PANNON Egyetem
TÁMOP-4.1.2.A/1-11/1-2011-0088

Magyarország a Kelet-Európai logisztika központja – Innovatív logisztikai képzés e-learning alapú fejlesztése

ÚJ SZÉCHENYI TERV

Probability theory and mathematical statistics for IT students

Dr. Mihálykóné dr. Orbán Éva

Veszprém, 2013. január 24.

A tananyag a TÁMOP-4.1.2.A/1-11/1-2011-0088 projekt keretében a Pannon Egyetem és a Miskolci Egyetem oktatói által készült.

H-8200 Veszprém, Egyetem u. 10.
H-8201 Veszprém, Pf. 158.
Telefon: (+36 88) 624-911
Fax: (+36 88) 624-751
Internet: www.uni-pannon.hu

Nemzeti Fejlesztési Ügynökség
www.ujszechenyiterv.gov.hu
06 40 638 638
MAGYARORSZÁG MEGÚJUL
A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

Contents

Introduction

a. Basic concepts and notations 2

b. Probability 10

c. Conditional probability and independence 31

d. Random variables 45

e. Numerical characteristics of random variables 69

f. Frequently used discrete distributions 89

g. Frequently used continuous distributions 111

h. Law of large numbers 140

i. Central limit theorem 156

j. Basic concepts of mathematical statistics 177

References 205

Tables 206

Introduction

The aim of this booklet is to introduce the students to the world of random phenomena. The real world is plenty of random things. Without striving to be complete, for example, think of the waiting time in the post office, the working time of a machine, the cost of the repair of an instrument, the event of insurance, stock market and rate of exchange, damages caused by computer viruses and so on. It is obvious that these random phenomena have economic significance as well; consequently their random behaviour has to be handled. The method is served by probability theory.

The concept of probability has been developingfor centuries. It originated in gambles, for example playing cards, games with dice but the idea and the methods developed can be applied to economic phenomena, as well. Since the medieval ages people realized that random phenomena have a certain type of regularity. Roughly spoken, although one can not predict what happens during one experiment but it can be predicted what happens during many experiments. The mentioned regularities are investigated and formed by formal mathematical apparatus. The axiomatic foundation of probability was published by Kolmogorov in 1933 and since then the theory of probability, as a branch of mathematics, has been growing incredibly. Nevertheless there are problems which are very simple to understand but very difficult to solve. Solving techniques require lots of mathematical knowledge in analysis, combinatorics, differential and integral equations. On the other hand computer technique is developing very quickly, as well; hence a large amount of random experiments can be performed. The behaviour of stochastic phenomena can be investigated experimentally, as well. Moreover, difficult probabilistic problems can be solved easily by simulation after performing a great amount of computations.

This booklet introduces the main definitions connected to randomness, highlights the concept of distribution, density function, expectation and dispersion. It investigates the most important discrete and continuous distributions and shows the connections among them. It leads the students from the properties of probability to the central limit theorem. Finally it ends with fundamentals of statistics preparing the reader for further statistical studies.

a. Basic concepts and notations

The aim of this chapter

The aim of this chapter is getting the reader acquainted with the concept of the outcome of an experiment, events, occurrence of an event, operations on events. We also introduce the σ algebra of events.

Preliminary knowledge

The applied mathematical apparatus: sets and set operations.

Content

a.1. Experiments, possible outcome, sample space, events

a.2. Operations on events

a.3. σ algebra of events

a.1. Experiments, possible outcome, sample space, events

The fundamental concept of probability theory is the experiment.
The experiment is the observation of a phenomenon.
This phenomenon can be an artificial one (caused by people) or natural phenomenon, as well. We do not care whether the experiment originates from man made or natural circumstances. We require that the observation could be repeated many times.
Now we list some experiments:

- Measuring the water level of a river.
- Measuring air pollution in a town.
- Measuring the falling time of a stone from a tower to the ground.
- Measuring the waiting time at an office.
- Measuring the amount of rainfall at a certain place.
- Counting the number of failures of a machine during a time period.
- Counting the number of complains connected to a certain product of a factory.
- Counting the infected files on a computer at a time dot.
- Counting the number of falling stars at night in August.
- Counting the number of heads if you flip 100 coins.
- Investigaing the result of flipping a coin.
- Investigating if there is an odd number among three rolls of dice.
- Investigating the energy consumption of a factory during a time period.
- Investigating the demand of circulation of banknotes at a bank machine.
- Investigating the working time of a part of a machine.
- Investigating the cost of the treatment of a patient in a hospital.
- Summing the daily income of a supermarket.
- Summing the amount of claims at an insurance company during a year.
- Listing the winning numbers of the lottery.

If one "measures", "counts", "investigates", "sums" and so on, one observes a phenomenon. In some cases the result of the observation is unique. These experiments are called deterministic experiment. In other cases the observation may end in more than one result. These experiments are called stochastic or random experiments. Probability theory deals with stochastic experiments.

If one performs an experiment (trial), he can consider what may happen. The possible results are called **possible outcome**s, or, in other words, **elementary events**. The set of possible outcomes will be called the **sample space.**
We denote a possible outcome by ω, and the sample space by Ω.
What is considered as a "possible outcome" of an experiment? It is optional. First, it depends on what we are interested in. If we flip a coin, we are interested if the result is head (H) or tail (T) but usually we are not interested in the number of turnings. We can also decide whether the result of a measurement should be an integer or a real number. What should be the unit of measurement? If you investigate the water level of a river, usually the most important thing is the danger of flood. Consequently low-medium-high might be enough as possible outcomes.
But possible outcomes are influenced by the things that are worth investigating to have such cases which are simple to handle. If we are interested in the number of heads during 100 flips, we have to decide whether we consider the order of heads and tails or it is unnecessary. Therefore, during a probabilistic problem the first task is to formulate possible outcomes and determine their set.
In the examples of the previous list, if we measure something, a possible result may be a nonnegative real number, therefore $\Omega = R_0^+$. If we count something, possible outcomes are nonnegative integers, therefore $\Omega = N$. If we investigate the result of a flip, the possible

outcomes are head and tail, so $\Omega = \{H, T\}$.This set does not contain numbers. The sample space may be an abstract set. If we list the winning numbers of the lottery (5 numbers are drawn out of 90), a possible outcome is $\omega_1 = \{1,2,3,4,5\}$, and another one is $\omega_2 = \{10,20,50,80,90\}$. Possible outcomes are sets themselves. Consequently, the sample space is a set of sets, which is an abstract set again.
If an experiment is performed, then one of its possible outcomes will be realized. If we repeat the experiment, the result of the observation is a possible outcome which might be different from the previous one. This is due to the random behaviour. After performing the trial we know its result, but before making the trial we are only able to consider the possible results.

In practice events are investigated: they either occur or not.
Events are considered as subsets of the sample space. That means, certain possible outcomes are contained in a fixed event, others are not in it. We say that the **event A occurs** during an experiment if the outcome in which the trial results is the element of the set A. If the outcome observed during the actual experiment is not in A, we say that A does not occur during the actual experiment. If the observed outcomes differ during the experiments, the event A may occur in one experiment and may not in another one.
This meaning coincides with the common meaning of occurrence. Let us consider some very simple examples.

E1. Roll a single six-sided dice. The possible outcomes are: 1 dot is on the upper face, 2 dots are on the upper face, ..., 6 dots are on the upper face. Briefly, $\Omega = \{1,2,3,4,5,6\}$. i=1,2,3,4,5,6 indicates the possible outcomes by the number of dots. Let $A \subset \Omega$. The elements of A are the odd dots on the face. If the result of the roll is $\omega_1 = 1$, then $\omega_1 \in A$. We say that A occurs during this experiment. On the other side, in common parlance we usually say that the result of the roll is an odd number. If the result of the experiment is $\omega_6 = 6$, then $\omega_6 \notin A$, A does not occur during this experiment. The result of roll is not odd. Although A is a set, A expresses the "sentence" that the result of the trial is odd. If the trial ends in showing up $\omega_6 = 6$, we say shortly that the result of the roll is "six".

E2. Measure the level of a river. $\Omega = R_0^+$. Suppose that if the level of the river is more than 800 cm, then there is a danger of flood. The sentence "there is a danger of flood" can be expressed by the event (set) $A = \{x \in R_0^+ : 800 < x\} \subset \Omega$. If the result of the measurement is ω=805 cm, then $\omega \in A$. A occurs, and indeed, there is a danger of flood. If the result is the measurement is $\omega = 650$ cm, then $\omega \notin A$. We say A does not occur, and really, there is no danger of flood in that case.

E3. Count complains connected to a certain type of product. Now $\Omega = N$. If "too much problems" means that the number of complains reaches a level, for example 100, then the sentence "too much problem" is the set $A = \{n \in N : 100 \le n\}$. If the number of complaints is $\omega = 160$, then $\omega \in A$.The event A occurs and there are too much complains. If the number of complains is $\omega = 86$, then $\omega \notin A$. A does not occur, and indeed, the result of the trial does not mean too much problems.

The event Ω is called **certain event**. It occurs for sure, as whatever the outcome of the experiment is, it is included in Ω , therefore Ω occurs.
The event $\emptyset$ (empty set) is called **impossible event**. It can not occur, since whatever the outcome of the experiment is, it is not the element of $\emptyset$.

Further examples of events:

E4. Flip a coin twice. Consider the order of the results of the two flips. Now $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$, where the outcome (H,T) represents that the first flip is head, the second one is tail.

The event "there is at least one head among the flips" is the set $A = \{(H,H),(H,T),((T,H)\}$.

The event "there is at most one head among the flips" is the set $B = \{(H,T),(T,H),(T,T)\}$.

The event "there is no head among the flips" is the set $C = \{(T,T)\}$.

The event "there is no tail among the flips" is the set $D = \{(H,H)\}$. The event "the first flip is tail" is the set $E = \{(T,H),(T,T)\}$.

The event "the flips are different" is the set $F = \{(H,T),(T,H)\}$.

The event "the flips are the same" is the set $G = \{(H,H),(T,T)\}$.

We note that the number of subsets of sample space Ω is $2^4 = 16$, consequently there are 16 events in this example including the certain and the impossible event as well.

E5. Roll a die twice. Take into consideration the order of the rolls. In that case

$$\Omega = \left\{ \begin{matrix} (1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),(3,1),(3,2),(3,3),(3,4),(3,5),(3,6), \\ (4,1),(4,2),(4,3),(4,4),(4,5),(4,6),(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),(6,1),(6,2),(6,3),(6,4),(6,5),(6,6) \end{matrix} \right\} =$$

The event "there is no 6 among the rolls" is

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

The event "the sum of the rolls is 6" is $B = \{(1,5),(2,4),(3,3),(4,2),(5,1)\}$.

The event " the maximum of the rolls is 3" is $C = \{(1,3),(2,3),(3,3),(3,2),(3,1)\}$.

The event "the minimum of the rolls is at most 5" is $D = \{(5,5),(5,6),(6,5),(6,6)\}$.

As the number of possible outcomes is $6 \cdot 6 = 36$, therefore the number of events is $2^{36} \approx 6.87 \cdot 10^{10}$.

E6. Pick one card from a deck of Hungarian "seasons" cards containing 32 playing cards. A deck of Hungarian cards contains 32 playing cards, 8 of them are reds, greens, pamkins and bells. The figures are ace, king, knight and knave, furthermore seven, eight, nine and ten. For example a playing card is ace of red, another one is ten of bells (see Fig a1).

Now, $\Omega = \left\{ \begin{matrix} \text{ace of reds, ace of greens, ace of nuts, ace of pamkins, knave of reds,....,} \\ \text{knight of reds,..., king of reds,...., ten of reds, seven of reds,...} \end{matrix} \right\}$.

The event "the picked card is a red" is

$A = \left\{ \begin{matrix} \text{ace of reds, knight of reds, knave of reds, king of reds,} \\ \text{ten of reds, nine of reds, eight of reds, seven of reds} \end{matrix} \right\}$.

The event "the picked card is ace" is

$B = \{\text{ace of reds, ace of greens, ace of nuts, ace of pumkins}\}$.

The event "the picked card is ace and red" is $C = \{\text{ace of reds}\}$.

Figure a1 Some cards from a deck of Hungarian "seasons" cards
http://www.wopc.co.uk/hungary/seasons.html

E7. Pick two cards from a deck of Hungarian "seasons" cards without replacing the chosen card. Do not take into consideration the order of the cards.
In this case the sample space is

$$\Omega = \left\{ \begin{array}{l} \{\text{ace of reds, ace of greens}\}, \{\text{ace of reds, knight of greens}\}, ..., \\ \{\text{seven of nuts, ten of pamkins}\}, ... \end{array} \right\},$$

containing all the sets of two different elements of cards.
The event "both cards are ace" is

$$A=\left\{\begin{matrix}\{\text{ace of reds, ace of greens}\},\{\text{ace of reds, ace of pamkins}\},\\ \{\text{ace of reds, ace of pamkins}\},\{\text{ace of greens, ace of nuts}\},\\ \{\text{ace of greens, ace of pamkins}\},\{\text{ace of pamkins, ace of nuts}\}\end{matrix}\right\}.$$

The event" both cards are reds" is
$B=\{\{\text{ace of reds, king of reds}\},\{\text{ace of reds, knight of reds}\},....\}$.

If we want to express the event the "first card is a red", it can not be expressed actually, because we do not consider the order of cards. If we want to express this event, we have to modify the sample space as follows:
$\Omega^{mod}=\{(\text{ace of reds, ace of greens}),(\text{ace of greens, ace of reds}),....\}$.

The outcome $(\text{ace of reds, ace of greens})$ means that the first card is the ace of reds; the second one is the ace of greens. The outcome $(\text{ace of greens, ace of reds})$ means that the first card is the ace of greens; the second one is the ace of reds. To clarify the difference, we emphasize that outcome $\{\text{ace of reds, ace of greens}\}$ means that one of the picked playing cards is the ace of reds, the other one is the ace of leaves. In the sample space Ω^{mod}, the event "first card is red" can be written easily. This is an example in which the formulation of the sample space depends on the question of the problem, not only on the trial.

E8. Choose a number from the interval $[0,1]$. In that case $\Omega=[0,1]$.

The event " the first digit of the number is 6" is $A=[0.6,0.7)$.

The event " the second digit is zero" is
$C=[0,0.01)\cup[0.1,0.11)\cup[0.2,0.21)\cup...\cup[0.9,0.91)$. The event "all the digitls of the number are the same" is $B=\left\{0,0.\dot{1},0.\dot{2},...,0.\dot{9}\right\}$.

In this example the number of all possible outcomes and the number of events are infinity.

a.2. Operations on events

As events are sets, the operations with events mean operations on sets. In this subsection we interpret the set operations by the terminology of events.

- **Union** (or sum) of events

First recall that the union of two or more sets contains all the elements of all the sets.
Let A and B be events, that is $A\subset\Omega$ and $B\subset\Omega$. Then $A\cup B\subset\Omega$ holds as well. $A\cup B$ occurs if $\omega\in A\cup B$ holds, consequently $\omega\in A$ or $\omega\in B$. If $\omega\in A$, then A occurs, if $\omega\in B$, then B occurs. Summarizing, the occurrence of $A\cup B$ means that either A or B occurs. At least one of them must occur. That means either A or B or both events occur. We emphasize that „OR" is not an exclusive choice but a concessive one. The union of events can be expressed by the word OR.

- **Intersection** (or product) of events

First recall that the intersection of two or more sets contains all the common elements of the sets.
Let A and B be events, that is $A\subset\Omega$ and $B\subset\Omega$. Now $A\cap B\subset\Omega$ holds, as well. $A\cap B$ occurs if $\omega\in A\cap B$ holds, consequently $\omega\in A$ and $\omega\in B$. If $\omega\in A$, then A occurs, if $\omega\in B$ then B occurs. Summarizing, occurrence of $A\cap B$ means that both A and B occur. The intersection of events can be expressed by the word AND.
Two events are called **mutually exclusive** if their intersection is the impossible event. That is if either of them holds the other one can not occur.

- **Difference** of two events

First recall that the difference of the sets A and B contains all of elements of A which are not contained by B.

Let A and B be events, that is $A \subset \Omega$ and $B \subset \Omega$. Then $A \setminus B \subset \Omega$ holds as well. $A \setminus B$ occurs if $\omega \in A \setminus B$ holds, consequently $\omega \in A$ and $\omega \notin B$. If $\omega \in A$ then A occurs. If $\omega \notin B$ then B does not occur. Summarizing, occurrence of $A \setminus B$ means that A occurs but B does not.

- **Complement of an event**

Note that the complement of a set A is the set of all the elements in Ω which are not in A. We denote it by $\overline{A}$.

Let A be an event, that is $A \subset \Omega$. Then $\overline{A} \subset \Omega$ holds, as well. $\omega \in \overline{A}$ holds, if $\omega \notin A$. If $\omega \notin A$, then A does not occur. Consequently, $\overline{A}$ can be expressed by the word NOT A.

<u>Remarks</u>

- Operations on events have all the properties of operations on sets: the union and intersection are commutative, associative, the union and intersection is distributive.
- Further often used equality is the following one:

$A \setminus B = A \cap \overline{B}$, and the de Morgan identities:

$\overline{A \cup B} = \overline{A} \cap \overline{B}$, and for infinitly many sets $\overline{\bigcup_{i=1}^{\infty} A_i} = \bigcap_{i=1}^{\infty} \overline{A_i}$

$\overline{A \cap B} = \overline{A} \cup \overline{B}$, and for infinitely many sets $\overline{\bigcap_{i=1}^{\infty} A_i} = \bigcup_{i=1}^{\infty} \overline{A_i}$.

Now we present some examples how to express complicated events by the help of simple ones and operations.

E1. Choose one from the students of Pannon University. Let A be the event that the student is a student of economics and let B be the event that the student lives in a student hostel. In this case the sample space is the set of all the students of the university, one of its subsets is the set of those students who are students of economics; another of its subsets is formed by the students living in a student hostel. If the chosen student belongs to the subset mentioned first, then the event A occurs. Actually, for example, the following events can be described by A, B and operations:

The chosen student is a student of economics but does not live in a student hostel: $A \cap \overline{B} = A \setminus B$.

He/she is not a student of economics and he does not live in a student hostel: $\overline{A} \cap \overline{B}$.

He/she is not a student of economics or does not live in a student hostel: $\overline{A} \cup \overline{B}$.

He/she is a student of economics or does not live in a student hostel: $A \cup \overline{B}$.

He/she is not a student of economics and he/she lives in a student hostel or he/she is a student of economics and does not live in a student hostel: $(A \setminus B) \cup (B \setminus A)$.

He/she is a student of economics and he/she lives in a student hostel or he/she is not a student of economics and he/she does not live in a student hostel: $(A \cap B) \cup (\overline{A} \cap \overline{B})$.

E2. In a machine two parts may fail: part x and part y. Let A be the event that part x fails, let B be the event that part y fails.

If both parts fail, then $A \cap B$ holds.

At least one of them fails: $A \cup B$ holds.

Part x fails but part y does not: $A \setminus B$ holds.

One of them fails: $(A \setminus B) \cup (B \setminus A)$ holds.

Neither of them fails: $\overline{A} \cap \overline{B}$ holds.

At least of them does not fail: $\overline{A} \cup \overline{B}$ holds.

We note that in this case the sample space can be defined as follows: $\Omega = \{(f,f),(f,n)(n,f),(n,n)\}$, and possible outcome (f,n) represents that part x fails and part y does not.

E3. Let us investigate the arrival time of a person to a meeting. Let us suppose that the arrival time is a point in $[-5,15]$. (-1 represents that he arrives 1 minute earlier than the scheduled time, 5 represents that he arrives 5 minutes late). Let A be the event that he is late, B the event that the difference of the scheduled time of meeting and the arrival time is less than 2 minutes (briefly "small difference"). Now $A=(0,15]$, $B=(-2,2)$.

The event that he is late but small difference is $A \cap B$.

He is not late or not small difference is: $\overline{A} \cup \overline{B}$.

Both events or neither of them hold: $(A \cap B) \cup (\overline{A} \cap \overline{B})$.

He is late but not small difference is: $A \cap \overline{B}$.

a.3. The σ algebra of events

Definition Let the set of all possible outcomes be fixed and denoted by Ω. The set $\mathcal{A}$ containing some of the subsets of Ω is called a σ algebra, if the following properties hold:

1. $\Omega \in \mathcal{A}$.
2. If $A \in \mathcal{A}$, then $\overline{A} \in \mathcal{A}$ holds, as well.
3. If $A_i \in \mathcal{A}$, $i = 1,2,3,...$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$ holds as well.

Remarks

- $\emptyset \in \mathcal{A}$ as $\emptyset = \overline{\Omega}$ and $\Omega \in \mathcal{A}$.
- Applying the properties of operations one can see that if $A_i \in \mathcal{A}$, then $\bigcap_{i=1}^{\infty} A_i \in \mathcal{A}$. For the proof, note that if $A_i \in \mathcal{A}$, then $\overline{A_i} \in \mathcal{A}$, consequently $\bigcup_{i=1}^{\infty} \overline{A_i} \in \mathcal{A}$. Therefore, $\overline{\bigcup_{i=1}^{\infty} \overline{A_i}} = \overline{\overline{\bigcap_{i=1}^{\infty} A_i}} = \bigcap_{i=1}^{\infty} A_i \in \mathcal{A}$.
- If $A \in \mathcal{A}$ and $B \in \mathcal{A}$, then $A \setminus B \in \mathcal{A}$ holds as well. For the proof, note that $A \setminus B = A \cap \overline{B}$. If $B \in \mathcal{A}$, then $\overline{B} \in \mathcal{A}$ holds as well, and $A \cap \overline{B} \in \mathcal{A}$ is also satisfied.

Strictly speaking, the elements of the σ algebra $\mathcal{A}$ are called events. The above properties express that if some sets are events, then their union, intersection, difference and complement are events, as well.

In probability theory we would like to determine the probability of events which characterize the relative frequency of their occurrence during many experiments.

b. Probability

The aim of this chapter

The aim of this chapter is getting the reader acquainted with the basic properties of probability. We present the relative frequency, introduce the axioms of probability and we derive the consequences of the axioms. Classical and geometric probability are also introduced and applied for sampling problems.

Preliminary knowledge

The applied mathematical apparatus: sets and set operations. Combinatorial counting problems. Co-ordinate geometry. Basic knowledge in any computer program language.

Content

b.1. Frequency, relative frequency

b.2. Axioms of probability

b.3. Consequences of axioms

b.4. Classical probability

b.5. Geometric probability

b.1. Frequency, relative frequency

The aim of probability theory is to characterize an event by a number which expresses its relative frequency. More precisely, let the events which occur frequently during many experiments be characterized by a “large” number. Moreover, let the events which are rare be characterized by a small number. If one performs n experiments and counts how many times the event A occurs, one gets the frequency of A denoted by $k_A(n)$. It is obvious, that $0 \le k_A \le n$. We are interested in the proportion of occurrences of A to the number of trials, so we have to divide $k_A(n)$ by n, that is to take the relative frequency, $\frac{k_A(n)}{n}$. It is easy to see that $0 \le \frac{k_A(n)}{n} \le 1$.

Moreover, $k_\Omega(n) = n$, therefore $\frac{k_\Omega(n)}{n} = 1$. If A and B are events for which $A \cap B = \emptyset$, then $k_{A\cup B}(n) = k_A(n) + k_B(n)$, consequently $\frac{k_{A\cup B}(n)}{n} = \frac{k_A(n)}{n} + \frac{k_B(n)}{n}$. The value of relative frequency depends on the actual series of experiments, hence it changes if we repeat the series of experiments again. During the centuries, people recognized that the relative frequency has a kind of stability. As if it had a limit. To present this phenomenon let us consider the following example.

Let the experiment be flipping a coin many times. Let A be the event that the result is a head during one flip.
In Table b.1, one can see the frequency and relative frequency of the event A as the function of the number of experiments (n).

Result of the trial	T	T	T	H	T	T	H	T	H	H
$k_A(n)$	0	0	0	1	1	1	2	2	3	4
n	1	2	3	4	5	6	7	8	9	10
$\frac{k_A(n)}{n}$	0	0	0	0.25	0.2	0.17	0.27	0.25	0.33	0.4

Table b.1 Frequency and relative frequency of heads as the function of the number of experiences

Draw the graph of relative frequency $\frac{k_A(n)}{n}$ as the function of n. We can see the graph in the following figures: Fig.b.1, Fig.b.2, Fig.b.3 show oscillations. On the top of all, if we performed the series of experiments once again, we presumably would get other results for relative frequencies. If we increase the number of experiments the graph changes. Although there are fluctuations at the beginning of the graph, later they disappear, the graph looks almost constant.

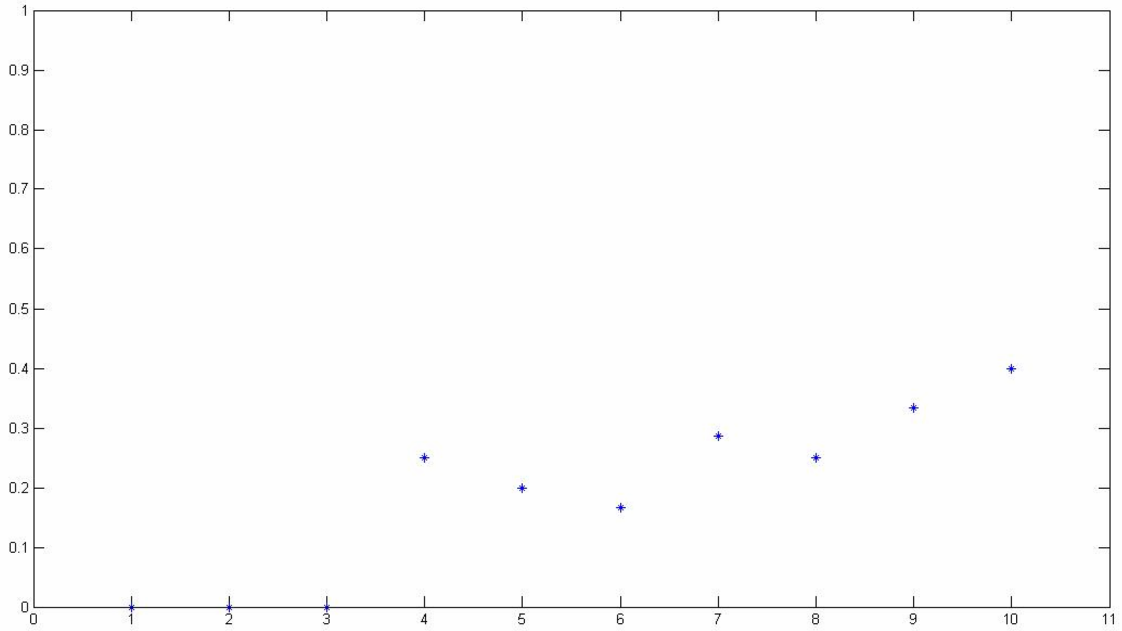


Fig.b.1 Relative frequency of heads as the function of the number of experiences (n=10)

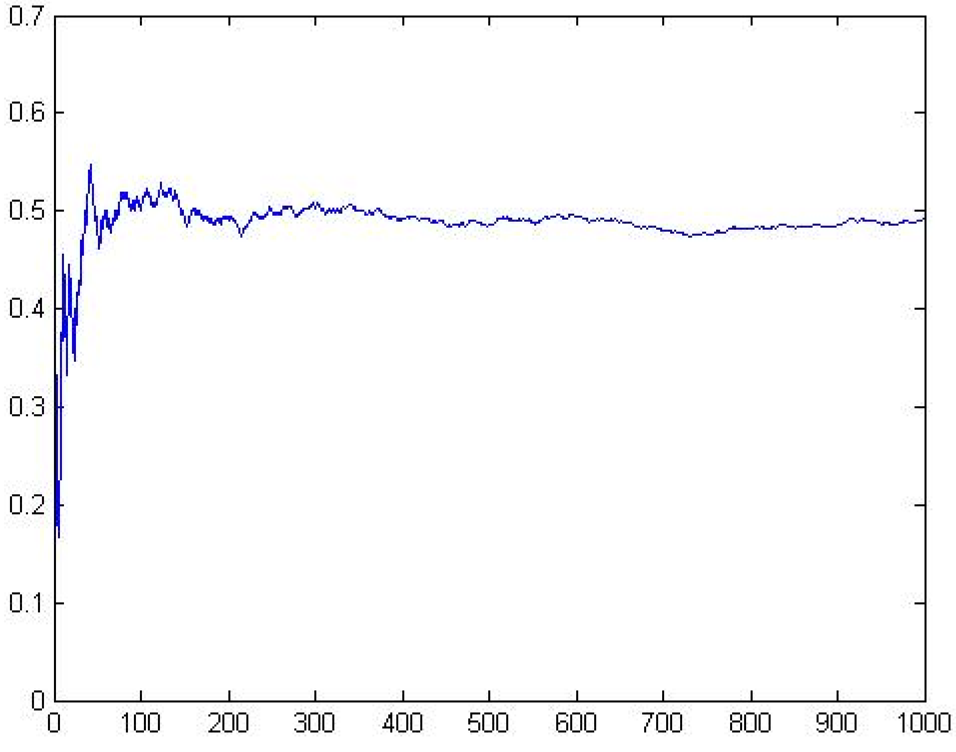


Fig.b.2 Relative frequency of heads as the function of the number of experiences (n=1000)

The mentioned phenomenon becomes more and more expressive if we increase the number of experiments, as Fig. b.3 shows, as well.

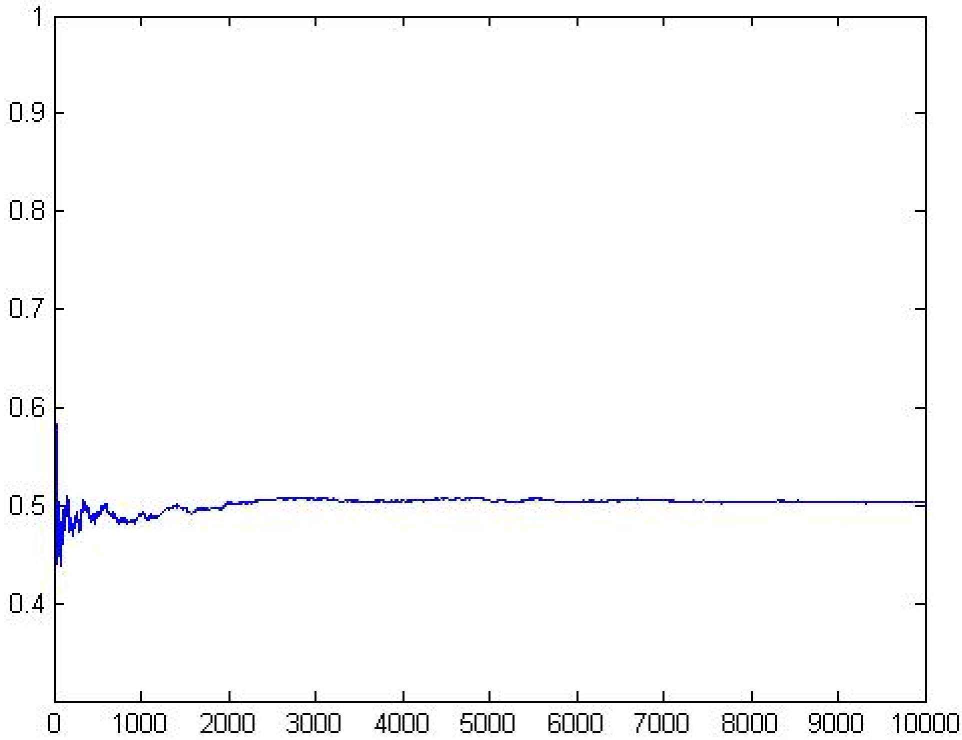


Fig.b.3 Relative frequency of heads as the function of the number of experiences (n=10000)

If we look at Fig.b.3 thoroughly, we can realize that for large values of experiments, the relative frequency is almost a constant function. Although fluctuations in the number of heads exist, they are small compared to the number of experiments. This phenomenon was mentioned during the centuries by the statement "relative frequency has a kind of stability". This phenomenon is expressed mathematically by the "law of large numbers".

b.2. Axioms of probability

If we would like to characterize the relative frequency by the probability, then the probability should have the same properties as the relative frequency. Therefore, we require the properties for probability presented previously for the relative frequency.

Definition Let $\mathcal{A}$ be a σ algebra. The function $P:\mathcal{A}\to R$ is called a **probability measure** if the following three requirements (axioms) hold:

I) $0 \le P(A)$.

II) $P(\Omega)=1$.

III) If $A_i \in \mathcal{A}$, $i=1,2,3,...$ **for which** $A_i \cap A_j = \emptyset$ $i \ne j$, **then** $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.

Remarks

- The above axioms I), II) and III) are called Kolmogorov's axioms of probability and were published in 1933.

- Probability measure maps the σ algebra of events to the set of real numbers. The elements of $\mathcal{A}$ (events) have probability. As P maps to R, $P(A)$ is a real number. The number $P(A)$ is called the **probability of the event A.**
- We define the probability by its properties. It means that every function is a probability measure that satisfies I), II) and III).
- Properties I), II) and III) correspond to the properties of relative frequency. The property $P(A) \le 1$ is not a requirement; it can be proved from the axioms. Additive property is presented for two events in the case of relative frequency, but it is required for countably infinitely many events in axiom III) in case of probability.
- Property I) expresses that the probability of any event is a nonnegative number.
- Property II) expresses that the probability of the certain event equals 1.
- Property III) expresses the additive property of the probability for countably infinitely many mutually exclusive events.
- As $\mathcal{A}$ is a σ algebra, property III) is well defined. If $A_i \in \mathcal{A}$, $i = 1,2,3,...$ hold, then $\left(\bigcup_{i=1}^{\infty} A_i\right) \in \mathcal{A}$ is also satisfied, consequently it has a probability.

If a function P satisfies axioms I), II) and III), then it satisfies many other properties, as well. These properties are called the consequences of axioms.

b3. Consequences of the axioms

We list the consequences of the axioms and we present their proofs. During this we do not use any heuristic evidences, we insist on strict mathematical inferences.

C1. $\boxed{\mathbf{P(\emptyset) = 0.}}$

$\emptyset = \emptyset \cup \emptyset \cup \emptyset \cup \ldots$ and $\emptyset \cap \emptyset = \emptyset$. That means that the impossible event can be written as the union of infinitely many pair-wise mutually exclusive events. Consequently, axiom III) can be applied and $P(\emptyset) = \sum_{i=1}^{\infty} P(\emptyset)$. Recalling that $\sum_{i=1}^{\infty} x_i = \lim_{n\to\infty} \sum_{i=1}^{n} x_i$, we can conclude that $P(\emptyset) = \sum_{i=1}^{\infty} P(\emptyset) = \lim_{n\to\infty} n \cdot P(\emptyset)$. If $0 < P(\emptyset)$ holds, then the limit is infinite, which is a contradiction, as $P(\emptyset)$ is a real number. If $P(\emptyset) = 0$, then $n \cdot P(\emptyset) = 0$ also holds for any value of n, therefore the limit is 0. In that case $0 = P(\emptyset) = \sum_{i=1}^{\infty} P(\emptyset) = 0$ holds, as well. Finally, $P(\emptyset)$ can not be negative, remember axiom I). Hence $P(\emptyset) = 0$ must be satisfied.

C2. (finite additive property) If $A_i \in \mathcal{A}$, $i = 1,2,\ldots,n$ and $A_i \cap A_j = \emptyset$, $i \ne j$, then

$$\boxed{\mathbf{P(A_1 \cup \ldots \cup A_n) = \sum_{i=1}^{n} P(A_i) = P(A_1) + \ldots + P(A_n)}}.$$

We trace this property to axiom III). Let $A_{n+1} = \emptyset$, $A_{n+2} = \emptyset, \ldots$. Now we have infinitely many events and $A_i \cap A_j = \emptyset$, $i = 1,2,\ldots$, $j = 1,2,\ldots$, $i \ne j$. If $i \le n$ and $j \le n$, this is our assumption, if $n < i$ or $n < j$ holds, then $A_i = \emptyset$ or $A_j = \emptyset$, consequently their intersection is the impossible event. Now axiom III) can be applied and $P(\bigcup_{i=1}^{n} A_i) = P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i) = \sum_{i=1}^{n} P(A_i) + P(\emptyset) + P(\emptyset) + \ldots$

As $P(\emptyset) = 0$, we get $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$ and the proof is completed.

C3. Let $A \in \mathcal{A}$ and $B \in \mathcal{A}$. **If $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$**.

This is the previous property for $n = 2$ with notation $A_1 = A$ and $A_2 = B$. We emphasize it because the additive property is frequently used in this form.

C4. Let $A \in \mathcal{A}$. **$P(\overline{A}) = 1 - P(A)$**

This connection is really very simple and it is frequently applied in the real world.

It can be proved as follows: $\Omega = A \cup \overline{A}$, and $A \cap \overline{A} = \emptyset$. Applying C3 we can see, that $P(\Omega) = P(A) + P(\overline{A})$. Taking into consideration axiom II) $P(\Omega) = 1$, we get $1 = P(A) + P(\overline{A})$. Rearranging the equality, it is easy to get C4. We mention that $\mathcal{A}$ is σ algebra, consequently if $A \in \mathcal{A}$ then $\overline{A} \in \mathcal{A}$, which means that $\overline{A}$ also has a probability.

