.$

Hence, Alice and her husband have

$$m + n - l = 5 + 7 - 2 = 10$$

different fruits.

Definition: Permutations

A permutation is any arrangement of the elements of a finite set in a definite order.

The number of permutations

The number of all permutations of $n -$ elements set is

$$P_n = n! = n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1$$

where $n!$ is called the factorial of n.

Example 9.4

In how many ways can we arrange five different books on a shelf.

Solution:

We must calculate the number of all permutations of 5 – elements set. This number is equal to

$$P_5 = 5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120.$$

Co-funded by the Erasmus+ Programme of the European Union

Exercise 9.1

In how many ways can seven different ships dock in a harbour?

Solution:

5040.

9.1.4 Permutations of n elements taken k at time

The number of sequences of length k without repetitions whose elements are taken from $n-$ elements set $(k \leq n)$ at once is

$$P(n,k) = n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!}.$$

Sequences of length k without repetitions whose elements are taken from $n-$ elements set at once are often called *permutations of* n *elements taken* k *at once* or *variation without repetition* or $k-$ *permutations of* n.

Example 9.5

In how many ways can five people be seated in a row of eleven chairs.

Solution:

The number N of permutations of 11 elements taken 5 at once should be calculated. Thus,

$n = 11,\ k = 5$ and

$$N = P(11,5) = \frac{11!}{(11-5)!} = \frac{11!}{6!} = 11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 = 55440.$$

Example 9.6

There are nine free places to dock in a harbour. In how many ways can six different ships dock there?

Solution:

The number N of permutations of 9 elements taken 6 at once should be calculated. Thus,

$n = 9,\ k = 6$ and

$$N = P(9,6) = \frac{9!}{(9-6)!} = \frac{9!}{3!} = 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 = 60480.$$

Co-funded by the
Erasmus+ Programme
of the European Union

9.1.5 Permutations with repetition

The number of different sequences of length k that can be formed from $n-$ elements set, when repetitions are allowed is

$$n^k.$$

Different sequences of length k that can be formed from $n-$ elements set are often called *permutations with repetitions* or *variations with repetitions.*

In the following we will assume that any arrangement of letters forms a word.

Example 9.7

How many three-letter words can be arranged from the letters a, b, c, d?

Solution:

The number N of different sequences (words) of length 3 taken from $4-$ element set $\{a, b, c, d\}$, when repetitions are allowed, should be calculated. The phrase "repetitions are allowed" means that each letter can be taken more than once to form a word. Thus,

$n = 4,\ k = 3$ and $N = 4^3 = 64.$

Example 9.8

How many four-letter words can be arranged from the letters a, b, c?

Solution:

The number N of different sequences (words) of length 4 taken from $3-$ element set $\{a, b, c\}$, when repetitions are allowed, should be calculated. Thus,

$n = 3,\ k = 4$ and $N = 3^4 = 81$

Remark In some problems, the formula for the number of all permutations with repetitions, is not easy to apply. It is easier to use the multiplication principle.

Example 9.9

In how many ways can five ships from a country A dock in three different harbours in a country B?

Co-funded by the Erasmus+ Programme of the European Union

Solution:

1st way:

- Let harbours in a country B be denoted by h1, h2, h3.
- Each ship has three possibilities to choose: to dock in h1, in h2 or in h3.
- Using the multiplication principle, we see that:
- two ships have $3 \cdot 3 = 9$ possibilities to choose,
- three ships have $3 \cdot 3 \cdot 3 = 27$ posibilities to choose,
- four ships have $3 \cdot 3 \cdot 3 \cdot 3 = 81$ posibilities to choose.
- Finally, five ships have

$$N = 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 = 3^5 = 243$$

possibilities to choose.

2nd way:

We have sequences of length $k = 5$ (five ships) that consist of numbers from $3 -$ elements set $\{1,2,3\}$ (harbours) $(n = 3)$. Repetitions are allowed since two or more ships can choose the same harbour. Thus,

$$N = 3^5 = 243.$$

Definition: Combinations

An $k -$ element subset of a set with n elements $(k \leq n)$ is called a combination of n elements taken k at once.

9.1.6 The number of combinations

The number of all possibilities to choose a subset of k elements from a set of n elements (the order being irrelevant) is

$$C(n,k) = \binom{n}{k} = \frac{n!}{k!\,(n-k)!},$$

where $\binom{n}{k}$ is binomial coefficient.

Example 9.10

Anne, Beth, Charlie and Donald are trying to win two tickets to the Bahamas.

a) What are all the combinations of winners?
b) What is a number of all possibilities to choose winners?

Co-funded by the Erasmus+ Programme of the European Union

c) If Anne and Charlie are chosen, does it matter who got the first, and who got the second ticket?

Solution:

a) All the combinations of winners are:

$$\{A, B\}, \{A, C\}, \{A, D\}, \{B, C\}, \{B, D\}, \{C, D\},$$

where A stands for Anne, B stands for Beth etc.

b) The number of all possibilities to choose winners is six since it is the number of all combinations of winners. On the other hand, by using the formula we get the same number.

$$N = C(4,2) = \binom{4}{2} = \frac{4!}{2!\,(4-2)!} = \frac{4!}{2!\,2!} = \frac{4 \cdot 3}{2 \cdot 1} = 6.$$

c) No, it does not matter since sets (subsets) have no order. We have $\{A, C\} = \{C, A\}$.

Remark

The following relations for binomial coefficients are useful in calculations:

- $\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!} = \frac{n(n-1)\cdots(k+1)}{(n-k)!}$
- $\binom{n}{k} = \binom{n}{n-k}$
- $\binom{n}{0} = \binom{n}{n} = 1$
- $\binom{n}{1} = \binom{n}{n-1} = n$

Exercise 9.2

1. Compute $\binom{4000}{3998}$.

2. How many subsets of a set with four thousand elements have two elements?

Solution:

1. 7 998 000

2. 7 998 000

Co-funded by the Erasmus+ Programme of the European Union

Example 9.11

How many different committees of four members can be formed out of a club with ten members?

Solution:

The committee is an $4-$ element subset of an $10-$ element set of club members. Thus, the number N of different committees is equal to $C(10,4)$. We have

$$N = C(10,4) = \binom{10}{4} = \frac{10\cdot 9\cdot 8\cdot 7}{4!} = 210.$$

9.2 EVENTS. PROBABILITY

Example 9.12

Suppose that we conduct an experiment by tossing two different fair coins. Let H represents heads and T represents tails. Then any possible outcome of the experiment is an element of the set

$$\Omega = \{(H,H),(H,T),(T,H),(T,T)\}.$$

Definition: : Sample space, elementary event

A *sample space* is a set Ω of elements that correspond one to one with the outcomes of an experiment. Each of the elements of Ω is called an *elementary event*.

Definition: Event

An *event* is any subset of a sample space.

Remark

The sample space Ω is an event (*a certain event*, *sure event*), any elementary event is an event, the empty set $\emptyset$ is an event (*an impossible event*).

Example 9.13

Each letter of the word *leopard* is written on a separate card and the cards are shuffled. List a sample space for the outcome of drawing one card.

Answer: $\Omega = \{l,e,o,p,a,r,d\}$ (also $\Omega = \{e,p,l,a,d,o,r\}$ since the order is not important).

Now suppose we are interested in whether the letter drawn is a vowel. We call the drawing of a vowel an event A. The event A is the occurrence of any of the outcomes e, o, a. It can be seen as the set $A = \{e, o, a\}$ which is a subset of the sample space Ω.

Definition: Classical definition of probability

Let Ω be a sample space of an experiment in which there are n possible outcomes, each equally likely. If an event A is a subset of Ω such that A contains k elements, then the probability of an event A, denoted by $P(A)$, is given by

$$P(A) = \frac{k}{n}.$$

Example 9.14

Suppose that we conduct an experiment by tossing two different six-sided dices. What is the probability that we get ten in total?

Solution:

The number of elementary events is $n = 6 \cdot 6 = 36$. Denote by A the event that we get the sum ten. We have

$$A = \{(6,4), (5,5), (4,6)\}$$

hence $k = 3$ and

$$P(A) = \frac{3}{36} = \frac{1}{12}.$$

Let Ω be now a finite or infinite set. We will call Ω a sample space. Denote by F the class of subsets of Ω. Elements of F will be called events. If Ω is finite, all its subsets are events. If Ω is not finite, we consider as events a large infinite class of Ω subsets.

Definition: Probability (axiomatic)

A probability of an event A is a real number $P(A)$ which satisfies the following three conditions:

1. $P(A) \geq 0$

2. $P(\Omega) = 1$

3. For every sequence $A_1, A_2, \ldots$ of mutually exclusive events it holds

$$P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \cdots.$$

Co-funded by the
Erasmus+ Programme
of the European Union

Properties of probability

1. $P(\emptyset) = 0$
2. $0 \leq P(A) \leq 1$
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
4. $A \subset B \Rightarrow P(A) \leq P(B)$
5. $P(\bar{A}) = 1 - P(A)$
6. $P(A \setminus B) = P(A) - P(A \cap B)$
7. $B \subset A \Rightarrow P(A \setminus B) = P(A) - P(B)$
8. $$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Co-funded by the Erasmus+ Programme of the European Union

9.3 GEOMETRIC PROBABILITY

As a sample space we consider a 1D, 2D or 3D set with a finite measure $m(\Omega)$ length, area or volume respectively. As events we consider subsets of Ω with finite measure.

Definition: geometric probability

Probability that a point x lying in a set Ω lies also in its subset A with a finite measure $m(A)$ is defined as the ratio

$$P(A) = \frac{m(A)}{m(\Omega)}.$$

Example 9.15

A circle is inscribed in a square. A point is selected at random from the area of the square. Calculate the probability that it lies inside the circle.

Solution:

Let Ω be a square and A be a circle of radius $r > 0$ that is inscribed in Ω. We have

$$m(A) = \pi r^2, \qquad m(\Omega) = 4r^2.$$

$$P(A) = \frac{\pi r^2}{4r^2} = \frac{\pi}{4}$$

Example 9.16

Two ships have arriving time between $\mathbf{2}$ pm and $\mathbf{3}$ pm. Both ships must dock on the same berth. Once docked, it takes each ship $\mathbf{10}$ minutes to restock and leave the dock. What is the probability that the ships won't have to wait for the berth?