C5. Let $A \in \mathcal{A}$ and $B \in \mathcal{A}$. **If $B \subset A$, then $P(A \setminus B) = P(A) - P(B)$**.

This formula expresses the probability of the difference of A and B with the help of the probabilities of A and B.

Note that $B \subset A$ implies the equality $A = (A \setminus B) \cup B$, moreover $(A \setminus B) \cap B = \emptyset$. Consequently C3 can be applied and results in $P(A) = P(A \setminus B) + P(B)$. Rearranging the formula we get C5.

C6. Let $A \in \mathcal{A}$ and $B \in \mathcal{A}$. **If $B \subset A$, then $P(B) \le P(A)$**.

Recall C5, and take into consideration axiom I). These formulas imply $0 \le P(A \setminus B) = P(A) - P(B).$ Non-negativity of $P(A) - P(B)$ means C6.

C7. Let $B \in \mathcal{A}$. **$P(B) \le 1$**.

This inequality is straightforward consequence of C6 with $A = \Omega$.

The formula expresses that the probability of any event is less than or equal to 1. This property coincides with the property of relative frequency $\frac{k_A(n)}{n} \le 1$.

C8. Let $A \in \mathcal{A}$ and $B \in \mathcal{A}$. **$P(A \setminus B) = P(A) - P(A \cap B)$**.

It is obvious that $A = (A \setminus B) \cup (A \cap B)$ and $(A \setminus B) \cap (A \cap B) = \emptyset$.

Using C3 they imply $P(A) = P(A \setminus B) + P(A \cap B)$. Subtracting $P(A \cap B)$ from both sides we get C8.

We emphasize that in this formula there is no extra condition on the events A and B, but C5 contains the condition $B \subset A$. Consequently C8 is a more general statement than C5. We mention that if $B \subset A$, then $A \cap B = B$, therefore in this case C5 coincides with C8.

C9. Let $A \in \mathcal{A}$ and $B \in \mathcal{A}$. **$P(A \cup B) = P(A) + P(B) - P(A \cap B)$**.

This formula expresses the probability of the union with the help of the probabilities of the events and the probability of their intersection.

To prove it, consider the identity $A \cup B = (A \setminus B) \cup B$. Now $(A \setminus B) \cap B = \emptyset$. Applying C3 we get $P(A \cup B) = P(A \setminus B) + P((B)$. Now C8 implies the identity

$P(A \cup B) = P(A) - P(A \cap B) + P(B)$ and the proof is completed.

We note that C9 does not require any assumption on the events A and B. C3 holds only for mutually exclusive events. If $A \cap B = \varnothing$, then $P(A \cap B) = 0$ and $P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B)$ coinciding with C3.
We emphasize that the probability is not an additive function. It is additive only in the case of mutually exclusive events.

C10. Let $A \in \mathcal{A}$ and $B \in \mathcal{A}$. $\boxed{\mathbf{P(A \cup B) \le P(A) + P(B)}}$.
This formula is a straightforward consequence of C9 taking into account that $0 \le P(A \cap B)$. If we do not subtract the nonnegative quantity $P(A \cap B)$ from $P(A) + P(B)$, we increase it, consequently C10 holds. We note that C10 is not an equality, it only gives an inequality for the probability of the union.

C11. Let $A \in \mathcal{A}, B \in \mathcal{A}$ and $C \in \mathcal{A}$.
Then,

$$\boxed{\mathbf{P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)}}.$$

This formula is generalization of C9 for three events.
It can be proved as follows. Let $X = A \cup B$ and $Y = C$. Now $A \cup B \cup C = X \cup Y$. Applying three times C9, first for X and Y, secondly for $A \cup B$ thirdly for $A \cap C$ and $B \cap C$ we get

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

We note that if $A \cap B = B \cap C = A \cap C = \varnothing$, then $A \cap B \cap C = \varnothing$, and $P(A \cap B) = P(A \cap C) = P(B \cap C) = P(A \cap C \cap B \cap C) = 0$. Hence in this case C11 is simplified to $P(A \cup B \cup C) = P(A) + P(B) + P(C)$ coinciding with C2.

C12. Let $A_i \in \mathcal{A}$, $i = 1,2,\dots,n$.

$$\boxed{P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{1 \le i < j \le n} P(A_i \cap A_j) + \sum_{1 \le i < j < k \le n} P(A_i \cap A_j \cap A_k) - \ldots + (-1)^{n+1} P(A_1 \cap \ldots \cap A_n)}$$

The formula can be proved by mathematical induction following the steps of the proof of C11 but we omit it.
It states that the probability of the union can be determined with the help of the probabilities of the events, the probabilities of the intersections of two, three,…., and all the events.

The relevance of the consequences is the following: if we check that the axioms are satisfied then we can use the formulas C1-C12, as well. With the help of them we can express the probabilities of “composite” events if we determine the probabilities of the “simple” events.

Now we present examples how to apply C1-C12, if we know the probability of some events. Further examples will be listed in the next subsection as well.

E1. In a factory two types of products are manufactured: Type I and Type II. Choosing one product, let A be the event that it is of Type I. According to quality, the products are divided into two groups: standard and substandard. Let B be the event that the chosen product is of standard quality. If we suppose that $P(A) = 0.7$, $P(B) = 0.9$ and $P(A \cap B) = 0.65$, give the probability of the following events:
The chosen product is of Type II.: $P(\overline{A}) = 1 - P(A) = 0.3$. (apply C4)
The chosen product is of substandard quality: $P(\overline{B}) = 1 - P(B) = 0.1$.(apply C4)

The chosen product is of Type I and it is of substandard quality: $P(A\cap\overline{B})=P(A\setminus B)=P(A)-P(A\cap B)=0.7-0.65=0.05$. (apply C8)

The chosen product is of Type II and it is of standard quality: $P(B\cap\overline{A})=P(B\setminus A)=P(B)-P(A\cap B)=0.9-0.65=0.25$. (apply C8)

The chosen product is of Type I or it is of standard quality: $P(A\cup B)=P(A)+P(B)-P(A\cap B)=0.7+0.9-0.65=0.95$. (apply C10)

The chosen product is of Type II or it is of substandard quality:
$P(\overline{A}\cup\overline{B})=P(\overline{A\cap B})=1-P(A\cap B)=1-0.65=0.35$.(apply de Morgan's identity and C4)

The chosen product is of Type II and it is of substandard quality: $P(\overline{A}\cap\overline{B})=P(\overline{A\cup B})=1-P(A\cup B)=1-0.95=0.05$. (apply de Morgan identity and C4)

The chosen product is of Type I and of standard quality or it is of Type II and of substandard quality.
$P((A\cap B)\cup(\overline{A}\cap\overline{B}))=P(A\cap B)+P(\overline{A}\cap\overline{B})-P((A\cap B)\cap(\overline{A}\cap\overline{B}))=0.65+0.05-0=0.7$.
(apply C10, and C1 as $(A\cap B)\cap(\overline{A}\cap\overline{B})=\emptyset$).

The chosen product is of Type I and of substandard quality or it is of Type II and of standard quality.
$P((A\cap\overline{B})\cup(\overline{A}\cap B))=P(A\cap\overline{B})+P(\overline{A}\cap B)-P((A\cap\overline{B})\cap(\overline{A}\cap B))=P(A\setminus B)+P(B\setminus A)=$
$P(A)-P(A\cap B)+P(B)-P(A\cap B)=0.7-0.65+0.9-0.65=0.3$. (apply C10, C8 and C1 taking into account that $(A\cap\overline{B})\cap(\overline{A}\cap B)=\emptyset$.)

E2. Choose a person from the population of a town. Let A be the event that the chosen person is unemployed, let B be the event that the chosen person can speak English fluently. If $P(A)=0.09$, $P(B)=0.25$ and $P(A\cap B)=0.02$, then determine the probability of the following events:

The chosen person is not unemployed: $P(\overline{A})=0.91$. (apply C4)

The chosen person can not speak English fluently and he is unemployed:
$P(\overline{B}\cap A)=P(A\setminus B)=P(A)-P(A\cap B)=0.09-0.02=0.07$. (apply C8)

The chosen person can speak English fluently and he is not unemployed:
$P(B\cap\overline{A})=P(B)-P(B\cap A)=0.25-0.02=0.23$. (apply C8)

The chosen person can not speak English fluently or he is unemployed: $P(\overline{B}\cup A)=P(\overline{B})+P(A)-P(\overline{B}\cap A)=1-0.25+0.09-0.07=0.77$ (apply C10 and C8)

The chosen person can speak English fluently or he is not unemployed:
$P(\overline{A}\cup B)=P(\overline{A})+P(B)-P(\overline{A}\cap B)=1-0.09+0.25-0.23=0.93$ (apply C10 and C8)

The chosen person is not unemployed or can not speak English fluently :
$P(\overline{A}\cup\overline{B})=P(\overline{A\cap B})=1-P(A\cap B)=1-0.02=0.98$ (apply de Morgan identity and C4)

The chosen person is not unemployed and can not speak English fluently:
$P(\overline{A}\cap\overline{B})=P(\overline{A\cup B})=1-P(A\cup B)=1-(P(A)+P(B)-P(A\cap B))=$
$1-(0.09+0.25-0.02)=0.68$ (apply de Morgan identity and C4 and C10)

E3. Gamble two types of races. Let A be the event that you win on the 1st race, let B be the event that you win on the 2nd race. Suppose $P(A)=0.01$, P(B)=0.03, $P(A\cap B)=0.002$. Determine the probability of the following events:

You win on the race of at least one type:
$P(A\cup B)=P(A)+P(B)-P(A\cap B)=0.01+0.03-0.002=0.038$ (apply C10)

You win on neither of them:
$P(\overline{A}\cap\overline{B})=P(\overline{A\cup B})=1-(P(A)+P(B)-P(A\cap B))=1-0.038=0.962$ (apply C4 and C10)

You do not win on at least one of them:

$P(\overline{A}\cup\overline{B})=P(\overline{A\cap B})=1-P(A\cap B)=1-0.002=0.998$ (apply de Morgan identity and C4)

You win on the 1st race but do not win on the 2nd race:

$P(A\cap\overline{B})=P(A)-P(A\cap B)=0.01-0.002=0.008$. (apply C8)

You win on 1st race or do not win on the 2nd race:

$P(A\cup\overline{B})=P(A)+P(\overline{B})-P(A\cap\overline{B})=0.01+(1-0.03)-(0.01-0.002)=0.978$.(Apply C10, C4 and C8)

You win on both of them or you win on neither of them:

$P((A\cap B)\cup(\overline{A}\cap\overline{B}))=P(A\cap B)+P(\overline{A}\cap\overline{B})-P(A\cap B\cap\overline{A}\cap\overline{B})=$

$0.002+0.962-0=0.964$. (apply C10 and de Morgan identity)

You win on one of them but not on the other one:

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

$P(A)-P(A\cap B)+P(B)-P(A\cap B)=0.01-0.002+0.03-0.002=0.036$ (apply C10)

b.4. Classical probability

In this subsection we present the often used classical probability. We prove that it satisfies axioms I), II) and III).

Definition Let Ω be a finite, nonempty set, $|\Omega|=n$. Let $\mathcal{A}=2^{\Omega}$, the set of all the subsets of Ω. The **classical probability** is defined as follows: $P(A):=\frac{|A|}{|\Omega|}$.

Theorem Classical probability satisfies axioms I), II) and III).

Proof First we note that $\mathcal{A}$ is a σ algebra, consequently P maps the elements of a σ algebra to the set of real numbers. Since $0\le|A|$ and $|\Omega|=n$, $P(A):=\frac{|A|}{|\Omega|}\ge 0$ is satisfied, as well. $P(\Omega):=\frac{|\Omega|}{|\Omega|}=1$.

Finally, if $A_i\subset\Omega$, $i=1,2,\ldots$ with $A_i\cap A_j=\varnothing$, $i\ne j$, then $A_i=\varnothing$ except for finitely many indices i, as Ω has only finitely many different subsets. If $A_i\ne\varnothing$ $i=1,2,...,k$, and $A_i\cap A_j=\varnothing$ $i\ne j$, then $\left|\bigcup_{i=1}^{k}A_i\right|=\sum_{i=1}^{k}|A_i|$, therefore $\frac{\left|\bigcup_{i=1}^{k}A_i\right|}{|\Omega|}=\frac{\sum_{i=1}^{k}|A_i|}{|\Omega|}$. We can conclude that $P(\bigcup_{i=1}^{k}A_i)=\frac{\left|\bigcup_{i=1}^{k}A_i\right|}{|\Omega|}=\frac{\sum_{i=1}^{k}|A_i|}{|\Omega|}=\sum_{i=1}^{n}P(A_i)$. If we supplement the events A_i by empty sets, neither the union nor the sum of the elements of the sets change. This means that axiom III) holds, as well.

Remarks

- In the case of classical probability $P(\{\omega\})=\frac{|\{\omega\}|}{|\Omega|}=\frac{1}{n}$, for any $\omega\in\Omega$. This formula

expresses that all outcomes have the same probability. Conversely, if $P(\{\omega\})=x$, for any $\omega\in\Omega$, then $1=P(\Omega)=P(\bigcup_{i=1}^{n}\omega_i)=n\cdot x$, which implies $x=\frac{1}{n}$. Furthermore,

$P(A)=P(\bigcup_{\omega\in A}\omega)=\sum_{\omega\in A}P(\{\omega\})=\sum_{\omega\in A}\frac{1}{n}=\frac{|A|}{|\Omega|}$. Consequently, if all the outcomes are equally probable, we can use the classical probability.

- In many cases, the number of elements of Ω and A can be determined by combinatorial methods.

Examples

E1. Roll a fair die once. Compute the probability that the result is odd, even, prime, can be divided by 3, prime and odd, prime or odd, prime but not odd.

A fair die is one for which each face appears with equal likelihood. The assumption "fair" contains the information that each outcome has the same chance, consequently we can apply classical probability. We usually suppose that the die is fair. If we do not assume it, we will emphasize it.

Returning to our example, $\Omega=\{1,2,3,4,5,6\}$. $|\Omega|=6$. $P(\{i\})=\frac{1}{6}$, i=1,2,3,4,5,6.

A=the result is odd $=\{1,3,5\}$, $|A|=3$, $P(A)=\frac{|A|}{|\Omega|}=\frac{3}{6}=0.5$.

B=the result is even= $\{2,4,6\}$, $B=3$, $P(B)=\frac{|B|}{|\Omega|}=\frac{3}{6}=0.5$.

C=the result is prime= $\{2,3,5\}$, $|C|=3$, $P(C)=\frac{|C|}{|\Omega|}=\frac{3}{6}=0.5$.

D=the result can be divided by 3= $\{3,6\}$, $|D|=2$, $P(D)=\frac{|D|}{|\Omega|}=\frac{2}{6}=0.333$.

E=the result is prime and odd $=\{3,5\}$, $|E|=2$, $P(E)=\frac{2}{6}=0.333$.

F=the result is prime or odd $=\{1,2,3,5\}$, $|F|=4$, $P(F)=\frac{4}{6}=0.667$.

G= the result is prime but not odd $=\{2\}$, $|G|=1$, $P(G)=\frac{1}{6}=0.167$.

We draw the attention that P(F) can be computed also in the following way: $F=C\cup A$, consequently $P(F)=P(C)+PA)-P(C\cap A)=\frac{3}{6}+\frac{3}{6}-\frac{2}{6}=\frac{4}{6}$.

Similarly, $G=C\cap\overline{A}=C\setminus A$, $P(C\setminus A)=P(C)\text{ - }P(A\cap C)=\frac{3}{6}-\frac{2}{6}=\frac{1}{6}$.

We note that these latest computations are unnecessary in this very simple example but can be very useful in complicated examples.

E2. Roll a fair die twice. Compute the probability of the following events: there is no six among the rolls, there is at least one six among the rolls, there is one six among the rolls, the sum of the rolls is 5, the difference of the rolls is 4, the two rolls are different.

$\Omega=\{(i,j):1\le i\le 6, 1\le j\le 6, i,j \text{ int egers}\}$. The outcome (i,j) can be interpreted as the result of the first roll and the result of the second roll. For example (1,1) denotes the outcome, when the first roll is 1, and the second roll is also 1. (3,1) denotes the outcome that the first roll is 3, the second one is 1. (1,3) means that the first roll is 1, and the second roll is 3, which differs from (3,1). If the die is fair, then (i,j) has the same probability as any other pair, whatever the values of i and j are (integers from 1 to 6). Consequently, each outcome has equal probability. $|\Omega|=6\cdot 6$.

A=there is no "six" among the rolls = $\{(1,1),(1,2),....(1,5),(2,1),...,(2,5),.....((5,1),(5,2),...(5,5)\}$.

$|A| = 5\cdot 5 = 25$, $P(A) = \frac{25}{36}$.

B= there is at least one „six" among the rolls

$= \{(1,6),(2,6),(3,6),(4,6),(5,6),(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}$. $|B| = 11$, $P(B) = \frac{11}{36}$.

Another way for solving this exercise if we realize that $B = \overline{A}$. Therefore, $P(B) = 1 - P(A) = 1 - \frac{25}{36} = \frac{11}{36}$.

C= there is one „six" among the rolls

$\{(1,6),(2,6),(3,6),(4,6),(5,6),(6,1),(6,2),(6,3),(6,4),(6,5)\}$. $|C| = 10$, $P(C) = \frac{10}{36} = 0.278$.

D=the sum of the rolls is 5 = $\{(1,4),(2,3),(3,2),\{4,1\}\}$. $|D| = 4$, $P(D) = \frac{4}{36} = \frac{1}{9} = 0.111$.

E=the difference between the two rolls is 4= $\{(1,5),(2,6),(6,2),(5,1)\}$. $|E| = 4$, $P(E) = \frac{4}{36} =$ 0.111.

F=the results of the rolls are different = $\{(1,2),(2,1),....(6,5),(5,6)\}$. $|F| = 30$, $P(F) = \frac{30}{36} = 0.833$.

Roughly spoken, the key of the solution is that we are able to list all the elements of the events and we can count them one by one.
Of course, if the number of possible outcomes is large, this way is impracticable.

E3. Roll a fair die repeatedly five times. Compute the probability of the following events: there is no „six" among the rolls, there is at least one „six" among the rolls, there is one „six" among the rolls, all the rolls are different, all the rolls are different and there is at least one „six" among the rolls, there is at least one „six" or all the rolls are different, there is at least one „six" and there are equal rolls.

$\Omega = \{(i_1, i_2, i_3, i_4, i_5): 1 \le i_j \le 6, \text{int egers}, j = 1,2,3,4,5\}$. Now i_1 denotes the result of the first roll, i_j denotes the result of the jth roll. If the die is fair, then all the outcomes are equally likely. $|\Omega| = 6\cdot 6\cdot 6\cdot 6\cdot 6 = 6^5 = 7776$.

A=there is no "„six"" among the rolls = $\{(i_1, i_2, i_3, i_4, i_5): 1 \le i_j \le 5, \text{int egers}, j = 1,2,3,4,5\}$.

$|A| = 5^5 = 3125$. $P(A) = \frac{3125}{7776} = 0.402$.

B=there is at least one „six" among the rolls = $\overline{A}$. $P(B) = 1 - P(A) = 1 - 0.402 = 0.598$.

C=there is exactly one "six" among the rolls $= \{(1,1,1,1,6),.(1,1,1,2,6),...,(6,5,5,5,5)\}$.

$|C| = \binom{5}{1}\cdot 1\cdot 5\cdot 5\cdot 5\cdot 5 = 3125$. $P(C) = \frac{3125}{7776} = 0.402$.

D=all the rolls are different = $\{(1,2,3,4,5),(1,2,3,4,6),...,(6,5,4,3,2)\}$. $|D| = 6\cdot 5\cdot 4\cdot 3\cdot 2 = 720$, $P(D) = \frac{720}{7776} = 0.093$.

E= all the rolls are different and there is at least one "six" among the rolls = $D \cap \overline{A} = D \setminus A$.

$P(E) = P(D) - P(A \cap D)$. As we need the value of $P(D \cap A)$, we have to compute it now. The set $D \cap A$ contains all the elements of Ω in which there is no "six" and the rolls are different.

$|A \cap D| = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$, $P(A \cap D) = \frac{120}{7776} = 0.015$. Finally, $P(E) = P(D) - P(A \cap D) = 0.093 - 0.015 = 0.078$.

F= there is at least one "six" or all the rolls are different = $\overline{A} \cup D$.

Applying $P(\overline{A} \cup D) = P(\overline{A}) + P(D) - P(\overline{A} \cap D)$ we get

$P(F) = P(\overline{A} \cup D) = 1 - P(A) + P(D) - \left(P(D) - P(D \cap A)\right) = 1 - 0.402 + 0.015 = 0.613$.

G=there is at least one "six" and there are equal rolls= $\overline{A} \cap \overline{D} = \overline{A \cup D}$.

$P(G) = P(\overline{A \cup D}) = 1 - P(A \cup D) = 1 - \left(P(A) + P(D) - P(A \cap D)\right) =$
$= 1 - \left(0.402 + 0.093 - 0.015\right) = 1 - 0.480 = 0.520$.

E4. Choose two numbers without replacement from a box containing the integer numbers 1,2,3,4,5,6,7,8,9. Compute the probability that both of them are odd, both of them are even, the sum of them is at least 15, one of them is less then 4 and the other is greater then 7, the difference of the numbers is 3.

If we take into consideration the order of drawn numbers, then the possible outcomes are (i_1, i_2) $i_1 \neq i_2$, $1 \le i_1 \le 9$, $1 \le i_2 \le 9$, i_1, i_2 are integers. $\Omega = \{(i_1, i_2): i_1 \neq i_2, 1 \le i_1 \le 9, 1 \le i_2 \le 9, \text{integers}\}$. $|\Omega| = 9 \cdot 8 = 72$. If we draw each number in the box with equal probability, all possible outcomes have the same chance. Consequently, classical probability can be applied. Now contract those outcomes which differ only in the order. For example, $(1,2)$ and $(2,1)$ can be contracted to $\{1,2\}$.

Actually, $\Omega^* = \{\{i_1, i_2\}: 1 \le i_1 < i_2 \le 9, \text{integers}\}$. As two possible outcomes were contracted, consequently each possible outcome (without order) has equal chance in this model as well. Roughly spoken, one can decide whether he/she wants to consider the order or not, classical probability can be applied in both cases. $\Omega^* = \binom{9}{2} = \frac{9!}{2! \cdot 7!} = \frac{9 \cdot 8}{2} = 36$.

Consider the event: both of them are even:
If we consider the order, then
A= $\{(2,4), (2,6), (2,8), (4,2), (4,6), (4,8), (6,2), (6,4), (6,8), (8,2), (8,4), (8,6)\}$

$|A| = 4 \cdot 3 = 12$, $P(A) = \frac{12}{72} = 0.167$.

If we do not consider the order, then

$A^* = \{\{2,4\}, \{2,6\}, \{2,8\}, \{4,6\}, \{4,8\}, \{6,8\}\}$. $|A^*| = \binom{4}{2} = 6$, $P(A^*) = \frac{6}{36} = 0.167$.

Finally, we can realize that we get the same result in both cases.
Both of them are odd:

$B = \{(1,3), (1,5), (1,7), (1,9), (3,1), \ldots, (9,7)\}$, $|B| = 5 \cdot 4 = 20$, $P(B) = \frac{20}{72} = 0.278$.

$B^* = \{\{1,3\}, \{1,5\}, \ldots, \{7,9\}\}$, $|B^*| = \binom{5}{2} = 10$, $P(B^*) = \frac{10}{36} = 0.278$.

The sum of them is at least 15:

$C = \{(6,9), (7,8), (7,9), (8,7), (8,9), (9,6), (9,7), (9,8)\}$, $|C| = 8$, $P(C) = \frac{8}{72} = 0.111$.

$C^* = \{\{6,9\}, \{7,8\}, \{7,9\}, \{8,9\}\}$, $|C^*| = 4$, $P(C^*) = \frac{4}{36} = 0.111$.

One of them is less than 4 and the other one is greater than 7:
$D = \{(1,8), (8,1), (1,9), (9,1), (2,8), (8,2), ((2,9), (9,2), (3,8), (8,3), (3,9), (9,3)\}$, $|D| = 12 = 2 \cdot 3 \cdot 2$,

$P(D)=\frac{12}{72}=0.167$.

$D^*=\{\{i_1,i_2\}:1\le i_1\le 3, 8\le i_2\le 9, \text{integers}\}$. $|D^*|=3\cdot 2=6$, $P(D^*)=\frac{6}{36}=0.167$.

The difference of the numbers is 3:

$E=\{(1,4),(4,1),(2,5),(5,2),(3,6),(6,3),(4,7),(7,4),(5,8),(8,5),(9,6),(6,9)\}$, $|E|=12$,

$P(E)=\frac{12}{72}=0.167$.

$E^*=\{\{1,4\},\{2,5\},\{3,6\},\{4,7\},\{5,8\},\{6,9\}\}$, $|E^*|=6$, $P(E^*)=\frac{6}{36}=0.167$.

E5. Pick 4 cards without replacement from a pack of French cards containing 13 of each of clubs (♣), diamonds (♦), hearts (♥) and spades (♠). Compute the probability that there is at least one spade or there is at least one heart, there is no spade or there is no heart, there is at least one spade but there is no heart, there are 2 spades, 1 heart and 1 other, there are more hearts than spades.

If we do not take into consideration the order of the cards picked, then $\Omega^*=\{\{\text{ace of hearts, 7 of diamonds, king of spades, 8 of spades}\},......\}$. $|\Omega^*|=\binom{52}{4}=270725$.

Actually the appropriate possible outcomes can not be listed and it is difficult to count them. The operations on the events and the consequences of axioms help us to answer the questions.

Let X^* be the event that there is no spade, Y^* the event that there is no heart among the picked cards. Now, $|X^*|=\binom{39}{4}=82251=|Y^*|$, $P(X^*)=P(Y^*)=\frac{82251}{270725}=0.304$.

A= there is at least one spade or there is at least one heart:

$A=\overline{X^*}\cup\overline{Y^*}=\overline{X^*\cap Y^*}$, consequently $P(A)=1-P(X^*\cap Y^*)$. We need the value of $P(X^*\cap Y^*)$. $X^*\cap Y^*$ means that there is no spade and at the same time there is no heart, therefore all of the cards picked are diamonds or clubs. $|X^*\cap Y^*|=\binom{26}{4}=14950$,

$P(X^*\cap Y^*)=\frac{14950}{270725}=0.055$, $P(A)=1-0.055=0.945$.

B=there is no spade or there is no heart:

$B=X^*\cup Y^*$,

$P(B)=P(X^*\cup Y^*)=\left(P(X^*)+P(Y^*)-P(X^*\cap Y^*)\right)=0.304+0.304-0.055=0.553$.

There is at least one spade but there is no heart:

C= $\overline{X^*}\cap Y^*=Y^*\setminus X^*$, $P(C)=P(Y^*)-P(X^*\cap Y^*)=0.304-0.055=0.249$.

D= there are 2 spades, 1 hearts and 1 other card.

$|D|=\binom{13}{2}\cdot\binom{13}{1}\binom{26}{1}=26364$, $P(D)=\frac{26364}{270725}=0.097$.

E=there are more spades than hearts = there is one of spade and there is no heart or there are 2 spades and 0 or 1 hearts or there are 3 spades and 0 or 1 heart or each card is a spade. These events are mutually exclusive therefore their probabilities can be summed up.

$$P(E)=\frac{\binom{13}{1}\cdot\binom{13}{0}\binom{26}{2}}{\binom{52}{4}}+\frac{\binom{13}{2}\cdot\left(\binom{13}{0}\binom{26}{2}+\binom{13}{1}\binom{26}{1}\right)}{\binom{52}{4}}+\frac{\binom{13}{3}\cdot\left(\binom{13}{0}\binom{26}{1}+\binom{13}{1}\binom{26}{0}\right)}{\binom{52}{4}}+\frac{\binom{13}{4}}{\binom{52}{4}}$$

.

The reader is kindly asked to compute it numerically.

b.5. Geometric probability

In this subsection we deal with geometric probability. It is important for understanding the concept of continuous random variable.

Definition Let Ω be a subset of R, R^2, R^3 or R^n, $4 \le n$, and let μ be the usual measure on the line, plane, space,… Let us assume that $\mu(\Omega) \ne 0$, and $\mu(\Omega) \ne \infty$. Let $\mathcal{A}$ be those subsets of Ω that have measure. Now the **geometric probability** is defined by $P(A) := \frac{\mu(A)}{\mu(\Omega)}$.

Remarks

- Axiom I) holds as $0 \le \mu(A)$, and $0 \le \mu(\Omega)$.
- Axiom II) is the consequence of the definition $P(\Omega) := \frac{\mu(\Omega)}{\mu(\Omega)} = 1$.
- Axiom III) follows from the measure-property of μ. For measures it holds that $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$ supposing $A_i \cap A_j = \emptyset,\ i \ne j$. Therefore, under the same assumption

$$P(\bigcup_{i=1}^{\infty} A_i) = \frac{\mu(\bigcup_{i=1}^{\infty} A_i)}{\mu(\Omega)} = \frac{\sum_{i=1}^{\infty} \mu(A_i)}{\mu(\Omega)} = \sum_{i=1}^{\infty} \frac{\mu(A_i)}{\mu(\Omega)} = \sum_{i=1}^{\infty} P(A_i).$$

- The usual measure on R, R^2, R^3 is the length, area and volume, respectively. Their concept can be generalized. Further knowledge on measures can be found in the book of Halmos.
- The definition $P(A) = \frac{\mu(A)}{\mu(\Omega)}$ expresses that the probability of an event is proportional to its measure. In the case of classical probability the "measure" is the number of the elements of Ω. Now the number of elements of Ω is infinity.
- If $\mu(\Omega) = 1$, then $P(A) = \mu(A)$. The consequences of the axioms are the frequently used properties of measure. See for example C8 and C9.
- The proof of the fact that the set of those subsets of Ω that have measure is a σ algebra requires lots of mathematical knowledge, we do not deal with it actually.
- Random numbers on computers are numbers chosen from the interval [0,1] by geometric probability approximately. That is, the probability that the number is situated in a subset of $[0,1]$ is proportional to the length of the subset. As the length of the interval $[0,1]$ equals 1, the probability coincides with the length of the set itself.

Examples

E1. Choose a point from the interval $[0, \pi]$ with geometric probability. Compute the probability that the second digit of the point equals 4.

$\Omega = [0, \pi]$, lenght is abbreviated by μ. $\mu(\Omega) = \pi$.

A= the second digit is 4 = $[0.04, 0.05) \cup [0.14, 0.15) \cup \ldots \cup [3.14, \pi]$.

$\mu(A) = 31 \cdot 0.01 + \pi - 3.14 = 0.3116$, $P(A) = \frac{\mu(A)}{\pi} = 0.0992$.

E2. Shot on a circle with radius R. The probability that the hit is situated in a subset of the circle is proportional to the area of the subset. Compute the probability that we have 10, 9 scores.

Ω is the circle with radius R. $\text{area}(\Omega)=\mu(\Omega)=R^2\cdot\pi$. Let A be the event that the hit is 10 scores. 10 scores means that the hit is inside the inner circle lined black, which is a circle with radius $\frac{R}{10}$. Consequently, $\mu(A)=\left(\frac{R}{10}\right)^2\pi$, $P(A)=\frac{\left(\frac{R}{10}\right)^2\pi}{R^2\pi}=\frac{1}{100}$.

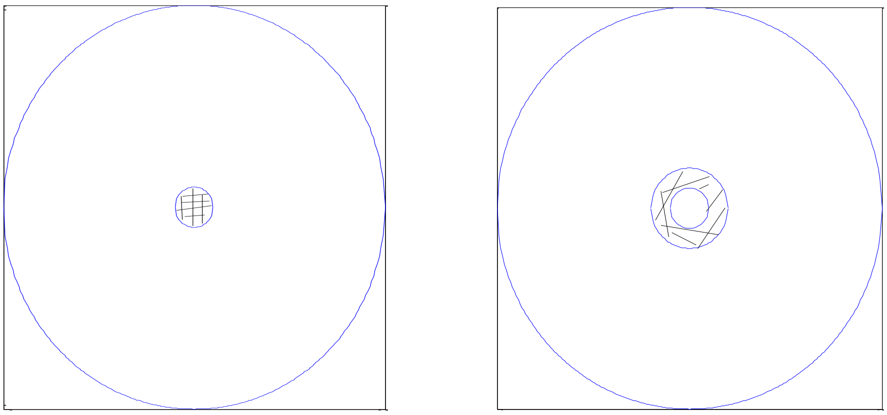


Fig.b.4 Events A and B

Let B be the event that the hit is 9 scores. It means that the hit is not in the inner part but in the following annulus. As the hits are between concentric circles, $\mu(B)=\left(\frac{2R}{10}\right)^2\pi-\left(\frac{R}{10}\right)^2\pi=\frac{3R^2}{100}\pi$. Consequently, $P(B)=\frac{3}{100}$.

Compute the probability that the distance of the hit and the centre of the circle equals $\frac{R}{2}$.

Let C be the event that the distance between the hit and centre of the circle equals $\frac{R}{2}$. The points whose distance from the centre equals $\frac{R}{2}$ are situated on the graph of the circle of radius $\frac{R}{2}$ drawn by red in Fig.b.5.The area of the curve is zero, as it can be covered by the annulus which is the difference of the open circle with radius $\frac{R}{2}+\Delta R$, and the open circle with radius $\frac{R}{2}$, for any positive value of ΔR.

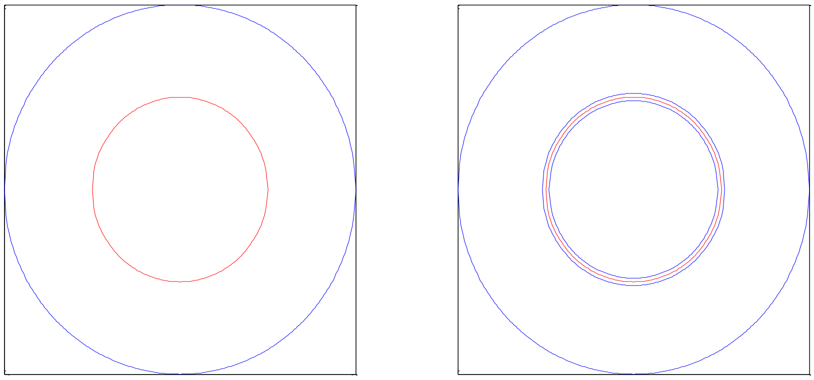


Fig.b.5 Event $C_{R/2}$ and event $\left\{\omega : \frac{R}{2} \le d(\omega,0) < \frac{R}{2} + \Delta R\right\}$

Consequently, $\mu(C) \le \left(\frac{R}{2} + \Delta R\right)^2 \cdot \pi - \left(\frac{R}{2}\right)^2 \cdot \pi = \left(R \cdot \Delta R + (\Delta R)^2\right)\pi$, which tends to zero if ΔR tends to zero. That implies that $\mu(C) = 0$. Therefore, $P(C) = \frac{\mu(C)}{R^2 \pi} = 0$.

We draw the attention to the fact that despite even though $C \ne \varnothing$, $P(C) = 0$ holds. Moreover, if we use the notation $C_x = \{Q : d(Q,O) = x\}$, then $P(C_x) = 0$, for any value of $0 \le x \le R$. Now $\Omega = \bigcup_{0 \le x \le R} C_x$ holds. Moreover, if $x \ne y$, then $C_x \cap C_y = \varnothing$. $P(\Omega) = 1$ but $P(\Omega) \ne \sum P(C_x)$.

The reason of this paradox is that the set $\{x : 0 \le x \le R\}$ is not finite and is not countable. This is a very important thing in order to understand the concept of continuous random variables.

E3. Choose two numbers independently of each other from the interval $[-1,1]$ with geometric probability. Compute the probability that the sum of the numbers is between 0.5 and 1.5.

To choose two numbers from the interval $[-1,1]$ with geometric probability independently of each other means to choose one point in the Cartesian coordinate system, namely from the square $[-1,1]x[-1,1]$ with geometric probability. If the first number equals x, the second number equals y, then let the two dimensional point be denoted by $Q(x, y)$. Roughly spoken, let the first number be put on the x axis, the second number be put on y axis. Now $\Omega = [-1,1]x[-1,1]$, $\mu(\Omega) = 4$. Let A be the event that the sum of the numbers is between 0.5 and 1.5. We seek the points $Q(x, y)$ for which $0.5 < x + y < 1.5$.

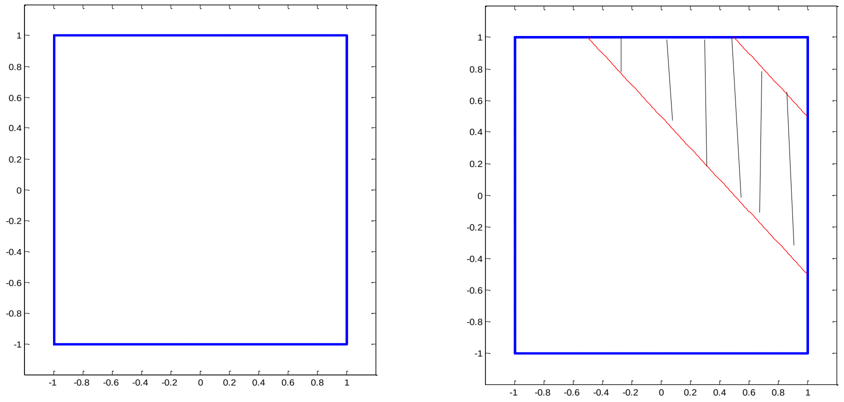


Fig.b.6. The set of all possible outcomes Ω and the set of appropriate points

These points are in the section between the red lines given by $x+y=0.5$ and $x+y=1.5$ presented in Fig.b.6.

$$\mu(A)=\frac{\left(\frac{3}{2}\right)^2}{2}-\frac{\left(\frac{1}{2}\right)^2}{2}=1,\ P(A)=\frac{1}{4}.$$

Compute the probability that the sum of the numbers equals 1.

Let B be the event that the sum of numbers equals 1. The points of B are the points of the line given by $x+y=1$ (see Fig.b.7)

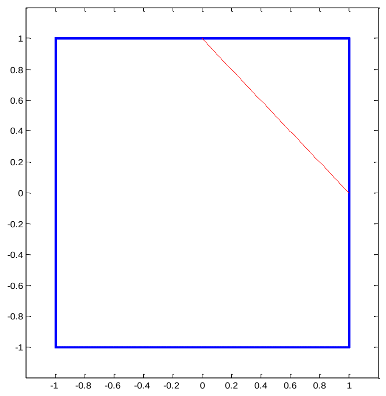


Fig.b.7. The set of points given by the equation $x+y=1$

$\mu(B)=0$, consequently, $P(B)=0$.

E4. Choose two numbers independently from each other with geometric probability from the interval $[0,1]$. Compute the probability that the square of the second number is less than the first one or the square of the first one is greater than the second one.

$\Omega=[0,1]x[0,1]$, $\mu(\Omega)=1$. We seek those points $Q(x,y)$ for which $y<x^2$ or $x<y^2$, that is $\sqrt{x}<y$. The appropriate points are bellow the curve given by $y=x^2$, furthermore above the curve given by $y=\sqrt{x}$ (see Figure b.8.)

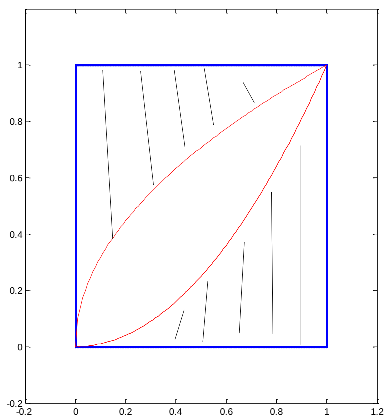


Fig.b.8. Those points for which $y < x^2$ or $x < y^2$ holds

If A is the set of appropriate points, then

$$\mu(A) = \int_0^1 x^2 dx + \int_0^1 \left(1 - \sqrt{x}\right) dx = \left[\frac{x^3}{3}\right]_0^1 + \left[x - \frac{\sqrt{x^3}}{\frac{3}{2}}\right]_0^1 = \frac{1}{3} + 1 - \frac{2}{3} = \frac{2}{3} = 0.667 \text{ and}$$

$$P(A) = \frac{0.667}{1} = 0.667 \,.$$

E5. Use the random number generator of your computer and generate N=1000, N=10000, N=100000, N=1000000 random numbers. Divide the interval [0,1] into 10 equal parts, and count the ratio of the random numbers situated in the sub-intervals $\left(\frac{i}{10}, \frac{i+1}{10}\right]$, i=0,1,2,…,9. Draw the figures!

Relative frequencies of random numbers being in the above intervals are shown in Figs.b.9. b.10. b11. and b.12. for the simulated random numbers N=1000, 10000, 100000, 1000000, respectively. The pictures show that by increasing the number of simulations, the relative frequencies become more and more similar, the random numbers are situated more and more uniformly. If the probability of being in the interval is really $\frac{1}{10}$, then relative frequencies are closer and closer to this probability.

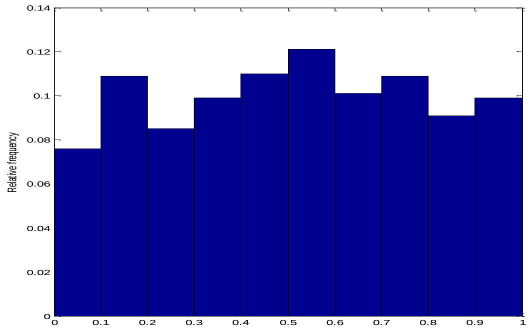


Figure b.9. Relative frequencies of random numbers in case of N=1000

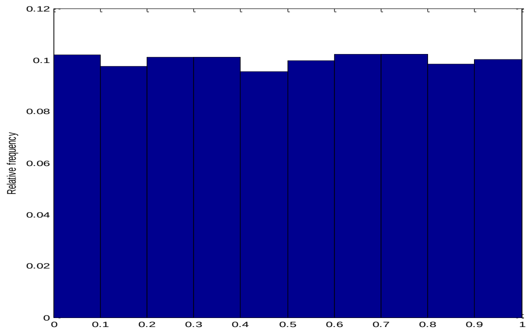


Figure b.10 Relative frequencies of random numbers in case of N=10000

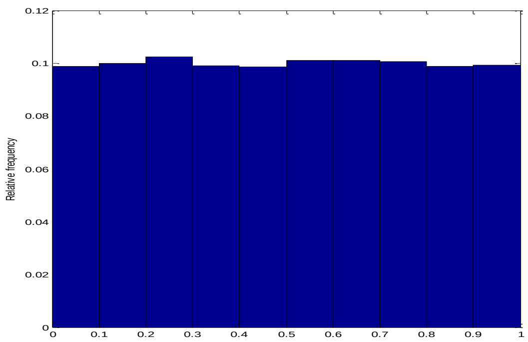


Figure b.11. Relative frequencies of random numbers in case of N=100000

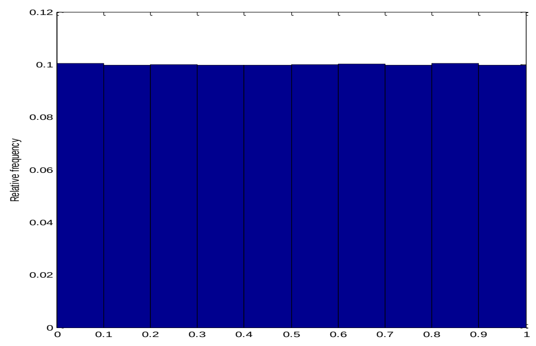


Figure b.12. Relative frequencies of random numbers in case of N=1000000

E6. Approximate the probability of event A in E3) by the relative frequency of the event A applying N=1000, 10000, 100000, 1000000 simulations. Give the difference between the approximate values and the exact probability.
First we mention that if a number is chosen from [0,1] with geometric probability, then its double is chosen from [0,2] with geometric probability and the double and minus 1 is chosen from the interval [-1,1] with geometric probability.

The relative frequencies of A and their differences from the exact probability 0.25 can be seen in Table b.2. One can realize that if the number of simulations increases, the difference decreases.

	N=1000	N=10000	N=100000	N=1000000
Relative frequency	0.2670	0.2584	0.2517	0.2502
Difference	0.0170	0.0084	0.0017	0.0002

Table b.2. Relative frequencies of the event and their differences from the exact probability

The relative frequencies of the event that the sum is in $\left(-2+\frac{i}{5},-2+\frac{i+1}{5}\right]$, $i=0,...,19$ can be seen in Figs.b.13,b.14. One can see that the shapes of the graphs are getting more and more similar to a roof.

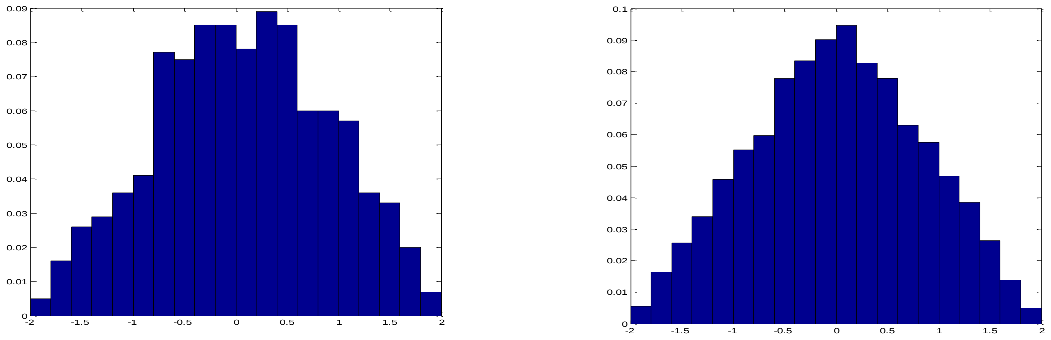


Fig.b.13. The relative frequencies of the event that the sum is in $\left(-2+\frac{i}{5},-2+\frac{i+1}{5}\right]$,

$i=0,...,19$ for N=1000 and 10000

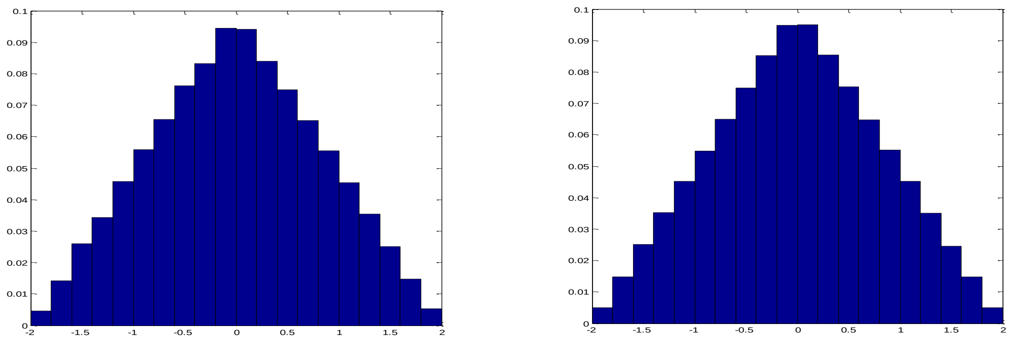


Fig.b.14. The relative frequencies of the event that the sum is in $\left(-2+\frac{i}{5},-2+\frac{i+1}{5}\right]$,

$i=0,...,19$ for N=10000 and 100000

c. Conditional probability and independence

The aim of this chapter

The aim of this chapter is to get the reader acquainted with the concept of conditional probability and its properties. We present the possibilities for computing non-conditional probabilities applying conditional ones. We also define the independence of events.

Preliminary knowledge

Properties of probability.

Content

c.1. Conditional probability.

c.2. Theorem of total probability and Bayes' theorem.

c.3. Independence of events

c.1. Conditional probability

In many practical cases we have some information. We would like to know the probability of an event and we know something. This "knowledge" has an effect on the probability of the event; it may increase or decrease the probability of its occurrence.
What is the essence of conditional probability? How can we express that we have some information?
Let Ω be the set of possible outcomes, $\mathcal{A}$ the set of events, let P be the probability. Let $A, B \in \mathcal{A}$. If we know that B occurs (this is our extra information), then the outcome which is the result of our experiment is an element of B. Our word is restricted to B. If A occurs, then the outcome is a common element of A and B, therefore it is in $A \cap B$. The probability of the intersection should be compared to the "measure" of the condition, i.e. $P(B)$. Naturally, $0 < P(B)$ has to be satisfied.
Definition The **conditional probability of the event** A **given B** is defined as $P(A|B) := \frac{P(A \cap B)}{P(B)}$, if $0 < P(B)$.

Remarks

- Notice that the definition of conditional probability implies the form $P(A \cap B) = P(A|B) \cdot P(B)$, called multiplicative formula.
- The generalization of the above form is the following statement: if $0 < P(A_1 \cap ... \cap A_{n-1} \cap A_n)$ holds, then

$P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 \cap A_2) \cdot ... \cdot P(A_n | A_1 \cap ... \cap A_{n-1})$.
It can be easily seen if we notice that $P(A_1) \cdot P(A_2 | A_1) = P(A_1 \cap A_2)$,
$P(A_3 | A_1 \cap A_2) \cdot P(A_1 \cap A_2) = P(A_1 \cap A_2 \cap A_3)$, and finally,
$P(A_n | A_1 \cap ... \cap A_{n-1}) \cdot P(A_1 \cap ... \cap A_{n-1}) = P(A_1 \cap ... \cap A_{n-1} \cap A_n)$.

- If we apply classical probability, then

$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{|A \cap B|}{|\Omega|}}{\frac{|B|}{|\Omega|}} = \frac{|A \cap B|}{|B|}$. Roughly spoken: there are some elements in B, these are our "new (restricted) world". Some of them are in A, as well. The ratio of the number of elements of A in our "new world" and the number of elements of the "new world" is the conditional probability of A.

Theorem Let the event B be fixed with $0 < P(B)$. The conditional probability given B satisfies the axioms of probability **I), II), III).**
Proof:
I) $0 \le P(A|B)$, as $0 \le P(A \cap B)$, and $0 < P(B)$.
II) $P(\Omega|B) = 1$, as $P(\Omega|B) := \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$.

III) If $A_i \in \mathcal{A}, i = 1,2,3,...$ $A_i \cap A_j = \emptyset$, $i \ne j$, then $P(\bigcup_{i=1}^{\infty} A_i | B) = \sum_{i=1}^{\infty} P(A_i | B)$.
The proof can be performed in the following way: notice that if $A_i \cap A_j = \emptyset$, then $(A_i \cap B) \cap (A_j \cap B) = \emptyset$ holds as well. Now

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

This theorem assures that we can conclude all of the consequences of axioms.We can state the following consequences corresponding to C1,…, C12 without any further proof.

- $P(\emptyset \mid B) = 0$.
- If $A_i \in \mathcal{A}, i = 1,2,..., n$ for which $A_i \cap A_j = \emptyset$, $i \neq j$, then $P(\bigcup_{i=1}^{n} A_i \mid B) = \sum_{i=1}^{n} P(A_i \mid B)$.
- If $C \subset A$, then $P(C \mid B) \leq P(A \mid B)$
- $P(A \mid B) \leq 1$.
- $P(\overline{A} \mid B) = 1 - P(A \mid B)$.
- $P(A \setminus C \mid B) = P(A \mid B - P(A \cap C \mid B)$.
- $P(A \cup C \mid B) = P(A \mid B) + P(C \mid B) - P(A \cap C \mid B)$.
- $P(A \cup C \mid B) \leq P(A \mid B) + P(C \mid B))$.
- $P(A \cup C \cup D \mid B) = P(A \mid B) + P(C \mid B) + P(D \mid B) -$

$P(A \cap C \mid B) - P(D \cap C \mid B) - P(A \cap D \mid B) + P(A \cap C \cap D \mid B)$.

- $P\left(\bigcup_{i=1}^{n} A_i \mid B\right) = \sum_{i=1}^{n} P(A_i \mid B) - \sum_{1 \leq i < j \leq n} P(A_i \cap A_j \mid B) +$

$+ \sum_{1 \leq i < j < k \leq n} P(A_i \cap A_j \cap A_k \mid B) - \ldots + (-1)^{n+1} P(A_1 \cap \ldots \cap A_n \mid B)$.

These formulas help us to compute conditional probabilities of "composite" events using the conditional probabilities of "simple" events.

<u>Examples</u>

E1. Roll a fair die twice. Given that there is at least one "six" among the results, compute the probability that the difference of the results equals 3.

Let A be the event that the difference is 3, B the event that there is at least one "six".

The first question is the conditional probability $P(B \mid A)$. By definition, $P(B \mid A) = \frac{P(A \cap B)}{P(A)}$.

$A \cap B = \{(6,3), (3,6)\}$, $P(A \cap B) = \frac{2}{36}$,

$A = \{(1,6), (2,6), , (3,6), (4,6), (5,6), (6,6), (6,1), (6,2), (6,3), (6,4), (6,5)\}$, $P(A) = \frac{11}{36}$.

$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{2}{36}}{\frac{11}{36}} = \frac{2}{11}$. Roughly spoken, our world is restricted to A, it contains 11 elements. Two of them have difference 3. If all possible elements are equally probable in the entire set Ω, then all possible outcomes are equally probable in A, as well. Consequently, the conditional probability is $\frac{2}{11}$.

Given that the difference of the results is 3, compute the probability that there is at least one "six".

The second question is the conditional probability $P(A|B)$. By definition, $P(A|B)=\frac{P(A\cap B)}{P(B)}$. $B=\{(1,4),(4,1),(2,5),(5,2),(3,6),(6,3)\}$, $P(B)=\frac{6}{36}$.

Consequently, $P(A|B)=\frac{P(A\cap B)}{P(B)}=\frac{\frac{2}{36}}{\frac{6}{36}}=\frac{1}{3}$.

Roughly spoken, our world is restricted to the set B. Two elements are appropriate among them. If all possible elements are equally probable in the entire set Ω, then all possible outcomes are equally probable in B, as well. Consequently the classical probability can be applied, which concludes that the conditional probability equals $\frac{2}{6}=\frac{1}{3}$.

E2. Roll a fair die 10 times, repeatedly. Given that there is at least one “six”, compute the probability that there is at least one “one”.

Let A be the event that there is no “six” among the results, and B the event that there is no “one” among the results. The question is the conditional probability $P(\overline{B}|\overline{A})$.

$$P(\overline{B}|\overline{A})=\frac{P(\overline{A}\cap\overline{B})}{P(\overline{A})}=\frac{P(\overline{A\cup B})}{P(\overline{A})}=\frac{1-P(A\cup B)}{1-P(A)}=\frac{1-\left(P(A)+P(B)-P(A\cap B)\right)}{1-P(A)}.$$

Now we can see that we have to compute the values $P(A)$, $P(B)$ and $P(A\cap B)$.

$$P(A)=\frac{5^{10}}{6^{10}}=0.161,\quad P(B)=\frac{5^{10}}{6^{10}}=0.161,\quad P(A\cap B)=\frac{4^{10}}{6^{10}}=0.017.$$

$$P(\overline{B}|\overline{A})=\frac{1-\left(P(A)+P(B)-P(A\cap B)\right)}{1-P(A)}=\frac{1-\left(0.161+0.161-0.017\right)}{1-0.161}=\frac{0.695}{0.839}=0.828.$$

E3. Choose two numbers independently in the interval [0,1] by geometrical probability. Given that the difference of the numbers is less than 0.3, compute the probability that the sum of the numbers is at least 1.5.

Let A be the event that the difference of the numbers is less than 0.3. The appropriate points in the square [0,1]x[0,1] are situated between the straight lines given by the equation $x-y=0.3$ and $y-x=0.3$. It is easy to see that $P(A)=1-0.7^2=0.51$. $A\cap B$ contains those points of A which are above the straight line given by x+y=1.5. This part is denoted by horizontal lines in Fig.c.1. The cross-points are $Q_1(0.6,0.9)$ and $Q_2(0.9,0.6)$. The area of the appropriate points is $\mu(A)=\left(\sqrt{0.3^2+0.3^2}\right)\cdot\left(\sqrt{0.1^2+0.1^2}\right)+\frac{0.3^2}{2}=0.06+0.045=0.105$, $P(A\cap B)=0.105$.

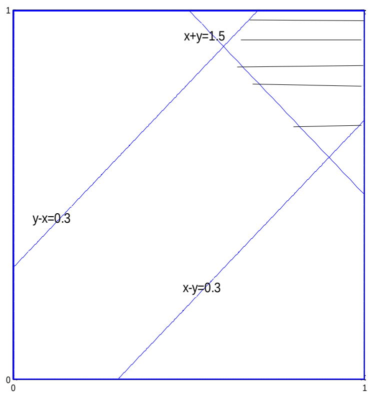


Fig.c.1. The points satisfying conditions $1.5\le x+y$ and $|x-y|<0.3$

$$P(B\,|\,A)=\frac{P(A\cap B)}{P(A)}=\frac{0.105}{0.51}=0.206\,.$$

E4. Order the numbers of the set $\{1,2,3,4,...,10\}$ and suppose that all arrangements are equally probable. Given that the number "1" is not on its proper place, compute the probability that the number 10 is on its proper place.

Let A_i be the event that the number "i" is in its proper place. The question is the conditional probability $P(A_{10}\,|\,\overline{A_1})$. Now

$$P(A_{10}\,|\,\overline{A_1})=\frac{P(A_{10}\cap\overline{A_1})}{P(\overline{A_1})}=\frac{P(A_{10}\setminus A_1)}{P(\overline{A_1})}=\frac{P(A_{10})-P(A_{10}\cap A_1)}{1-P(A_1)}\,.$$

We can see that we need the values $P(A_1)$, $P(A_{10})$ and $P(A_{10}\cap A_1)$.

$\Omega=\{(i_1,i_2,...,i_{10}):1\le i_j\le 10,\text{integers},\ j=1,2,...,10,\ i_j\ne i_k\ \text{if}\ j\ne k\}$, for example $(1,2,3,4,5,6,7,8,9,10)$, $(5,2,3,4,7,9,10,8,1,7)$ and so on. $|\Omega|=10!=3628800$.

$A_1=\{(1,i_2,....,i_{10}):2\le i_j\le 10,\text{integers},\ j=2,...,10,\ i_j\ne i_k\ \text{if}\ j\ne k\}$, $|A_1|=9!$, $P(A_1)=\frac{9!}{10!}=0.1$.

Similarly, $P(A_{10})=\frac{9!}{10!}=0.1$.

$A_1\cap A_{10}=\{(1,i_2,....,10):1\le i_j\le 10,\text{integers},\ j=2,...,9,\ i_j\ne i_k\ \text{if}\ j\ne k\}$, $|A_1\cap A_{10}|=8!$ as the numbers 1 and 10 have to be on their proper places, $P(A_1\cap A_{10})=\frac{8!}{10!}=\frac{1}{10\cdot 9}=0.011$.

Therefore, $$P(A_{10}\,|\,\overline{A_1})=\frac{P(A_{10})-P(A_{10}\cap A_1)}{1-P(A_1)}=\frac{\frac{1}{10}-\frac{1}{10\cdot 9}}{\frac{9}{10}}=\frac{\frac{8}{10\cdot 9}}{\frac{9}{10}}=\frac{8}{81}=0.099\,.$$

E5. Order the numbers of the set $\{1,2,3,4,...,10\}$ and suppose that all arrangements are equally probable. Given that the number "1" is not on its proper place, compute the probability that the number "10" or the number "5" is on its proper place.

Let A_i the event that the number "i" is on its proper place. The question is the conditional probability $P(A_{10}\cup A_5\,|\,\overline{A_1})$. Recall the properties of conditional probability, namely

$$P(A_{10}\cup A_5\,|\,\overline{A_1})=P(A_{10}\,|\,\overline{A_1})+P(A_5\,|\,\overline{A_1})-P(A_{10}\cap A_5\,|\,\overline{A_1})\,.$$

We can realize that the conditional probabilities $P(A_{10}\,|\,\overline{A_1})$, $P(A_5\,|\,\overline{A_1})$ and $P(A_{10}\cap A_5\,|\,\overline{A_1})$ are needed. $P(A_{10}\,|\,\overline{A_1})$ was computed in the previous example, and $P(A_5\,|\,\overline{A_1})$ can be computed in the same way.

$$P(A_{10}\cap A_5\,|\,\overline{A_1})=\frac{P(A_{10}\cap A_5\cap\overline{A_1})}{P(\overline{A_1})}=\frac{P(A_{10}\cap A_5)-P(A_{10}\cap A_5\cap A_1)}{1-P(A_1)}\,.$$

$$A_{10}\cap A_5\cap A_1=\left\{\begin{matrix}(1,i_2,i_3,i_4,5,i_6,i_7,i_8,i_9,10):2\le i_j\le 4,6\le i_j\le 9,\\ \text{integers},\ j=2,3,4,6,7,8,9,\ i_j\ne i_k\ \text{if}\ j\ne k\end{matrix}\right\}$$

$|A_{10}\cap A_5\cap A_1|=7!$ as the numbers „1", „10" and „5" are on their proper places.

Consequently, $P(A_{10}\cap A_5\cap A_1)=\frac{7!}{10!}=\frac{1}{10\cdot 9\cdot 8}=0.001$, and

$$P(A_{10}\cap A_5\,|\,\overline{A_1})=\frac{P(A_{10}\cap A_5)-P(A_{10}\cap A_5\cap A_1)}{1-P(A_1)}=\frac{\frac{1}{10\cdot 9}-\frac{1}{10\cdot 9\cdot 8}}{1-\frac{1}{10}}=\frac{\frac{7}{10\cdot 9\cdot 8}}{\frac{9}{10}}=$$

$$=\frac{7}{81\cdot 8}=0.011\,.$$

Now

$$P(A_{10}\cup A_5\,|\,\overline{A_1})=P(A_{10}\,|\,\overline{A_1})+P(A_5\,|\,\overline{A_1})-P(A_{10}\cap A_5\,|\,\overline{A_1})=\frac{8}{81}+\frac{8}{81}-\frac{7}{81\cdot 8}=\frac{121}{81\cdot 8}=0.187\,.$$

E6. Pick 4 cards without replacement from a package containing 52 cards. Given that there are no hearts or there are no spades, compute the probability that there are no hearts and there are no spades.

Let A be the event that there are no hearts, B the event that there are no spades. The question is the conditional probability $P(A\cap B\,|\,A\cup B)$.

$$P(A\cap B\,|\,A\cup B)=\frac{P((A\cap B)\cap(A\cup B))}{P(A\cup B)}=\frac{P(A\cap B)}{P(A\cup B)}\text{, as }(A\cap B)\subset(A\cup B).$$

We have to compute the probabilities $P(A\cap B)$ and $P(A\cup B)$. This latter one requires $P(A)$, $P(B)$ and $P(A\cap B)$. As the sampling is performed without replacement we do not have to consider the order of the cards.

$\Omega=\{\{i_1,i_2,i_3,i_4\}: i_j \text{ are the cards from the package}, i_j\ \ j=1,2,3,4 \text{ are different if } j\neq k\}$.

$$|\Omega|=\binom{52}{4},\ |A|=|B|=\binom{39}{4},\ |A\cap B|=\binom{26}{4},$$

$$P(A)=\frac{\binom{39}{4}}{\binom{52}{4}}=P(B)=0.304\ ,\ P(A\cap B)=\frac{\binom{26}{4}}{\binom{52}{4}}=0.055\ ,$$

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)=0.553\,.$$

$$P(A\cap B\,|\,A\cup B)=\frac{P(A\cap B)}{P(A\cup B)}=\frac{0.055}{0.553}=0.099\,.$$

E7. Pick 4 cards without replacement from a package containing 52 cards. Compute the probability that the first card is a heart, the second and the third cards are diamonds and the fourth one is a spade.

Let A be the event that the first card is a heart, B the event that the second one is a diamond, C the event that the third card is a diamond and D the event that the last one is a spade. The question is $P(A\cap B\cap C\cap D)$. Applying the generalized form of the multiplicative rule, we can write that $P(A\cap B\cap C\cap D)=P(A)\cdot P(B|A)\cdot P(C|A\cap B)\cdot P(D|A\cap B\cap C)$. Notice that the conditional probabilities $P(B|A)$, $P(C|A\cap B)$, $P(D|A\cap B\cap C)$ can be computed by the following arguments. If we know that the first card is a heart, then the package contains at the 2nd draw 51 cards and 13 are diamonds of them. The third and last ones can be any cards, consequently $P(B\,|\,A)=\frac{13}{51}$. If we know that the first card is a heart and the second one is a diamond, then the package contains 50 cards at the third draw and 12 are diamonds of them. The last one can be any card, consequently $P(C\,|\,A\cap B)=\frac{12}{50}$.

Finally, if we know that the first card is a heart, the second and third ones are diamonds, then the package contains 49 cards at the last picking and 13 are spades among them. Consequently, $P(D\,|\,A\cap B\cap C)=\frac{13}{49}$. As $P(A)=\frac{13}{52}$, $P(A\cap B\cap C\cap D)=\frac{13}{52}\cdot\frac{13}{51}\cdot\frac{12}{50}\cdot\frac{13}{49}=0.004$.

We present the following "simple" solution as well. As the question is connected to the order of pickings, we have to consider the order of the picked cards.

$\Omega=\{(i_1,i_2,i_3,i_4): i_j \text{ are the cards from the package}, i_j\ \ j=1,2,3,4 \text{ are different if } j\neq k\}$.

$|\Omega|=52\cdot 51\cdot 50\cdot 49$. If the first draw is a heart, then we have 13 possibilities at the first draw. If the second card is a diamond, then we have 13 possibilities at the second picking. If the third card is a diamond again we have only 12 possibilities at the third picking, as the previous draw eliminates one of the diamond cards. Finally, if the last card is a spade, we have 13 possibilities at the last picking. Consequently, $|A\cap B\cap C\cap D|=13\cdot 13\cdot 12\cdot 13$,

$P(A\cap B\cap C\cap D)=\frac{13\cdot 13\cdot 12\cdot 13}{52\cdot 51\cdot 50\cdot 49}$, which is exactly the same as we have got by applying the multiplicative rule.

c.2. Theorem of total probability, Bayes' theorem

In the examples of the previous section the conditional probabilities were computed from unconditional ones. The last example was solved by two methods. One of them has applied conditional probabilities for determining an unconditional one. The law of total probability applies conditional probabilities for computing unconditional (total) probabilities. To do this, we need a partition of the sample space Ω.

Suppose that Ω, $\mathcal{A}$, and P are given.

Definition The set of events $B_1, B_2,..., B_n \in \mathcal{A}$ is called a **partition** of Ω, if $\Omega=\bigcup_{i=1}^{n} B_i$ and $B_i\cap B_j=\emptyset$, $i\neq j$, $1\le i\le n$, $1\le j\le n$.

We note that a partition cuts the set of possible outcomes into some mutually exclusive events. Every possible outcome belongs to an event and none of them can belong to two events.

Theorem (Law of total probability) Let $B_1, B_2,..., B_n \in \mathcal{A}$ be a partition of Ω, and assume that $0<P(B_i)$, $i=1,2,...,n$. Then for any event $A\in\mathcal{A}$ the following equality holds

$$P(A)=\sum_{i=1}^{n} P(A\,|\,B_i)P(B_i).$$

Proof: As $0<P(B_i)$, the conditional probabilities are well defined.

$$P(A)=P(A\cap\Omega)=P\left(A\cap\left(\bigcup_{i=1}^{n} B_i\right)\right)=P\left(\bigcup_{i=1}^{n}(A\cap B_i)\right).$$

Notice that if $B_i\cap B_j=\emptyset$, then $(A\cap B_i)\cap(A\cap B_j)=\emptyset$. Therefore the unioned events are mutually exclusive and the probability of the union is the sum of the probabilities.

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

Recalling the multiplicative rule $P(A\cap B_i)=P(A\,|\,B_i)\cdot P(B_i)$ we get

$$P(A)=\sum_{i=1}^{n} P(A\,|\,B_i)P(B_i).$$

An inverse question can be asked in the following way: if we know that A occurs, compute the probability that B_i occurs. The answer can be given by Bayes' theorem as follows:

<u>Theorem</u> (Bayes' theorem) Let $B_1, B_2, ..., B_n \in \mathcal{A}$ be a partition of Ω, and assume $0 < P(B_i)$, $i = 1,2,...,n$. Then for any event $A \in \mathcal{A}$ with $0 < P(A)$, the following holds:
$$P(B_i \mid A) = \frac{P(A \mid B_i) \cdot P(B_i)}{P(A)} = \frac{P(A \mid B_i) \cdot P(B_i)}{\sum_{i=1}^{n} P(A \mid B_i) P(B_i)}, \ i = 1,2,...,n.$$

<u>Proof</u> $P(B_i \mid A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(A \mid B_i) \cdot P(B_i)}{P(A)} = \frac{P(A \mid B_i) \cdot P(B_i)}{\sum_{i=1}^{n} P(A \mid B_i) P(B_i)}$.

<u>Remarks</u>

- Notice that the unconditional probability is the weighted sum of the conditional probabilities.
- The law of total probability is worth applying when it is easy to compute conditional probabilities.
- The construction of the partition is sometimes easy, in other cases it can be difficult. The main point is to be able to compute conditional probabilities.
- The theorem can be proved for countably many sets B_i, $i = 1,2,...$, as well.
- Bayes' theorem can be interpreted as the probability of „reasons". If A occurs, what is the probability that its „reason" is B_i, $i = 1,2,3,...$

<u>Examples</u>

E1. In a factory, there are three shifts. 45% of all products are manufactured by the morning shift, 35% of all products are manufactured by the afternoon shift, 20% are manufactured by the evening shift. A product manufactured by the morning shift is substandard with probability 0.04, a product manufactured by the afternoon shift is substandard with probability 0.06, and a product manufactured by the evening shift is substandard with probability 0.08. Choose a product from the entire set of products. Compute the probability that the chosen product is substandard.

Let B_1 be the event that the chosen product was produced by the morning shift, let B_2 be the event that the chosen product was produced by the afternoon shift and let B_3 be the event that the chosen product was produced by the evening shift. B_1, B_2, B_3 is a partition of the entire set of all products. Let S be the event that the chosen product is substandard. Now, $P(S \mid B_1) = 0.04$, $P(S \mid B_2) = 0.06$, $P(S \mid B_3) = 0.08$. Furthermore, $P(B_1) = 0.45$, $P(B_2) = 0.35$, $P(B_3) = 0.2$. Applying the law of total probability we get $P(S) = P(S \mid B_1) \cdot P(B_1) + P(S \mid B_2) \cdot P(B_2) + P(S \mid B_3) \cdot P(B_3) =$ $0.04 \cdot 0.45 + 0.06 \cdot 0.35 + 0.08 \cdot 0.2 = 0.055$.

If the chosen product is substandard, compute the probability that it was produced by the morning shift. If the chosen product is substandard, which shift produced it most probably?

$$P(B_1 \mid S) = \frac{P(S \mid B_1) \cdot P(B_1)}{P(S)} = \frac{0.04 \cdot 0.45}{0.055} = 0.327.$$

$$P(B_2 \mid S) = \frac{P(S \mid B_2) \cdot P(B_2)}{P(S)} = \frac{0.06 \cdot 0.35}{0.055} = 0.382.$$

$$P(B_3 \mid S) = \frac{P(S \mid B_3) \cdot P(B_3)}{P(S)} = \frac{0.08 \cdot 0.2}{0.055} = 0.291.$$

If the chosen product is substandard, the second shift is the most probable, as a „reason".

This example draws the attention to the differences between the conditional probabilities $P(S|B_1)$ and $P(B_1|S)$, $P(S|B_2)$ and $P(B_2|S)$, $P(S|B_3)$ and $P(B_3|S)$. Although the maximal value among $P(S|B_1)$, $P(S|B_2)$ and $P(S|B_3)$ is the first conditional probability, the maximal value among $P(B_1|S)$, $P(B_2|S)$ and $P(B_3|S)$ is the second one. $P(S|B_1)$ is the ratio of the substandard products among the products produced by the morning shift, $P(B_1|S)$ is the ratio of the products produced by morning shift among all substandard products. These ratios have to be strictly distinguished.

E2. People are divided into three groups on the basis of their qualification: people with higher, intermediate and elementary degree. We investigate the adults. 25% of all adults have elementary, 40% of all adults have intermediate and the rest of people have higher degree. A person having elementary degree is unemployed with probability 0.18, a person having intermediate degree is unemployed with probability 0.12 and a person having higher degree is unemployed with probability 0.05. Choose a person from the adults. Compute the probability that he/she is unemployed.

Let B_1 be the event that the chosen person has elementary degree, B_2 be the event that the chosen person has intermediate degree, B_3 be the event that the chosen person has higher degree. B_1, B_2, B_3 is a partition of the entire set of Ω. Let E be the event that the chosen person is unemployed. $P(B_1) = 0.25$, $P(B_2) = 0.4$ and $P(B_3) = 0.35$, furthermore $P(E|B_1) = 0.18$, $P(E|B_2) = 0.12$, $P(E|B_3) = 0.05$. Applying the law of total probability we get $P(E) = P(E|B_1)\cdot P(B_1) + P(E|B_2)\cdot P(B_2) + P(E|B_3)\cdot P(B_3) =$
$0.18\cdot 0.25 + 0.12\cdot 0.4 + 0.05\cdot 0.35 = 0.1105$.

If the chosen person is not unemployed compute the probability that he has elementary/intermediate/ higher degree.

$$P(B_1|\overline{E}) = \frac{P(\overline{E}|B_1)\cdot P(B_1)}{P(\overline{E})} = \frac{(1-P(E|B_1))\cdot P(B_1)}{1-P(E)} = \frac{0.82\cdot 0.25}{1-0.1105} = 0.230 .$$

$$P(B_2|\overline{E}) = \frac{P(\overline{E}|B_2)\cdot P(B_2)}{P(\overline{E})} = \frac{(1-P(E|B_2))\cdot P(B_2)}{1-P(E)} = \frac{0.88\cdot 0.4}{1-0.1105} = 0.396 .$$

$$P(B_3|\overline{E}) = \frac{P(\overline{E}|B_3)\cdot P(B_3)}{P(\overline{E})} = \frac{(1-P(E|B_3))\cdot P(B_3)}{1-P(E)} = \frac{0.95\cdot 0.35}{1-0.1105} = 0.374 .$$

We draw the attention to the fact that $P(\overline{E}|B_1) = 1 - P(E|B_1)$ according to the properties of conditional probability.