Solution:

As a unit of time, we set 10 minute. A time interval between 2 and 3 pm we denote by $[0,6]$. Let x be the time when the first boat will arrive, and y be the time when the second boat will arrive. We have $x, y \in [0,6]$ and $|x - y| \geq 1$.

Let $\Omega = [0,6] \times [0,6]$ be a sample space. The area of Ω is $\mu(\Omega) = 36$.

The event that both ships won't have to wait can be represented by the set

$$A = \{(x, y) \in \Omega : |x - y| \geq 1\}.$$

Since

$$A = \{(x, y) \in \Omega : y \leq x - 1 \text{ or } y \geq x + 1\},$$

the area of the set A is

$$\mu(A) = \frac{1}{2} \cdot 5 \cdot 5 + \frac{1}{2} \cdot 5 \cdot 5 = 25.$$

Co-funded by the Erasmus+ Programme of the European Union

From the geometric probability formula, we have

$$P(A) = \frac{m(A)}{m(\Omega)} = \frac{25}{36} \approx 0.6944.$$

There is a 69.44% chance that the ships won't have to wait for the berth.

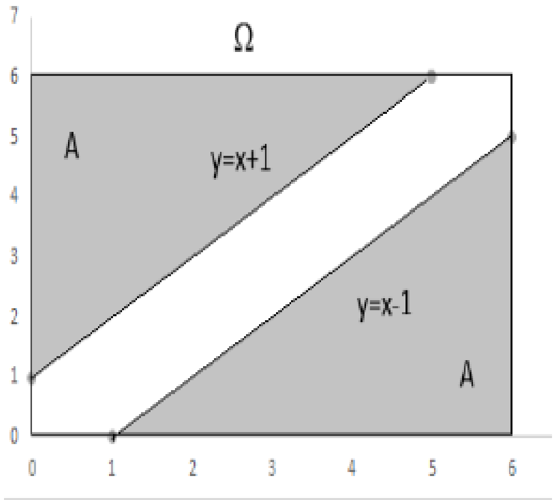

Figure 9.1. Illustration for Example 9.16

9.4 COMPLETE SYSTEM OF EVENTS. BAYES' RULE

Definition: Independent events

Two events A and B are called *independent* (statistically independent) if

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

Remark

Physically independent experiments are assumed statistically independent, e.g., successive throws of a coin, a die, etc.

Example 9.17

Suppose that two balls are drawn with replacement (the first ball is replaced before the second is drawn) at random from a bag containing 4 red and 3 black balls. What is the probability that both balls selected are red?

Solution:

Let A be the event of drawing a red ball the first time, and B the event of drawing a red ball the second time. The sample space Ω consists of all ordered pairs (x, y), where both x and y denote elements of a set of 7 outcomes (4 red, 3 black).

The multiplication rule implies that Ω contains $7 \cdot 7 = 49$ elements. Since $A \cap B$ consists of the ordered pairs of the form (red, red), there is $4 \cdot 4 = 16$ elements in $A \cap B$. Therefore,

$$P(A \cap B) = \frac{16}{49} = \frac{4}{7} \cdot \frac{4}{7} = P(A)P(B).$$

Hence, events A and B are independent.

Example 9.18

Suppose that two balls are drawn without replacement (the first ball is not replaced before the second is drawn) at random from a bag containing 4 red and 3 black balls.

Any one of 7 balls may be selected on the first draw. One of only 6 balls may be chosen on the second draw (no replacement).

Therefore, there are $7 \cdot 6 = 42$ elements in Ω (ordered pairs without repetitions).

Let A be the event of drawing a red ball the first time, and B the event of drawing a red ball the second time.

Co-funded by the Erasmus+ Programme of the European Union

There is of course $4 \cdot 6 = 24$ elements of the set A and $6 \cdot 4 = 24$ elements of the set B. This gives

$$P(A) = P(B) = \frac{24}{42} = \frac{4}{7}.$$

We have $4 \cdot 3 = 12$ elements of $A \cap B$. Therefore, we have

$$P(A \cap B) = \frac{12}{42} = \frac{2}{7} \neq \frac{4}{7} \cdot \frac{4}{7} = P(A)P(B).$$

Hence, events A and B are dependent.

Remark

Since in Example 9.18 $P(A \cap B) = \frac{12}{42} = \frac{4}{7} \cdot \frac{3}{6}$ and $P(A) = \frac{4}{7}$ we can interpret the second factor $\frac{3}{6}$ as the probability that the second ball drawn is red under the condition that the first ball drawn was red. Denote this by $P(A|B)$.

The last remark suggests a general law

Conditional probability formula

Let $P(A)$ denote the probability of an event A, and $P(B|A)$ denote the probability of an event B after event A has occurred. If $P(A \cap B)$ is the probability that A and B occur, then

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

If $P(A) > 0$ we have *conditional probability formula*

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

or if $P(B) > 0$

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

Remark

In the Example 9.18 $P(A|B) = \frac{3}{6} = \frac{1}{2}$ denotes the probability of an event A after event B has occurred. In the example the probability that the first ball drown was red under the condition that the second ball drawn is red.

$P(A|B)$ is called "the conditional probability of A given B", or "the probability of A under the condition B".

Remark:

If A and B are independent events, then $P(A|B) = P(A)$ if $P(B) > 0$.

Co-funded by the
Erasmus+ Programme
of the European Union

Exercise 9.3

Show that if A and B are independent events, then

A and $\bar{B}$ are independent events.

Solution:

$$\begin{aligned} P(A \cap \bar{B}) &= P\big(A \cap (\Omega \setminus B)\big) = P(A \setminus B) = P(A) - P(A \cap B) = P(A) - P(A)P(B) \\ &= P(A)\big(1 - P(B)\big) = P(A)P(\bar{B}) \end{aligned}$$

Exercise 9.4

Show that if A and B are independent events, then $\bar{A}$ and $\bar{B}$ are independent events.

Solution:

$$\begin{aligned} P(\bar{A} \cap \bar{B}) &= P\big((\Omega \setminus A) \cap (\Omega \setminus B)\big) = P\big(\Omega \setminus (A \cup B)\big) = 1 - P(A \cup B) \\ &= 1 - P(A) - P(B) + P(A \cap B) = 1 - P(A) - P(B) + P(A)P(B) \\ &= \big(1 - P(A)\big)\big(1 - P(B)\big) = P(\bar{A})P(\bar{B}). \end{aligned}$$

Bayes' rule

Suppose that $P(B) > 0$. Then

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

We say that Ω is partitioned into cases $E_1, \dots, E_n$ if events $E_1, \dots, E_n$ are mutually disjoint sets and $\Omega = E_1 \cup \dots \cup E_n$.

Total Probability Formula

Suppose that the sample space Ω is partitioned into $E_1, \dots, E_n$ cases.

Then for any event A

$$P(A) = P(A|E_1)P(E_1) + \cdots + P(A|E_n)P(E_n).$$

Example 9.19

Three identical bowls are labelled $1, 2, 3$. First bowl contains 3 red and 3 blue marbles. Second bowl contains 4 red and 2 blue marbles. Third bowl contains 1 red and 5 blue. First, a bowl is randomly selected, and then a marble is randomly selected from the bowl.

What is the probability that a marble selected is blue?

Co-funded by the Erasmus+ Programme of the European Union

Solution:

E_i – an event that bowl number i is selected. We have $P(E_i) = \frac{1}{3}$.

A – an event that blue marble is selected. We have

$$P(A|E_1) = \frac{1}{2}, \qquad P(A|E_2) = \frac{1}{3}, \qquad P(A|E_3) = \frac{5}{6}$$

and

$$P(A) = P(A|E_1)P(E_1) + P(A|E_2)P(E_2) + P(A|E_3)P(E_3) = \frac{1}{2}\cdot\frac{1}{3} + \frac{1}{3}\cdot\frac{1}{3} + \frac{5}{6}\cdot\frac{1}{3} = \frac{5}{9}.$$

Given that the selected marble was blue, what is the probability that bowl 2 was selected?

Solution:

Using Bayes' rule, we get

$$P(E_2|A) = \frac{P(A|E_2)P(E_2)}{P(A)} = \frac{\frac{1}{3}\cdot\frac{1}{3}}{\frac{5}{9}} = \frac{1}{5}.$$

Example 9.20

A certain disease has an incidence rate of 1%. Suppose that for some diagnostic test the false negative rate is 10% and false positive rate is 2%. Compute the probability that a person, chosen at random from the population, who tests positive actually has the disease.

Solution:

Define the events for the person in question:
D: has the disease, T: test indicate disease (test is positive)
The complementary events are:
$\bar{D}$: has not the disease, $\bar{T}$: test does not indicate disease (test is negative).
Since disease has an incidence rate of 1%, we have

$$P(D) = 0.01$$

and for the complementary event

$$P(\bar{D}) = 1 - P(D) = 0.99.$$

Since the false negative rate is 10%, we have

$$P(\bar{T}|D) = 0.1.$$

Since the false positive rate is 2%, we have

$$P(T|D) = 0.02$$

We have to compute $P(D|T)$.

We use Bayes' rule

Co-funded by the Erasmus+ Programme of the European Union

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)},$$

where

$$P(T|D) = 1 - P(\overline{T}|D) = 1 - 0.1 = 0.9.$$

By the total probability formula

$$P(T) = P(T|D)P(D) + P(T|\overline{D})P(\overline{D}) = 0.9 \cdot 0.01 + 0.02 \cdot 0.99 = 0.0288.$$

Finally,

$$P(D|T) = \frac{0.9 \cdot 0.01}{0.028} = 0.3125.$$

Only 31.25% of randomly selected people whose test is positive has in fact a disease.
The result seems surprising. The low incidence of the disease in the population is a critical variable here.

Exercise 9.5

In the above example compute the probability that a person, chosen at random from the population, who tests negative actually has not the disease.