E3. Pick two cards without replacement from a package of cards containing 52 cards. Compute the probability that the second card is a heart.

If we knew that the first card is a heart or not, the conditional probabilities of the event “second draw is a heart” could be easily computed. Consequently the unconditional probability can be also computed with the help of the conditional probabilities.

Let B_1 be the event that the first card is a heart and $B_2 = \overline{B_1}$. Now B_1 and B_2 form a partition of the entire set of Ω. Let A be the event that the second draw is a heart. Now, $P(A|B_1) = \frac{12}{51}$, $P(A|B_2) = \frac{13}{51}$, furthermore $P(B_1) = \frac{13}{52}$, $P(B_2) = \frac{39}{52}$. Applying the law of total probability we get

$$P(A) = P(A|B_1)\cdot P(B_1) + P(A|B_2)\cdot P(B_2) = \frac{12}{51}\cdot\frac{13}{52} + \frac{13}{51}\cdot\frac{39}{52} = \frac{13\cdot(12+39)}{51\cdot 52} = \frac{13}{52} = 0.25 .$$

Given that the second draw is a heart compute the probability that the first one is not a heart.

$$P(B_2 \mid A) = \frac{P(A \mid B_2) \cdot P(B_2)}{P(A)} = \frac{\frac{13}{51} \cdot \frac{3}{4}}{0.25} = \frac{39}{51}.$$

Given that the second draw is not a heart compute the probability that the first one is a heart.

$$P(B_1 \mid \overline{A}) = \frac{P(\overline{A} \mid B_1) \cdot P(B_1)}{P(\overline{A})} = \frac{\frac{39}{51} \cdot \frac{1}{4}}{\frac{3}{4}} = \frac{13}{51}$$, taking into account that $P(\overline{A} \mid B_1) = 1 - P(A \mid B_1)$.

c3. Independence of events

The conditional probability of an event may differ from the unconditional one. It may be greater or smaller than the unconditional probability, and in some cases they can be equal, as well. Let us consider the following very simple examples.

Roll two fair dies. Let A be the event that the sum of the rolls is 7, let B be the event that the difference of the rolls is at least 4, let be C the event that the difference of the rolls is 0, finally let D be the event that the first roll is 1. Now $P(A) = \frac{6}{36}$, $P(B) = \frac{6}{36}$, $P(C) = \frac{6}{36}$, $P(D) = \frac{6}{36}$.

One can easily see that $P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{\frac{2}{36}}{\frac{6}{36}} = \frac{1}{3} > P(B)$,

$P(C \mid A) = \frac{P(C \cap A)}{P(A)} = \frac{P(\emptyset)}{\frac{1}{6}} = 0 < P(C)$, $P(D \mid A) = \frac{P(D \cap A)}{P(A)} = \frac{\frac{1}{36}}{\frac{1}{6}} = P(D)$. This latter case is the case when the information contained in A does not change the chance of D. It can be checked that $P(A \mid D) = \frac{P(A \cap D)}{P(D)} = \frac{1}{6} = P(A)$ also holds, which means that the information in D does not change the chance of A either. The relation is symmetric. Similarly,

$P(A \mid B) = \frac{P(B \cap A)}{P(B)} = \frac{1}{3} > P(A)$ and $P(A \mid C) = \frac{P(A \cap C)}{P(C)} = 0 < P(A)$.

<u>Definition</u> The events $A, B \in \mathcal{A}$ are called **independent** if $P(A \cap B) = P(A) \cdot P(B)$.

Now we prove that this definition is a generalization of the previous concept.

<u>Theorem</u> Let A and B be events for which $0 < P(A)$ and $0 < P(B)$. Then A and B are independent if and only if $P(A \mid B) = P(A)$ and/or $P(B \mid A) = P(B)$.

<u>Proof</u> Recalling the definition of conditional probability, we can write that $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$ and $P(B \mid A) = \frac{P(B \cap A)}{P(A)}$. If A and B are independent, then, by definition, $P(A \cap B) = P(A) \cdot P(B)$. Dividing by P(A) and P(B) we get the equalities

$\frac{P(A \cap B)}{P(A)} = P(B)$ and $\frac{P(A \cap B)}{P(B)} = P(A)$, respectively. Conversely, $\frac{P(A \cap B)}{P(A)} = P(B)$ implies $P(A \cap B) = P(A) \cdot P(B)$, and so does $\frac{P(A \cap B)}{P(B)} = P(A)$.

Remarks

- The definition of independence is symmetric.
- The definition of independence is valid even in the case of $0 = P(A)$ or $P(B) = 0$.
- If $0 = P(A)$ or $P(B) = 0$, then A and B are independent. Take into consideration that $P(A \cap B) \le P(A)$, $P(A \cap B) \le P(B)$, consequently $P(A \cap B) \le \min(P(A), P(B)) = 0$. Therefore, $P(A \cap B) = 0 = P(A) \cdot P(B)$.
- Independent events are strongly different from mutually exclusive events. If A and B are mutually exclusive, then $A \cap B = \varnothing$, $P(A \cap B) = 0$. $P(A) \cdot P(B) = 0$ implies $P(A) = 0$ or $P(B) = 0$. If A and B are mutually exclusive and $P(A) \ne 0 \ne P(B)$ holds, then A and B can not be independent. Roughly spoken, if A and B are mutually exclusive and one of them occurs, the other one can not occur. Occurrence of A is a very important piece of information with respect to B.
- In the example presented at the beginning of the subsection the events A and D are independent but the events A and B are not. Neither are A and C.
- The independence of A and B means that the “weight” of A in the entire set equals the “weight” of A in B.

Examples

E1. Roll a fair die 5 times. Let A be the event that all rolls are different let B the event that there is no “six” among the rolls. Are the events A and B independent?

Applying our knowledge on sampling with replacement it is easy to see that $P(A) = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2}{6^5} = 0.093$, $P(B) = \frac{5^5}{6^5} = 0.42$, $P(A \cap B) = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{6^5} = 0.015$. As $P(A \cap B) \ne P(A) \cdot P(B)$, A and B are not independent. If we know that there is no “six” among the rolls then we can “feel” that the chance that all the rolls are different has been decreased. We have only five numbers to roll instead of six ones.

E2. There are N balls in a box, M of them are white, N-M are red. Pick n balls from the urn with replacement. Let A be the event that the first one is red, let B the event that the last one is white. Are the events A and B independent?

Recalling the results in connection with sampling with replacement, $P(A) = \frac{(N-M) \cdot N^{n-1}}{N^n} = \frac{N-M}{N} = 1 - \frac{M}{N}$, $P(B) = \frac{N^{n-1} \cdot M}{N^n} = \frac{M}{N}$,

$P(A \cap B) = \frac{(N-M) \cdot N^{n-2} \cdot M}{N^n} = \frac{(N-M) \cdot M}{N^2} = \left(1 - \frac{M}{N}\right) \cdot \frac{M}{N}$. As $P(A \cap B) = P(A) \cdot P(B)$, A and B are independent.

Roughly spoken, the result of the first picking does not effect the result of the last picking, it does not increase and does not decrease the chance of picking a white ball.

E3. There are N balls in an urn, M of them are white N-M are red. Pick 2 balls from the urn without replacement. Let A be the event that the first one is red, let B the event that the second one is white. Are the event A and B independent?

Recalling the results in connection with sampling without replacement, we can write $P(A \cap B) = \frac{(N-M) \cdot M}{N \cdot (N-1)}$, $P(A) = \frac{(N-M) \cdot (N-1)}{N \cdot (N-1)} = \frac{(N-M)}{N}$. $P(B)$ can be computed by the

help of the theorem of total probability as follows: $P(B)=P(B\,|\,A)\cdot P(A)+P(B\,|\,\overline{A})\cdot P(\overline{A})=\dfrac{M}{N-1}\cdot\dfrac{N-M}{N}+\dfrac{M-1}{N-1}\cdot\dfrac{M}{N}=\dfrac{M\cdot(N-M+M-1)}{(N-1)\cdot N}=$
$\dfrac{M}{N}$. As $P(A\cap B)\neq P(A)\cdot P(B)$, A and B are not independent.
Roughly spoken, if we know that the first draw is red, the chance of the second one being white has been increased. The reason is that the relative number of white balls in the urn has increased.

E4. People are grouped into three groups on the basis of their qualification: people with higher, intermediate and elementary degree. We investigate the adults. 25% of all adults have elementary, 40% of all adults have intermediate and the rest of people have higher degree. A person having elementary degree is unemployed with probability 0.18, a person having intermediate degree is unemployed with probability 0.12 and a person having higher degree is unemployed with probability 0.05. Choose a person from the adults. Are the events A="the chosen person is unemployed" and $B_1=$"the chosen person has higher degree" independent?
Recalling the law of total probability we get $P(A)=0.1105$, but $P(A\,|\,B_1)=0.05$. As $P(A\,|\,B_1)\neq P(E)$, A and B_1 are not independent. If somebody has higher degree, the probability of the event that he is unemployed has decreased. The ratio of the unemployed people in the whole population is higher than the ration of the unemployed among people having higher degree.

E5. Roll a fair die 3 times. Let A be the event that the sum of the rolls is at least 17, and let B be the event that all the rolls are the same. Are A and B independent?
Taking the condition into account, the sum of the rolls can be 17 and 18. If the sum is 17 then we roll two "six"s and one "five". if the sum is 18, then we have three "six"-s. $P(A)=\dfrac{3\cdot 1\cdot 1\cdot 1}{6^3}+\dfrac{1}{6^3}=\dfrac{4}{6^3}$. There are four elements in A. One of them satisfies that all of the rolls are the same, consequently $P(B\,|\,A)=\dfrac{1}{4}$. Finally, $P(B)=\dfrac{6\cdot 1\cdot 1}{6^3}=\dfrac{1}{36}$. Now we can see that $P(B\,|\,A)\neq P(B)$, therefore A and B are not independent.

Theorem If the events A and B are independent, then A and $\overline{B}$, furthermore $\overline{A}$ and $\overline{B}$ are independent, as well.
Proof

$$P(A\cap\overline{B})=P(A\setminus B)=P(A)-P(A\cap B)=P(A)-P(A)\cdot P(B)=P(A)(1-P(B))=P(A)\cdot P(\overline{B}).$$

$$P(\overline{A}\cap\overline{B})=P(\overline{A\cup B})=1-P(A\cup B)=1-\big(P(A)+P(B)-P(A\cap B)\big)=$$
$$1-\big(P(A)+P(B)-P(A)\cdot P(B)\big)=\big(1-P(A)\big)\big(1-P(B)\big).$$

Now let us consider the independency of more than two events.
Definition The events $A_i\;\; i\in I$ are called **pairwise independent** if any two of them are independent, that is $P(A_j\cap A_k)=P(A_j)\cdot P(A_k)\;\; j,k\in I,\, j\neq k$.
Definition The events $A_i\;\; i\in I$ are called **independent**, if for any finite set of different indices $\{i_1,i_2,\ldots,i_n\}$ the equality $P(A_{i1}\cap A_{i2}\cap\ldots\cap A_{in})=P(A_{i1})\cdot P(A_{i2})\cdot\ldots\cdot P(A_{in})$ holds.

Remarks

- If the number of elements of the set of indices equals 2, the above property expresses the pairwise independence.
- Pairwise independence of events does not imply independence of the events. We construct the following example in which pairwise independence holds but

$P(A\cap B\cap C)\neq P(A)\cdot P(B)\cdot P(C)$. Let $\Omega=\{1,2,3,4\}$, $P(\{i\})=\frac{1}{4}$, $i=1,2,3,4$. Let $A=\{1,2\}$, $B=\{1,3\}$, $C=\{1,4\}$. Now $P(A)=P(B)=P(C)=\frac{2}{4}=0.5$,

$A\cap B=B\cap C=A\cap C=\{1\}$, $P(A\cap B)=P(B\cap C)=P(A\cap C)=P(\{1\})=\frac{1}{4}=0.25$.

Consequently, $P(A\cap B)=P(A)\cdot P(B)$, $P(A\cap C)=P(A)\cdot P(C)$, $P(B\cap C)=P(B)\cdot P(C)$. It means that A, B and C are pairwise independent. But $P(A\cap B\cap C)=P(\{1\})=0.25\neq P(A)\cdot P(B)\cdot P(C)=\frac{1}{8}$.

Definition Experiments are called **independent** if the events connected to them are independent. In more detail for two experiments: if $\mathcal{A}_1$ is the set of events connected to an experiment, $\mathcal{A}_2$ is the set of events connected to another experiment, then for any $A\in\mathcal{A}_1$ and $B\in\mathcal{A}_2$ the events A and B are independent. The experiments characterized by the set of events $\mathcal{A}_i$, $i\in I$ are independent if for any $A_i\in\mathcal{A}_i$ the events A_i are independent.

Remarks

- Sampling with replacement can be considered as a sequence of independent experiments. If the first draw is the first experiment, the second draw is the second experiment and so on, the events connected to different draws are independent.
- If we do sampling without replacement, then the consecutive draws are not independent experiments, as E3) in the previous subsection illustrates.

Examples

E6. Fill two lotteries (90/5) independently. Compute the probability that at least one of them is bull's-eye.

Let A be the event that the first lottery is bull's-eye, let B the event that the second one is bull's-eye. The question is $P(A\cup B)$. $P(A)=\frac{1}{\binom{90}{5}}$, $P(B)=\frac{1}{\binom{90}{5}}$,

$P(A\cap B)=P(A)\cdot P(B)=\frac{1}{\binom{90}{5}}\cdot\frac{1}{\binom{90}{5}}$. Applying $P(A\cup B)=P(A)+P(B)-P(A\cap B)$ we get

$$P(A\cup B)=\frac{2}{\binom{90}{5}}-\frac{1}{\binom{90}{5}}\cdot\frac{1}{\binom{90}{5}}=4.6\cdot 10^{-8}.$$

E7. Fill 10 million lotteries independently. Compute the probability that at least one of them is bull's-eye.

Let A_i be the event that the ith experiment is bull's-eye. The question is $P(A_1\cup\ldots\cup A_{10^7})$. Instead of it, let us first consider its complement. $P(\overline{A_1\cup\ldots\cup A_{10^7}})=P(\overline{A_1}\cap\overline{A_2}\cap\ldots\cap\overline{A_{10^7}})$. As the experiments are independent, the probability of the intersection of the events connected to them is the product of the probabilities. Therefore

$$P(\overline{A_1} \cap \overline{A_2} \cap ... \cap \overline{A_{10^7}}) = P(\overline{A_1}) \cdot ... \cdot P(\overline{A_{10^7}}) = \left(1 - \frac{1}{\binom{90}{5}}\right)^{10^7} = 0.796 .$$

Consequently, $P(A_1 \cup ... \cup A_{10^7}) = 1 - 0.796 = 0.204$.

E8. How many lotteries are filled independently, if the probability that there is at least one bull's-eye among them equals 0.5?

Let A_i $i = 1,2,...,n$ be the event that the ith experiment is bull's-eye. The question is the value of n if $P(A_1 \cup ... \cup A_n) = 0.5$. Following the argument of the previous example E7

$$P(\overline{A_1 \cup ... \cup A_n}) = P(\overline{A_1} \cap \overline{A_2} \cap ... \cap \overline{A_n}) = \left(1 - \frac{1}{\binom{90}{5}}\right)^{n} = 1 - 0.5 = 0.5 .$$

Taking the logarithm of both sides, we get

$n \cdot \log(1 - \frac{1}{\binom{90}{5}}) = \log 0.5$, $n = \frac{\log 0.5}{\log(1 - \frac{1}{\binom{90}{5}})} = 30463322$, which is much more than the half of possible fillings. But if you fill 30 million lotteries the probability that there are identical fillings is almost 1. If you fill them independently, it may happen that the first one and the second one contain the same numbers crossed.

d. Random variable

The aim of this chapter

This chapter aims to get the reader acquainted with the concept of random variables as random valued functions. We introduce the concept of distribution, cumulative distribution function and probability density function. We present how to the use cumulative distribution function to express probabilities. We introduce the concept of independent random variables.

Preliminary knowledge

Properties of probability. Analysis, taking derivative and integrate.

Content

d.1. Random variables as random valued functions.

d.2. Cumulative distribution function.

d.3. Continuous random variable.

d.4. Independent random variables.

d.1. Random variables as random valued functions

In this section we introduce the concept of random variables as random valued functions.
We suppose that Ω, $\mathcal{A}$ and P are given.
First we introduce a simple definition and later, after presenting lots of examples, we make it mathematically exact.

Definition The function $\xi:\Omega\rightarrow R$ is called a **random variable**.

Remarks

- Random variables map the set of possible outcomes to the set of real numbers. The values of random variables are numbers. If we know the result of the experiment, we know the actual value of the random variable. Before we perform the experiment, we do not know the actual outcome; hence we do not know the value of the function. "Randomness" is hidden in the outcome.
- Although we do not know the value of the function, we know the possible outcomes and the values assigned to them. In analysis, these values are called the image of the function. We will call them possible values of the random variable.
- If we know the possible values of the function, we can presumably compute the probabilities belonging to these possible values. That is we can compute the probability that the function takes this value. Additional refinement is needed to enable us to do this in all cases.
- As the elements of Ω are not real numbers in some cases, the function ξ may not be drawn in a usual Cartesian frame.

Examples

E1. Flip a coin. If the result is heads we gain 10 HUF, if the result is tail we pay 5 HUF. Let ξ be the money we get/pay during a game.

$\Omega=\{H,T\}$, $\mathcal{A}=2^{\Omega}$, P is the classical probability. $\xi:\Omega\rightarrow R$, $\xi(H)=10$, $\xi(T)=-5$. The possible values of ξ are 10 and -5, and $P(\xi=10)=P(\{H\})=0.5$, $P(\xi=-5)=P(\{T\})=0.5$. Before performing the experiment we do not know the value of our gain, but we can state that it can be 10 or -5 and both values are taken with probability 0.5.

E2. Roll a fair die. We gain the square of the result. Let ξ be the gain playing one game.

$\Omega=\{1,2,3,4,5,6\}$, $\mathcal{A}=2^{\Omega}$, P is the classical probability. $\xi:\Omega\to R$, $\xi(i)=i^2$. $\xi(1)=1^2=1$, $\xi(2)=2^2=4$, $\xi(3)=3^2=9$, $\xi(4)=4^2=16$, $\xi(5)=5^2=25$, $\xi(6)=6^2=36$. Moreover, $P(\xi=i^2)=P(\{i\})=\frac{1}{6}$. Summarizing, the possible values of ξ are 1,4,9,16,25,36, and the probabilities belonging to them are $\frac{1}{6}$. Before we roll the die we do not know how much money we gain, but we can state that it may be 1,4,9,16,25 or 36, and all of them have probability $\frac{1}{6}$.

E3. Roll a fair die twice. Let ξ be the sum of the rolls.

$\Omega=\{(1,1),(1,2),.....(6,6)\}$, $\mathcal{A}=2^{\Omega}$, P is the classical probability. $\xi:\Omega\to R$, $\xi((i,j))=i+j$. For example, $\xi((1,1))=2$, $\xi((2,5))=7$, $\xi((6,6))=12$. The possible values of ξ are 2,3,4,5,6,7,8,9,10,11,12.

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

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

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

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

$P(\xi=8)=P(\{(2,6,(3,5),(4,4),(5,3),(6,2)\})=\frac{5}{36}$, $P(\xi=9)=P(\{(3,6),(4,5),(5,4),(6,3)\})=\frac{4}{36}$,

$P(\xi=10)=P(\{(4,6),(5,5),(6,4)\})=\frac{3}{36}$, $P(\xi=11)=P(\{(5,6),(6,5)\})=\frac{2}{36}$,

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

We mention that the sets $B_i=\{\omega:\xi(\omega)=i\}$ $i=2,3,...,12$ are mutually exclusive and the union of them is Ω. They form a partition. Consequently, the sum of the probabilities belonging to the possible values equals 1.

E4. Choose two numbers without replacement from the set $\{0,1,2,3,4\}$. Let ξ be the minimum of the chosen numbers.

Actually, $\Omega=\{\{i_1,i_2\}:0\le i_1<i_2\le 4,\text{int egers}\}$, $\xi:\Omega\to R$, $\xi(\{i_1,i_2\})=\min\{i_1,i_2\}$, $|\Omega|=\binom{5}{2}=10$. $\xi(\{0,4\})=0$, $\xi(\{2,3\})=2$ and so on. The possible values of ξ are 0,1,2,3

and $P(\xi=0)=P(\{0,1\},\{0,2\},\{0,3\},\{0,4\})=\frac{4}{10}$, $P(\xi=1)=P(\{1,2\},\{1,3\},\{1,4\})=\frac{3}{10}$,

$P(\xi=2)=P(\{2,3\},\{2,4\})=\frac{2}{10}$, $P(\xi=3)=P(\{3,4\})=\frac{1}{10}$.

E5. Pick two numbers with replacement from the set $\{0,1,2,3,4\}$. Let ξ be the minimum of the picked numbers.

Actually, $\Omega=\{(i_1,i_2):0\le i_1,i_2\le 4, \text{int egers}\}$, $\xi:\Omega\to R$, $\xi((i_1,i_2))=\min\{i_1,i_2\}$, $|\Omega|=5\cdot 5=25$. $\xi((0,4))=0$, $\xi((3,3))=3$ and so on. The possible values of ξ are $0,1,2,3,4$

and $P(\xi=0)=P(\{(0,0),(0,1),(0,2),(0,3),(0,4),(1,0),(2,0),(3,0),(4,0)\})=\frac{9}{25}$,

$P(\xi=1)=P((1,1),(1,2),(1,3),(1,4),(4,1),(4,2),(4,3))=\frac{7}{25}$,

$P(\xi=2)=P((2,2),(2,3),(2,4),(3,2),(4,2))=\frac{5}{25}$, $P(\xi=3)=P((3,3),(3,4),(4,3))=\frac{3}{25}$,

$P(\xi=4)=P((4,4))=\frac{1}{25}$.

E6. Choose two numbers with replacement of the set $\{0,1,2,3,4\}$. Let ξ be their difference.

Actually, the elements of the sample space are as in the previous example, but the mappings differ. $\xi((1,1))=0$, $\xi((4,1))=3$, and so on. The possible values of ξ are $0,1,2,3,4$ and

$P(\xi=0)=P(\{(0,0),(1,1)(2,2),(3,3),(4,4)\})=\frac{5}{25}$,

$P(\xi=1)=P(\{(0,1),(1,0)(2,1),(1,2),(3,2),(2,3),(3,4),(4,3)\})=\frac{8}{25}$,

$P(\xi=2)=P(\{(0,2),(2,0)(3,1),(1,3),(4,2),(2,4)\})=\frac{6}{25}$,

$P(\xi=3)=P(\{(0,3),(3,0),(1,4),(4,1)\})=\frac{4}{25}$, $P(\xi=4)=P(\{(0,4),(4,0)\})=\frac{2}{25}$.

E7. Fire into a circle with radius R and suppose that the probability that the hit is situated in a subset of the circle is proportional to the area of the subset. Let ξ be the distance of the hit from the centre of the circle.

Actually, Ω is the circle and $\mathcal{A}$ are those subsets of the circle which have area. If Q is a point of the circle, then $\xi(Q)=d(O,Q)$. Possible values of ξ are the points of the interval $[0,R]$. $P(\xi=0)=P(\{O\})=\frac{\mu(O)}{R^2\pi}=0.$ $P(\xi=R)=\frac{\mu_R}{R^2\pi}$, where μ_R is the area of the boundary of the circle with radius R, which equals 0. $P(\xi=R)=0.$ If $0<x<R$, then

$P(\xi = x) = \frac{\mu_x}{R^2 \pi}$, where μ_x is the area of the boundary of the circle with radius x, which equals 0, as well. Consequently, all possible values have probability 0.

E8. Choose two numbers independently from the interval [0,1] by geometric probability. Let ξ be their difference.

Now, $\Omega = [0,1]x[0,1]$, which is a square. $\mu(\Omega) = 1$. The possible values of ξ are the points of $[0,1]$. Actually, $P(\xi = 0) = \frac{\mu(\{Q(x,y): x = y\} \cap \Omega)}{1}$. The area of the line given by the equation $x = y$ in the square equals 0, consequently, $P(\xi = 0) = 0$. $P(\xi = 1) = \frac{\mu(\{(1,0),(0,1)\})}{1} = 0.$ Generally, If $0 < u < 2$, then $P(\xi = u) = \frac{t(\{Q(x,y): |x - y| = u\} \cap \Omega)}{1}$. The set $\{Q(x,y): |x - y| = u\}$ consists of the points of the lines given by x-y=u and y-x=u, and the area of the two lines equals 0. Therefore $P(\xi = u) = 0$.

Remarks

- The common feature of E1, E2,…,E6 is that the set of the possible values are finite.
- Another common feature of E1, E2,…,E6 is that if x_i is a possible value of ξ, then $P(\xi = x_i) \neq 0$.
- If the possible values of ξ are denoted by $x_1,..., x_n$, then the sets $B_i = \{\omega : \xi(\omega) = x_i\}$ form a partition of Ω. Consequently, $\sum_{i=1}^{n} P(\xi = x_i) = \sum_{i=1}^{n} P(B_i) = P(\Omega) = 1$.
- The common feature of E7 and E8 is that the set of possible values is uncountably infinite and if x is a possible value then $P(\xi = x) = 0$. Nevertheless, $P(\cup\{\omega : \xi(\omega) = x\}) = 1$. If $B_x = \{\omega : \xi(\omega) = x\}$, and $B_y = \{\omega : \xi(\omega) = y\}$, then $B_x \cap B_y = \varnothing$, if $x \neq y$. If the set of possible values were countable, then $P(\bigcup_{i=1}^{\infty}\{\omega : \xi(\omega) = x_i\}) = \sum_{i=1}^{\infty} P(\{\omega : \xi(\omega) = x_i\}) = 0$ would hold.
- In E7 and E8 the probabilities $P(\xi < x)$ are worth investigating instead of $P(\xi = x)$, if the set $\{\omega : \xi(\omega) < x\}$ has probability, i.e. $\{\omega : \xi(\omega) < x\} \in \mathcal{A}$. This requirement is included in the mathematically correct definition of random variables.

<u>Definition</u> The function $\xi:\Omega\to R$ is called a **random variable,** if for any $x\in R$ $\{\omega:\xi(\omega)<x\}\in\mathcal{A}$.

<u>Definition</u> The function $\xi:\Omega\to R$ is called a **discrete random variable,** if the set $\mathrm{Im}(\xi)$ is finite or countably infinite. Those values in $\mathrm{Im}\,\xi$ for which $P(\xi=x)\neq 0$, are the called the possible values.

<u>Definition</u> The d**istribution of the discrete random variable** ξ is the set of the possible values together with the probabilities belonging to them. We denote it by $\xi\sim\begin{pmatrix} x_1, & x_2, & . & . & x_n \\ p_1, & p_2, & . & . & p_n \end{pmatrix}$ or in the infinite case $\xi\sim\begin{pmatrix} x_1, & x_2, & . & . & . \\ p_1, & p_2, & . & . & . \end{pmatrix}$ with $p_i=P(\xi=x_i)$.

<u>Remarks</u>

- The definition of a discrete random variable can be more general as well. In many cases ξ is called a discrete random variable, if there is countable subset C of $\mathrm{Im}\,\xi$, for which $\sum_{x\in C}P(\xi=x)=1$. This means that the set $\mathrm{Im}\,\xi$ may be uncountable, but the values outside C have probability zero together, that is $P(\bigcup_{x\notin C}\{\omega:\xi(\omega)=x\})=0$.

- If $\{\omega:\xi(\omega)<x\}\in\mathcal{A}$, then $\{\omega:\xi(\omega)=x\}=\bigcap_{n=1}^{\infty}\left\{\omega:x\le\xi(\omega)<x+\frac{1}{n}\right\}=$ $\bigcap_{n=1}^{\infty}\left(\left\{\omega:\xi(\omega)<x+\frac{1}{n}\right\}\setminus\{\omega:\xi(\omega)<x\}\right)\in\mathcal{A}$, as $\mathcal{A}$ is σ algebra. Consequently, $P(\{\omega:\xi(\omega)=x\})$ is well defined.

- In examples E1.,…, E6. in the previous subsection, the distributions of random variables ξ are given, namely:

In E1. $\xi\sim\begin{pmatrix} 10, & -5 \\ 0.5, & 0.5 \end{pmatrix}$.

In E2. $\xi\sim\begin{pmatrix} 1 & 4 & 9 & 16 & 25 & 36 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}$.

In E3. $\xi=\begin{pmatrix} 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ \frac{1}{36} & \frac{2}{36} & \frac{3}{36} & \frac{4}{36} & \frac{5}{36} & \frac{6}{36} & \frac{5}{36} & \frac{4}{36} & \frac{3}{36} & \frac{2}{36} & \frac{1}{36} \end{pmatrix}$.

In E4. $\xi\sim\begin{pmatrix} 0 & 1 & 2 & 3 \\ 0.4 & 0.3 & 0.2 & 0.1 \end{pmatrix}$.

In E5. $\xi \sim \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 0.36 & 0.28 & 0.2 & 0.12 & 0.04 \end{pmatrix}$

In E6. $\xi \sim \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 0.2 & 0.32 & 0.24 & 0.16 & 0.08 \end{pmatrix}$.

- The examples in E7. and E8. are not discrete random variables even in the generalized sense of definition.

d.2. Cumulative distribution function

As the probabilities $P(\xi = x)$ are not always appropriate for characterizing random variables, the probability $P(\xi < x)$ is investigated. This probability depends on the value of x. If we consider this probability as the function of x, we get a real-real function. This function is called the cumulative distribution function.

Definition Let ξ be a random variable. The **cumulative distribution function** of ξ is defined as $F: R \to R$ $\quad F_\xi(x) = P(\xi < x) = P(\{\omega : \xi(\omega) < x\})$.

Remarks

- If the random variable ξ is fixed, then the index is omitted.
- As F is a real-real function, it can be represented in the usual Cartesian frame.

Examples

Determine the cumulative distribution functions of the random variables in E1, E2, E6, E7, and E8 in subsection d.1.

E1. $\quad \xi \sim \begin{pmatrix} 10, & -5 \\ 0.5, & 0.5 \end{pmatrix}$.

It can be seen easily that if $x \le -5$, then $P(\xi < x) = P(\emptyset) = 0$.

If $-5 < x \le 10$, then $P(\xi < x) = P(\xi = -5) = P(\{T\}) = 0.5$.

If $10 < x$, then $P(\xi < x) = P(\Omega) = 1$.

Summarizing $F(x) = P(\xi < x) = \begin{cases} 0 & \text{if } x \le -5 \\ 0.5 & \text{if } -5 < x \le 10 \\ 1 & \text{if } 10 < x \end{cases}$.

The graph of this function can be seen in Fig. d.1.

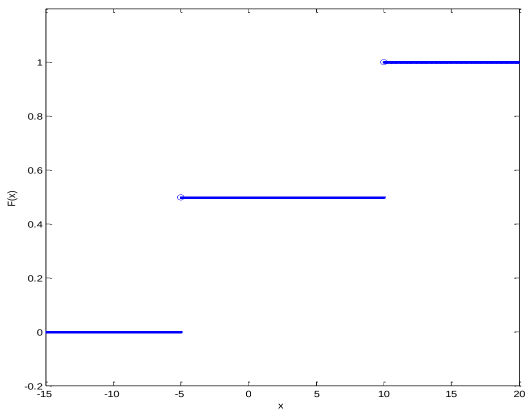


Figure d.1. The cumulative distribution function of the random variable $\xi \sim \begin{pmatrix} 10, & -5 \\ 0.5, & 0.5 \end{pmatrix}$

E2. $\xi \sim \begin{pmatrix} 1 & 4 & 9 & 16 & 25 & 36 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}$.

$$F(x) = \begin{cases} 0 \text{ if } x \le 1 \\ \frac{1}{6} \text{ if } 1 < x \le 4 \\ \frac{2}{6} \text{ if } 4 < x \le 9 \\ \frac{3}{6} \text{ if } 9 < x \le 16 \\ \frac{4}{6} \text{ if } 16 < x \le 25 \\ \frac{5}{6} \text{ if } 25 < x \le 36 \\ 1 \text{ if } 36 < x \end{cases}$$

E6. $\xi \sim \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 0.2 & 0.32 & 0.24 & 0.16 & 0.08 \end{pmatrix}$

$$F(x) = \begin{cases} 0 \text{ if } x \le 0 \\ 0.2 \text{ if } 0 < x \le 1 \\ 0.52 \text{ if } 1 < x \le 2 \\ 0.76 \text{ if } 2 < x \le 3 \\ 0.92 \text{ if } 3 < x \le 4 \\ 1 \text{ if } 4 < x \end{cases}$$

E7. If $0 < x \le R$,

then $F(x) = P(\xi < x) = P(\{Q \in \Omega : d(O,Q) < x\}) = \frac{\mu(x)}{R^2\pi} = \frac{x^2\pi}{R^2\pi} = \frac{x^2}{R^2}$, where $\mu(x)$ is the area of the circle with radius x. Of course, if $x \le 0$, then $P(\xi < x) = P(\emptyset) = 0$, and if $R < x$, then $P(\xi < x) = P(\Omega) = 1$. Summarizing,

$$F(x) = \begin{cases} 0 & \text{if } x \le 0 \\ \frac{x^2}{R^2} & \text{if } 0 < x \le R \\ 1 & \text{if } R < x \end{cases}.$$

which can be seen in Fig.d.2.

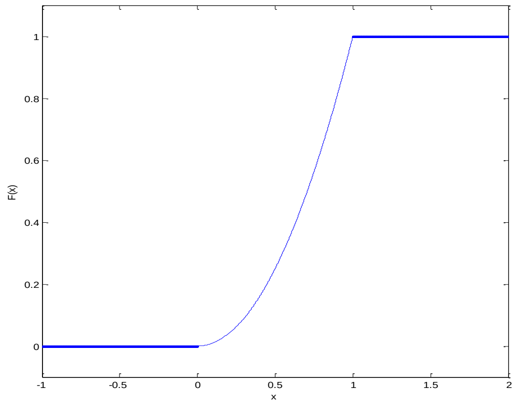


Figure d.2. The cumulative distribution function of the random variable presented in E7

E8. $F(u) = P(\xi < u) = P(\{Q(x,y) : |x - y| < u\})$ if $0 < u < 1$.

Recall that $|x - y| < u$ means, that $x - u < y$ if $y < x$, and $y < x + u$ if $x < y$.

Those points for which $|x - y| < u$ are situated between the straight lines given by the equations $y - x = u$ and $x - y = u$.The area of the appropriate points can be computed by subtracting the area of the two triangles from the area of the square. The area of a triangle is $\frac{(1-u)^2}{2}$. Consequently, $P(\xi < u) = 1 - (1-u)^2$ if $0 < u \le 1$. It is obvious that if $u \le 0$, then $P(|x - y| < u) = P(\emptyset) = 0$, and if $1 < u$, then $P(|x - y| < u) = P(\Omega) = 1$.

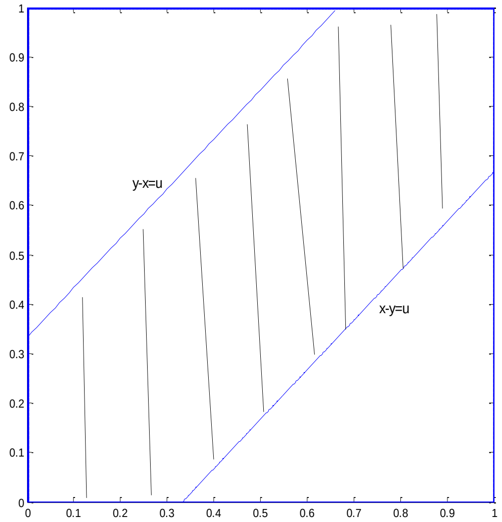


Figure d.3. Appropriate points for the $|x-y|<u$ with $u=0.35$

Summarizing, $F(u)=\begin{cases} 0 & \text{if } u \le 0 \\ 1-(1-u)^2 & \text{if } 0<u\le 1 \\ 1 & \text{if } 1<u \end{cases}$.

The graph of the cumulative distribution function cumulative can be seen in Fig. d.4.

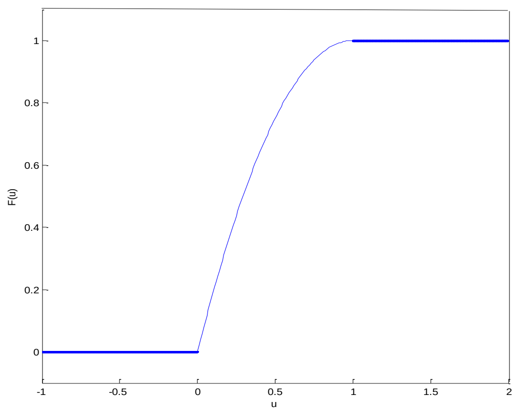


Figure d.4. The cumulative distribution function of the random variable presented in E8

The graphs of the cumulative distribution functions presented have common features and differences, as well. The most conspicuous difference is in continuity, namely the cumulative distribution functions of E1, E2, E6 have discontinuity in jumps, while the cumulative distribution functions of E7. and E8. are continuous. The common features are that they are all increasing functions with values between 0 and 1.