Solution:

0.99897

Co-funded by the
Erasmus+ Programme
of the European Union

9.5 DISCRETE PROBABILITY DISTRIBUTIONS

Definition: random variable

A real valued function X defined on the sample space Ω is called a *random variable*.

Remark:

We write $X: \Omega \to \mathbb{R}$ and $X(\omega)$ is a value of X for the event ω in Ω.

Example 9.21

Suppose that we are tossing a coin. Then, sample space is $\Omega = \{H, T\}$, where H represents head and T represents tail. We define a random variable $X: \Omega \to \mathbb{R}$ by

$$X(H) = 1 \quad \text{and} \quad X(T) = 0.$$

Example 9.22

Suppose that we are rolling a dice. Then, sample space is $\Omega = \{1,2,3,4,5,6\}$, where the numbers represent the numbers on the dice. We define a random variable $X: \Omega \to \mathbb{R}$ by

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

We simply say that X is a number of dots rolled.

Remark

If a coin, we are tossing in Example 9.21 is fair the probability that we get a head is $\frac{1}{2}$ and the probability that we get a tail is $\frac{1}{2}$. We write this:

$$P(H) = P(\{\omega \in \Omega: X(\omega) = 1\}) = \frac{1}{2}$$

and

$$P(T) = P(\{\omega \in \Omega: X(\omega) = 0\}) = \frac{1}{2}.$$

For simplicity we will write $P(X = x) = P(\{\omega \in \Omega: X(\omega) = x\})$.

Similarly, we understand the notation: $P(X < x),\ P(X > x),\ P(X \le x),\ P(X \ge x)$.

Definition: cumulative distribution function

The cumulative distribution function (distribution function) of a random variable $\boldsymbol{X}$ is defined by $F(x) = P(X \le x)$ for every real number x.

Co-funded by the Erasmus+ Programme of the European Union

Properties of the cumulative distribution function

1. For every real x: $0 \leq F(x) \leq 1$
2. $\lim_{x\to-\infty} F(x) = 0, \lim_{x\to\infty} F(x) = 1$
3. $\lim_{x\to a^+} F(x) = F(a)$
4. F is nondecreasing

Definition: discrete random variable

A random variable is called *discrete* if it takes on with positive probability either a finite or at most a countably infinite set of discrete values i.e., there exists a finite sequence $x_1, \dots, x_n$ such that

(1)

$$P(X = x_i) = p_i, \quad \sum_{i=1}^{n} p_i = 1, \quad p_1, \dots, p_n \text{ positive numbers}$$

or there exists an infinite sequence $x_1, x_2, \dots$ such that

(2)

$$P(X = x_i) = p_i, \quad \sum_{i=1}^{\infty} p_i = 1, \quad p_1, p_2, \dots \text{ positive numbers.}$$

Definition: probability distribution

The formula (1) or resp. (2) is called probability distribution of discrete random variable.

Moreover, (1) can be written in the form

$X = x_i$	x_1	x_2	…	x_n
$P\,(X = x_i)$	p_1	p_2	…	p_n

or simply

x_i	x_1	x_2	…	x_n
p_i	p_1	p_2	…	p_n

Remark: In Example 9.21 if the coin is fair, we have

x_i	0	1
p_i	0.5	0.5

Co-funded by the Erasmus+ Programme of the European Union

Definition: Expected value

The *expected value* (*expectation* or *mean value)* of a discrete random variable X is defined by

$$E(X) = \sum_i x_i P(X = x_i).$$

Linearity of the expected value

If $E(X) < \infty$ and $E(y) < \infty$, then for any reals a, b

$$E(aX + bY) = aE(X) + bE(Y).$$

Expected value of a function of X.

Let $Y = g(X)$, where g is a function of discrete random variable X. Then Y is a discrete random variable with expectation

$$E(Y) = \sum_i g(x_i) P(X = x_i).$$

Definition: Variance, standard deviation

The *variance* of a discrete random variable X is defined by

$$V(X) = E\big(X - E(X)\big) = E(X^2) - E(X)^2.$$

The *standard deviation* of X is defined by

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

Example 9.23

A random variable X has the following distribution

$X = x_i$	-2	0	2	3
$P(X = x_i)$	0.2	0.4	0.1	0.3

Compute: $E(X)$, $E(X^2)$, $V(X)$ and $\sigma(X)$. Sketch the graph of a cumulative distribution function.

Solution:

$$E(X) = (-2) \cdot (0.2) + 0 \cdot (0.4) + 2 \cdot (0.1) + 3 \cdot (0.3) = 0.7$$
$$E(X^2) = (-2)^2 \cdot (0.2) + 0^2 \cdot (0.4) + 2^2 \cdot (0.1) + 3^2 \cdot (0.3) = 3.9$$

Co-funded by the
Erasmus+ Programme
of the European Union

$V(X) = 3.9 - (0.7)^2 = 3.41$
$\sigma(X) = \sqrt{3.41} \approx 1.8466$

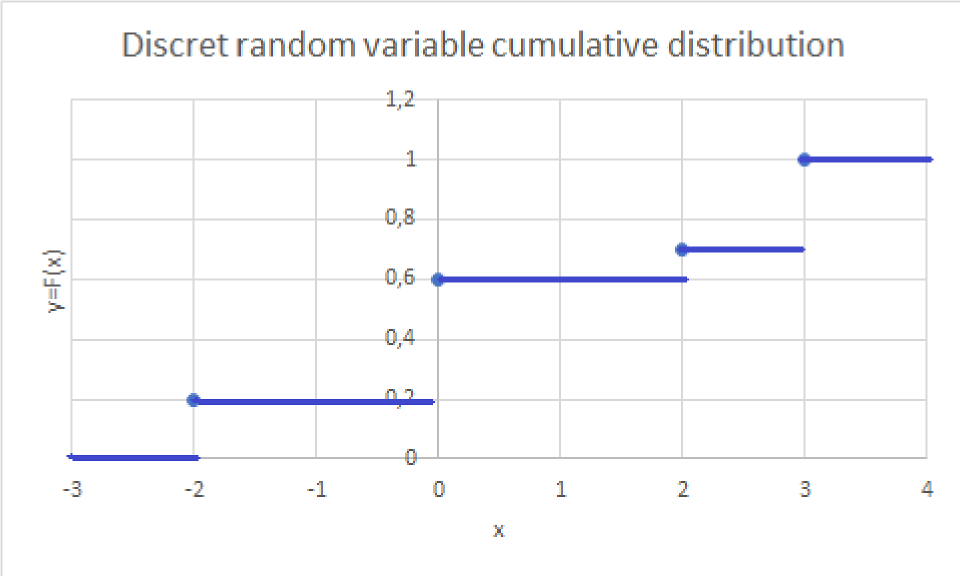

Figure 9.2. *Illustration to Example 9.23*

Remark

A random variable $Y = X^2$ in the above example has the following distribution:

$Y = y_i$	0	4	9
$P(Y = y_i)$	0.4	0.3	0.3

$$P(Y = 0) = P(X^2 = 0) = P(X = 0) = 0.4$$

$$P(Y = 4) = P(X^2 = 4) = P(X = 2 \text{ or } X = -2) = P(X = -2) + P(X = 2) = 0.2 + 0.1 = 0.3.$$

Note that events $\{X = -2\}$ and $\{X = 2\}$ are disjoint, hence the probability of its sum is equal to the sum of its probabilities.

$$P(Y = 9) = P(X^2 = 9) = P(X = 3) = 0.3$$

Note that $P(X = -3) = 0$.

By using the distribution of $Y = X^2$ we can compute $E(X^2)$ the other way as follows,

$$E(X^2) = E(Y) = 0 \cdot 0.4 + 4 \cdot 0.3 + 9 \cdot 0.3 = 3.9.$$

Co-funded by the
Erasmus+ Programme
of the European Union

Example 9.24

Suppose that we are rolling two dice. We define a random variable X as the sum of the numbers obtained. Find the probability distribution of X. Compute expected value, variance and standard deviation of a random variable X.

Solution:

The sample space of the experiment is

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

The number of elements of Ω is $\mathbf{36}$.

Using the natural definition of probability, we have:

$P(X=2) = P(\{(1,1)\}) = \frac{1}{36}$

$P(X=3) = P(\{(1,2),(2,1)\}) = \frac{2}{36} = \frac{1}{18}$

$P(X=4) = P(\{(1,3),(2,2),(3,1)\}) = \frac{3}{36} = \frac{1}{12}$

$P(X=5) = P(\{(1,4),(2,3),(3,2),(4,1)\}) = \frac{4}{36} = \frac{1}{9}$

$P(X=6) = P(\{(1,5),(2,4),(3,3),(4,2),(5,1)\}) = \frac{5}{36}$

$P(X=7) = P(\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}) = \frac{6}{36} = \frac{1}{6}$

$P(X=8) = P(\{(2,6),(3,5),(4,4),(5,3),(6,2)\}) = \frac{5}{36}$

$P(X=9) = P(\{(3,6),(4,5),(5,4),(6,3)\}) = \frac{4}{36} = \frac{1}{9}$

$P(X=10) = P(\{(4,6),(5,5),(6,4)\}) = \frac{3}{36} = \frac{1}{12}$

$P(X=11) = P(\{(5,6),(6,5)\}) = \frac{2}{36} = \frac{1}{18}$

$P(X=12) = P(\{(6,6)\}) = \frac{1}{36}$.

The probability distribution of X written in a table is:

$X = x_i$	2	3	4	5	6	7	8	9	10	11	12
$P(X = x_i)$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

The expected value $E(X)$ of X is computed as:

$$E(X) = 2 \cdot \frac{1}{36} + 3 \cdot \frac{1}{18} + 4 \cdot \frac{1}{12} + 5 \cdot \frac{1}{9} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{1}{6} + 8 \cdot \frac{5}{36} + 9 \cdot \frac{1}{9} + 10 \cdot \frac{1}{12} + 11 \cdot \frac{1}{18} + 12 \cdot \frac{1}{36} = 7$$

The variance of X is computed in two steps. First, we compute the expected value of X^2.

$$E(X^2) = 2^2 \cdot \frac{1}{36} + 3^3 \cdot \frac{1}{18} + 4^2 \cdot \frac{1}{12} + 5^2 \cdot \frac{1}{9} + 6^2 \cdot \frac{5}{36} + 7^2 \cdot \frac{1}{6} + 8^2 \cdot \frac{5}{36} + 9^2 \cdot \frac{1}{9} + 10^2 \cdot \frac{1}{12} + 11^2 \cdot \frac{1}{18} + 12^2 \cdot \frac{1}{36} = \frac{329}{6} \approx 54.83$$

Finally, we compute the variance,

$$V(X) = E(X^2) - E(X)^2 \approx 54.83 - 7^2 = 5.83.$$

The standard deviation is

$$\sigma(X) = \sqrt{V(X)} \approx 2.42.$$

Co-funded by the
Erasmus+ Programme
of the European Union

9.6 BINOMIAL DISTRIBUTION AND POISSON DISTRIBUTION

Definition: Bernoulli experiment with n trials

The rules for a *Bernoulli experiment*:

1. The experiment is repeated a fixed number of times (n times).
2. Each trial has only two possible outcomes, ``success" and ``failure". The possible outcomes are the same for each trial.
3. The probability of success remains the same for each trial. We use p for the probability of success (on each trial) and $q = 1 - p$ for the probability of failure.
4. The trials are independent (the outcome of previous trials has no influence on the outcome of the next trial).

We are interested in the random variable X where X stands for the number of successes. Note the possible values of X are $0,1,2,\dots,n$.

Our next goal is to calculate the probability distribution for the random variable X, where X counts the number of successes in a Bernoulli experiment with n trials.

Definition: binomial distribution

If X is the number of successes in a Bernoulli experiment with n independent trials, where the probability of success is p in each trial and the probability of failure is then $q = 1 - p$, then

$$P(X = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0,1,2,\dots,n.$$

Example 9.25

A basketball player takes 4 independent free throws. Every time, the probability of scoring a point is 0.6, and each point scored is recorded. Here each free throw is a trial and trials are assumed to be independent. Each trial has two outcomes: point (success) or no point (failure). The probability of success is $p = 0{,}6$ and the probability of failure is $q = 1 - p = 0{,}4$. We are interested in the variable X which counts the number of successes in 4 trials. This is an example of a Bernoulli experiment with $n = 4$ trials.

Solution:

$n = 4, \qquad p = 0.6, \qquad q = 0.4$

$$P(X = 0) = \binom{4}{0} \cdot (0.6)^0 \cdot (0.4)^4 = (0.4)^4 = 0.0256$$

$$P(X = 1) = \binom{4}{1} \cdot (0.6)^1 \cdot (0.4)^3 = 4 \cdot 0.6 \cdot (0.4)^3 = 0.1536$$

Co-funded by the
Erasmus+ Programme
of the European Union

$$P(X=2)=\binom{4}{2}\cdot(0.6)^2\cdot(0.4)^2=6\cdot(0.6)^2\cdot(0.4)^2=0.3456$$

$$P(X=3)=\binom{4}{3}\cdot(0.6)^3\cdot(0.4)^1=4\cdot(0.6)^3\cdot0.4=0.3456$$

$$P(X=4)=\binom{4}{4}\cdot(0.6)^4\cdot(0.4)^0=(0.6)^4=0.1296$$

Finally

x_i	0	1	2	3	4
p_i	0.0256	0.1536	0.3456	0.3456	0.1296

Definition: Poisson distribution

The *Poisson distribution* with parameter $\lambda > 0$ is given by

$$P(X=k)=e^{-\lambda}\frac{\lambda^k}{k!},\qquad k=0,1,2,3,\dots$$

The main property of Poisson distribution is that if X follows a Poisson distribution with parameter $\lambda > 0$ then $E(X)=V(X)=\lambda$.

Remark

The Poisson distribution is a probability distribution for the number of events that occur randomly and independently in a fixed interval of time (or space). If the mean number of events per interval is λ then the probability of observing k events in this interval is given by the Poisson distribution.

Example 9.26

Suppose that the average number of some events in a fixed time interval is 2. Find the probability that we will observe today five events in this time interval.

Solution:

The random variable X follows a Poisson distribution with mean 2, find $P(X=5)$. Since

$\lambda=2$, we have

$$P(X=5)=e^{-2}\frac{2^5}{5!}\approx 0.03609.$$

9.7 CONTINUOUS PROBABILITY DISTRIBUTIONS

Definition: continuous random variable

We call a random variable X *continuous*, if there exists a continuous (almost everywhere) nonnegative function f such that for any real number a

$$P(X \leq a) = \int_{-\infty}^{a} f(x)dx.$$

The function f is called the *probability density function* (*density function*) of a random variable X.

Remark

If a given function is continuous except for a finite number of points then it is continuous almost everywhere.

If $\boldsymbol{f}$ is a probability density function, then

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

Remark:

The *cumulative distribution function* (distribution function) of a continuous random variable $\boldsymbol{X}$ is defined by

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

for every real number $\boldsymbol{x}$.

Cumulative distribution function of continuous variable and probability

1. $P(X \geq a) = P(X > a) = 1 - F(a)$
2. $P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b) = F(b) - F(a)$
3. $P(X = a) = 0$

Example 9.27

Given is a probability density function of a random variable X,

$$f(x) = \begin{cases} 1, & x \in [0,1] \\ 0, & elswhere. \end{cases}$$

Find the cumulative distribution function F of a random variable X.

Solution:

We will use the formula for the cumulative distribution function. We must consider three cases:

1. $x < 0$: $\qquad F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} 0dt = 0$

2. $0 \leq x \leq 1$: $F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{0} 0\, dt + \int_{0}^{x} 1\, dt = 0 + x = x$

3. $x > 1$: $F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{0} 0dt + \int_{0}^{1} 1dt + \int_{1}^{x} 0dt = 0 + 1 + 0 = 1$

Finally, a distribution function is given by

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

Its graph is given below.

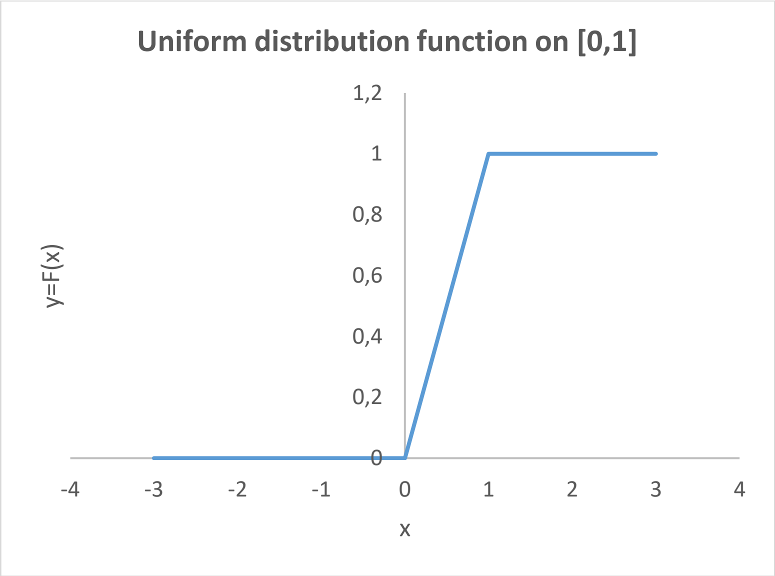

Figure 9.3. *Illustration for Example 9.27*

Definition: uniform random variable (continuous uniform distribution)

Uniform (rectangular) random variable X on the interval $[a, b]$ is a variable with the probability density

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}.$$

9.7.1 Cumulative distribution function of a uniform random variable

The cumulative distribution function of a uniform random variable $\boldsymbol{X}$ is given by

$$F(x) = \begin{cases} 0, & x < a \\ \dfrac{1}{b-a}, & a \leq x \leq b. \\ 1, & x > b \end{cases}$$

Co-funded by the
Erasmus+ Programme
of the European Union

Exercise 9.6

Find the cumulative distribution function of a uniform random variable X on the interval $[1,4]$.

Solution:

Since $a = 1,\ b = 4$, we have $b - a = 4 - 1 = 3$ and

$$F(x) = \begin{cases} 0, & x < 1 \\ \frac{1}{3}, & x \in [1,4] \\ 1, & x > 4. \end{cases}$$

Definition: Expected value

The *expected value* (*expectation or mean value*) of a continuous random variable X is defined by

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

where $\boldsymbol{f}$ is a probability density function of $\boldsymbol{X}$.

Similarly, as for discrete variables we have a linearity of $E(X)$.

If $E(X) < \infty$ and $E(y) < \infty$, then for any reals a, b

$$E(aX + bY) = aE(X) + bE(Y).$$

Expected value of a function of X

Let $Y = g(X)$, where g is a continuous function of a continuous random variable X with probability density function f. Then Y is a continuous random variable and it holds

$$E(Y) = \int_{-\infty}^{\infty} g(x) f(x) dx.$$

Similarly, as for discrete variables we define:

The *variance* of X

$$V(X) = E\big(X - E(X)\big) = E(X^2) - E(X)^2.$$

The *standard deviation* of X

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

The standard deviation of X is also denoted by σ and variance by σ^2.

Example 9.28

Compute $E(X)$, $E(X^2)$, $V(X)$, $\sigma(X)$ for uniform random variable on the interval $[0,1]$.

Solution:

The density is given by

$f(x) = 1$ for $x \in [0,1]$ and $f(x) = 0$ elsewhere.

Thus, we have

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 x dx = \frac{1}{2}$$
$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^1 x^2 dx = \frac{1}{3}$$
$$V(X) = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}$$
$$\sigma(X) = \sqrt{\frac{1}{12}} \approx 0.2887.$$

Co-funded by the Erasmus+ Programme of the European Union

9.8 NORMAL DISTRIBUTION

Definition: Normal distribution

A *normal (Gaussian) distribution* is a continuous distribution with density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < \infty.$$

It is denoted by $N(\mu, \sigma)$, where μ is mean and σ is standard deviation.