Let us first consider the property of cumulative distribution functions. First we note that $0 \le F(x) \le 1$ for any $x \in R$, as $F(x)$ is a probability.

Theorem Let ξ be a random variable and let $F: R \to R$ be its cumulative distribution function. Then F satisfies the followings:

A) F is a monotone increasing function, that is, in case of $x < y$ $F(x) \le F(y)$.

B) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

C) F is continuous function from the left.

Remark

- The proof of the previous properties can be executed on the basis of the properties of probabilities but we omit it.

The above properties characterize cumulative distribution functions, namely

Theorem If the function $F: R \to R$ satisfies the properties A) B) and C), then there exist a sample space Ω, a σ algebra $\mathcal{A}$ and a probability measure P, furthermore a random variable ξ whose cumulative distribution function is the function F.

Cumulative distribution functions are suitable for expressing the probability that the value of the random variable ξ is situated in a fixed interval. We list these probabilities with explanation in the following theorem:

Theorem

a) $\boxed{P(\xi \in (-\infty, a)) = P(\xi < a) = F(a)}$ by the definition of cumulative distribution function.

b) $\boxed{P(\xi \in [a, \infty)) = P(\xi \ge a) = 1 - F(a)}$.

$P(\xi \in [a, \infty)) = P(\xi \ge a) = P(\overline{\xi < a}) = P(\overline{\{\omega : \xi(\omega) < a\}}) = 1 - F(a)$.

c) $\boxed{P(\xi \in (-\infty, a]) = P(\xi \le a) = F(a) + P(\xi = a)}$

$P(\xi \le a) = P(\{\omega : \xi(\omega) < a\} \cup \{\omega : \xi(\omega) = a\}) = P(\{\omega : \xi(\omega) < a\}) + P(\{\omega : \xi(\omega) = a\})$

$= F(a) + P(\xi = a)$.

d) $\boxed{P(\xi \in (a,\infty)) = P(\xi > a) = 1 - F(a) - P(\xi = a)}$.

$P(\xi > a) = P(\{\omega : \xi(\omega) > a\}) = P(\{\omega : \xi(\omega) \geq a\}) - P(\{\omega : \xi(\omega) = a\}) = 1 - F(a) - P(\xi = a)$.

e) $\boxed{P(\xi \in [a,b)) = P(a \leq \xi < b) = F(b) - F(a)}$

$P(\xi \in [a,b)) = P(a \leq \xi < b) = P(\{\omega : \xi(\omega) < b\}) - P(\{\omega : \xi(\omega) < a\} = F(b) - F(a)$.

Note that $\{\omega : \xi(\omega) < a\} \subset \{\omega : \xi(\omega) < b\}$, consequently the probability of the difference is the difference of probabilities.

f) $\boxed{P(\xi \in [a,b]) = P(a \leq \xi \leq b) = F(b) - F(a) + P(\xi = b)}$

$P(a \leq \xi \leq b) = P(\{\omega : a \leq \xi(\omega) < b\} \cup \{\omega : \xi(\omega) = b\}) = P(\{\omega : a \leq \xi(\omega) < b\}) + P(\{\omega : \xi(\omega) = b\})$
$F(b) - F(a) + P(\xi = b)$. We note that $\{\omega : a \leq \xi(\omega) < b\} \cap \{\omega : \xi(\omega) = b\} = \varnothing$, consequently the probability of the union equals the sum of the probabilities.

g) $\boxed{P(\xi \in (a,b)) = P(a < \xi < b) = F(b) - F(a) - P(\xi = a)}$

$P(a < \xi < b) = P(\{\omega : a \leq \xi(\omega) < b\} \setminus \{\omega : \xi(\omega) = a\}) = F(b) - F(a) - P(\{\omega : \xi(\omega) = a\})$.

h) $\boxed{P(\xi \in (a,b]) = P(a < \xi \leq b) = F(b) - F(a) - P(\xi = a) + P(\xi = b)}$-

$P(a < \xi \leq b) = P(\{\omega : a < \xi(\omega) < b\} \cup \{\omega : \xi(\omega) = b\}) =$

$P(\{\omega : a < \xi(\omega) < b\}) + P(\{\omega : \xi(\omega) = b\}) = F(b) - F(a) - P(\xi = a) + P(\xi = b)$.

i) $\boxed{P(\xi = a) = \lim_{\Delta a \to 0+} F(a + \Delta a) - F(a)}$.

$$P(\xi = a) = P(\bigcap_{n=1}^{\infty}\left\{\omega : a \leq \xi(\omega) < a + \frac{1}{n}\right\}) = \lim_{n\to\infty} P(\left\{\omega : a \leq \xi(\omega) < a + \frac{1}{n}\right\}) = \lim_{n\to\infty}\left(F(a + \frac{1}{n}) - F(a)\right)$$

$$= \lim_{n\to\infty}\left(F(a + \frac{1}{n})\right) - F(a).$$

Remarks

- $\lim_{n\to\infty}\left(F(a + \frac{1}{n})\right) - F(a)$ is the value of the jump of the cumulative distribution function at „a”.
- If F is continuous at “a”, then $\lim_{\Delta a \to 0+} F(a + \Delta a) = F(a)$, consequently $P(\xi = a) = 0$.
- If F is continuous on R, then $P(\xi = x) = 0$ for any $x \in R$. Examples for this case were presented in E7 and E8. Further examples can be given with the help of geometric probability.

Examples

E9. Let the lifetime of a machine be a random variable which has cumulative distribution function $F(x)=\begin{cases}0, & \text{if } x\le 0\\ \dfrac{e^{x}-e^{-x}}{e^{x}+e^{-x}}, & \text{if } 0<x\end{cases}$.

Prove that $F(x)$ is a cumulative distribution function.

To prove that $F(x)$ is a cumulative distribution function it is necessary and sufficient to check the properties A), B) and C).

A) For checking the monotone increasing property, take the derivative.

$$F'(x)=\frac{\left(e^{x}-(-e^{-x})\right)\cdot\left(e^{x}+e^{-x}\right)-\left(e^{x}-e^{-x}\right)\cdot\left(e^{x}+(-1)\cdot e^{-x}\right)}{\left(e^{x}+e^{-x}\right)^{2}}=\frac{4}{\left(e^{x}+e^{-x}\right)^{2}}>0 \quad \text{if} \quad 0<x,$$

consequently the function F is monotone increasing for positive values. As at x=0 the function is continuous and it is constant for negative values, then it is increasing for all values of x.

B) $\lim\limits_{x\to-\infty}F(x)=\lim\limits_{x\to-\infty}0=0$ and $\lim\limits_{x\to\infty}F(x)=1$. $\lim\limits_{x\to\infty}\dfrac{e^{x}-e^{-x}}{e^{x}+e^{-x}}=\lim\limits_{x\to\infty}\dfrac{1-e^{-2x}}{1+e^{-2x}}=1.$

C) $\lim\limits_{x\to0+}\dfrac{e^{x}-e^{-x}}{e^{x}+e^{-x}}=\dfrac{0}{1}=0=\lim\limits_{x\to0-}0$, consequently F is continuous at $x=0$, and it is continuous at any point x. Therefore F is continuous from the left.

Compute the probability that the lifetime of the machine is less than 1 unit.

$$P(\xi<1)=F(1)=\frac{e^{1}-e^{-1}}{e^{1}+e^{-1}}=0.762 .$$

Compute the probability that the lifetime of the machine is between 1 and 2 unit.

$$P(1\le\xi<2)=F(2)-F(1)=\frac{e^{2}-e^{-2}}{e^{2}+e^{-2}}-\frac{e^{1}-e^{-1}}{e^{1}+e^{-1}}=0.964-0.762=0.202$$

Compute the probability that the lifetime is between 2 and 3 unit.

$$P(2\le\xi<3)=F(3)-F(2)=\frac{e^{3}-e^{-3}}{e^{3}+e^{-3}}-\frac{e^{2}-e^{-2}}{e^{2}+e^{-2}}=0.995-0.964=0.031$$

Compute the probability that the lifetime is at least 3 unit.

$$P(3\le\xi)=1-F(3)=1-\frac{e^{3}-e^{-3}}{e^{3}+e^{-3}}=0.005 .$$

Compute the probability that the lifetime of the machine equals 3.

$P(\xi=x)=0$, as the cumulative distribution function of the lifetime is continuous at $x=3$.

At least how much time is the lifetime of the machine with probability 0.9?

$x=?$ $P(\xi\ge x)=0.9$. $1-F(x)=0.9\Rightarrow F(x)=0.1$.

$\frac{e^x - e^{-x}}{e^x + e^{-x}} = 0.1$. Substitute $e^x = y$, we have to find the solution of the following equation: $\frac{y - \frac{1}{y}}{y + \frac{1}{y}} = 0.1$. $\frac{y^2 - 1}{y^2 + 1} = 0.1 \Rightarrow 0.9y^2 = 1.1$. Consequently, $y^2 = \frac{1.1}{0.9} = 1.222$, $y = \pm\sqrt{1.222} = \pm 1.105$. As $y = e^x$, $0 < y$ holds. $e^x = 1.105$ implies $x = \ln 1.105 = 0.100$.

Finally, at most how much time is the lifetime of the machine with probability 0.9?

$x = ?$ $P(\xi \le x) = 0.9$. $P(\xi \le x) = P(\xi < x) + P(\xi = x) = F(x) = 0.9$. Substitute $e^x = y$, we have to find the solution of the following equation: $\frac{y - \frac{1}{y}}{y + \frac{1}{y}} = 0.9$. Following the above steps we get $y = \sqrt{\frac{1.9}{0.1}}$, and $x = \ln\sqrt{\frac{1.9}{0.1}} = 1.472$.

<u>Definition:</u> The random variables ξ and η are called **identically distributed** if $F_\xi(x) = F_\eta(x)$ holds for any value $x \in R$.

<u>Example</u>

E10. $\Omega_1 = \{H, T\}$, $\mathcal{A}_1 = 2^\Omega$, P is the classical probability, $\xi(H) = -1$, $\xi(T) = 1$.

$\Omega_2 = \{1,2,3,4,5,6\}$, $\mathcal{A}_2 = 2^\Omega$, P is the classical probability, $\eta(i) = -1$ if i is odd, $\eta(i) = 1$ if i is even. Now, ξ and η are identically distributed random variables, as

$$F_\xi(x) = F_\eta(x) = \begin{cases} 0 & \text{if } x \le -1 \\ 0.5 & \text{if } -1 < x \le 1 \\ 1 & \text{if } 1 < x \end{cases}.$$

We draw the attention to the fact that the distribution functions may be equal even if the mappings are different.

<u>Theorem</u> If ξ and η are discrete and identically distributed then they have common possible values and $P(\xi = x_i) = P(\eta = x_i)$, $i = 1,2,3,...$

<u>Proof</u> If the random variables have common distribution functions, then the jumps of the cumulative distribution functions are at the same places. This concludes in common possible values. Furthermore, the values of the jumps are equal, as well. Recalling that the jump

equals the probability belonging to the possible value, this means that the random variables take the possible value with the same probability. Consequently, they have the same distribution.

d.3. Continuous random variable

Now we turn our attention to those random variables which have continuous cumulative distribution function.

Definition The random variable ξ is called a **continuous random variable** if its cumulative distribution function is the integral function of a piecewise continuous function, that is there exists a $f: R \to R$ piecewise continuous (continuous except from finitely many points) for which $F(x) = \int_{-\infty}^{x} f(t)dt$. The function f is called the **probability density function** of ξ.

Remarks

- The integral is a Riemann integral.
- It is a well-known fact in analysis that the integral function is continuous at any point, and at the points where f is continuous F is differentiable and $F'(x) = f(x)$.
- If f is changed at a point, its integral function does not change. Consequently the probability density function of a random variable is not unique. Consequently, we can define it at some points arbitrarily. It is typically the case at the endpoints of intervals when f has discontinuity.
- The name "probability density function" can be explained by the followings:

$\frac{P(a \le \xi < a + \Delta a)}{\Delta a}$ expresses the probability that ξ is situated in the neighbourhood of the point "a" relative to the length of the interval. It is a kind of density of being at the neighbourhood of "a". As

$P(a \le \xi < a + \Delta a) = F(a + \Delta a) - F(a)$, $\frac{P(a \le \xi < a + \Delta a)}{\Delta a} = \frac{F(a + \Delta a) - F(a)}{\Delta a}$.

If $0 \le \Delta a \to 0$, then $\lim_{\Delta a \to 0+} \frac{P(a \le \xi < a + \Delta a)}{\Delta a} = \lim_{\Delta a \to 0+} \frac{F(a + \Delta a) - F(a)}{\Delta a} = F'(a) = f(a)$, supposing that the limit exists.

- $F(a + \Delta a) - F(a) \approx F'(a) \cdot \Delta a = f(a) \cdot \Delta a$, therefore where the probability density function has large values there the random variable takes its values with high probability, if the length of the interval is fixed. If the probability density function is zero in the interval $[a, b]$, then the random variable takes its values in $[a, b]$ with probability zero.

- If the cumulative distribution function is a jump function, then at the points of jumps the derivatives do not exist. On the open intervals, when the cumulative distribution function is constant, the derivative takes value zero, consequently there is no sense to take the derivative of the cumulative distribution function.
- We note that there exist random variables which are neither discrete nor continuous. They can be a "mix" of discrete and continuous random variables, their cumulative distribution function is strictly monotone increasing continuous function in some intervals and has jumps at some points. These random variables are out of the frame of this booklet.

<u>Examples</u>

E1. In the example given in E7 in subsection d.1., the probability density function is the following:

$$f(x) = F'(x) = \begin{cases} \frac{2x}{R^2} & \text{if } 0 < x < R \\ 0 & \text{otherwise} \end{cases}.$$

We note that at $x = 0$ the function F is differentiable, and the derivative equals 0.At $x = R$ the function F is not differentiable. The graph of the probability density function for $R = 1$ can be seen in Fig. d.5.

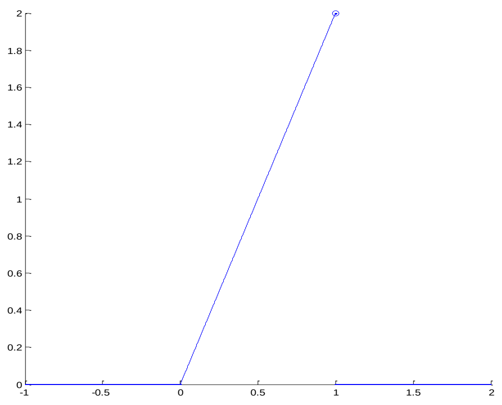


Figure d.5. The probability density function of the random variable given in E7.

E2. The probability density function of E8. in subsection d.1. is

$$f(u) = F'(u) = \begin{cases} 2 - 2u & \text{if } 0 \le u \le 1 \\ 0 & \text{otherwise} \end{cases}.$$

The graph of $f(u)$ can be seen in Fig.d.6.

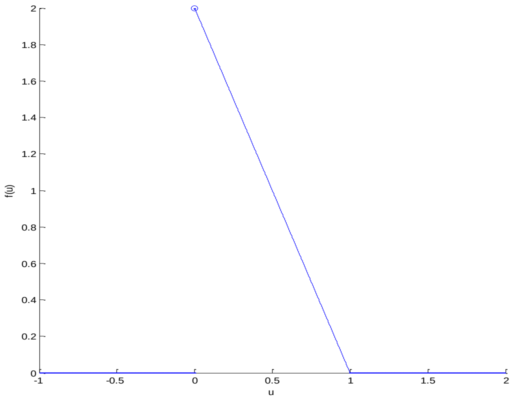


Figure d.6. The probability density function of the random variable given in E8.

E3. The probability density function of E9. in the previous subsection

$$f(x) = F'(x) = \begin{cases} \dfrac{4}{\left(e^{x} + e^{-x}\right)^{2}}, & \text{if } 0 < x \\ 0 & \text{otherwise} \end{cases}.$$

This function can be seen in Fig.d.7.

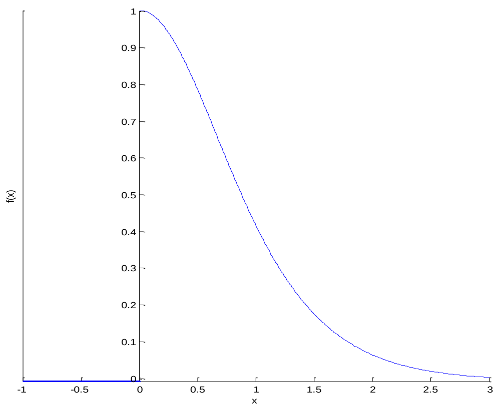


Figure d.7. The probability density function of the random variable given in E9.

The above probability density function takes large values in the interval $[0,1]$ and small values in $[2,3]$ and indeed, $P(0 \le \xi < 1) = 0.762 > P(2 \le \xi < 3) = 0.031$.

Now let us investigate the general properties of density functions.

Theorem If ξ is a continuous random variable, with probability density function f, then

D) $0 \le f(x)$ except from "some" points and

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

Proof F is a monotone increasing function, consequently, its derivative is nonnegative, when the derivative exists. If we choose the values of f to be negative when the derivative does not exist, these points can belong to the set of exceptions. Usually we choose the values of f at these points to be zero. On the other hand, by the definition of the improper integral $\int_{-\infty}^{\infty} f(x)dx = \lim_{x\to\infty} F(x) - \lim_{x\to-\infty} F(x) = 1 - 0$.

The properties D) and E) characterize the probability density functions, namely

Theorem If the function $f: R \to R$ satisfies properties D) and E) then there exist a sample space Ω, a σ algebra $\mathcal{A}$ and a probability measure P, furthermore a continuous random variable ξ whose probability density function is the function f.

Remarks

- If the random variables ξ and η have the same probability density functions, then they have the same cumulative distribution functions as well, therefore they are identically distributed.
- If the random variables ξ and η have the same cumulative distribution functions, then there derivatives also are equal at the points when the derivatives exist. At the points when the derivatives do not exist we can define the probability density functions arbitrarily, but only some points have this property. Consequently, if the continuous random variables ξ and η are identically distributed, then they essentially have the same probability density functions.
- If we would like to express the probability that the continuous random variable ξ takes its values in an interval, we can write the following:

$$\boxed{P(\xi < x) = P(\xi \le x) = F(x)},$$

$$\boxed{P(\xi \ge x) = P(\xi > x) = 1 - F(x)},$$

$$\boxed{P(a \le \xi < b) = P(a \le \xi \le b) = P(a < \xi < b) = P(a < \xi \le b) = F(b) - F(a)}.$$

The reason for this is the fact that the cumulative distribution function of a continuous random variable is continuous at any point, consequently it takes any (given) value with probability 0. Hence we do not have to consider the endpoints of the interval.

Now, we can express the probability of taking values in an interval with the help of the probability density function.

Theorem If the continuous random variable ξ has probability density function f, then $P(a \le \xi \le b) = \int_a^b f(t)dt$.

Proof Applying the formula concerning the cumulative distribution function and the properties of integrals we get

$$\boxed{P(a \le \xi \le b) = F(b) - F(a) = \int_{-\infty}^{b} f(t)dt - \int_{-\infty}^{a} f(t)dt = \int_a^b f(t)dt}.$$

Remarks

- As the integral of a nonnegative function equals the area under the function, the above formula states that the probability of taking values in the interval $[a,b]$ equals the area under the probability density function in $[a,b]$. For example, in the case of the random variable given by the probability density function $f(x) = \begin{cases} 0.5\sin x & \text{if } 0 \le x \le \pi \\ 0 & \text{otherwise} \end{cases}$, the probability of taking values between $\frac{\pi}{6}$ and $\frac{5\pi}{6}$ can be seen in Fig.d.8. It is the area between the two red lines.

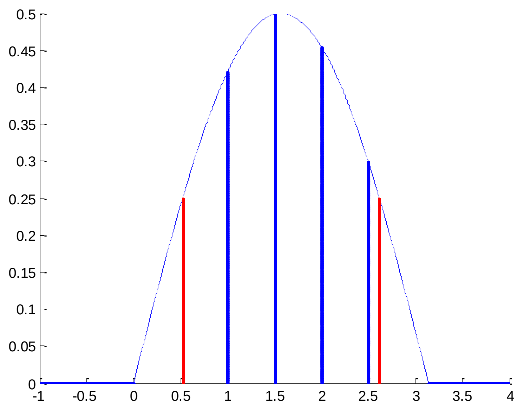


Figure d.8. Probability expressed by the area between the two read lines

Example

E4. Let the error of a measurement be a random variable ξ with probability density function $f(x)=\begin{cases}0.5\cdot e^{x} & \text{if } x<0\\ 0.5\cdot e^{-x} & \text{if } 0\le x\end{cases}$.

The graph of this function can be seen in Fig.d.9.

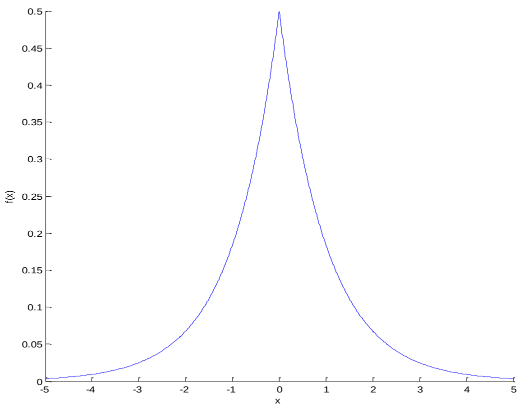


Figure d.9. The probability density function given by f

Prove that f is probability density function.

To do this, check properties D) and E). As exponential functions take only positive values, the inequality $0<f(x)$ holds. Moreover,

$$\int_{-\infty}^{\infty} f(x)dx=\int_{-\infty}^{0}0.5e^{x}dx+\int_{0}^{\infty}0.5e^{-x}dx=0.5\left[e^{x}\right]_{-\infty}^{0}+0.5\left[-e^{-x}\right]_{0}^{\infty}=0.5\left(1-\lim_{x\to-\infty}e^{x}\right)+$$

$$+0.5\left(\lim_{x\to\infty}-e^{-x}-(-1)\right)=0.5+0.5=1.$$

Determine the cumulative distribution function of ξ.

$$F(x)=\int_{-\infty}^{x} f(t)dt=\begin{cases}0.5\cdot e^{x} & \text{if } x\le 0\\ 1-0.5\cdot e^{-x} & \text{if } 0<x\end{cases}.$$

The detailed computations are the following:

If $x\le 0$, then $\int_{-\infty}^{x} f(t)dt=\int_{-\infty}^{x}0.5e^{t}dt=\left[0.5\cdot e^{t}\right]_{-\infty}^{x}=0.5e^{x}-\lim_{x\to-\infty}0.5e^{x}=0.5\cdot e^{x}-0=0.5e^{x}$.

If $0<x$, then

$$\int_{-\infty}^{x} f(t)dt=\int_{-\infty}^{0}0.5e^{t}dt+\int_{0}^{x}0.5\cdot e^{-t}dt=\left[0.5\cdot e^{t}\right]_{-\infty}^{0}+\left[0.5(-e^{-x})\right]_{0}^{x}=0.5+(-0.5e^{-x}-(-0.5))$$

$1-0.5\cdot e^{-x}$.

Compute the probability that the error of the measurement is less than -2.

$P(\xi < -2) = F(-2) = 0.5e^{-2} = 0.068$.

Compute the probability that the error of the measurement is less than 1.

$P(\xi < 1) = F(1) = 1 - 0.5 \cdot e^{-1} = 0.816$.

Compute the probability that the error of the measurement is between -1 and 3.

$P(-1 < \xi < 3) = F(3) - F(-1) = 1 - 0.5 \cdot e^{-3} - 0.5 \cdot e^{-1} = 0.975 - 0.184 = 0.791$.

Compute the probability that the error of the measurement is more than 1.5.

$P(1.5 < \xi) = 1 - F(1.5) = 1 - \left(1 - 0.5 \cdot e^{-1.5}\right) = 0.112$.

Now we ask the inverse question: at most how much is the error with probability 0.9?

We want to find the value x for which $P(\xi \le x) = 0.9$.

Taking into account that $P(\xi \le x) = P(\xi < x) = F(x)$, we seek the value x for which $F(x) = 0.9$. Namely, we would like to determine the cross point of the function F and the line $y = 0.9$, as shown in Fig.d.10.

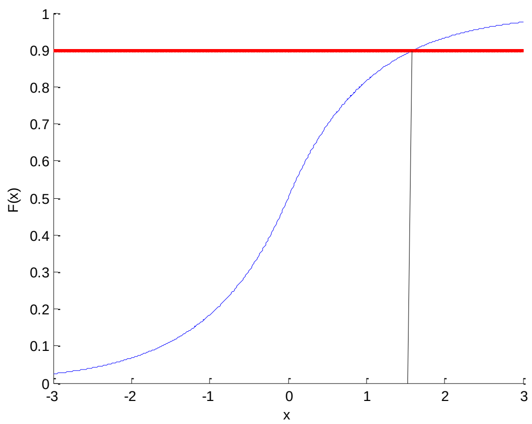


Figure d.10.The cumulative distribution function of ξ and the level 0.9

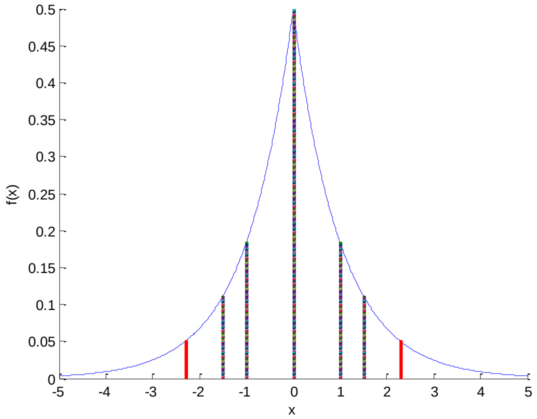
$F(0) = 0.5$, consequently x is positive. For positive values of x $F(x) = 1 - 0.5 \cdot e^{-x}$.

Consequently, $1 - 0.5e^{-x} = 0.9$. This implies $0.5 \cdot e^{-x} = 0.1$, $e^{-x} = 0.2$, $x = -\ln 0.2 = 1.61$.

Give an interval symmetric to 0 in which the value of the error is situated with probability 0.9.

Now we have to determine the value x for which $P(-x < \xi < x) = 0.9$. This means that $F(x) - F(-x) = 0.9$. Substituting the formula concerning $F(x)$ we get $1 - 0.5 \cdot e^{-x} - 0.5e^{-x} = 1 - e^{-x} = 0.9$. This equality implies $e^{-x} = 0.1$, $x = -\ln 0.1 = 2.3$.

In Fig.d.11, the area between the two red lines equals 0.9.

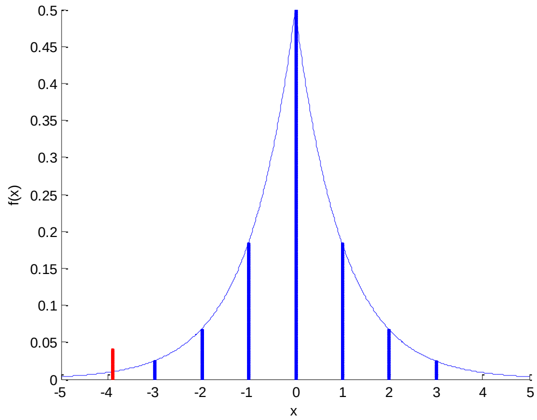


d.11. The probability expressed by the area between the two read lines

At least how much is the error of the measurement with probability 0.99?

Now we would like to determine the value x for which $P(x \le \xi) = 0.99$.

$P(x \le \xi) = 1 - F(x)$, therefore $F(x) = 0.01$. As $F(0) = 0.5$, x is negative. Now we can write the equality $0.5e^{x} = 0.01$, $x = \ln\dfrac{0.01}{0.5} = -3.91$. As Fig.d.12. shows, the area under the density function from the red line to infinity equals 0.99.



d.12.The probability expressed by the area above the red line

d.4. Independent random variables

In this subsection we define independence of random variables.

Definition The random variables ξ and η are called **independent**, if for any values of $x \in R$ and $y \in R$ the events $\{\xi < x\}$ and $\{\eta < y\}$ are independent, that is $P(\xi < x \cap \eta < y) = P(\xi < x) \cdot P(\eta < y)$. For more than two variables, the random variables ξ_i, $i = 1,2,,\ldots$ are called **independent**, if for any value of j, any indices $i_1, i_2,\ldots, i_j \in \{1, 2, 3,\ldots\}$ and any value of x_{ik}, $k = 1,2,.., j$

$P(\xi_{i1} < x_{i1} \cap \ldots \cap \xi_{ij} < x_{ij}) = P(\xi_{i1} < x_{i1}) \cdot P(\xi_{i2} < x_{i2}) \cdot \ldots \cdot P(\xi_{ij} < x_{ij})$.

The independence of random variables is defined by the independence of events connected to them.

The following theorems can be stated:

Theorem If ξ and η are discrete random variables, the distributions of them are $\xi \sim \begin{pmatrix} x_1 & x_2 & . & . & . \\ p_1 & p_2 & . & . & . \end{pmatrix}$ and $\eta \sim \begin{pmatrix} y_1 & y_2 & . & . & . \\ q_1 & q_2 & . & . & . \end{pmatrix}$, then ξ and η are independent if and only if for any $i = 1,2,.$ and $j = 1,2..$ the equality $P(\xi = x_i \cap \eta = y_j) = P(\xi = x_i) \cdot P(\eta = y_j) = p_i \cdot q_j$ holds.

Theorem Let ξ and η be continuous random variables with probability density functions $f(x)$ and $g(y)$, respectively. ξ and η are independent if and only if for any $x \in R$ and $y \in R$ where the $P(\xi < x \cap \eta < y)$ is differentiable, there the following equality holds: $\frac{\partial^2 P(\xi < x, \eta < y)}{\partial x \partial y} = f(x) \cdot g(y)$.

Examples

E1. Flip a coin twice. Let ξ be the number of heads, and let η be the difference between the number of head and tails. Now, $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$. $\xi((H,H)) = 2$, $\xi((T,T)) = 0$, $\xi((H,T)) = 1$, $\xi((T,H)) = 1$. Therefore, $\xi \sim \begin{pmatrix} 0 & 1 & 2 \\ 0.25 & 0.5 & 0.25 \end{pmatrix}$. Moreover, $\eta((H,H)) = 2 = \eta((T,T))$, and $\eta((H,T)) = 0 = \eta((T,H))$.

$\eta \sim \begin{pmatrix} 0 & 2 \\ 0.5 & 0.5 \end{pmatrix}$. $P(\xi = 0 \cap \eta = 0) = P(\emptyset) = 0 \neq P(\xi = 0) \cdot P(\eta = 0) = 0.125$. consequently ξ and η are not independent.

E2. Choose one point Q from the circle with radius 1 by geometric probability. Put the circle into the Cartesian frame and let the centre be the point O(0,0). Let ξ be the distance of the point Q from the centre O(0,0) of the circle, and η be the angle of the vector $\vec{OQ}$. Now, $0 \le \xi \le 1$, $0 \le \eta \le 2\pi$. $P(\xi < x) = \frac{x^2 \cdot \pi}{\pi} = x^2$, if $0 \le x \le 1$.

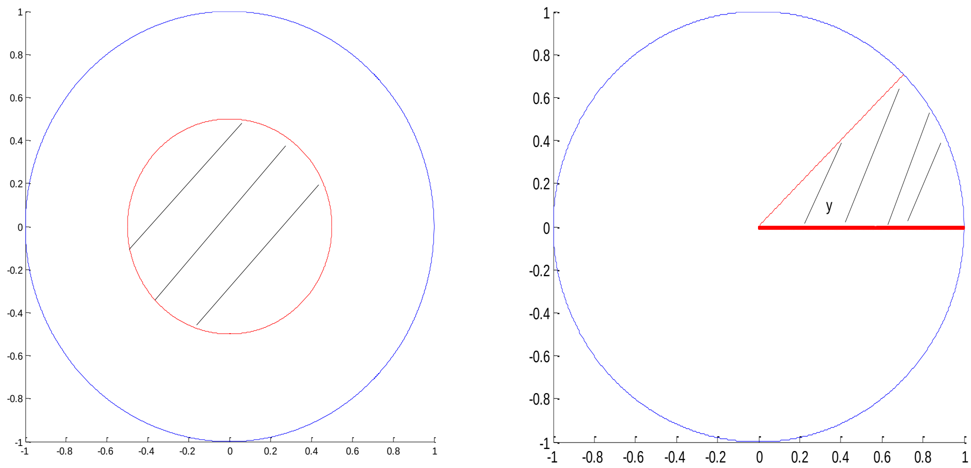


Figure d.13. Appropriate points for $\{\xi < x\}$ and for $\{\eta < y\}$

$$P(\eta < y) = \frac{\frac{y \cdot \pi}{2\pi}}{\pi} = \frac{y}{2\pi}, \ 0 \le y \le 2\pi.$$

Furthermore, $P(\xi < x \cap \eta < y) = \frac{x^2 \cdot \pi \cdot \frac{y}{2\pi}}{\pi} = \frac{\frac{x^2 \cdot y}{2}}{\pi}$, if $0 \le x \le 1$, $0 \le y \le 2\pi$.

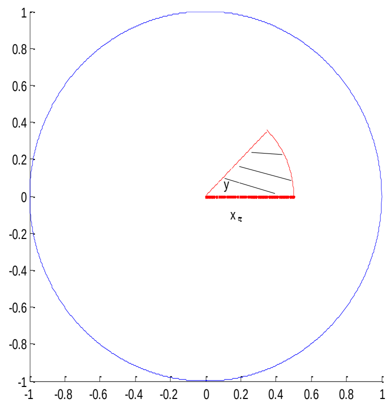


Figure d.14. Appropriate points for $\{\xi < x\} \cap \{\eta < y\}$

These together imply that $P(\xi < x \cap \eta < y) = P(\xi < x) \cdot P(\eta < y)$, if $0 \le x \le 1$, $0 \le y \le 2\pi$.

For the values outside [0,1]x[0,2π] one can check the equality easily, consequently ξ and η are independent.

e. Numerical characteristics of random variables

The aim of this chapter

In the previous chapter random variables were characterized by functions, such as the cumulative distribution function or the probability density function. This chapter aims to get the reader acquainted with the numerical characteristics of random variables. These numbers contain less information than cumulative distribution functions do but they are easier to be interpreted. We introduce the expectation, dispersion, mode and median. Beside the definitions, main properties are also presented.

Preliminary knowledge

Random variables, computing series and integrals. Improper integral.

Content

e.1. Expectation.

e.2. Dispersion and variance.

e.3. Mode.

e.4. Median.

e.1. Expectation

The cumulative distribution function of a random variable contains all the information about the random variable but it is not easy to know and handle it. This information can be condensed more or less into some numbers. Although we lose information during this concentration, these numbers carry important information about the random variable, consequently they are worth dealing with.

First of all we present a motivational example. Let us imagine the following gamble: we throw a die once and we gain the square of the result (dots on the surface). How much money is worth paying for a gamble, if after many rounds we would like get more money than we have paid. About some values one can easily decide: for example 1 is worth paying but 40 is not. Other values, for example 13, are not obvious. Let us follow a heuristic train of thought. Let the price of a round be denoted by x, and let the number of rounds be n. Now, the frequency of “one”, “two”, “three”, “four”, “five”, “six” are k_1, k_2,…, k_6, respectively. The money we get together equals

$1^2 \cdot k_1 + 2^2 \cdot k_2 + 3^2 \cdot k_3 + 4^2 \cdot k_4 + 5^2 \cdot k_5 + 6^2 \cdot k_6$.

The money we pay for gambling is $n \cdot x$. We get more money than we pay if the following inequality holds: $n \cdot x < 1^2 \cdot k_1 + 2^2 \cdot k_2 + 3^2 \cdot k_3 + 4^2 \cdot k_4 + 5^2 \cdot k_5 + 6^2 \cdot k_6$. Dividing by n, we get $x < 1^2 \cdot \frac{k_1}{n} + 2^2 \cdot \frac{k_2}{n} + 3^2 \cdot \frac{k_3}{n} + 4^2 \cdot \frac{k_4}{n} + 5^2 \cdot \frac{k_5}{n} + 6^2 \cdot \frac{k_6}{n}$. $\frac{k_i}{n}$ i=1,2,…,6 expresses the relative frequency of the result "i". If they were about the probabilities of the result "i", then $\frac{k_i}{n} \approx \frac{1}{6}$ and the right hand side of the previous inequality equals $1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6} = \frac{91}{6} = 15\frac{1}{6}$. Therefore, if $x < 15\frac{1}{6}$ then the money we get after many rounds is more than what we paid, in the opposite case it is less than we what paid. The heuristic is $\frac{k_i}{n} \approx \frac{1}{6}$, which has not been proved yet in this booklet, but it will be done in the chapter h.

How can the value $\frac{91}{6}$ be interpreted? If we define the random variable ξ as the gain during one round, then ξ is a discrete random variable with the following distribution: $\xi \sim \begin{pmatrix} 1 & 4 & 9 & 16 & 25 & 36 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}$. The right hand side of the inequality for x is the weighted sum of the possible values of ξ and the weights are the probabilities belonging to the possible values. This motivates the following definition:

Definition Let ξ be a discrete random variable with finitely many possible values. Let the distribution of ξ be $\xi \sim \begin{pmatrix} x_1 & x_2 & . & . & x_n \\ p_1 & p_2 & . & . & p_n \end{pmatrix}$. Then the **expectation** of ξ is defined as

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

Let ξ be a discrete random variable with infinitely many possible values. Let $\xi \sim \begin{pmatrix} x_1 & x_2 & . & . & . \\ p_1 & p_2 & . & . & . \end{pmatrix}$. Then the **expectation** of ξ is defined as $E(\xi) = \sum_{i=1}^{\infty} x_i \cdot p_i$, if the series is absolutely convergent, that is $\sum_{i=1}^{\infty} |x_i| \cdot p_i < \infty$.

Let ξ be a continuous random variable with probability density function f. Then the **expectation** of ξ is defined as $E(\xi) = \int_{-\infty}^{\infty} x \cdot f(x)dx$ supposing that the improper integral is absolutely convergent, that is $\int_{-\infty}^{\infty} |x| \cdot f(x)dx < \infty$.

Remarks

- If the discrete random variable has only finitely many values, then its expectation exists.
- If $\sum_{i=1}^{\infty} |x_i| \cdot p_i = \infty$ or $\int_{-\infty}^{\infty} |x| \cdot f(x)dx = \infty$, then, by definition, the expectation does not exist.
- $\sum_{i=1}^{\infty} |x_i| \cdot p_i < \infty$ implies $\sum_{i=1}^{\infty} x_i \cdot p_i < \infty$. Similarly, $\int_{-\infty}^{\infty} |x| \cdot f(x)dx < \infty$ implies $\int_{-\infty}^{\infty} x \cdot f(x)dx < \infty$.
- The expectation of a random variable is finite, if it exists.
- $\sum_{i=1}^{\infty} x_i \cdot p_i$ can be convergent even if it is not absolutely convergent. But in this case if the series is rearranged, the sum can change. Therefore the value of the sum may depend

on the order of the members, which is undesirable. This can not happen, if the series is absolutely convergent.

- The expectation may not be an element of the set of possible values. For example, if the random variable takes values -1 and 1 with probability 0.5 and 0.5, then expectation is $-1\cdot 0.5+1\cdot 0.5=0$.

Examples

E1. We gamble. We roll a die twice and we gain the difference of the results. Compute the expectation of the gain.

Let ξ be the difference of the results. The distribution of ξ can be given as follows:

$$\xi \sim \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ \frac{6}{36} & \frac{10}{36} & \frac{8}{36} & \frac{6}{36} & \frac{4}{36} & \frac{2}{36} \end{pmatrix}.$$

Now $E(\xi)=\sum_{i=1}^{6} x_i p_i = 0\cdot\frac{6}{36}+1\cdot\frac{10}{36}+2\cdot\frac{8}{36}+3\cdot\frac{6}{36}+4\cdot\frac{4}{36}+5\cdot\frac{2}{36}=1.94$.

E2. We gamble. We roll a die n times and we gain the maximum of the results. Compute the expectation of the gain.

Let ξ be the maximum of the results. The distribution of ξ can be given as follows: Possible values are 1,2,3,4,5,6. and $P(\xi=1)=\left(\frac{1}{6}\right)^n$, $P(\xi=2)=\left(\frac{2}{6}\right)^n-\left(\frac{1}{6}\right)^n$,

$P(\xi=3)=\left(\frac{3}{6}\right)^n-\left(\frac{2}{6}\right)^n$,

$P(\xi=4)=\left(\frac{4}{6}\right)^n-\left(\frac{3}{6}\right)^n$, $P(\xi=5)=\left(\frac{5}{6}\right)^n-\left(\frac{4}{6}\right)^n$, $P(\xi=6)=1-\left(\frac{5}{6}\right)^n$

$$E(\xi)=\sum_{i=1}^{6} x_i\cdot p_i = 1\cdot\left(\frac{1}{6}\right)^n+2\cdot\left(\left(\frac{2}{6}\right)^n-\left(\frac{1}{6}\right)^n\right)+3\cdot\left(\left(\frac{3}{6}\right)^n-\left(\frac{2}{6}\right)^n\right)+4\cdot\left(\left(\frac{4}{6}\right)^n-\left(\frac{3}{6}\right)^n\right)+$$

$$5\cdot\left(\left(\frac{5}{6}\right)^n-\left(\frac{4}{6}\right)^n\right)+6\cdot\left(1-\left(\frac{5}{6}\right)^n\right)=6-\left(\left(\frac{5}{6}\right)^n+\left(\frac{4}{6}\right)^n+\left(\frac{3}{6}\right)^n+\left(\frac{2}{6}\right)^n+\left(\frac{1}{6}\right)^n\right).$$

E3. Flip a coin repeatedly. The gain is 10^n if a head appears first at the nth game. Compute the expectation of the gain.

Let ξ be the gain. Now the possible values of ξ are 10, 100, 1000,…. and $P(\xi = 10^n) = \left(\frac{1}{2}\right)^n$. $E(\xi) = \sum_{i=1}^{\infty} x_i \cdot p_i = \sum_{i=1}^{\infty} 10^i \cdot \left(\frac{1}{2}\right)^i = \sum_{i=1}^{\infty} 5^i = \infty$, consequently the expectation does not exist.

E4. Flip a coin repeatedly. The gain is 10^n if a head appears first at the nth game supposing $n \le m$ and 10^m, if the we do not get a head until the m*th* game.(The bank is able to pay at most a given sum, which is a reasonable assumption.) Compute the expectation of the gain.

Let ξ be the gain. Now the possible values of ξ are 10, 100, 1000,…., 10^m. $P(\xi = 10^n) = \left(\frac{1}{2}\right)^n$, if $n < m$ and $P(\xi = 10^m) = \left(\frac{1}{2}\right)^{m-1}$.

$$E(\xi) = \sum_{i=1}^{m} x_i \cdot p_i = \sum_{i=1}^{m-1} 10^i \cdot \left(\frac{1}{2}\right)^i + 10^m \cdot \left(\frac{1}{2}\right)^{m-1} = \sum_{i=1}^{m-1} 5^i + 10 \cdot 5^{m-1} =$$

$5 \cdot \frac{5^{m-1} - 1}{4} + 10 \cdot 5^{m-1} = 11.25 \cdot 5^{m-1} - 1.25$, consequently the expectation exists.

E5. We compare the expectation of a random variable and the average of the results of many experiences. We make computer simulations, we generate random numbers in the interval $[0,1]$ by geometric probability. Let the random number be denoted by ξ. Let $\eta = [6 \cdot \xi] + 1$. Now the possible values of η are 1,2,3,4,5,6,7 and $P(\eta = 1) = P([6 \cdot \xi] = 0) = P(0 \le \xi < \frac{1}{6}) = \frac{1}{6}$, $P(\eta = 2) = P([6 \cdot \xi] = 1) = P(\frac{1}{6} \le \xi < \frac{2}{6}) = \frac{1}{6}$, …,

$P(\eta = 6) = P([6 \cdot \xi] = 5) = P(\frac{5}{6} \le \xi < 1) = \frac{1}{6}$, finally, $P(\eta = 7) = P([6\xi] = 6) = P(\xi = 1) = 0$.

Therefore, the distribution of η equals the distribution of the random variable which is equal to the number of dots on the surface of a fair die. If we take the square of this random variable, we get our motivating example presented at the beginning of this subsection.

Now repeating the process many times, and taking the average of the numbers 1,4,…,36 , we get the following results in Table e.1. Recall that the expectation of the gain equals 15.1667. The larger the number of simulations, the smaller the difference between the average and the expectation.

Numbers of simulations	100	1000	10000	100000	100000	10000000
Average	13.94	15.130	15.0723	15.1779	15.1702	15.1646

Difference	1.2267	0.0367	0.0944	0.0112	0.0035	0.0021

Table e.1. Averages and their differences from the expectation in case of simulation numbers $n = 100,...,10^7$

E6. Recall the example presented in E7. in subsection d.1. Compute the expectation of the distance between the chosen point and the centre of the circle.
Let ξ be the distance. The probability density function of ξ, as presented in subsection d.3. is the following: $f(x) = \begin{cases} \frac{2x}{R^2} & 0 \le x \le R \\ 0 & \text{otherwise} \end{cases}$.

Now $E(\xi) = \int_{-\infty}^{\infty} x \cdot f(x)dx = \int_0^R \frac{x \cdot 2x}{R^2} dx = \left[\frac{2x^3}{3R^2}\right]_0^R = \frac{2}{3} R$.

E7. Recall the example presented in E8. in subsection d.1. Compute the expectation of the distance between the chosen points.
The probability density function of ξ as presented in subsection d.3. is the following: $f(x) = \begin{cases} 2(1-x) & \text{if } x \le 0 \le 1 \\ 0 & \text{otherwise} \end{cases}$.

$E(\xi) = \int_{-\infty}^{\infty} x \cdot f(x)dx = \int_0^1 x \cdot 2(1-x)dx = \left[x^2 - \frac{2x^3}{3}\right]_0^1 = 1 - \frac{2}{3} = \frac{1}{3}$.

E8. Compute the approximate value of the above expectation. Generate two random numbers in $[0,1]$ by geometric probability, compute their difference and take the average of all differences. Repeating this process many times, we get the following results:

Numbers of simulations	100	1000	10000	100000	100000	10000000
Average	0.3507	0.3325	0.3323	0.3328	0.3331	0.3333
Difference	0.0174	0.0008	0.0010	0.0005	0.0002	0.00007

Table e.2. Differences of the approximate and the exact expectation in case of different numbers of simulations

E9. Choose two numbers in the interval $[0,1]$ independently by geometric probability. Let ξ be the sum of them. Now one can prove that the probability density function is

$$f(x)=\begin{cases} x & \text{if } 0\le x\le 1 \\ 2-x & \text{if } 1\le x\le 2 \\ 0 & \text{otherwise} \end{cases}.$$

Now

$$E(\xi)=\int_{-\infty}^{\infty} x\cdot f(x)dx=\int_0^1 x\cdot x\,dx+\int_0^1 x\cdot(2-x)\,dx=\left[\frac{x^3}{3}\right]_0^1+\left[x^2-\frac{x^3}{3}\right]_1^2=\frac{1}{3}+4-\frac{8}{3}-1+\frac{1}{3}=1$$

.

It is also possible to solve this problem by simulation. Generating two random numbers, summing them up and averaging the sums one can see the following:

Numbers of simulations	100	1000	10000	100000	100000	10000000
Average	0.9761	1.0026	0.99995	1.0001	0.9999	1
Difference	0.0239	0.0026	0.00005	0.0001	0.0001	0.00001

Table e.3. Differences of the approximate and the exact expectation in case of different numbers of simulations

Properties of the expectation

Now we list some important properties of the expectation. When it is easy to do, we give some explanation, as well. Let ξ and η be random variables, and suppose that $E(\xi)$ and $E(\eta)$ exist. Let $a,b,c\in R$.

1. If ξ and η are identically distributed, then $E(\xi)=E(\eta)$.

If ξ and η are discrete, then they have common possible values and $P(\xi=x_i)=P(\eta=x_i)$, consequently the weighted sums are equal, as well. If ξ and η are continuous random variables, then they have common probability density function, consequently the improper integrals are equal.

2. If $\xi=c$ or $P(\xi=c)=1$, then $E(\xi)=c\cdot 1=1$.
3. If $0\le\xi$, then $0\le E(\xi)$ holds.

If ξ is discrete, then all the possible values of ξ are nonnegative, therefore so is the weighted sum, as well. If ξ is a continuous random variable, then $0\le\xi$ implies that its probability density function is zero for negative x values. Consequently, $E(\xi)=\int_{-\infty}^{\infty} x\cdot f(x)dx=\int_0^{\infty} x\cdot f(x)dx$, which must be nonnegative.

4. $E(\xi+\eta)=E(\xi)+E(\eta)$.

The additive property is difficult to prove using elementary analysis, but it follows from the general properties of integral.

5. $E(a\cdot\xi+b)=a\cdot E(\xi)+b$.

If ξ is discrete, then the possible values of $a\cdot\xi+b$ are „a” times the possible values of ξ plus b, therefore so is their weighted sum. If ξ is continuous, then $F_{a\xi+b}(x)=P(a\xi+b<x)=P(\xi<\frac{x-b}{a})=F(\frac{x-b}{a})$ supposing $0<a$. Taking the derivative $f_{a\xi+b}(x)=\frac{1}{a}\cdot f(\frac{x-b}{a})$,

$$E(a\xi+b)=\int_{-\infty}^{\infty}x\cdot f_{a\xi+b}(x)dx=\int_{-\infty}^{\infty}x\cdot f_{a\xi+b}(x)dx=\int_{-\infty}^{\infty}x\cdot\frac{1}{a}\cdot f(\frac{x-b}{a})dx=\int_{-\infty}^{\infty}(ay+b)\cdot f(y)dy=$$

$a\int_{-\infty}^{\infty}y\cdot f(y)dy+b\cdot\int_{-\infty}^{\infty}f(y)dy=aE(\xi)+b$. A similar argument can be given for negative value of “a” as well. If $a=0$ holds, then $E(a\cdot\xi+b)=b=aE(\xi)+b$.

6. If $a\le\xi\le b$, then $a\le E(\xi)\le b$.

As $a\le\xi$, $0\le\xi-a$ holds, therefore $0\le E(\xi-a)=E(\xi)-a$, which implies $a\le E(\xi)$. A similar argument can be given for the upper bound.

7. If $\xi\le\eta$, that is $\xi(\omega)\le\eta(\omega)$ for any $\omega\in\Omega$, then $E(\xi)\le E(\eta)$.

Consider that $\xi\le\eta$ implies $0\le\eta-\xi$, consequently $0\le E(\eta-\xi)=E(\eta)-E(\xi)\Rightarrow E(\xi)\le E(\eta)$. We point out that it is not enough that the possible values of ξ are less than the possible values of η, respectively. For example, $\xi\sim\begin{pmatrix}1 & 4\\ 0.1 & 0.9\end{pmatrix}$, $\eta\sim\begin{pmatrix}2 & 5\\ 0.8 & 0.2\end{pmatrix}$. Now $E(\xi)=1\cdot 0.1+4\cdot 0.9=3.7$, $E(\eta)=2\cdot 0.8+5\cdot 0.2=3.6$, that is $E(\eta)<E(\xi)$.

8. Let ξ_i i=1,2,…, n be independent identically distributed random variables with expectation $E(\xi_i)=m$. Then $E(\sum_{i=1}^{n}\xi_i)=n\cdot m$.

This is the straightforward consequence of the above properties, namely $E(\sum_{i=1}^{n}\xi_i)=\sum_{i=1}^{n}E(\xi_i)=n\cdot m$.

9. Let ξ_i i=1,2,…,n be independent identically distributed random variables with expectation $E(\xi_i) = m$. Then $E\left(\frac{\sum_{i=1}^{n}\xi_i}{n}\right) = m$.

It follows from $\frac{\sum_{i=1}^{n}\xi_i}{n} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$.

10. If ξ and η are independent random variables and $E(\xi\cdot\eta)$ exists, then $E(\xi\cdot\eta) = E(\xi)\cdot E(\eta)$. The proof of this statement is outside this booklet.

11. If ξ is a discrete random variable with distribution $\xi \sim \begin{pmatrix} x_1 & x_2 & . & . & . \\ p_1 & p_2 & . & . & . \end{pmatrix}$, $g: I \to R$ is a function for which $\{x_1, x_2, ...\} \subset I$, furthermore $\sum_{i=1}^{\infty}|g(x_i)|p_i < \infty$, then $E(g(\xi)) = \sum_{i=1}^{\infty} g(x_i)p_i$.

Now $g(\xi): \Omega \to R$ is a random variable and its possible values are $g(x_i)$, and $P(g(\xi)) = g(x_i)) = \sum_{j:g(x_j)=g(x_i)} p_j = q_i$. This implies the equality $E(g(\xi)) = \sum_{i=1}^{\infty} g(x_i)p_i$ .Especially, if $g(x) = x^2$, then $E(\xi^2) = \sum_{i=1}^{\infty} x_i^{\ 2}\cdot p_i$.

12. If ξ is a continuous random variable with probability density function f, $g: I \subset R$ for which $Im(\xi) \subset I$ and $\int_{-\infty}^{\infty}|g(x)|f(x)dx < \infty$, then $E(g(\xi) = \int_{-\infty}^{\infty} g(x)\cdot f(x)dx$. Especially, if $g(x) = x^2$, then $E(g(\xi)) = E(\xi^2) = \int_{-\infty}^{\infty} x^2 f(x)dx$.

<u>Examples</u>

E9. The latest property provides a possibility for computing integrals by computer simulation. Since the expectation is an integral, and the expectation is around the average of many values of the random variables, we can compute the average and it can be used for the approximation of the integral. For example, if we want compute the integral $\int_0^1 \sin x dx$, then it can be interpreted as an expectation. Namely, let ξ be a random variable with probability density function $f(x)=\begin{cases}1 & \text{if } 0\le x\le 1\\ 0 & \text{otherwise}\end{cases}$, and

$E(\sin\xi)=\int_{-\infty}^{\infty}\sin x\cdot f(x)dx=\int_0^1 \sin x dx$. If ξ is a random number chosen by geometric probability, then $F(x)=P(\xi<x)=\begin{cases}0 & \text{if } x\le 0\\ x & \text{if } 0<x\le 1\\ 1 & \text{if } 1<x\end{cases}$, and $f(x)=F'(x)=\begin{cases}1 & \text{if } 0\le x\le 1\\ 0 & \text{otherwise}\end{cases}$.

Consequently, generating random numbers, substituting them into the function sinx, and taking their average we get an approximate value for the integral. This is a simple algorithm. We point out that the statement that expectation is about the average of many experiments has not been proved yet in this booklet. It will be done using the law of large numbers in chapter h. The following Table e.4. presents some results:

Numbers of simulations	100	1000	10000	100000	100000	10000000
Averages	0.4643	0.4548	0.4588	0.4586	0.4596	0.4597
Difference	0.0046	0.0049	0.001	0.0011	0.0011	0.00002

Table e.4. Differences of the approximate and the exact value of the integral in case of different numbers of simulations

E10. The additive property of the expectation helps us to simplify computations. For example, consider the following example. Roll a die twice. Let η be the sum of the results. Now, one can check that

$$\eta\sim\begin{pmatrix}2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12\\ \frac{1}{36} & \frac{2}{36} & \frac{3}{36} & \frac{4}{36} & \frac{5}{36} & \frac{6}{36} & \frac{5}{36} & \frac{4}{36} & \frac{3}{36} & \frac{2}{36} & \frac{1}{36}\end{pmatrix},$$

and

$$E(\eta)=\sum_{i=1}^{11}x_ip_i=2\cdot\frac{1}{36}+3\cdot\frac{2}{36}+4\cdot\frac{3}{36}+5\cdot\frac{4}{36}+6\cdot\frac{5}{36}+7\cdot\frac{6}{36}+8\cdot\frac{5}{36}++9\cdot\frac{4}{36}+$$

$10\cdot\frac{3}{36}+11\cdot\frac{2}{36}+12\cdot\frac{1}{36}=7$. Another method is the following: $\eta=\xi_1+\xi_2$ where ξ_1 is the result of the first throw and ξ_2 is the result of the second throw. Now ξ_1 and ξ_2 are identically distributed random variables and $\xi_1\sim\begin{pmatrix}1 & 2 & 3 & 4 & 5 & 6\\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6}\end{pmatrix}$.

$$E(\xi_1)=\sum_{i=1}^{6}x_ip_i=1\cdot\frac{1}{6}+2\cdot\frac{1}{6}+3\cdot\frac{1}{6}+4\cdot\frac{1}{6}+5\cdot\frac{1}{6}+6\cdot\frac{1}{6}=3.5=E(\xi_2),$$

consequently, $\eta = \xi_1 + \xi_2 = E(\xi_1 + \xi_2) = 2 \cdot 3.5 = 7$.

e.2. Dispersion and variance

The expectation is a kind of average. It is easy to construct two different random variables which have the same expectation. For example, $\xi_1 \sim \begin{pmatrix} -1 & 1 \\ 0.5 & 0.5 \end{pmatrix}$ and $\xi_2 \sim \begin{pmatrix} -2 & -1 & 0 & 1 & 2 \\ 0.1 & 0.3 & 0.225 & 0.25 & 0.125 \end{pmatrix}$. They both have the same expectation, namely zero.

The measure of the average distance from the expectation can be an important information, as well. As $E(\xi - E(\xi)) = E(\xi) - E(E(\xi)) = 0$, therefore it is not appropriate to characterize the distance from the average. The reason is that the negative and positive differences balance out. This phenomenon disappears if we take $E(|\xi - E(\xi)|)$. But if we use the square instead of absolute value, the signs disappear again and, on the top of all the small differences become smaller, large differences become larger. Squaring punishes large differences but does not punish small ones. Consequently, it is worth investigating $E\left((\xi - E(\xi))^2\right)$ instead of $E(|\xi - E(\xi)|)$, if it exists.

<u>Definition</u> Let ξ be a random variable with expectation $E(\xi)$. The **variance** of ξ is defined as $E^2((\xi - E(\xi))^2) = D^2(\xi)$, if it exists.

<u>Definition</u> Let ξ be a random variable with expectation $E(\xi)$. The **dispersion** of ξ is defined as $D(\xi) = \sqrt{D^2(\xi)}$, if $D^2(\xi)$ exists.

<u>Remarks</u>

- As $0 \le (\xi - E(\xi))^2$, so is its expectation. Therefore its square root is well-defined.
- By definition, dispersion of a random variable is a nonnegative number. It is the square root of the averaged square differences.
- It is easy to construct such a random variable which has expectation but does not have dispersion. We will do it in this subsection, after proving the rule for its calculation.
- Another name of the dispersion is the standard deviation.

<u>Theorem</u> If ξ is a random variable with expectation $E(\xi)$, and $E(\xi^2)$ exists, then $D^2(\xi) = E(\xi^2) - (E(\xi))^2$.

<u>Proof</u> Applying the properties of expectation

$$D^2(\xi)=E((\xi-E(\xi))^2)=E(\xi^2-2\xi E(\xi)+(E(\xi))^2)=E(\xi^2)-2E(\xi)E(\xi)+E((E(\xi))^2)$$
$$=E(\xi^2)-2(E(\xi))^2+(E(\xi))^2=E(\xi^2)-(E(\xi))^2.$$

Remarks

- $D^2(\xi)=\sum_{i=1}^{\infty}x_i^2p_i-\left(\sum_{i=1}^{\infty}x_ip_i\right)^2$ if ξ is discrete, and

$D^2(\xi)=\int_{-\infty}^{\infty}x^2f(x)dx-\left(\int_{-\infty}^{\infty}xf(x)dx\right)^2$ if ξ is a continuous random variable.

- If ξ and η are identically distributed random variables, then $D(\xi)=D(\eta)$
- In case of a discrete random variable with infinitely many possible values, $E(\xi^2)=\sum_{i=1}^{\infty}x_i^2p_i$. If the series is not (absolutely) convergent, then $\sum_{i=1}^{\infty}x_i^2p_i=\infty$.
- In case of a continuous random variable with probability density function f, $E(\xi^2)=\int_{-\infty}^{\infty}x^2f(x)dx$. If the improper integral is not (absolutely) convergent, then $\int_{-\infty}^{\infty}x^2f(x)dx=\infty$.
- If $E(\xi^2)$ does not exist, neither does $D^2(\xi)$. $\sum_{i=1}^{\infty}x_i^2p_i=\infty$ implies $\sum_{i=1}^{\infty}(x_i-c)^2p_i=\infty$ and $\int_{-\infty}^{\infty}x^2f(x)dx=\infty$ implies $\int_{-\infty}^{\infty}(x-c)^2f(x)dx=\infty$ for any value of c.
- It can be proved that if $E(\xi^2)$ exists, then so does $E(\xi)$.
- Let ξ be a continuous random variable with probability density function $f(x)=\begin{cases}0 & \text{if } x\le 1\\ \frac{2}{x^3} & \text{if } 1<x\end{cases}$. Then the expectation of the random variable is

$$E(\xi)=\int_{-\infty}^{\infty}x\cdot f(x)dx=\int_{1}^{\infty}x\cdot\frac{2}{x^3}dx=2\cdot\int_{1}^{\infty}\frac{1}{x^2}dx=2\cdot\left[-\frac{1}{x}\right]_1^{\infty}=2\left(\left(\lim_{x\to\infty}-\frac{1}{x}\right)-(-1)\right)=2(0+1)=2$$

.

. Consequently, $E(\xi)$ exists, but $D(\xi)$ does not.

Example

E1. Roll a die twice. Let ξ be the maximum of the results. Compute the dispersion of ξ. First we have to determine the distribution of ξ. It is easy to see that $\xi \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{36} & \frac{3}{36} & \frac{5}{36} & \frac{7}{36} & \frac{9}{36} & \frac{11}{36} \end{pmatrix}$.

$$E(\xi) = 1\cdot\frac{1}{36} + 2\cdot\frac{3}{36} + 3\cdot\frac{5}{36} + 4\cdot\frac{7}{36} + 5\cdot\frac{9}{36} + 6\cdot\frac{11}{36} = 4.472 .$$

$$E(\xi^2) = 1^2\cdot\frac{1}{36} + 2^2\cdot\frac{3}{36} + 3^2\cdot\frac{5}{36} + 4^2\cdot\frac{7}{36} + 5^2\cdot\frac{9}{36} + 6^2\cdot\frac{11}{36} = \frac{791}{36} = 21.972 .$$

Applying the above theorem,

$$D(\xi) = \sqrt{E(\xi^2) - (E(\xi))^2} = \sqrt{21.972 - 4.472^2} = \sqrt{1.973} = 1.405$$

E2. Choose two numbers from the interval [0,1] independently by geometrical probability. Let ξ be the difference of the two numbers. Compute the dispersion of ξ. Recall from E8. in subsection d3 that the probability density function of ξ is $f(x) = \begin{cases} 2-2x & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$.

We need $E(\xi)$ and $E(\xi^2)$.

$$E(\xi) = \int_{-\infty}^{\infty} x\cdot f(x)dx = \int_0^1 x\cdot(2-2x)dx = \left[x^2 - \frac{2x^3}{3}\right]_0^1 = \frac{1}{3}.$$

$$E(\xi^2) = \int_{-\infty}^{\infty} x^2\cdot f(x)dx = \int_0^1 x^2\cdot(2-2x)dx = \left[\frac{2x^3}{3} - \frac{2x^4}{4}\right]_0^1 = \frac{1}{6}.$$

$$D(\xi) = \sqrt{\frac{1}{6} - \left(\frac{1}{3}\right)^2} = \sqrt{\frac{1}{18}} = 0.236 .$$

Now we list the most important properties of the variance and the dispersion. As variance and dispersion are in close connection, we deal with their properties together.

<u>Properties of the variance and dispersion</u>

Let ξ and η be random variables with dispersion $D(\xi)$ and $D(\eta)$, respectively, and let a, b, c be constant values.

1. If $\xi = c$, then $D^2(\xi) = D(\xi) = 0$. It is obvious, as $E(\xi) = c$, $(\xi - E(\xi))^2 = 0$, and $E(0) = 0$.

2. If $D(\xi)=0$, then $P(\xi=c)=1$. Consequently, dispersion being zero characterizes the constant random variable.

3. $D^2(a\xi+b)=a^2D^2(\xi)$ and $D(a\xi+b)=|a|D(\xi)$.

Consider, that $E(a\xi+b)=aE(\xi)+b$,

$E((a\xi+b-(aE(\xi)+b))^2)=E(a^2(\xi-E(\xi))^2)=a^2E((\xi-E(\xi))^2)=a^2D^2(\xi)$.

$D(a\xi+b)=\sqrt{a^2D^2(\xi)}=|a|D(\xi)$.

4. Let ξ be a random variable with dispersion $D(\xi)$. Now the value of $g(c)=E((\xi-c)^2)$ is minimal if $c=E(\xi)$. Consider that $g(c)=E((\xi-c)^2)=c^2-2cE(\xi)+(E(\xi))^2$ is a quadratic polynomial of c. Moreover, the coefficient of c^2 is positive, therefore the function has a minimum value. If we take its derivative, $g'(c)=-2c+2E(\xi)$. It is zero if and only if $c=E(\xi)$ which implies our statement.

5. If ξ is a random variable for which $a\le\xi\le b$ holds, then its dispersion exists. If it is denoted by $D(\xi)$, then $D(\xi)\le\frac{b-a}{2}$.

If ξ is discrete, then $E(\xi^2)=\sum_{i=1}^{\infty}x_i^2p_i\le\max\{a^2,b^2\}\cdot\sum_{i=1}^{\infty}p_i=\max\{a^2,b^2\}<\infty$. If ξ is continuous, then $E(\xi^2)=\int_{-\infty}^{\infty}x^2f(x)dx\le\max\{a^2,b^2\}\cdot\int_{-\infty}^{\infty}f(x)dx=\max\{a^2,b^2\}<\infty$, which proves the existence of dispersion. Applying the properties of expectation we can write for any value of $x\in R$,

$D^2(\xi)=E((\xi-E(\xi))^2)\le E((\xi-x)^2)\le(a-x)^2P(\xi<x)+(b-x)^2P(\xi\ge x)=$

$=(a-x)^2-(a-x)^2P(\xi\ge x)+(b-x)^2P(\xi\ge x)=(a-x)^2+(b-a)(b+a-2x)P(\xi\ge x)$.

Substituting $x=\frac{a+b}{2}$, $b+a-2x=0$, $(a-x)^2=\left(\frac{b-a}{2}\right)^2$. We get that $D^2(\xi)\le\frac{(b-a)^2}{4}$, therefore $D(\xi)\le\frac{b-a}{2}$. We note that in case of $\xi\sim\begin{pmatrix}a & b\\ 0.5 & 0.5\end{pmatrix}$, $D(\xi)=\frac{b-a}{2}$. Consequently, the inequality can not be sharpened.

6. If ξ and η are independent, then $D^2(\xi+\eta)=D^2(\xi)+D^2(\eta)$ and $D(\xi+\eta)=\sqrt{D^2(\xi)+D^2(\eta)}$.

$$D^2(\xi+\eta)=E((\xi+\eta-E(\xi)-E(\eta))^2)=E((\xi-E(\xi))^2)+E((\eta-E(\eta))^2)+2\cdot E((\xi-E(\xi)\cdot(\eta-E(\eta)))$$

Recall that if ξ and η are independent then $E(\xi\cdot\eta)=E(\xi)E(\eta)$, therefore $E((\xi-E(\xi)\cdot(\eta-E(\eta))=E((\xi-E(\xi))\cdot E((\eta-E(\eta))=0$.

We would like to emphasize that the dispersions can not be summed, only the variances. Namely, it is important to remember, that $D(\xi+\eta)\neq D(\xi)+D(\eta)$. This fact has very important consequences when taking the average of random variables.

7. Let ξ_i i=1, 2, …, n be independent identically distributed random variables with dispersion $D(\xi_i)=\sigma$. Then $D^2(\sum_{i=1}^{n}\xi_i)=n\cdot\sigma^2$ and $D(\sum_{i=1}^{n}\xi_i)=\sqrt{n}\cdot\sigma$. This is the straightforward consequence of the above properties, namely $D^2(\sum_{i=1}^{n}\xi_i)=\sum_{i=1}^{n}D^2(\xi_i)=n\cdot\sigma^2$.

8. Let ξ_i i=1,2,,…,n be independent identically distributed random variables with dispersion $D(\xi_i)=\sigma$. Then $D^2(\frac{\sum_{i=1}^{n}\xi_i}{n})=\frac{\sigma^2}{n}$ and $D(\frac{\sum_{i=1}^{n}\xi_i}{n})=\frac{\sigma}{\sqrt{n}}$. This is again the straightforward consequence of properties 3 and 7.

e.3. Mode

Expectation is the weighted average of the possible values and it may not be the element of the set of possible values. A very simple example is the random variable taking values 0 and 1 with probabilities 0.5. In that case the distribution of ξ is given by $\xi\sim\begin{pmatrix}0 & 1\\ 0.5 & 0.5\end{pmatrix}$, $E(\xi)=0\cdot 0.5+1\cdot 0.5=0.5$, and 0.5 is not among the possible values of ξ. Mode is in the set of the possible values and the most probable value among them.

Definition Let ξ be a discrete random variable with distribution $\xi\sim\begin{pmatrix}x_1 & x_2 & . & . & x_n & . & .\\ p_1 & p_2 & . & . & p_n & . & .\end{pmatrix}$. The **mode** of ξ is x_k, if $p_i\le p_k$, $i=1,2,3,...$.

Definition Let ξ be a continuous random variable with probability density function $f(x)$. The **mode** of ξ is x if f has its local maximum at x, and the maximum value is not zero.

Remark

- The mode of a discrete random variable exists. If it has finitely many values then the maximum of a finite set exists. If it has infinitely many values, then only 0 may have infinitely many probabilities in its neighbourhood. The remaining part of the probabilities is a finite set, it must have a maximum value, and the index belonging to it defines the mode.

- The mode of a discrete random variable may not be unique. For example, consider $\xi \sim \begin{pmatrix} 0 & 1 \\ 0.5 & 0.5 \end{pmatrix}$. Now both possible values have equal likelihood.

- The mode of a continuous random variable is a more complicated case, as the probability density functions may be changed at any point and the distribution of the random variable does not change. Consequently we usually only deal with the mode of such continuous random variables which have a continuous probability function on finitely many subintervals. We consider the maximum of these functions in the inner parts of the subintervals, and they are the modes. Consequently, mode of a continuous random variable may not exist, see for example the following probability density function: $f(x) = \begin{cases} e^{-x} & 0 < x \\ 0 \text{ if } & x \le 0 \end{cases}$. It has its maximum value at zero, at the endpoint of the interval $[0, \infty)$ and no other maximum value exists.

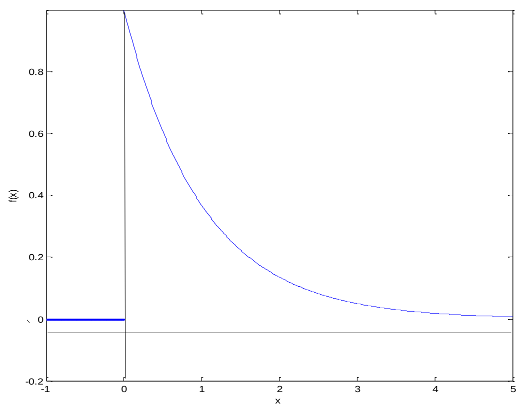


Figure e.1. Probability density function without a local maximum

- The mode of a continuous random variable may be unique, see for example

$$f(x) = \begin{cases} \dfrac{e^{-x} + x^3 e^{-x}}{7} & \text{if } 0 < x \\ 0 \text{ if } & x < 0 \end{cases}.$$

The graph of this probability density function can be seen in Fig.e.2.

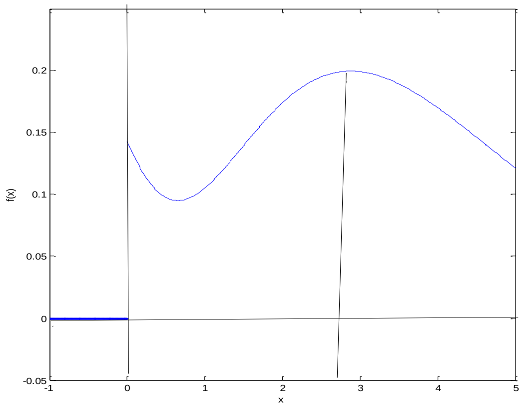


Figure e.2. Probability density function with a unique local maximum

The maximum can be determined by taking the derivative of $f(x)$ and finding where the derivative equals zero. Namely,

$$f'(x)=\begin{cases}(1/7)\cdot(-e^{-x}+3x^2e^{-x}-x^3e^{-x}) & \text{if } 0<x\\ 0 & \text{if } x\le 0\end{cases}.$$

$f'(x)=0$ implies $-e^{-x}+3x^2e^{-x}-x^3e^{-x}=0$ which means that $-1+3x^2-x^3=0$ .It is satisfied at x=2.8794 and x=0.6527. At x=0.6527 the function takes its minimum, at x=2. 8794 the function takes its maximum. Consequently, the mode is 2.8794.

- The mode of a continuous random variable may not be unique. If the probability density function of the random variable is $f(x)=\frac{1}{2\sqrt{2\pi}}(e^{\frac{-x^2}{2}}+e^{\frac{-(x-5)^2}{2}})$, it has two maximum values, one of them is about zero, the other one is about 5. Consequently, two modes exist.