Definition: Standard normal distribution

A *standard normal (Gaussian) distribution* is a normal (Gaussian) distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$. Probability density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \qquad -\infty < x < \infty.$$

It is denoted by $N(0,1)$.

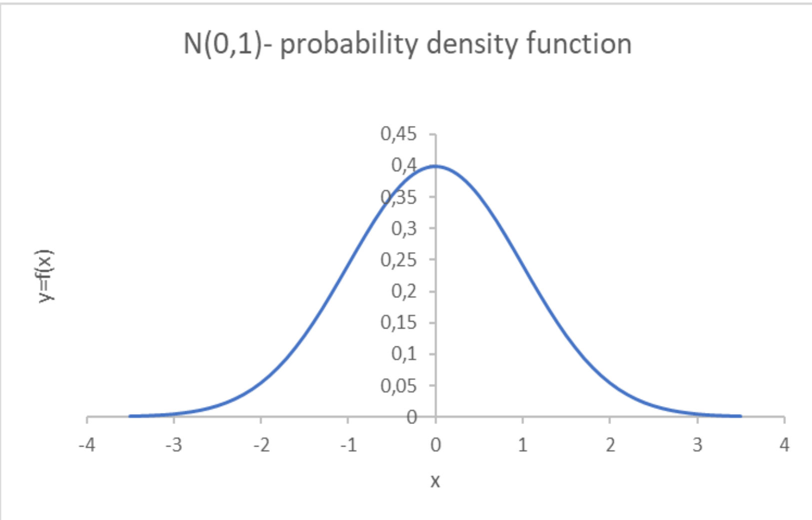

Figure 9.4: Standard normal distribution

Co-funded by the Erasmus+ Programme of the European Union

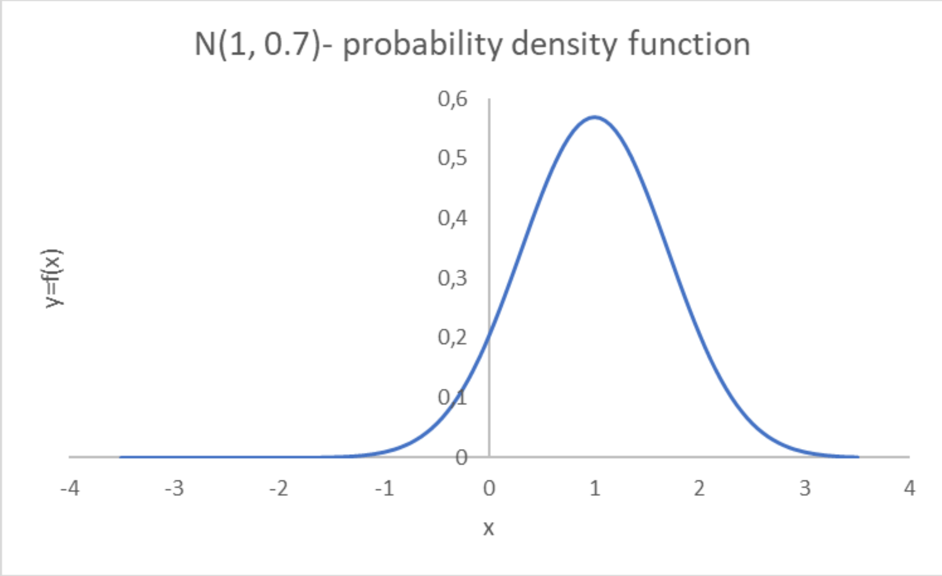

Figure 9.5: *Normal distribution* $\mu = 1, \sigma = 0.7$

Remark:

The graph of normal distribution density function is called Gaussian curve.

The *cumulative distribution function* of a normal random variable is given by

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt, \qquad -\infty < x < \infty.$$

This integral is not elementary.

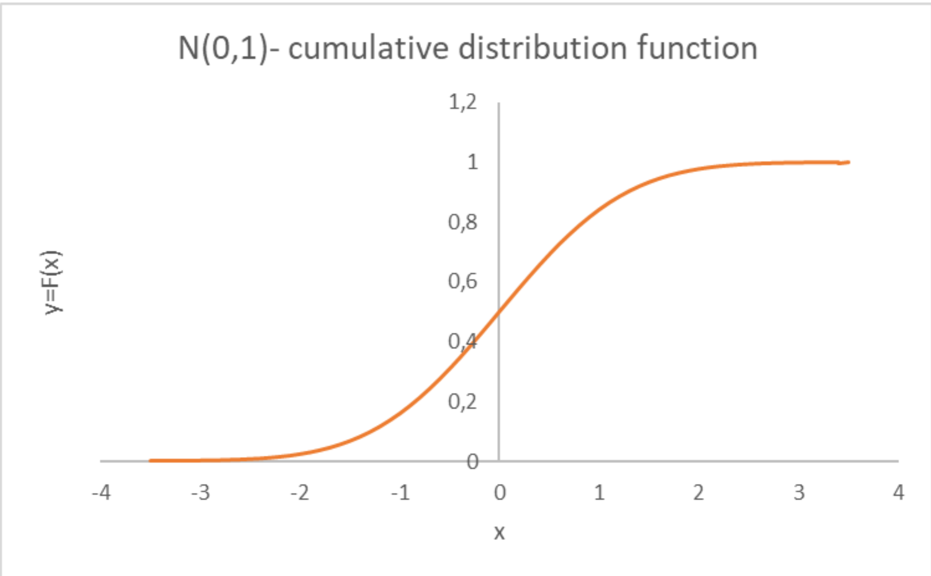

Figure 9.6. *Standard normal cumulative distribution function*

Co-funded by the Erasmus+ Programme of the European Union

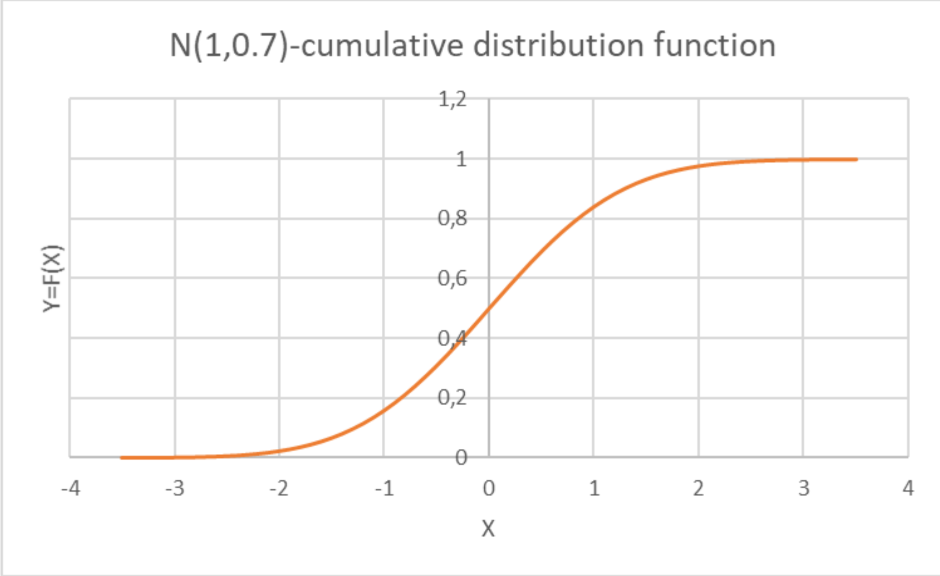

Figure 9.7. *Normal cumulative distribution function* $\mu = 1, \sigma = 0.7$

Example 9.29

A variable X has a normal distribution with a mean of 20 and a standard deviation of 3. One score is randomly sampled.

1. What is the probability that it is above 15?

2. What is the probability that it is in the interval $[18,22]$?

Solution:

1. A variable X has a distribution $N(20,3)$. We must compute $P(X > 15)$. Let F be the cumulative distribution function of X.

$$P(X > 15) = 1 - P(X \leq 15) = 1 - F(15) \approx 1 - 0.0478 = 0.9522$$

We find $F(15)$ by calculating NORM.DIST(15;20;3;1) in Excel or NORMDIST(15;20;3;1) in OpenOffice Calc.

2. $P(18 \leq X \leq 22) = F(22) - F(18) \approx 0.7475 - 0.2525 = 0.4950$

Exercise 9.7

Suppose that herring lengths is normally distributed with the mean of 8 inches and the standard deviation of 1.5 inches. A fishing vessel can catch 10 tons of herrings daily. Estimate:

1. How many of the fish are longer than 11 inches?
2. How many would you expect to be shorter than 6 inches?
3. How many have a length between 8 and 10 inches?

Co-funded by the
Erasmus+ Programme
of the European Union

Solution:

1. $\approx 0.0228 \cdot 10\ t = 0.228\ t$
2. $\approx 0.0912 \cdot 10\ t = 0.912\ t$
3. $\approx 0.4088 \cdot 10\ t = 4.088\ t$

Co-funded by the
Erasmus+ Programme
of the European Union

9.9 STATISTICS, BASIC DEFINITIONS AND NOTATIONS

Statistics is the science of collecting, analysing, presenting, and interpreting data. A population includes all the elements from a set of data.

A sample consists of one observation drawn from the population (a small part of population).

Studying a population, we are interested in collecting information about different characteristics of the subject (like their length, or weight, or age) in a sample. Those characteristics are called variables.

There are two main branches of statistics: descriptive and inferential. Descriptive statistics is used to say something about a set of information that has been gathered from a small part. Inferential statistics is used to make predictions or comparisons about a larger group (a population) using information gathered about a small part of that population.

Other distinctions are made between data types. Discrete data consist of integer numbers and usually describes a number of objects. For instance, one study might describe how many children different families own.

Measured data, in contrast to discrete data, are continuous, and thus may take on any real value. For example, the amount of time a group of children spent playing computer games would be measured data, since they could spend any number of hours playing.

For a given random sample of data:

$$x_1, x_2, \dots, x_n$$

we define a few numbers that characterise a given sample.

Definition: Sample mean

The *sample mean* is defined by

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

Definition: Sample variance

The *sample variance* is defined by

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Co-funded by the
Erasmus+ Programme
of the European Union

Remark:

The sample variance is a measure of how items in a sample are dispersed about their mean.