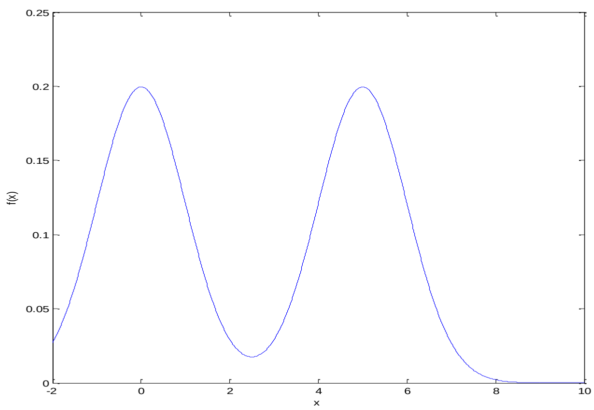


Figure e.3. Probability density function with two local maximums

e.4. Median

The mode is the most likely value of the random variable, the median is the middle one. Namely, the random variable takes values with equal chance under and below the median. More precisely, the probability of taking values at most the median and at least the median, both are at least 0.5.

Definition ξ is a random variable. The **median** of ξ is the value y, if $0.5 \le P(\xi \le y)$ and $0.5 \le P(y \le \xi)$.

Remark

- If ξ is a continuous random variable with cumulative distribution function $F(x)$, then the median of ξ is the value y for which $F(y) = 0.5$ holds. The inequality $0.5 \le P(y \le \xi) = 1 - F(y)$ implies $F(y) \le 0.5$. Taking into account that ξ is a continuous random variable, $P(\xi \le y) = P(\xi < y) = 0.5$, therefore $0.5 \le P(\xi \le y) = F(y)$. Consequently, $F(y) = 0.5$. As the function F is continuous, and it tends to 0 if x tends to $-\infty$ and it tends to 1 if x tends to infinity, the median of a continuous random variable exists, but may not be unique.
- Let ξ be a discrete random variable. The median of ξ is the value y for which $F(y) \le 0.5$ and $0.5. \le F(y+)$ $0.5 \le P(y \le \xi) = 1 - F(y)$ implies $F(y) \le 0.5$, and $0.5 \le P(\xi \le y) = \lim_{a \to y+} F(a) = F(y+)$ is the second inequality.

Examples

E1. Consider a random variable with cumulative distribution function

$$F(x) = \begin{cases} 0 \quad \text{if } x \le 0 \\ 1 - (1-x)^2 \quad \text{if } 0 < x \le 1 \\ 1 \quad \text{if } 1 < x \end{cases}.$$

Determine the median of the random variable.

We have to find the cross point of $F(x)$ and $y = 0.5$. As the function takes the value 0.5 when its argument is in $[0,1]$, we have to solve the equation $1 - (1-x)^2$ =0.5. It implies the equality $2x - x^2 = 0.5,$ therefore $x_1 = 0.293,$ and $x_2 = 1.707$. This last number is not in the interval $[0,1]$, consequently the median is 0.293. As a checking, $F(0.293) = 1 - (1 - 0.293)^2 = 0.5001$.

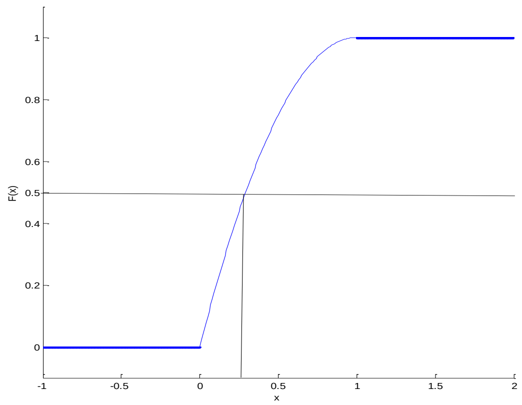


Figure e.5. The cross point of the cumulative distribution function and the line $y = 0.5$

E2. Let ξ be a discrete random variable with distribution $\xi \sim \begin{pmatrix} 0 & 1 & 5 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$.

Determine the median of ξ.

$$F(x) = \begin{cases} 0 & \text{ha } x \le 0 \\ \frac{1}{3} & \text{ha } 0 < x \le 1 \\ \frac{2}{3} & \text{ha } 1 < x \le 5 \\ 1 & \text{ha } 5 < x \end{cases}.$$

Now $F(x) \ne 0.5$. $P(\xi \le 1) = \frac{2}{3}$, $P(1 \le \xi) = \frac{2}{3}$, both of them are greater than 0.5. No other value of x satisfies this property. Consequently the unique median is 1.

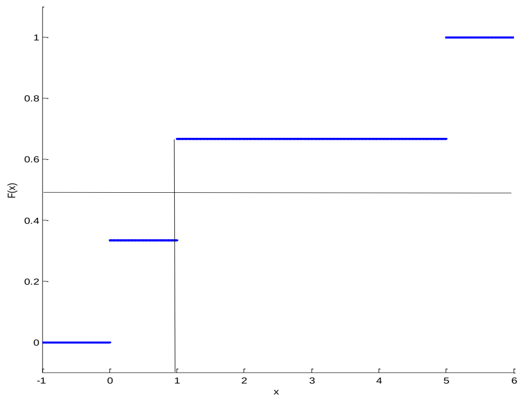


Figure e.6. The cumulative distribution function of the random variable ξ and the line $y = 0.5$

Median equals the argument when the cumulative distribution function jumps the level 0.5.

E3. Let ξ be a discrete random variable with distribution $\xi \sim \begin{pmatrix} 2 & 5 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$.

Determine the median of ξ.

Now $F(x) = \begin{cases} 0 & \text{if } x \le 2 \\ \frac{1}{2} & \text{if } 2 < x \le 5 \\ 1 & \text{if } 5 < x \end{cases}$, and $F(x)$ takes value 0.5 in the interval $(2,5]$. $P(\xi \le 2) = 0.5$, $P(\xi \ge 2) = 0.5$, consequently $x = 2$ is a median. Moreover, $P(\xi \le x) = P(\xi \ge x) = 0.5$ holds for any value of $[2,5)$. Therefore, they are all medians. Usually the middle of the interval (actually 3.5) is used for the value of the median.

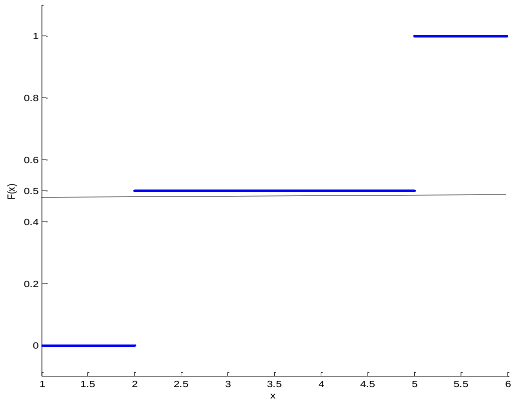


Figure e.7. The cumulative distribution function of the random variable ξ and the line $y = 0.5$

f. Frequently used discrete distributions

The aim of this chapter

In the previous chapters we have got acquainted with the concept of random variables. Now we investigate some frequently used types. We compute their numerical characteristics, and study their main properties as well. We highlight their relationships.

Preliminary knowledge

Random variables and their numerical characteristics. Computing numerical series and integrals. Sampling.

Content

f.1. Characteristically distributed random variables.

f.2. Uniformly distributed discrete random variables.

f.3. Binomially distributed random variables.

f.4. Hypergeometrically distributed random variables.

f.5. Poisson distributed random variables.

f.6. Geometrically distributed random variables.

f.1. Characteristically distributed random variables

First we deal with a very simple random variable. It is usually used as a tool in solving problems. Let Ω, $\mathcal{A}$, and P be given.

Definition The random variable ξ is called a **characteristically distributed random variable** with parameter $0 \le p \le 1$, if it takes only two values, namely 0 and 1, furthermore $P(\xi = 1) = p$ and $P(\xi = 0) = 1 - p$. Briefly written, $\xi \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$.

Example

E1. Let $A \in \mathcal{A}$, $P(A) = p$. Let us define $\xi : \Omega \to R$ as follows: $\xi(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$. Now ξ is a characteristically distributed random variable with parameter p.

In terms of event, ξ equals 1 if A occurs and ξ equals zero if it does not. Therefore ξ characterizes the occurrence of event A. It is frequently called an indicator random variable of the event A, and denoted by $\mathbf{1}_A$.

Numerical characteristics of characteristically distributed random variables:

Expectation

$E(\xi) = p$, which is a straightforward consequence of $E(\xi) = \sum_{i=1}^{2} x_i \cdot p_i = 1 \cdot p + 0 \cdot (1 - p) = p$.

Dispersion

$D(\xi) = \sqrt{p \cdot (1 - p)}$. As a proof, recall that $D^2(\xi) = E(\xi^2) - (E(\xi))^2$.

$E(\xi^2) = \sum_{i=1}^{2} x_i^2 \cdot p_i = 1^2 \cdot p + 0^2 \cdot (1 - p) = p$, consequently, $D^2(\xi) = p - p^2 = p(1 - p)$. This implies the formula $D(\xi) = \sqrt{p(1 - p)}$.

Mode

There are two possible values, namely 0 and 1. The most likely of them is 1, if $0.5 < p$, 0, if $p < 0.5$ and both of them, if $p = 0.5$.

Median

If $p < 0.5$, then $0.5 \le P(\xi \le 0) = 1 - p$ and $0.5 \le P(0 \le \xi) = 1$. Consequently, the median equals 0.

If $0.5 < p$, then $0.5 \le P(\xi \le 1) = 1$ and $0.5 \le P(1 \le \xi) = p$. Consequently, the median equals 1.

If $p = 0.5$, then $P(\xi \le x) = 0.5$ and $P(\xi \ge x) = 0.5$ for any value of $(0,1)$. Moreover, $P(\xi \le 0) = 0.5$, $P(\xi \ge 0) = 1$, and $P(\xi \le 1) = 1$ and $P(\xi \ge 1) = 0.5$. This means that any point of $[0,1]$ is a median.

Theorem If A and B are independent events, then $\mathbf{1_A}$ and $\mathbf{1_B}$ are independent random variables.

Proof $P(\mathbf{1_A}=1\cap\mathbf{1_B}=1)=P(A\cap B)=P(A)\cdot P(B)=P(\mathbf{1_A}=1)\cdot P(\mathbf{1_B}=1)$.

$P(\mathbf{1_A}=1\cap\mathbf{1_B}=0)=P(A\cap\overline{B})=P(A)\cdot P(\overline{B})=P(\mathbf{1_A}=1)\cdot P(\mathbf{1_B}=0)$.

$P(\mathbf{1_A}=0\cap\mathbf{1_B}=1)=P(\overline{A}\cap B)=P(\overline{A})\cdot P(B)=P(\mathbf{1_A}=0)\cdot P(\mathbf{1_B}=1)$.

$P(\mathbf{1_A}=0\cap\mathbf{1_B}=0)=P(\overline{A}\cap\overline{B})=P(\overline{A})\cdot P(B)=P(\mathbf{1_A}=0)\cdot P(\mathbf{1_B}=0)$.

f.2. Uniformly distributed discrete random variables

The second type of discrete random variables applied frequently is a uniformly distributed random variable. In this subsection we deal with discrete ones.

Definition The discrete random variable ξ is called a uniformly **distributed random variable,** if it takes finitely many values, and the probabilities belonging to the possible values are equal. Shortly written, $\xi\sim\begin{pmatrix} x_1 & x_2 & . & . & x_n \\ p_1 & p_2 & . & . & p_n \end{pmatrix}$, $p_i=p_j, i=1,2,...,n, \quad j=1,2,...,n.$

Remarks

- As $1=\sum_{i=1}^{n}p_i=np_1$, $p_1=p_2=...=p_n=\frac{1}{n}$. $\xi=\begin{pmatrix} x_1 & x_2 & . & . & x_n \\ \frac{1}{n} & \frac{1}{n} & . & . & \frac{1}{n} \end{pmatrix}$.
- There is no uniformly distributed discrete random variable if the set of possible values contains infinitely many elements. This is the straightforward consequence of the condition $1=\sum_{i=1}^{\infty}p_i$. With notation $P(\xi=x_i)=p$, if $p=0$ then $\sum_{i=1}^{\infty}0=0$, if $0<p$, $\sum_{i=1}^{\infty}p=\infty$.

Numerical characteristics of uniformly distributed random variables:

Expectation

$$E(\xi)=\sum_{i=1}^{n}x_i\frac{1}{n}=\overline{x}.$$

Dispersion

$D(\xi)=\sqrt{\frac{\sum_{i=1}^{n}x_i^2}{n}-\left(\frac{\sum_{i=1}^{n}x_i}{n}\right)^2}$, which can be computed by substituting into the formula concerning the dispersion.

Mode

All of the possible values have the same chance, all of them are modes.

Median

$x_{\frac{n+1}{2}}$ if n is odd, and $\frac{x_{\frac{n}{2}}+x_{\frac{n}{2}+1}}{2}$ if n is even.

Example

E1. Throw a die, and let ξ be the square of the result. Actually, $\xi \sim \begin{pmatrix} 1 & 4 & 9 & 16 & 25 & 36 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}$. As all possible values have the same chance, ξ is a uniformly distributed random variable. Note that there is no requirement for the possible values.

f.3. Binomially distributed random variables

After the above simple distributions we consider a more complicated one.

Definition The random variable ξ is called a **binomially distributed random variable** with parameters $2 \le n$ and $0 < p < 1$, if its possible values are $0,1,2,...,n$ and $P(\xi = k) = \binom{n}{k} \cdot p^k (1-p)^{n-k}$, $k = 0,1,2,...,n$.

Remark

- It is obvious that $0 \le P(\xi = k) = \binom{n}{k} \cdot p^k (1-p)^{n-k}$. Furthermore, the binomial theorem implies that $\sum_{k=0}^{n} P(\xi = k) = \sum_{k=0}^{n} \binom{n}{k} \cdot p^k (1-p)^{n-k} = 1$. Recalling that $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$, and substituting $a = p$ and $b = 1-p$, we get $a + b = p + 1 - p = 1$.

Theorem If ξ_i $i = 1,2,...,n$ are independent characteristically distributed random variables with parameter $0 < p < 1$, then $\eta = \sum_{i=1}^{n} \xi_i$ is a binomially distributed random variable with parameters n and p.

Proof Recall that $\xi_i \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$. Their sum can take any integer from 0 to n.

$$P(\sum_{i=1}^{n} \xi_i = 0) = P(\xi_1 = 0 \cap \xi_2 = 0 \cap ... \cap \xi_n = 0) = P(\xi_1 = 0) \cdot P(\xi_2 = 0) \cdot ... \cdot P(\xi_n = 0) = (1-p)^n$$

$$P(\sum_{i=1}^{n}\xi_i=1)=n\cdot P(\xi_1=1\cap\xi_2=0\cap...\cap\xi_n=0)=P(\xi_1=1)\cdot P(\xi_2=0)\cdot...\cdot P(\xi_n=0)=n\cdot p\cdot(1-p)^{n-1}$$

.

The factor n is included because the event A can occur at any experiment, not only at the first one.

$$P(\xi_1=1\cap\xi_2=1\cap...\cap\xi_k=1\cap\xi_{k+1}=0\cap...\cap\xi_n=0)=$$

$$P(\xi_1=1)\cdot P(\xi_2=1)\cdot...\cdot P(\xi_k=1)\cdot P(\xi_{k+1}=0)\cdot...\cdot P(\xi_n=0)=p^k\cdot(1-p)^{n-k}$$

If the event A occurs k times, then the indices of experiments when A occurs can be chosen in $\binom{n}{k}$ ways, consequently, $P(\sum_{i=1}^{n}\xi_i=k)=\binom{n}{k}\cdot p^k\cdot(1-p)^{n-k}$.

<u>Theorem</u> Repeat a trial n times, independently of each other. Let A be an event with probability $P(A)=p,\ 0<p<1$. Let ξ be the number of times the event A occurs during the n independent experiments. Then ξ is a binomially distributed random variable with parameters n and p.

Proof:

Let $1_A{}^i=\begin{cases}1 \text{ if A occurs at the ith experiment}\\ 0 \text{ if A does not occur at the ith experiment}\end{cases}$.

Considering that the experiments are independent, so are $\mathbf{1_A}^i$, i=1,2,…,n.

As $\xi=\sum_{i=1}^{n}\mathbf{1_A}^i$, ξ is the sum of n independent indicator random variables, consequently, ξ is binomially distributed random variable.

<u>Examples</u>

E1. Throw a fair die n times. Let ξ be the number of "6"-s. Then ξ is a binomially distributed random variable with parameters n and $p=\frac{1}{6}$.

E2. Flip a coin n times. Let ξ be the number of heads. Then ξ is binomially distributed random variable with parameters n and $p=\frac{1}{2}$.

E3. Throw a fair die n times. Let ξ be the number of even numbers. Then ξ is a binomially distributed random variable with parameters n and $p=\frac{1}{2}$. We note that the random variable being in this example is identically distributed with the random variable presented in E2.

E4. Draw 10 cards with replacement from a pack of French cards. Let ξ be the number of diamonds among the picked cards. Then ξ is a binomially distributed random variable with parameters $n=10$, $p=\frac{13}{52}$.

E5. Draw 10 cards with replacement from a pack of French cards. Let ξ be the number of aces among the picked cards. Then ξ is a binomially distributed random variable with parameters $n=10$, $p=\frac{4}{52}$.

E6. There are N balls in an urn, M of them are red, N-M are white. Pick n with replacement from them. Let ξ be the number of red balls among the chosen ones. ξ is a binomially distributed random variable with parameters n and $p=\frac{M}{N}$. ($2\le N$, $1\le M$, $1\le N-M$, $2\le n$)

Numerical characteristics of binomially distributed random variables

Expectation

$E(\xi)=np$, which is a straightforward consequence of $E(\xi)=E(\sum_{i=1}^{n}\mathbf{1}_A^i)=\sum_{i=1}^{n}E(\mathbf{1}_A^i)=\sum_{i=1}^{n}p=np$.

Dispersion

$D(\xi)=\sqrt{np\cdot(1-p)}$.

As an explanation consider that, as 1_A^i $(i=1,2,...,n)$ are independent,

$D^2(\xi)=D^2(\sum_{i=1}^{n}\mathbf{1}_i)=nD^2(1_i)=n\cdot p$. This implies $D(\xi)=\sqrt{np\cdot(1-p)}$.

Mode

If $(n+1)p$ is integer, then there are two modes, namely $(n+1)\cdot p$ and $(n+1)p-1$.

If $(n+1)p$ is not integer, then there is a unique mode, namely $[(n+1)\cdot p]$.

As an explanation, investigate the ratio of probability of consecutive possible values.

$$\frac{P(\xi=k)}{P(\xi=k-1)}=\frac{\binom{n}{k}p^k\cdot(1-p)^{n-k}}{\binom{n}{k-1}p^{k-1}\cdot(1-p)^{n-(k-1)}}=\frac{\frac{n!}{k!\cdot(n-k)!}}{\frac{n!}{(k-1)!\cdot(n-k+1)!}}\cdot\frac{p}{1-p}=\frac{n-k+1}{k}\cdot\frac{p}{1-p},$$

$k=1,2,...,n$.

$1<\frac{P(\xi=k)}{P(\xi=k-1)}$ implies that $P(\xi=k-1)<P(\xi=k)$, that is the probabilities are growing.

$\frac{P(\xi=k)}{P(\xi=k-1)}<1$ implies that $P(\xi=k)<P(\xi=k-1)$, that is the probabilities are decreasing.

$\frac{P(\xi=k)}{P(\xi=k-1)}=1$, then $P(\xi=k)=P(\xi=k-1)$.

$1<\frac{n-k+1}{k}\cdot\frac{p}{1-p}$ holds, if only if $k<(n+1)p$. $\frac{n-k+1}{k}\cdot\frac{p}{1-p}<1$ holds, if and only if $(n+1)\cdot p<k$, and $\frac{n-k+1}{k}\cdot\frac{p}{1-p}=1$ holds if and only if $k=(n+1)\cdot p$. This is satisfied only in the case if $(n+1)p$ is an integer. Therefore, if $(n+1)p$ is not integer, then, up to $k=[(n+1)p]$, the probabilities are growing, after that the probabilities are decreasing.

Consequently, the most probable value is $[(n+1)p]$. If $(n+1)p$ is an integer, then $P(\xi=k)=P(\xi=k-1)$, consequently there are two modes, namely $(n+1)p$ and $(n+1)p-1$.

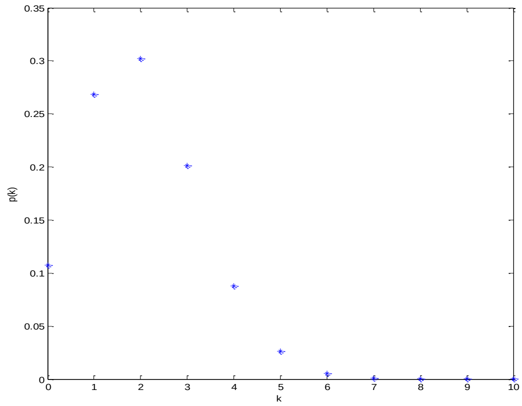


Figure f.1. Probabilities of possible values of a binomially distributed random variable with parameters $n=10$ and $p=0.2$

Without proof we can state the following theorem:

Theorem
If ξ_1 is a binomially distributed random variable with parameters n_1 and p, ξ_2 is a binomially distributed random variable with parameters n_2 and p, furthermore they are independent, then $\xi_1+\xi_2$ is also binomially distributed with parameters n_1+n_2 and p.

As an illustration, if ξ_1 is the number of "six"-es if we throw a fair die repeatedly n_1 times, ξ_2 is the number of "six" -es if we throw a fair die n_2 times, then $\xi_1+\xi_2$ is the number of "six" -es if we throw a fair die n_1+n_2 times, which is also binomially distributed random variable.

Theorem
If ξ_n is a sequence of binomially distributed random variables with parameters n and q_n, furthermore $n\cdot q_n=\lambda$, and k is a fixed value, then $P(\xi_n=k)=\binom{n}{k}(q_n)^k(1-q_n)^{n-k}\to\frac{\lambda^k}{k!}e^{-\lambda}$, if $n\to\infty$.

Proof
Substitute $q_n=\frac{\lambda}{n}$,

$$P(\xi_n=k)=\binom{n}{k}(q_n)^k(1-q_n)^{n-k}=\frac{n!}{k!\cdot(n-k)!}\frac{\lambda^k}{n^k}\left(1-\frac{\lambda}{n}\right)^{n-k}=$$

$$\frac{n(n-1)(n-2)....(n-k+1)}{n^k}\left(1-\frac{\lambda}{n}\right)^{-k}\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n.$$

Separating the factors,

$\frac{n(n-1)(n-2)....(n-k+1)}{n^k}=\frac{n-1}{n}\cdot\frac{n-2}{n}\cdot...\cdot\frac{n-k+1}{n}\to 1$, if $n\to\infty$, as each factor tends to 1, and k is fixed.

Similarly, $\left(1-\frac{\lambda}{n}\right)^{-k}\to 1$, if $n\to\infty$.

Since $\left(1+\frac{x}{n}\right)^{n}\to e^{x}$ if $n\to\infty$, consequently, $\left(1-\frac{\lambda}{n}\right)^{n}\to e^{-\lambda}$, if $n\to\infty$.

Summarizing, $P(\xi_n=k)=\binom{n}{k}(q_n)^k(1-q_n)^{n-k}\to\frac{\lambda^k}{k!}e^{-\lambda}$ supposing $n\to\infty$.

Example

E7. There are 10 balls and 5 boxes. We put the balls into the boxes, one after the other. We suppose that all the balls fall into any box with equal chance, independently of the other balls. Compute the probability that there is no ball in the first box. Compute the probability that there is one ball in the first box. Compute the probability that there are two balls in the first box. Compute the probability that there are at most two balls in the first box. Compute the probability that there are at least two balls in the first box. Compute the expectation of the balls in the first box. How many balls are in the first box most likely?

Let η be the number of balls in the first box. η is a binomially distributed random variable with parameters $n=10$ and $p=\frac{1}{5}$. We can give an explanation of this statement as follows: we repeat 10 times the experiment that we put a ball into a box. We consider whether the ball falls into the first box or not. If η is the number of balls in the first box, then η is the number of occurrences of the event $A=$"actual ball has fallen into the first box". It is easy to see that $P(A)=\frac{1}{5}$. Therefore, the possible values of η are 0,1,2,…,10, and the probabilities are $P(\eta=k)=\binom{10}{k}\left(\frac{1}{5}\right)^k\left(1-\frac{1}{5}\right)^{10-k}$, $k=0,1,2,...,10$.

If we calculate the probabilities, we get

$P(\eta=0)=\binom{10}{0}\left(\frac{1}{5}\right)^0\left(1-\frac{1}{5}\right)^{10}=0.1074$, $P(\eta=1)=\binom{10}{1}\left(\frac{1}{5}\right)^1\left(1-\frac{1}{5}\right)^{9}=0.2684$,

$P(\eta=2)=\binom{10}{2}\left(\frac{1}{5}\right)^2\left(1-\frac{1}{5}\right)^{8}=0.3020$, $P(\eta=3)=\binom{10}{3}\left(\frac{1}{5}\right)^3\left(1-\frac{1}{5}\right)^{7}=0.2013$,…,

$P(\eta=10)=\binom{10}{10}\left(\frac{1}{5}\right)^{10}\left(1-\frac{1}{5}\right)^{0}=10^{-7}$.

$$\eta\sim\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 0.1074 & 0.2884 & 0.3020 & 0.2013 & 0.08808 & 0.0264 & 0.0055 & 0.0007 & 10^{-4} & 10^{-5} & 10^{-7} \end{pmatrix}.$$

Returning to our questions, the probability that there is no ball in the first box is

$P(\eta=0)=\binom{10}{0}\left(\frac{1}{5}\right)^0\left(1-\frac{1}{5}\right)^{10}=0.1074$.

The probability that there is one ball in the first box equals $P(\eta=1)=\binom{10}{1}\left(\frac{1}{5}\right)^1\left(1-\frac{1}{5}\right)^9=0.2684$.

The probability that there are two balls in the first box is $P(\eta=2)=\binom{10}{2}\left(\frac{1}{5}\right)^2\left(1-\frac{1}{5}\right)^8=0.3020$.

The probability that there are at most two balls in the first box is $P(\eta\le 2)=P(\eta=0)+P(\eta=1)+P(\eta=2)=0.1074+0.2684+0.3020=0.6778$.

The probability that there are at least two balls in the first box can be computed as $P(2\le\eta)=P(\eta=2)+P(\eta=3)+...+P(\eta=10)=0.3020+0.2013+0.0088+...+10^{-7}=0.6242$, or in a simpler way,

$P(2\le\eta)=1-\left(P(\eta=0)+P(\eta=1)\right)=1-(0.1074+0.2684)=1-0.3758=0.6242$.

The expectation of the number of balls in the first box is $E(\eta)=10\cdot\frac{1}{5}=2$, which coincides with the mode, $[(n+1)p]=\left[11\cdot\frac{1}{5}\right]=2$.

E8. There are 10 balls and 5 boxes, 100 balls and 50 boxes, 1000 balls and 500 boxes, 10^n balls and $10^n/2$ boxes, $n=1,2,3,...$. Balls are put into the boxes and all the balls fall into any box with equal probability. Let us denote by $\xi_n=\eta_{10^n}$ the number of balls in the first box. Let k be fixed and investigate the probabilities $P(\xi_n=k)$. Compute the limit of these probabilities.

Referring to the previous example, ξ_n is a binomially distributed random variable with parameters 10^n and $q(n)=\frac{2}{10^n}$. The product of the two parameters equals $10^n\cdot\frac{2}{10^n}=2$ always, consequently, $P(\xi_n=k)\to\frac{2^k}{k!}e^{-2}$, if $n\to\infty$.

In details,

	ξ_1 $(10, \frac{1}{5})$	ξ_2 $(100, \frac{1}{50})$	ξ_3 $(1000, \frac{1}{500})$	ξ_3 $(10000, \frac{1}{5000})$	.	.	$\frac{2^k}{k!}e^{-\lambda}$
k=0	0.1074	0.1326	0.1351	0.1353	.	.	0.1353
k=1	0.2684	0.2706	0.2707	0.2707	.	.	0.2707
k=2	0.3020	0.2734	0.2709	0.2707	.	.	0.2707
k=3	0.2013	0.1823	0.1806	0.1805	.	.	0.1804

Table f.1. The probabilities of being k balls in a box in case of different parameters of total number of balls and boxes

We can see that the probabilities computed by the binomial formula are close to their limits, if the number of experiments is large (for example 10000). Consequently, the probabilities of binomially distributed random variables can be approximated by the formula $\frac{\lambda^k}{k!}e^{-\lambda}$, called Poisson probabilities.

f.4. Hypergometrically distributed random variables

After sampling with replacement, we deal with sampling without replacement, as well. The random variable which handles the number of specified elements in the sample if the sampling has been performed without replacement is a hypergeometrically distributed random variable.

Definition The random variable ξ is called a **hypergeometrically distributed random variable** with parameters $2 \le N$, $1 \le S \le N-1$ and $1 \le n$, $n \le S$, $n \le N-S$ integers, if its possible values are $0,1,2,...,n$ and $P(\xi = k) = \frac{\binom{S}{k} \cdot \binom{N-S}{n-k}}{\binom{N}{n}}$, $k = 0,1,2,...,n$.

Example

E1. We have N products, S of them have a special property, $N-S$ have not. We choose n ones from them without replacement. Let ξ be the number of products with the special property in the sample. Then, the possible values of ξ are $0,1,2,3,...,n$, and the probabilities (referring to the subsection of classical probability) are

$$P(\xi = k) = \frac{\binom{S}{k} \cdot \binom{N-S}{n-k}}{\binom{N}{n}}.$$

Remarks

- The previous example shows that the sum of the probabilities $\frac{\binom{S}{k} \cdot \binom{N-S}{n-k}}{\binom{N}{n}}$ equals 1.

The events „there are k products with the special property in the sample" k=0,1,2,…n form a partition of the sample space, consequently the sum of their probabilities equals 1.

- Similarly to the binomially distributed random variable, actually, ξ can also be written as a sum of indicator random variables, but these random variables are not independent.

Numerical characteristics of hypergeometrically distributed random variables:

Expectation

$E(\xi) = n\frac{S}{N}$. This formula can be computed using the definition of expectation as follows:

$$E(\xi)=\sum_{k=0}^{n}k\cdot\frac{\binom{S}{k}\cdot\binom{N-S}{n-k}}{\binom{N}{n}}=$$

$$\sum_{k=0}^{n}k\cdot\frac{\frac{S(S-1)(S-2)...(S-k+1)}{k!}\cdot\frac{(N-S)\cdot(N-S-1)\cdot..\cdot(N-S-(n-k)+1)}{(n-k)!}}{\frac{N!}{n!(N-n)!}}=$$

$$\sum_{k=1}^{n}\frac{\frac{S(S-1)(S-2)...(S-k+1)}{(k-1)!}\cdot\frac{(N-S)\cdot(N-S-1)\cdot...\cdot(N-S-(n-k)+1)}{(n-k)!}}{\frac{N!}{n!(N-n)!}}=$$

$$\sum_{k=1}^{n}n\frac{S}{N}\frac{\binom{S-1}{k-1}\binom{N-1-(S-1)}{n-1-(k-1)}}{\binom{N-1}{n-1}}=\sum_{k-1=0}^{n-1}n\frac{S}{N}\frac{\binom{S-1}{k-1}\binom{N-1-(S-1)}{n-1-(k-1)}}{\binom{N-1}{n-1}}=$$

$$\sum_{j=0}^{n-1}n\frac{S}{N}\frac{\binom{S-1}{j}\binom{N-1-(S-1)}{n-1-j}}{\binom{N-1}{n-1}}$$

Taking into account that $\sum_{j=0}^{n-1}\frac{\binom{S-1}{j}\binom{N-1-(S-1)}{n-1-j}}{\binom{N-1}{n-1}}=1$, we get the closed form for the expectation presented.

Dispersion

$D(\xi)=\sqrt{n\frac{S}{N}\cdot(1-\frac{S}{N})\left(1-\frac{n-1}{N-1}\right)}$. We do not prove this formula, because it requires too much computation.

Mode

$\left[\frac{(S+1)(n+1)}{N+2}\right]$ if $\frac{(S+1)(n+1)}{N+2}$ is not an integer and there are two modes, namely $\frac{(S+1)(n+1)}{N+2}$ and $\frac{(S+1)(n+1)}{N+2}-1$ if $\frac{(S+1)(n+1)}{N+2}$ is an integer.

Similarly to the way applied to the binomially distributed random variable we investigate the ratio $\frac{P(\xi=k)}{P(\xi=k-1)}$. Writing it explicitly and simplifying we get

$$\frac{P(\xi=k)}{P(\xi=k-1)}=\frac{\dfrac{\binom{S}{k}\cdot\binom{N-S}{n-k}}{\binom{N}{n}}}{\dfrac{\binom{S}{k-1}\cdot\binom{N-S}{n-k+1}}{\binom{N}{n}}}=\frac{S-k+1}{k}\cdot\frac{n-k+1}{N-S-n+k}$$. In order to know for which indices the probabilities are growing and or decreasing we have solve to the inequalities $1<\frac{S-k+1}{k}\cdot\frac{n-k+1}{N-S-n+k}$, $\frac{S-k+1}{k}\cdot\frac{n-k+1}{N-S-n+k}<1$, $\frac{S-k+1}{k}\cdot\frac{N-S-n+k+1}{n-k+1}=1$.

After some computations we get that

$1<\frac{S-k+1}{k}\cdot\frac{n-k+1}{N-S-n+k}$ holds if and only if $k<\frac{(S+1)(n+1)}{N+2}$,

$\frac{S-k+1}{k}\cdot\frac{n-k+1}{N-S-n+k}<1$ holds if and only if $\frac{(S+1)(n+1)}{N+2}<k$

$1=\frac{S-k+1}{k}\cdot\frac{n-k+1}{N-S-n+k}$ holds if and only if $k=\frac{(S+1)(n+1)}{N+2}$. This equality can be satisfied if $\frac{(S+1)(n+1)}{N+2}$ is an integer. Consequently, the mode is unique and it equals $\left[\frac{(S+1)(n+1)}{N+2}\right]$, if $\frac{(S+1)(n+1)}{N+2}$ is not an integer and there are two modes, namely $\frac{(S+1)(n+1)}{N+2}$ and $\frac{(S+1)(n+1)}{N+2}-1$ if $\frac{(S+1)(n+1)}{N+2}$ is an integer.

<u>Theorem</u>

Let $N\to\infty$, $S\to\infty$, $\frac{S}{N}=p$, and let k ,n be fixed integer values.

Then $\frac{\binom{S}{k}\cdot\binom{N-S}{n-k}}{\binom{N}{n}}\to\binom{n}{k}p^{k}(1-p)^{n-k}$.

<u>Proof</u>

$$\frac{\binom{S}{k}\cdot\binom{N-S}{n-k}}{\binom{N}{n}}=\frac{S(S-1)...(S-k+1)}{k!}\cdot\frac{(N-S)(N-S-1)...(N-S-n+k+1)}{(n-k)!}\cdot\frac{n!}{N(N-1)...(N-n+1)}$$

.

The number of factors in the numerator is $k+n-k=n$ and so is in the denominator. Taking into account that $\frac{n!}{k!(n-k)!}=\binom{n}{k}$, and $\frac{S}{N}=p$, $\frac{S-1}{N-1}=\frac{\frac{S}{N}-\frac{1}{N}}{1-\frac{1}{N}}\to p$ if $N\to\infty$,

$$\frac{S-k+1}{N-k+1}=\frac{\frac{S}{N}-\frac{k}{N}+\frac{1}{N}}{1-\frac{k}{N}+\frac{1}{N}}\to p \text{ if } N\to\infty \text{, furthermore } \frac{N-S}{N-k}=\frac{1-\frac{S}{N}}{1-\frac{k}{N}}\to 1-p,$$

$$\frac{(N-S-n+k+1)}{N-n+1}=\frac{1-\frac{S}{N}-\frac{n}{N}-\frac{k}{N}+\frac{1}{N}}{1-\frac{n}{N}+\frac{1}{N}}\to 1-p \text{ if } N\to\infty .$$

The number of factors tending to p equals k, the number of multipliers tending to 1-p equals n-k, consequently $\frac{\binom{S}{k}\cdot\binom{N-S}{n-k}}{\binom{N}{n}}\to\binom{n}{k}p^k(1-p)^{n-k}$.

Remark

- The meaning of the previous theorem is the following: if the number of elements is large and we choose a sample of a few elements, then the probabilities of having k elements with a special property in the sample is approximately the same if we take the sample with or without replacement.

Example

E1. There are 100 products, 60 of them are of first quality, 40 of them are substandard. Choose 10 of them with/ without replacement. Let ξ be the number of substandard products in the sample if we take the sample with replacement. Let η be the number of substandard products in the sample if we take the sample without replacement. Determine the distribution, expectation, dispersion, and mode of both random variables.