Definition: Sample standard deviation

The sample *standard deviation* is a square root of the variance

$$s = \sqrt{s^2}.$$

Example 9.30

Sample:

$$3,\ 3,\ 3,\ 3,\ 3$$

mean:

$$\bar{x} = \frac{1}{5}(3 + 3 + 3 + 3 + 3) = 3$$

variance:

$$s^2 = \frac{1}{5}[(3-3)^2 + (3-3)^2 + (3-3)^2 + (3-3)^2 + (3-3)^2] = 0$$

standard deviation:

$$s = \sqrt{0^2} = 0$$

There is no dispersion of data around the mean.

Example 9.31

Sample:

$$3,\ 2,\ 4,\ 3,\ 3$$

mean:

$$\bar{x} = \frac{1}{5}(3 + 2 + 4 + 3 + 3) = 3$$

variance:

$$s^2 = \frac{1}{5}[(3-3)^2 + (2-3)^2 + (4-3)^2 + (3-3)^2 + (3-3)^2] = \frac{2}{5} = 0.4$$

standard deviation:

$$s = \sqrt{\frac{2}{5}} = 0.63$$

There is some dispersion of the data around the mean.

Co-funded by the
Erasmus+ Programme
of the European Union

Example 9.32

sample:

$$5,\ 2,\ 4,\ 2,\ 2$$

mean:

$$\bar{x} = \frac{1}{5}(5+2+4+2+2) = 3$$

variance:

$$s^2 = \frac{1}{5}[(5-3)^2 + (2-3)^2 + (4-3)^2 + (2-3)^2 + (2-3)^2] = \frac{8}{5} = 1.6$$

Standard deviation

$$s = \sqrt{\frac{8}{5}} = 1.26$$

There is a significant dispersion of the data around the mean.

Remark

We will define one more characteristic of the sample.

Standard deviation and variance are positive unless all elements in the sample are the same. In this case are always equal to zero.

Definition: Estimator of variance

The *estimator of variance* is given by

$$\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

The *standard deviation* of the estimator of variance is given by

$$\hat{s} = \sqrt{\hat{s}^2}.$$

Example 9.33

Sample:

$$5,\ 2,\ 4,\ 2,\ 2$$

mean:

$$\bar{x} = \frac{1}{5}(5+2+4+2+2) = 3$$

estimator of variance:

$$\hat{s}^2 = \frac{1}{4}[(5-3)^2 + (2-3)^2 + (4-3)^2 + (2-3)^2 + (2-3)^2] = \frac{8}{4} = 2$$

standard deviation of the estimator of variance:

$\hat{s} = \sqrt{2} = 1.41$

Co-funded by the Erasmus+ Programme of the European Union

9.10 TESTING OF STATISTICAL HYPOTHESES. STATISTICAL CONCLUSION ERRORS

Definition: Null and alternative hypothesis

- The *null* hypothesis is a statement to be tested in an experiment. It is labelled H_0.
- The *alternative hypothesis* is a statement that is contradictory to H_0. It is labelled H_a.

There are two option for a decision: reject H_0 or do not reject H_0.

Definition: Type I and Type II errors

- A *Type I* error occurs when we reject H_0 but in fact H_0 is true (incorrect decision).
- A *Type II* error occurs when we do not reject H_0 but in fact H_0 is false.

The probability of Type I error is denoted by α. The probability of Type II error is denoted by β.

Remark:
Type I error and Type II are of course incorrect decisions.

The remaining two decision are correct:

- Not to reject H_0 when, in fact, H_0 is true (correct decision).
- To reject H_0 when, in fact, H_0 is false (correct decision).

We will introduce two tests in which we want to verify the hypothesis that a population mean is $\boldsymbol{\mu_0}$. In first (Model I) we assume that standard deviation of the population $\boldsymbol{\sigma}$ is known.

In the second test (Model II) we do not know the value of $\boldsymbol{\sigma}$. In both cases we assume that a population follows a normal distribution $\boldsymbol{N(\mu,\sigma)}$.

Model I of the test is the following.

Suppose that a population follows a normal distribution $\boldsymbol{N\,(\mu,\sigma)}$ and that the standard deviation of the population $\boldsymbol{\sigma}$ is known. Null hypothesis is $\boldsymbol{H_0: \mu = \mu_0}$, where $\boldsymbol{\mu_0}$ is a predicted mean value for entire population. From a random sample $\boldsymbol{x}$ we want to check hypothesis $\boldsymbol{H_0: \mu = \mu_0}$ against the alternative hypothesis $\boldsymbol{H_1: \mu \neq \mu_0}$. We assume some preconceived or pre-set a "significance level" $\boldsymbol{\alpha}$ of the test. A pre-set $\boldsymbol{\alpha}$ is the probability that we reject $\boldsymbol{H_0}$ but in fact $\boldsymbol{H_0}$ is true i.e., the probability of a Type I error. If $\boldsymbol{\alpha}$ is not given, the accepted standard is to set $\boldsymbol{\alpha = 0.05}$.

We verify the null hypothesis H_0 as follows:

First, we compute a mean $\bar{x}$ of the sample x. Next, we compute the value of random variable u with standard normal distribution $N(0,1)$ using the formula:

Co-funded by the Erasmus+ Programme of the European Union

$$u=\frac{\bar{x}-\mu_0}{\sigma}\sqrt{n}.$$

From the standard normal distribution $N(0,1)$ table or using inverse normal distribution calculator (see for instance:

http://onlinestatbook.com/2/calculators/inverse_normal_dist.html),

or using some popular spreadsheet like Excel, OpenOffice Calc, we find such critical value u_α that

$$P(|u|\geq u_\alpha)=\alpha.$$

The set of values of u defined by inequality $|u|\geq u_\alpha$ is a critical area of this test i.e if we get such value u that $|u|\geq u_\alpha$ the hypothesis H_0 must be rejected. If $|u|<u_\alpha$, we cannot reject hypothesis H_0.

The test described above is sometimes called a two-sided test since we have a two-sided critical area i.e.

$|u|\geq u_\alpha$ means the same that $u\in(-\infty,-u_\alpha]\cup[u_\alpha,\infty)$.

This occurs whenever we have the alternative hypothesis of the form $H_a\colon \mu\neq\mu_0$.

In some cases, however, we must perform one-sided tests.

When the alternative hypothesis is of the form $H_a\colon \mu<\mu_0$ we have a left-sided critical area i.e. $u\leq u_\alpha$. In this case we derive u_α such that

$$P(u\leq u_\alpha)=\alpha.$$

When the alternative hypothesis is of the form $H_a\colon \mu>\mu_0$ we have a right-sided critical area i.e. $u\geq u_\alpha$. In this case we derive u_α such that

$$P(u\geq u_\alpha)=\alpha.$$

Remark We also perform Model I if a population follows a normal distribution $N(\mu,\sigma)$ and standard deviation is unknown, but the sample is large i.e., it contains several dozen elements. In this case we set $s=\sigma$.

Example 9.34

In 81 randomly picked manufacturing companies, material cost used in the production of one particular product was measured. The mean of these values is $\bar{x}=540$ € and $s=150$ €. On a significance level $\alpha=0.05$, verify the hypothesis, that the average material cost of producing that product is 600 €.

Solution:

The null hypothesis H_0 is that the average material cost of producing the product is 600 € i.e.,

$$H_0: \mu_0 = 600.$$

Since the sample is large, we can assume that $s = \sigma = 150$. We apply Model I.

Compute

$$u = \frac{540 - 600}{150}\sqrt{81} = -3.6.$$

We have

$$P(|u| \geq u_\alpha) = 0.05$$

for $u_\alpha = 1.96$ (see table, spreadsheet or some inverse normal distribution calculator i.e., http://onlinestatbook.com/2/calculators/inverse_normal_dist.html (option: outside since the test is two-tailed).

We see that

$$|-3.6| = 3.6 \geq 1.96 = u_\alpha.$$

Hence on the significance level $\alpha = 0.05$ the hypothesis H_0 must be rejected in favour of the alternative hypothesis that the average material cost of producing considered product is not equal to 600 €.

Model II of the test is the following.

Suppose that a population follows a normal distribution $\boldsymbol{N(\mu, \sigma)}$ and that the mean μ and the standard deviation of the population σ is unknown. A small sample was selected at random from the population. Null hypothesis is $\boldsymbol{H_0: \mu = \mu_0}$. From a random sample $\boldsymbol{x}$ we want to check $\boldsymbol{H_0}$ against the alternative hypothesis $\boldsymbol{H_1: \mu \neq \mu_0}$. We assume some a "significance level" $\boldsymbol{\alpha}$ of the test.

We verify the null hypothesis H_0 as follows:

First, we compute a sample mean $\bar{x}$ and a sample standard deviation s

or estimator of the standard deviation $\hat{s}$. Next, we compute the value of a random variable t using the formula:

$$t = \frac{\bar{x} - \mu_0}{s}\sqrt{n - 1} = \frac{\bar{x} - \mu_0}{\hat{s}}\sqrt{n}.$$

Co-funded by the
Erasmus+ Programme
of the European Union

Under condition that H_0 holds random variable t follows student's $t-$ distribution with $n-1$ degrees of freedom.

From the $t-$ distribution table or using inverse normal distribution calculator (see for instance: https://www.statology.org/inverse-t-distribution-calculator) or using some popular spreadsheet like Excel, OpenOffice Calc,

we find such critical value t_α that

$$P(|t| \geq t_\alpha) = \alpha.$$

The set of values of t defined by inequality $|t| \geq t_\alpha$ is a critical area of this test i.e., if we get from the formula such value t that $|t| \geq t_\alpha$ the hypothesis H_0 must be rejected in favour of H_1. If $|t| < t_\alpha$ we cannot reject the hypothesis H_0.

The test described above is of course two-sided (see Model I) since we have a two-sided critical area i.e.

$$|t| \geq t_\alpha$$

This occurs whenever we have the alternative hypothesis of the form $H_1: \mu \neq \mu_0$.

In some cases, however, we have to perform one-sided tests.

When the alternative hypothesis is of the form H_1: $\mu < \mu_0$ we have a left-sided critical area i.e. $t \leq t_\alpha$. In this case we derive t_α such that

$$P(t \leq t_\alpha) = \alpha.$$

When the alternative hypothesis is of the form $H_1: \mu > \mu_0$ we have a right-sided critical area i.e., $t \geq t_\alpha$. In this case we derive t_α such that

$$P(t \geq t_\alpha) = \alpha.$$

Example 9.35

A machine produces metal plates of certain dimensions, with their nominal thickness being $0.04\ mm$. After randomly selecting and measuring 25 plates, their average thickness was

$\bar{x} = 0.037\ mm$ and $\hat{s} = 0.005\ mm$. Is it possible to conclude that produced plates are thinner than $0.04\ mm$? Assume the significance level $\alpha = 0.01$.

Solution:

The null hypothesis H_0 is that, the average thickness of metal plates is not less than 0.04 mm,

$$H_0: \mu_0 \geq 0.04 \text{ and } H_1: \mu_0 < 0.04.$$

Co-funded by the Erasmus+ Programme of the European Union

Compute

$$t = \frac{0.037 - 0.04}{0.005}\sqrt{25} = -3$$

We use $t -$ distribution with 24 degree of freedom and test is left-sided.

From the $t -$ distribution table or using inverse normal distribution calculator (see for instance: https://www.statology.org/inverse-t-distribution-calculator) or using some popular spreadsheet like Excel, OpenOffice Calc,

we get $t_\alpha = -2.492.$

We have $t = -3 < -2.492 = t_\alpha.$

Hence t belong to a critical area. We must reject H_0 in favour of H_1. Produced plates are thinner than $0.04\ mm.$

Co-funded by the
Erasmus+ Programme
of the European Union

9.11 REGRESSION, CORRELATION

Suppose that we have two data samples of different statistical features of some population

$$x:\ x_1, x_2, \dots, x_n \qquad y:\ y_1, y_2, \dots, y_n$$

We are interested in the following question: Is there any relation between these two features of the population?

Definition: Covariance

Covariance of x, y is defined by

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Definition: Pearson's correlation coefficient

Pearson's correlation coefficient is defined by

$$r = \frac{cov(x,y)}{s(x)s(y)}$$

where $s(x)$ is the standard deviation of x and $s(y)$ is the standard deviation of y.

Remark:

The coefficient r has a value between -1 and 1.

The value 1 means that there is a total positive linear correlation between x and y, 0 that there is no linear correlation between x and y, and -1 means that there is a total negative linear correlation between x and y.

Example 9.36

For

$$x:\ 3, 3, 4, 5, 5 \qquad y:\ 5, 7, 6, 4, 8$$

$$\bar{x} = 4, \qquad \bar{y} = 6$$

$$cov(x,y) =$$

$$\frac{1}{5}\big((3-4)(5-6) + (3-4)(7-6) + (4-4)(6-6) + (5-4)(4-6) + (5-4)(8-6)\big)$$

$$= \frac{1}{5}(1 - 1 + 0 - 2 + 2) = 0$$

$$r = \frac{0}{s(x)s(y)} = 0.$$

Co-funded by the Erasmus+ Programme of the European Union

Example 9.37

For

$$x:\ \ 3,3,4,5,5 \qquad y:\ \ 5,8,6,6,10$$

$$\bar{x} = 4, \qquad \bar{y} = 7$$

$$cov(x,y) =$$

$$\frac{1}{5}\big((3-4)(5-7) + (3-4)(8-7) + (4-4)(6-7) + (5-4)(6-7)$$
$$+ (5-4)(10-7)\big) = \frac{1}{5}(2-1+0-1+3) = \frac{3}{5}$$

$$s^2(x) =$$

$$\frac{1}{5}((3-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (5-4)^2)$$

$$= \frac{1}{5}(1+1+0+1+1) = \frac{4}{5}$$

$$s(x) = \sqrt{\frac{4}{5}} = \frac{2}{\sqrt{5}}.$$

$$s^2(y) =$$

$$\frac{1}{5}((5-7)^2 + (8-7)^2 + (6-7)^2 + (6-7)^2 + (10-7)^2)$$

$$= \frac{1}{5}(4+1+1+1+9) = \frac{16}{5}$$

$$s(y) = \sqrt{\frac{16}{5}} = \frac{4}{\sqrt{5}}.$$

$$r = \frac{\frac{3}{5}}{\frac{2}{\sqrt{5}} \cdot \frac{4}{\sqrt{5}}} = \frac{3}{8} = 0.375.$$

9.11.1 Linear regression

Suppose that we have two samples:

$$x:\ x_1, x_2, \dots, x_n, \qquad y:\ y_1, y_2, \dots, y_n.$$

By using *linear regression*, we model a relationship between two variables x and y.

One of the variables (x) is called independent (or explanatory) variable. The second variable (y) is called dependent (or response) variable.

The relation between x and y is modelled using a linear function

$$y = ax + b$$

whose unknown parameters a, b are estimated from the data.

To find a, b we use the "least squares" method. This method builds the line which minimizes the squared distance of each point from this line. We call this line a line of best fit.

We must solve the following problem

$$\sum_{i=1}^{n} (y_i - ax_i - b)^2 \to \min, \ \ a, b = ?$$

It is possible to demonstrate that such minimizing problem has always solution a, b given by

$$a = \frac{cov(x, y)}{s^2(x)}, \ \ b = \bar{y} - a\bar{x},$$

where

$$cov(x, y) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

is covariance of x, y,

$$s^2(x) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

is sample variation of x and

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

are sample means of x and y respectively.

Co-funded by the
Erasmus+ Programme
of the European Union

Example 9.38

Consider data

$$x:\ \ 3,5,2,2,1,4,6,1 \qquad y:\ \ 3,4,3,4,2,5,4,3.$$

Figure 9.8. *Regression function* $y = \frac{1}{3}x - \frac{5}{2}$.

We have

$$cov(x,y) = 1,\ \ s^2(x) = 3,\ \ \bar{x} = 3,\ \ \bar{y} = \frac{7}{2}$$

$$a = \frac{cov(x,y)}{s^2(x)} = \frac{1}{3},\qquad b = \bar{y} - a\bar{x} = \frac{7}{2} - \frac{1}{3}\cdot 3 = \frac{5}{2}.$$

A linear regression is $y = \frac{1}{3}x + \frac{5}{2}$. We can see its graph in the figure.

Exercise 9.8

Find a linear regression function for data:

$$x:\ \ 1,2,4 \qquad y:\ \ 0,2,1.$$

Solution:

$$y = \frac{3}{14}x + \frac{1}{2}.$$

Co-funded by the
Erasmus+ Programme
of the European Union

9.12 CONNECTIONS AND APPLICATIONS

9.12.1 Reliability - application of Poisson distribution.

The probability of k failures in a unit of time is given by

$$f(k) = e^{-\lambda}\frac{\lambda^k}{k!},$$

where λ is average number of failures (expected number of failures) in a unit of time.

A function

$$f(k;\lambda,t) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}$$

is a probability of k failures in time t.

Note that a probability of 0 failures in time t is given by

$$f(0;\lambda,t) = e^{-\lambda t}\frac{(\lambda t)^0}{0!} = e^{-\lambda t}.$$

Definition: Reliability function

Reliability function $R(t)$ is a probability of zero failures at time interval of length t. It is given by

$$R(t) = \begin{cases} 1, & t < 0 \\ e^{-\lambda t}, & t \geq 0 \end{cases}.$$

Probability of no more than m failures at time interval of length $t > 0$ is given by

$$R(t,m) = \sum_{k=0}^{m} e^{-\lambda t}\frac{(\lambda t)^k}{k!}.$$

Exercise 9.9

Some complex system has an average failure rate $\lambda = 0.002$ transistor failures per hour. What is the reliability for a 30 days period if the number of transistor failures cannot exceed 1?

Solution:

$\lambda = 0.002$
$t = 30 \cdot 24 = 720$
$m \leq 1$
$\lambda t = 1.44$

We have to compute $R(720,1)$.

$R(720,1) = e^{-1.44} + e^{-1.44} \cdot 1.44 \approx 0.58$

Co-funded by the
Erasmus+ Programme
of the European Union

9.13 EXERCISES

Exercise 9.10

1. How many two-digit odd natural numbers greater than 30 are there?

2. How many three-digit even natural numbers less than 500 are there?

3. In how many ways can we arrange the letters of the word $MONTANA$?

4. There are seven harbours in a country A. In how many ways can four different ships dock there?

5. In how many ways can six coins be hidden in four boxes.
6. Six points lie on the circumference of a circle. How many of inscribed triangles can be drawn having these points as vertices?

7. If the letters of word *algebra* are placed at random in a row, what is the probability that two successive letters will be a.

8. If the letters of word *about* are placed at random in a row, what is the probability that three successive letters will be vowels.

9. A couple wants to have three children. We suppose that the gender of the child is equally likely. Give the probabilities that:
   a) the couple has at least one boy,
   b) there is no girl older than a boy,
   c) the couple has exactly one girl.

10. Two ships A and B have arriving time between $\mathbf{1}$ pm and $\mathbf{5}$ pm. Both ships must dock on the same berth. Once docked, it takes each ship $\mathbf{30}$ minutes to restock and leave the dock. What is the probability
    a) that the ships won't have to wait for the berth?
    b) that the ship A won't have to wait for the berth?
    c) that the ship B will have to wait for the berth?

11. Suppose that two balls are drawn without replacement (the first ball is not replaced before the second is drawn) at random from a bag containing 4 red and 3 black balls. Let A be the event of drawing a red ball the first time, and B the event of drawing a red ball the second time. Are the events A and B independent?

12. Three identical bowls are labelled $1, 2, 3$. First bowl contains 3 red and 3 blue marbles. Second bowl contains 4 red and 2 blue marbles. Third bowl contains 1 red and 5 blue marbles. A bowl is randomly selected, and a marble is randomly selected from the

Co-funded by the Erasmus+ Programme of the European Union

bowl. a) What is the provability that a marble selected is blue? b) Given that a marble selected is red, what is the probability that bowl 2 was selected?

13. Suppose that the probability that John will solve a certain problem is $\frac{2}{3}$, that Mary will solve it is $\frac{3}{4}$ and that Bill will solve it is $\frac{1}{2}$. What is the probability
   d) that at least one person will solve it?
   e) that Mary and Bill will solve it but John will not?
   f) that John and Mary will solve it but Bill will not?
   g) that at least two people will solve it?

14. A box contains three coins: two regular coins and one fake two-headed coin $(P(H) = 1)$. You pick a coin at random and toss it.
   a) What is the probability that it lands heads up?
   b) You pick a coin at random and toss it and get heads. What is the probability that it is the two-headed coin?

15. You toss a fair coin three times:
   a) What is the probability of three heads?
   b) What is the probability that you observe exactly one heads?
   c) Given that you have observed at least one heads, what is the probability that you observe at least two heads?

16. Articles coming through an inspection line are visually inspected by two successive inspectors. When a defective article comes through the inspection line, the probability that it gets by the first inspector is 0.1. The second inspector will miss five out of ten of the defective items that get past the first inspector. What is the probability that a defective item gets by both inspectors?

17. In an exam, two reasoning problems, 1 and 2, are asked. 35% students solved problem 1 and 15% students solved both the problems. How many students who solved the first problem will also solve the second one?

18. Out of 50 people in a group, 35 smoke in which there are 20 males and 15 do not smoke in which there are 10 females. What is the probability that if the person taken at random is a male then he is a smoker?

19. A certain disease has an incidence rate of 2%. Suppose that for some diagnostic test the false negative rate is 1% and false positive rate is 1%. Compute the probability that a person, chosen at random from the population:
   a) who tests positive actually has the disease.
   b) who tests negative actually has not the disease

20. The probability distribution for a random variable X is given in table.

Co-funded by the Erasmus+ Programme of the European Union

x_i	-3	-1	0	2	3
p_i	0.1	0.4	0.2	0.2	0.1

Find the mean, variance, and standard deviation of X.

21. Five out of hundred men and two out of 1000 women are color-blind persons. From a group of the same number of men and women one person was taken at random. This person was found to be a colour-blind. What is the probability that it was a male?

\

22. Three identical bowls are labelled $1, 2, 3$. First bowl contains 3 red and 4 blue and 3 black marbles. Second bowl contains 6 red and 2 blue and 2 black marbles. Third bowl contains 2 red and 5 blue and 3 black marbles. A bowl is randomly selected, and a marble is randomly selected from the bowl.

    a) What is the provability that a marble selected is black?
    b) Given that a marble selected is black, what is the probability that bowl 3 was selected?
    c) Given that a marble selected is blue, what is the probability that bowl 1 was selected?

23. Three identical bowls are labelled $1, 2, 3$. First bowl contains 3 red and 4 blue and 3 black marbles. Second bowl contains 6 red and 2 blue and 2 black marbles. Third bowl contains 2 red and 5 blue and 3 black marbles. A bowl is randomly selected, and two marbles are randomly selected without replacement from the bowl.

    a) What is the provability that both marbles selected are blue?

    b) Given that both marbles selected are blue, what is the probability that bowl 2 was selected?

    c) Given that a marble selected are blue and black, what is the probability that bowl 2 was selected?

24. A supervisor in a factory has three men and three women working for him. He wants to choose two workers for a special job. He decides to select the two workers at random. Let Y denote the number of women in his selection. Find the probability distribution for Y.

25. Each of three balls are randomly placed into one of three bowls. Find the probability distribution for Y = the number of empty bowls.

26. A balanced coin is tossed three times. Let Y equal the number of heads observed.
    a) Calculate the probabilities associated with $Y = 0, 1, 2$, and 3.

Co-funded by the
Erasmus+ Programme
of the European Union

b) Construct a probability distribution table.

c) Find the expected value and standard deviation of Y.

27. An insurance company issues a one-year $\$2000$ policy insuring against an occurrence A that historically happens to 1 out of every 100 owners of the policy. How much should the company charge for the policy if it requires that the expected profit per policy be $\$75$?

28. A basketball player takes 4 independent freethrows with a probability of 0.7 of getting a basket on each shot. Let Y = the number of baskets he gets. Find the probability distribution for a random variable Y. Find the probability that he gets at least 3 baskets.

29. Suppose that a radio contains six transistors, two of which are defective. Three transistors are selected at random, removed from the radio, and inspected. Let X equal the number of defectives observed. Find the probability distribution for X.

30. Suppose that two balls are drawn with replacement (the first ball is replaced before the second is drawn) at random from a bag containing 5 red and 3 black balls. Let X is equal to the number of black balls drawn. Find the probability distribution for a random variable X .

31. Suppose that two balls are drawn with no replacement at random from a bag containing 5 red and 3 black balls. Let X is equal to the number of black balls drawn. Find the probability distribution for a random variable X .

32. The probability distribution for a random variable X is given in table.

x_i	0	2	4	5	6
p_i	$\frac{1}{3}$	$\frac{1}{12}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{12}$

a) Find the cumulative distribution function of X. Sketch the graph.

b) Calculate: i) $P(2 \leq X \leq 4)$ ii) $P(0 < X < 4)$ iii) $P(X > 1)$.

33. A football player takes three independent penalties with a probability $\frac{4}{5}$ of scoring a goal on each shot. Let X be the number of goals he scores. Find the probability distribution for a random variable X. Find the expected number of goals.

Co-funded by the Erasmus+ Programme of the European Union

34. A variable X has a normal distribution with a mean of 10 and a standard deviation of 2. One score is randomly sampled. What is the probability that it is between 11 and 12?

35. Sea depth was measured in 5 independent trials, and the results were (in meters): $862, 870, 876, 866, 871.$ Knowing that the distribution of measurements is normal with standard deviation of $5\ m$, and the significance level being $\alpha = 0{,}05$, verify the hypothesis that the average sea depth in that area is $870m.$

36. In a biochemical experiment, lifespan of certain organisms was measured. Distribution of that time can be assumed as normal. 8 measurements were taken, and the results were (in hours): $4.7;\ 5.3;\ 4.0;\ 3.8;\ 6.2;\ 5.5;\ 4.5;\ 6.0.$ Assuming the significance level $\alpha = 0{,}05$, assess the hypothesis that the average lifespan of these organisms is 4.0 hours.

37. Find a linear regression function for data: $x = -1, 1, 2, \quad y = 1, -2, 2.$

38. Some complex system has an average failure rate $\lambda = 0.005$ lamp failures per hour. What is the reliability for a 60 days period if the number of lamp failures cannot exceed 2?

Solutions:

1. 35
2. 200
3. 1260
4. 720
5. 4096
6. 20
7. $\frac{2}{7}$
8. $\frac{1}{10}$
9. a) $\frac{7}{8}$ b) $\frac{1}{2}$ c) $\frac{3}{8}$
10. a) ≈ 0.7656 b) ≈ 0.8828 c) ≈ 0.1172
11. not
12. a) $\frac{5}{9}$ b) $\frac{1}{2}$
13. a) $\frac{23}{24}$ b) $\frac{1}{8}$ c) $\frac{1}{4}$ d) $\frac{17}{24}$
14. a) $\frac{2}{3}$ b) $\frac{1}{2}$
15. a) $\frac{1}{8}$ b) $\frac{3}{8}$ c) $\frac{4}{7}$
16. 0.05
17. 42.8%

18. $\frac{4}{5}$
19. a) ≈ 0.6689 b) ≈ 0.9998
20. $0,\ 3,\ \sqrt{3}$
21. $\frac{25}{26}$
22. a) $\frac{4}{15}$ b) $\frac{3}{8}$ c) $\frac{4}{11}$
23. a) $\frac{17}{135}$ b) $\frac{1}{17}$ c) $\frac{4}{393}$
24. $P(Y=0)=\frac{1}{5},\ P(Y=1)=\frac{3}{5},\ P(Y=2)=\frac{1}{5}.$
25. $P(Y=0)=\frac{2}{9},\ P(Y=1)=\frac{2}{3},\ P(Y=2)=\frac{1}{9}.$
26. a) $P(Y=0)=\frac{1}{8},\ P(Y=1)=\frac{3}{8},\ P(Y=2)=\frac{3}{8},\ P(Y=3)=\frac{1}{8}$ c) $E(Y)=1.5,$ $V(Y)=0.75.$
27. \$95
28. $P(Y=0)=0.0081, P(Y=1)=0.0756, P(Y=2)=0.2646, P(Y=3)=0.4116,$ $P(Y=4)=0.2401,\ P(Y\geq 3)=0.6517.$
29. $P(X=0)=\frac{1}{5},\ P(X=1)=\frac{3}{5},\ P(X=2)=\frac{1}{5}.$
30. $P(X=0)=\frac{25}{64},\ P(X=1)=\frac{15}{32},\ P(X=2)=\frac{9}{64}.$
31. $P(X=0)=\frac{5}{14},\ P(X=1)=\frac{15}{28},\ P(X=2)=\frac{3}{28}.$
32. a) $F(x)=0\ \ x<0,\ F(x)=\frac{1}{3}\ \ 0\leq x<2,\ F(x)=\frac{5}{12}\ \ 2\leq x<4,\ F(x)=\frac{3}{4}\ \ 4\leq x<5,\ F(x)=\frac{11}{12}\ \ 5\leq x<6,\ F(x)=1\ \ x\geq 6.$ b) i) $\frac{5}{12}$ ii) $\frac{1}{12}$ iii) $\frac{2}{3}.$
33. $P(X=0)=\frac{1}{125},\ P(X=1)=\frac{12}{125},\ P(X=2)=\frac{48}{125},\ P(X=3)=\frac{64}{125}.$ $E(X)=2.4.$
34. 0.1499
35. $|u|=0.447<1.96=u_\alpha,\ H_0$ cannot be rejected
36. $|t|=3.17>2.365=t_\alpha,\ H_0$ must be rejected
37. $y\cong 0.0714x+0.2857$
38. ≈ 0.0255

Co-funded by the Erasmus+ Programme of the European Union