ξ is a binomially distributed random variable with parameters $n=10$, $p=\frac{40}{100}$. This means, that the possible values of ξ are 0,1,2,3,…,10, and $P(\xi=k)=\binom{10}{k}0.4^k 0.6^{10-k}$. η is a hypergeometrically distributed random variable with parameters $N=100$, $S=40$, $n=10$. Therefore the possible values of η are 0,1,2,3,..,10 and $P(\eta=k)=\frac{\binom{40}{k}\cdot\binom{60}{10-k}}{\binom{100}{10}}$. To compare the probabilities in case of choosing with and without replacement we write them in the following Table f.2.

k	0	1	2	3	4	5	6	7	8	9	10
$P(\xi=k)$	0.006	0.040	0.121	0.215	0.251	0.201	0.111	0.042	0.010	0.001	0.0001
$P(\eta=k)$	0.004	0.0034	0.115	0.220	0.264	0.208	0.108	0.037	0.008	0.001	0.00004

Table f.2.Probabilities of the numbers of substandard products in the sample in case of sampling with and without replacement

It can be seen that there are very small differences between the appropriate probabilities, therefore it is almost the same if we take the sample with or without replacement.

$E(\xi) = 10 \cdot 0.4 = 4$, $E(\eta) = 10 \cdot \frac{40}{100} = 4.$

$D(\xi) = \sqrt{10 \cdot 0.4 \cdot 0.6} = 1.55$, $D(\eta) = \sqrt{10 \cdot \frac{40}{100} \cdot \frac{60}{100} \cdot \left(1 - \frac{9}{99}\right)} = 1.48$.

The mode of ξ and η are the same values, namely 4, as it can be seen in the Table f.1., or applying the formula $[(n+1) \cdot p] = [11 \cdot 0.4] = 4$, or $\left[\frac{(S+1)(n+1)}{N+2}\right] = \left[\frac{41 \cdot 11}{102}\right] = [4.42] = 4$, respectively.

E2. There are N balls in a box, S are red, N-S are white. Choose 10 from them without replacement. Compute the probability that there are 4 red balls in the sample if the total number of balls are $N_1 = 10$, $N_2 = 100$, $N_3 = 1000$, $N_4 = 10000$, $N_4 = 100000$, and $S_1 = 4$, $S_2 = 40$, $S_3 = 400$, $S_4 = 4000$, $S_5 = 40000$. Notice that $\frac{S_i}{N_i} = p = 0.4$ is constant.

N	10	100	1000	10000	100000	limit
$P(\eta_N = 4)$	1	0.26431	0.25209	0.25095	0.25084	0.25082

Table f.3. Probabilities of 4 red balls in the sample in case of different numbers of total balls

One can follow the convergence in Table f.3. very easily on the basis of the computed probabilities. We emphasize that both of the values n and k are fixed.

f.5. Poisson distributed random variables

After investigating sampling without replacement, we return to the limit of probabilities of binomially distributed random variables.

Definition The random variable ξ is called a **Poisson distributed random variable** with parameter $0 < \lambda$, if its possible values are $0,1,2,....$, and $P(\xi = k) = \frac{\lambda^k}{k!} e^{-\lambda}$, k=0,1,2,…

Remarks

- $0 < \frac{\lambda^k}{k!} e^{-\lambda}$ holds obviously, furthermore $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \cdot \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \cdot e^{\lambda} = 1$.
- The last theorem of subsection f.3. states that the limit of the distribution of binomially distributed random variables is a Poisson distribution.

Numerical characteristics of Poisson distributed random variables

Expectation

$E(\xi) = \lambda$. This formula can be proved as follows:

$$E(\xi)=\sum_{k=0}^{\infty}k\cdot\frac{\lambda^k}{k!}e^{-\lambda}=e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!}=\lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}=\lambda e^{-\lambda}\sum_{i=0}^{\infty}\frac{\lambda^i}{i!}=\lambda e^{-\lambda}e^{\lambda}=\lambda .$$

<u>Dispersion</u>

$D(\xi)=\sqrt{\lambda}$. Recall that $D^2(\xi)=E(\xi^2)-\left(E(\xi)\right)^2$.

$$E(\xi^2)=\sum_{k=0}^{\infty}k^2\cdot\frac{\lambda^k}{k!}e^{-\lambda}=\sum_{k=1}^{\infty}k^2\cdot\frac{\lambda^k}{k!}e^{-\lambda}=\sum_{k=1}^{\infty}k\cdot\frac{\lambda^k}{(k-1)!}e^{-\lambda}=e^{-\lambda}\cdot\lambda\sum_{k=1}^{\infty}k\cdot\frac{\lambda^{k-1}}{(k-1)!}=$$

$$e^{-\lambda}\cdot\lambda\left(\sum_{k=1}^{\infty}(k-1)\cdot\frac{\lambda^{k-1}}{(k-1)!}+\sum_{k=1}^{\infty}\cdot\frac{\lambda^{k-1}}{(k-1)!}\right)=e^{-\lambda}\cdot\lambda^2\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}+e^{-\lambda}\cdot\lambda\sum_{k=1}^{\infty}\cdot\frac{\lambda^{k-1}}{(k-1)!}=$$

$e^{-\lambda}\cdot\lambda^2\cdot e^{\lambda}+e^{-\lambda}\cdot\lambda\cdot e^{\lambda}=\lambda^2+\lambda$. Therefore $D^2(\xi)=E(\xi^2)-\left(E(\xi)\right)^2=\lambda^2+\lambda-\lambda^2=\lambda$.

Finally, $D(\xi)=\sqrt{D^2(\xi)}=\sqrt{\lambda}$.

<u>Mode</u>

There is a unique mode, namely $[\lambda]$ if λ is not an integer and there are two modes, namely λ and $\lambda-1$ if λ is an integer.

Similarly to the way applied in the previous subsections, we investigate the ratio $\frac{P(\xi=k)}{P(\xi=k-1)}$. Writing it explicitly and simplifying we get $\frac{\frac{\lambda^k}{k!}e^{-\lambda}}{\frac{\lambda^{k-1}}{(k-1)!}e^{-\lambda}}=\frac{\lambda}{k}$. The inequality $1<\frac{\lambda}{k}$, holds if and only if $k<\lambda$, the inequality $\frac{\lambda}{k}<1$, holds if and only if $\lambda<k$, and $1=\frac{\lambda}{k}$, holds if and only if $k=\lambda$. This can be achieved only in the case if λ is integer. Summarizing, for the values of k which are less than λ the probabilities are growing, for the values of k greater than λ the probabilities are decreasing, consequently the mode is $[\lambda]$. The same probability appears at $\lambda-1$, if λ is an integer.

<u>Examples</u>

E1. The number of faults in some material is supposed to be a Poisson distributed random variable. In a unit volume material there are 2.3 faults on average. Compute the probability that there are at most 3 faults in a unit volume material. How much volume contains at least 1 fault with probability 0.99?

Let ξ_1 be the number of faults in a unit volume of material. Now the possible values of ξ_1 are $0,1,2,...,k,...$ and $P(\xi_1=k)=\frac{\lambda^k}{k!}e^{-\lambda}$. The parameter λ equals the expectation, hence $P(\xi_1=k)=\frac{2.3^k}{k!}e^{-2.3}$. Now, $P(\xi_1\le 3)=P(\xi_1=0)+P(\xi_1=1)+P(\xi_1=2)+P(\xi_1=3)=$

$$\frac{2.3^0}{0!}e^{-2.3}+\frac{2.3^1}{1!}e^{-2.3}+\frac{2.3^2}{2!}e^{-2.3}+\frac{2.3^3}{3!}e^{-2.3}=0.799 .$$

Compute the probability that there are at least 3 faults in a unit volume material.
$P(\xi_1 \geq 3) = 1 - (P(\xi_1 = 0) + P(\xi_1 = 1) + P(\xi_1 = 2)) = 1 - (\frac{2.3^0}{0!}e^{-2.3} + \frac{2.3^1}{1!}e^{-2.3} + \frac{2.3^2}{2!}e^{-2.3}) = 0.404$.
How many faults are most likely in a unit volume material?
$\lambda = 2.3$ is not integer, consequently there is a unique mode, namely $[2.3] = 2$.
The probabilities are included in the following Table f.5. and can be seen in Fig.f.2.

k	0	1	2	3	4	5	6	7	8	9
$P(\xi_1 = k)$	0.100	0.230	0.203	0.117	0.0538	0.0206	0.0068	0.0019	0.0005	0.0001

Table f.5. Probabilities belonging to the possible values in case of Poisson distribution with parameter $\lambda = 2.3$

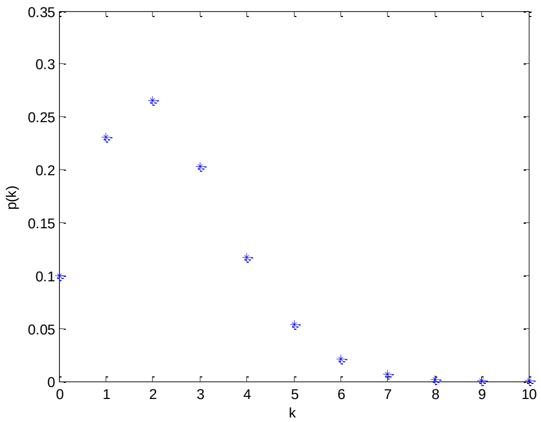


Figure f.2. Probabilities belonging to the possible values in case of Poisson distribution with parameter $\lambda = 2.3$

How many faults are most likely in a 10 unit volume material?
Let ξ_{10} be the number of faults in a 10 unit volume. ξ_{10} is also a Poisson distributed random variable with parameter $\lambda^* = 10 \cdot 2.3 = 23$. As λ^* is integer, two modes exist, namely $\lambda^* = 23$ and $\lambda^* - 1 = 22$. It is easy to see that
$P(\xi_{10} = 22) = \frac{(\lambda^*)^{22}}{22!}e^{-\lambda^*} = \frac{(23)^{22}}{22!}e^{-23} = \frac{(23)^{23}}{23!}e^{-23} = P(\xi_{10} = 23)$.
How much volume contains at least one fault with probability 0.99?
Let x denote the unknown volume and ξ_x the number of faults in the material of volume x. We want to know x if we know that $P(1 \leq \xi_x) = 0.99$. Taking into account that $P(1 \leq \xi_x) = 1 - P(\xi_x = 0)$, $P(1 \leq \xi_x) = 0.99$ implies $P(\xi_x = 0) = 0.01$. ξ_x is a Poisson distributed random variable with parameter $\lambda_x = x \cdot 2.3$, consequently $\frac{(x \cdot 2.3)^0}{0!}e^{-2.3x} = 0.01$. As $(x \cdot 2.3)^0 = 1$, $0! = 1$, we get $e^{-2.3x} = 0.01$. Taking the logarithm of both sides, we get $-2.3x = \ln 0.1$, therefore $x = \frac{\ln 0.01}{-2.3} = 2.003 \approx 2$.

E2. The number of viruses arriving to a computer is a Poisson distributed random variable. The probability that there is no file with viruses for 10 minutes equals 0.7. How many viruses arrive to the computer during 12 hours most likely?
Let ξ_{10} be the number of viruses arriving to our computer during a 10 minutes period. We do not know the parameter of ξ_{10}, but we know that $P(\xi_{10}=0)=0.7$. Since ξ_{10} is a Poisson distributed random variable with parameter λ, therefore $P(\xi_{10}=0)=\frac{\lambda^0}{0!}e^{-\lambda}=0.7$.
This implies $\lambda=-\ln 0.7=0.357$.
If ξ_{720} is the number of viruses arriving to the computer during 12 hours, ξ_{720} is also a Poisson distributed random variable with parameter $\lambda^*=12\cdot 6\cdot 0.357=25.68$, consequently there is a unique mode, $[25.68]=25$.

<u>Theorem</u> If ξ is a Poisson distributed random variable with parameter λ_1, η is a Poisson distributed random variable with parameter λ_2 furthermore they are independent, then $\xi+\eta$ is also a Poisson distributed random variable with parameter $\lambda_1+\lambda_2$.
<u>Proof</u>
As ξ is a Poisson distributed random variable with parameter λ_1, the possible values of ξ are $0,1,2,3,...$ and $P(\xi=i)=\frac{(\lambda_1)^i}{i!}e^{-\lambda_1}$. As η is a Poisson distributed random variable with parameter λ_2, the possible values of η are $0,1,2,3,...$ and $P(\eta=j)=\frac{(\lambda_2)^j}{j!}e^{-\lambda_2}$. It is obvious that the possible values of $\xi+\eta$ are $0,1,2,3,...$. We prove that $P(\xi+\eta=k)=\frac{(\lambda_1+\lambda_2)^k}{k!}e^{-(\lambda_1+\lambda_2)}$.
First, investigate $P(\xi+\eta=0)$.

$$P(\xi+\eta=0)=P(\xi=0\cap\eta=0)=P(\xi=0)\cdot P(\eta=0)=$$
$$\frac{(\lambda_1)^0}{0!}e^{-\lambda_1}\cdot\frac{(\lambda_2)^0}{0!}e^{-\lambda_2}=e^{-(\lambda_1+\lambda_2)}=\frac{(\lambda_1+\lambda_2)^0}{0!}e^{-(\lambda_1+\lambda_2)}.$$
Similarly,
$$P(\xi+\eta=1)=P(\xi=1\cap\eta=0)+P(\xi=0\cap\eta=1)=P(\xi=1)\cdot P(\eta=0)+P(\xi=0)\cdot P(\eta=1)=$$
$$\frac{(\lambda_1)^1}{1!}e^{-\lambda_1}\cdot\frac{(\lambda_2)^0}{0!}e^{-\lambda_2}+\frac{(\lambda_1)^0}{0!}e^{-\lambda_1}\cdot\frac{(\lambda_2)^1}{1!}e^{-\lambda_2}=\frac{(\lambda_1+\lambda_2)^1}{1!}e^{-(\lambda_1+\lambda_2)}$$
coinciding with the requirement.
In general,
$$P(\xi+\eta=k)=\sum_{i=0}^{k}P(\xi=i\cap\eta=k-i)=\sum_{i=0}^{k}P(\xi=i)\cdot P(\eta=k-i)=\sum_{i=0}^{k}\frac{(\lambda_1)^i}{i!}e^{-\lambda_1}\frac{(\lambda_2)^{k-i}}{(k-i)!}e^{-\lambda_2}=$$

$$e^{-(\lambda_1+\lambda_2)}\sum_{i=0}^{k}\frac{(\lambda_1)^i}{i!}\frac{(\lambda_2)^{k-i}}{(k-i)!}=\frac{e^{-(\lambda_1+\lambda_2)}}{k!}\sum_{i=0}^{k}\binom{k}{i}(\lambda_1)^i(\lambda_2)^{k-i}=\frac{e^{-(\lambda_1+\lambda_2)}}{k!}(\lambda_1+\lambda_2)^k.$$

E3. The number of people served in an office is a Poisson distributed random variable. There are two attendants in the office and the number of people served by the first one and the second one are independent random variables. The average number served by them during an hour is 3 and 2.5, respectively. Compute the probability that together they serve more than 4 people during an hour.

Let ξ_1 and ξ_2 be the numbers of people served by the attendants, respectively. ξ_1 is a Poisson distributed random variable with parameter $\lambda_1 = 3$, ξ_2 is a Poisson distributed random variable with parameter $\lambda_2 = 2.5$, and according to the assumption, they are independent. The total number of people served by them is $\xi_1 + \xi_2$. Applying the previous theorem, $\xi_1 + \xi_2$ is also a Poisson distributed random variable with parameter $\lambda = \lambda_1 + \lambda_2 = 5.5$ Consequently,

$$P(4 < \xi_1 + \xi_2) = 1 - \left(P(\xi_1 + \xi_2 = 0) + P(\xi_1 + \xi_2 = 1) + P(\xi_1 + \xi_2 = 2) + P(\xi_1 + \xi_2 = 3) + P(\xi_1 + \xi_2 = 4)\right) =$$

$$1 - \left(\frac{5.5^0}{0!}e^{-5.5} + \frac{5.5^1}{1!}e^{-5.5} + \frac{5.5^2}{2!}e^{-5.5} + \frac{5.5^3}{3!}e^{-5.5} + \frac{5.5^4}{4!}e^{-5.5}\right) = 1 - 0.358 = 0.642\ .$$

Given that they serve 5 people together, compute the probability that the first attendant serves 3 and the second one serves two clients.

The second question can be written as follows: $P(\xi_1 = 3 \cap \xi_2 = 2 \mid \xi_1 + \xi_2 = 5) =?$

Recall that the conditional probability is given by $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$. Consequently,

$$P\left((\xi_1 = 3 \cap \xi_2 = 2) \mid (\xi_1 + \xi_2 = 5)\right) = \frac{P\left((\xi_1 = 3 \cap \xi_2 = 2) \cap (\xi_1 + \xi_2 = 5)\right)}{P(\xi_1 + \xi_2 = 5)}.$$

The event $\{\xi_1 + \xi_2 = 5\}$ is the consequence of $\{\xi_1 = 3 \cap \xi_2 = 2\}$, therefore their intersection is the event $\{\xi_1 = 3 \cap \xi_2 = 2\}$. Now, taking into consideration the independence of the random variables ξ_1 and ξ_2 we get

$$P\left((\xi_1 = 3 \cap \xi_2 = 2) \mid (\xi_1 + \xi_2 = 5)\right) = \frac{P(\xi_1 = 3 \cap \xi_2 = 2)}{P(\xi_1 + \xi_2 = 5)} = \frac{P(\xi_1 = 3) \cdot P(\xi_2 = 2)}{P(\xi_1 + \xi_2 = 5)} = \frac{\frac{3^3}{3!}e^{-3} \cdot \frac{2.5^2}{2!}e^{-2.5}}{\frac{5.5^5}{5!}e^{-5.5}}$$

$$= \binom{5}{3}\left(\frac{3}{5.5}\right)^3\left(1 - \frac{3}{5.5}\right)^2 = 0.615\ .$$

f.6. Geometrically distributed random variables

At the end of this section we deal with geometrically distributed random variables.

In this case we perform independent experiments until a fixed event occurs. We finish the experiments when the event occurs first. Actually we do not know the number of experiments in advance.

Definition The random variable ξ is called a **geometrically distributed random variable** with parameter $0 < p < 1$, if its possible values are $1,2,3,\ldots,k,\ldots.$ and $P(\xi = k) = p(1-p)^{k-1}$, k=1,2,3,…

Remarks

- The above probabilities are really nonnegative, and their sum equals 1. It can be seen easily if we apply the formula concerning the sum of infinite geometric series, namely $\sum_{i=1}^{\infty} x^i = \frac{1}{1-x}$, if $|x|<1$ holds.

$$P(\xi = k) = \sum_{k=1}^{\infty} P(\xi = k) = \sum_{k=1}^{\infty} p(1-p)^{k-1} = p\sum_{k=1}^{\infty}(1-p)^{k-1} = p\sum_{k=0}^{\infty}(1-p)^{k} = p\cdot\frac{1}{1-(1-p)} = 1 .$$

- The quantities $p(1-p)^{k-1}$ form a geometric series, this is the reason of the name.
- Do not confuse this discrete random variable with the geometric probability presented in the first chapter.

Theorem We repeat an experiment until a fixed event A occurs, $0 < P(A) < 1$. Suppose that the experiments are independent. Let ξ be the number of necessary experiments. Then, ξ is a geometrically distributed random variable with parameter $p = P(A)$.

Proof Let A_i denote that the event A occurs at the ith experiment. Now, the values of ξ can be 1,2,3,… any positive integer. $\xi = 1$ means that the event A occurs at the first experiment, therefore $P(\xi = 1) = P(A_1) = p$. $\xi = 2$ means that the event A does not occur at the first experiment, but it does at the second experiment, that is $P(\xi = 2) = P(\overline{A_1} \cap A_2) = P(\overline{A_1})\cdot P(A_2) = (1-p)p$, which meets the requirements. In general, $\xi = k$ means, that the event A does not occur at the 1st, 2nd, …,(k-1)th experiments, but it occurs at the kth one. Hence $P(\overline{A_1} \cap \overline{A_2} \cap ... \cap \overline{A_{k-1}} \cap A_k) = P(\overline{A_1})P(\overline{A_2})P(\overline{A_3})...P(\overline{A_{k-1}})P(A_k) = (1-p)^{k-1}\cdot p$, which is the statement needed to be proved.

Numerical characteristics of geometrically distributed random variables

Expectation

$E(\xi) = \frac{1}{p}$. This formula can be proved as follows:

$E(\xi) = \sum_{i=1}^{\infty} x_i p_i = \sum_{k=1}^{\infty} k\cdot p(1-p)^{k-1} = p\sum_{k=1}^{\infty} k(1-p)^{k-1}$. $k(1-p)^{k-1}$ is similar to derivative. If we investigate the function $\sum_{k=1}^{\infty} kx^{k-1}$ for values $|x|<1$, then $\sum_{k=1}^{\infty} kx^{k-1} = \sum_{k=1}^{\infty}\left(x^k\right)' = \left(\sum_{k=1}^{\infty} x^k\right)' = \left(\frac{x}{1-x}\right)' = \frac{1}{(1-x)^2}$. Substituting $x = 1-p$, we get $\sum_{k=1}^{\infty} k(1-p)^{k-1} = \frac{1}{(1-(1-p))^2} = \frac{1}{p^2}$. This implies the formula

$$E(\xi) = p\sum_{k=1}^{\infty} k(1-p)^{k-1} = \frac{p}{p^2} = \frac{1}{p} .$$

Dispersion

$D(\xi)=\frac{\sqrt{1-p}}{p}$. We do not prove this formula. It can be proved similarly to the previous statement, but it requires more computation.

Mode

There is a unique mode, namely it is always 1. This is the straightforward consequence of the fact that the ratio of consecutive probabilities is $\frac{P(\xi=k)}{P(\xi=k-1)}=\frac{p(1-p)^{k-1}}{p(1-p)^{k-2}}=1-p<1$. This implies that the probabilities are decreasing, therefore the first one is the greatest.

Example

E1. We throw a die until we succeed in "six". Compute the probability that at most 6 throws are needed.

Let ξ be the number of necessary throws. ξ is a geometrically distributed random variable with parameter $\frac{1}{6}$. This means that the possible values of ξ are 1,2,3,... and $P(\xi=k)=\left(\frac{5}{6}\right)^{k-1}\cdot\left(\frac{1}{6}\right)$.

$$P(\xi\le 6)=\sum_{i=1}^{6}P(\xi=i)=\frac{1}{6}+\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)+\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^2+\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^3+\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^4+\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^5=\frac{1}{6}\frac{\left(\frac{5}{6}\right)^6-1}{\frac{5}{6}-1}=$$

$$1-\left(\frac{5}{6}\right)^6=0.665 .$$

In general, $P(\xi\le n)=\frac{1}{6}\frac{\left(\frac{5}{6}\right)^n-1}{\frac{5}{6}-1}=1-\left(\frac{5}{6}\right)^n$.

Compute the probability that more than 10 throws are needed.

According to the previous formula, $P(\xi>10)=\left(\frac{5}{6}\right)^{10}=0.1615$.

At most how many throws are needed with probability 0.9?

The question is to find the value of n for which $P(\xi\le n)=0.9$. As $P(\xi\le n)=1-\left(\frac{5}{6}\right)^n$, we have to solve the equality $1-\left(\frac{5}{6}\right)^n=0.9$. This implies $\left(\frac{5}{6}\right)^n=0.1$, that is $n\ln(\frac{5}{6})=\ln 0.1$.

Computing the value of n we get $n=\frac{\ln 0.1}{\ln\frac{5}{6}}=12.63$. But we expect an integer value for n, hence we have to decide whether n=12 or n=13 is appropriate.

$P(\xi \le 12) = 1 - \left(\frac{5}{6}\right)^{12} = 0.888$, which is less than the required probability 0.9. $P(\xi \le 13) = 1 - \left(\frac{5}{6}\right)^{13} = 0.907$, which is larger than the requirement. Exactly 0.9 can not be achieved, the series skip over this level, as it can be seen in Fig. f.3.

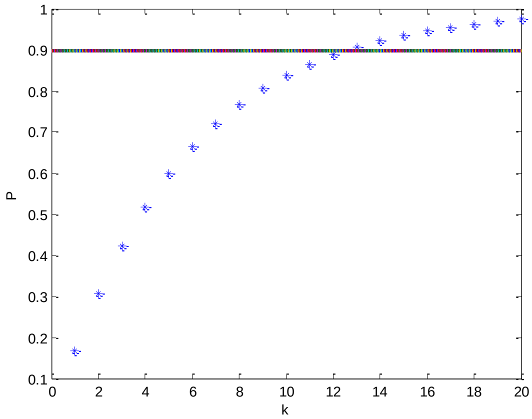


Figure f.3. The probabilities $P(\xi \le k)$ and the level y=0.9

The probabilities $P(k) = P(\xi = k)$ are presented in Fig.f.4.

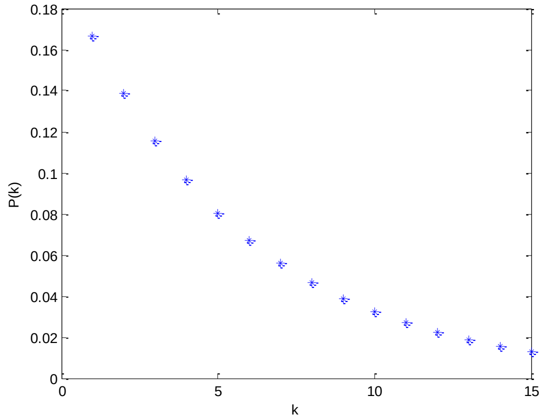


Figure f.4. The probabilities $P(\xi = k)$

Which is the most probable value of the throws? The most probable value of ξ equals 1, the probability belonging to them is $\frac{1}{6}$. All of the probabilities belonging to other value are smaller than $\frac{1}{6}$. We draw the attention to $P(\xi \ne 1) = \frac{5}{6}$, which is much more than the probability belonging to the value 1.

<u>Theorem</u> If ξ is a geometrically distributed random variable, then for any nonnegative integer values of n and m the following equality holds: $P(\xi > m+n \mid \xi > n) = P(\xi > m)$.

<u>Proof</u> Recall that $P(\xi > k) = (1-p)^k$. Applying the definition of conditional probability, $P(\xi > m+n \mid \xi > n) = \dfrac{P((\xi > m+n) \cap (\xi > n))}{P(\xi > n)}$. $\{\xi > m+n\}$ implies $\{\xi > n\}$, consequently the intersection is $\{\xi > m+n\}$. Therefore

$P(\xi > m+n \mid \xi > n) = \dfrac{P(\xi > m+n)}{P(\xi > n)} = \dfrac{(1-p)^{m+n}}{(1-p)^n} = (1-p)^m$, which coincides with $P(\xi > m)$.

<u>Remarks</u>

- The property $P(\xi > m+n \mid \xi > n) = P(\xi > m)$ is the so called forever young property. If we do not succeed until n, the probability that we will not succeed until further m experiments is the same that the probability that we do not succeed until m. Everything begins as if we were at the starting point.
- One can also prove that the forever young property implies the geometric distribution in the set of positive integer valued random variables. Consequently, this property is a pivotal property.
- $P(\xi > m+n \mid \xi > n) = P(\xi > m)$ implies the formula $P(\xi \le m+n \mid \xi > n) = P(\xi \le m)$ as well. As an explanation recall that $P(\overline{A} \mid B) = 1 - P(A \mid B)$. $P(\xi \le m+n \mid \xi > n) = 1 - P(\xi > m+n \mid \xi > n) = 1 - P(\xi > m) = P(\xi \le m)$.

<u>Example</u>

E2. At an exam there are 10 tests. The candidate gives it back if the test is not from the first three tests. Compute the probability that the candidate will succeed in 4 experiments.

Let ξ be the number of bids. ξ is a geometrically distributed random variable with parameter $p = \dfrac{3}{10}$.

$P(\xi \le 4) = P(\xi = 1) + P(\xi = 2) + P(\xi = 3) + P(\xi = 4) =$

$0.3 + 0.3 \cdot 0.7 + 0.3 \cdot 0.7^2 + 0.3 \cdot 0.7^3 = 1 - 0.7^4 = 0.760$

At most how many bids does he need with probability 0.95?

n=? $P(\xi \le n) = 0.95$. $P(\xi \le n) = 1 - 0.7^n = 0.95$, which implies n=8.4. Consequently, the candidate needs at most 9 bids until the hit.

If he does not succeed up to the 5th experiment, compute the probability that he succeeds until the 8th one.

The question can be easily answered by applying the forever young property as follows:

$P(\xi \le 8 \mid \xi > 5) = P(\xi \le 3) = P(\xi = 1) + P(\xi = 2) + P(\xi = 3) = 0.3 + 0.3 \cdot 0.7 + 0.3 \cdot 0.7^2 = 0.657$.

g. Frequently used continuous distributions

The aim of this chapter

In chapter d. we have dealt with continuous random variables. Now we investigate some frequently used types. We compute their numerical characteristics, study their main properties and we present their relationships with some discrete distributions, as well. We derive new random variables from normally distributed random variables. These are often used in statistics.

Preliminary knowledge

Random variables and their numerical characteristics. Probability density function. Partial integration.

Content

g.1. Uniformly distributed random variables.

g.2. Exponentially distributed random variables.

g.3. Normally distributed random variables.

g.4. Further random variables derived from normally distributed ones.

g.1. Uniformly distributed random variables

In this chapter we deal with some frequently used continuous random variable. We defined them with the help of their probability density function.

First we deal with a very simple continuous random variable. Let Ω, $\mathcal{A}$, and P be given and ξ be a random variable.

Definition The random variable ξ is called a **uniformly distributed random variable with parameters a, b (a<b),** if its probability density function is $f(x) = \begin{cases} c & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$.

Remarks

- As the area under the probability density function equals 1, $c = \frac{1}{b-a}$. This value is positive, consequently all the values of the probability density function are nonnegative.
- The constant value of the probability density function express that all the values of the interval $[a, b]$ are equally probable.
- A uniformly distributed random variable with parameter a, b (a<b) is often called a uniformly distributed random variable in $[a, b]$
- The graph of the probability density function of the uniformly distributed random variable with parameters $a = -1$, $b = 5$ can be seen in Fig.g.1.

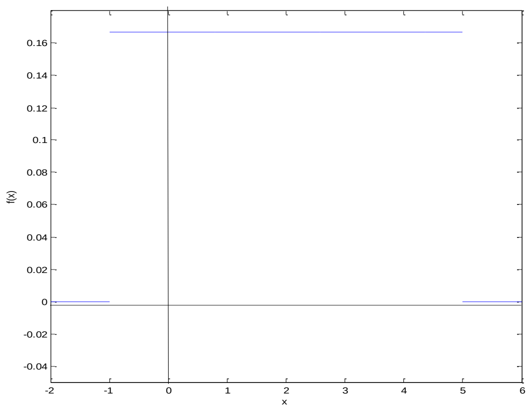


Figure g.1. The probability density function of a uniformly distributed random variable with parameters a=-1, b=5

Theorem
The cumulative distribution function of a uniformly distributed random variable in $[a, b]$ is

$$F(x) = \begin{cases} 0 & \text{if } x \le a \\ \frac{x-a}{b-a} & \text{if } a < x \le b \\ 1 & \text{if } b < x \end{cases}.$$

Proof

Recall the relationship $F(x)=\int_{-\infty}^{x} f(t)dt$ between the probability density function and the cumulative distribution function presented in section d.

If $x<a$, then $F(x)=\int_{-\infty}^{x} f(t)dt=\int_{-\infty}^{x} 0dt=0$.

If $a\le x\le b$, then $F(x)=\int_{-\infty}^{x} f(t)dt=\int_{-\infty}^{a} 0dt+\int_{a}^{x}\frac{1}{b-a}dt=0+\frac{1}{b-a}\cdot[t]_a^x=\frac{x-a}{b-a}$.

Finally, if $b<x$, then $F(x)=\int_{-\infty}^{x} f(t)dt=\int_{-\infty}^{a} 0dt+\int_{a}^{b}\frac{1}{b-a}dt+\int_{b}^{x} 0dt=0+1+0=1$.

The graph of the cumulative distribution function of a uniformly distributed random variable with parameters a=-1 and b=5 is presented in Fig.g.2.

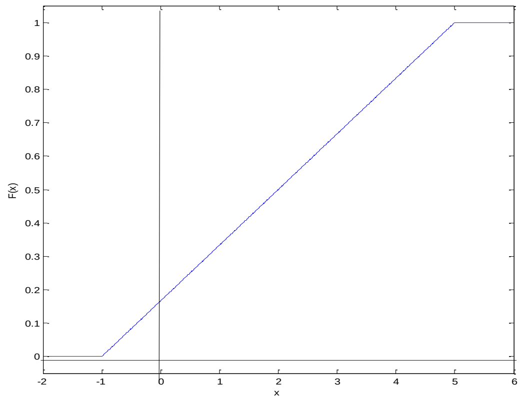


Figure g.2. The cumulative distribution function of the uniformly distributed random variable with parameters a=-1, b=5

Remarks

- Let ξ be a uniformly distributed random variable in the interval $[a,b]$ and $a<c<d<b$. Then $P(c<\xi<d)=F(d)-F(c)=\frac{d-a}{b-a}-\frac{c-a}{b-a}=\frac{d-c}{b-a}$. The probability the value of ξ in the interval (c,d) is proportional to the length of the interval (c,d).
- Choose a number from the interval $[a,b]$ with geometric probability. Let ξ be the chosen number. Then ξ is a uniformly distributed random variable in the interval $[a,b]$.

As justification consider that $P(\xi<x)=P(\emptyset)=0$, if $x<a$, $P(\xi<x)=P(a\le\xi<x)=\frac{x-a}{b-a}$, if $a\le x\le b$ and $P(\xi<x)=P(\Omega)=1$, if $b<x$. $F(x)=P(\xi<x)$, and $f(x)=F'(x)=\frac{1}{b-a}$ if $a<x<b$, and 0 if $x<a$ or $b<x$. At the endpoints $x=a$ and $x=b$ the cumulative distribution function is not differentiable, we can define the probability density

function in anyway. Defining $f(a)=\frac{1}{b-a}=f(b)$, f equals to the density function in the definition.

- The random number generator of computers usually generates approximately uniformly distributed random variables in $[0,1]$.

Numerical characteristics of uniformly distributed random variables:

Expectation

$E(\xi)=\frac{a+b}{2}$, which is a straightforward consequence of

$E(\xi)=\int_{-\infty}^{\infty} x\cdot f(x)dx=\int_a^b x\cdot\frac{1}{b-a}dx=\frac{1}{b-a}\int_a^b xdx=\frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b=\frac{b^2-a^2}{2(b-a)}=\frac{b+a}{2}$. Note that this value is the midpoint of the interval $[a,b]$.

Dispersion

$D(\xi)=\frac{b-a}{\sqrt{12}}$. As a proof, recall that $D^2(\xi)=E(\xi^2)-(E(\xi))^2$.

$E(\xi^2)=\int_{-\infty}^{\infty} x^2 f(x)dx=\int_a^b x^2\frac{1}{b-a}dx=\frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b=\frac{b^3-a^3}{3(b-a)}=\frac{b^2+ab+a^2}{3}$.

$D^2(\xi)=E(\xi^2)-(E(\xi))^2=\frac{b^2+ab+a^2}{3}-\left(\frac{a+b}{2}\right)^2=\frac{b^2-2ab+b^2}{12}=\frac{(b-a)^2}{12}$.

Consequently, $D(\xi)=\sqrt{\frac{(b-a)^2}{12}}=\frac{|b-a|}{\sqrt{12}}=\frac{b-a}{\sqrt{12}}$.

Mode

All of the values of the interval $[a,b]$ have the same chance, consequently, all the points of (a,b) are modes.

Median

$\frac{a+b}{2}$. We have to find the value of y for which $F(y)=0.5$. As neither 0 nor 1 equal 0.5, the following equality has to hold: $\frac{y-a}{b-a}=0.5$. This implies $y-a=0.5(b-a)$. After rearranging it, we get $y=\frac{b+a}{2}$.

Example

E1. Let ξ be uniformly distributed random variable in $[2,10]$. Compute the probability that the value of the random variable is between 5 and 8.

The cumulative distribution function of ξ is given by $F(x)=\begin{cases}0 & \text{if } x\le 2\\ \dfrac{x-2}{8} & \text{if } 2<x\le 10\\ 1 & \text{if } 10<x\end{cases}$,

which is a useful tool to compute probabilities.

$P(5<\xi<8)=F(8)-F(5)=\frac{8-2}{8}-\frac{5-2}{8}=\frac{3}{8}=0.375$.

Compute the probability that the value of the random variable is less than 5.

$P(\xi<5)=F(5)=\frac{5-2}{8}=\frac{3}{8}=0.375$.

Compute the probability that the value of the random variable is greater than 8.

$P(8<\xi)=1-F(8)=1-\frac{8-2}{8}=\frac{2}{8}=0.25$.

Compute the probability that the value of the random variable is greater than the half of its expectation and less than the double of the expectation.

$E(\xi)=\frac{2+10}{2}=6$, $P(3<\xi<12)=F(12)-F(3)=1-\frac{3-2}{8}=\frac{7}{8}$.

At most how much is the value of the random variable with probability 0.9?

x=? for which $P(\xi\le x)=0.9$. $P(\xi\le x)=F(x)$, we have to solve $\frac{x-2}{8}=0.9$. This implies $x=9.2$.

At least how much is the value of the random variable with probability 0.9?

x=? for which $P(\xi\ge x)=0.9$. $P(x\le\xi)=1-F(x)$, we have to solve $1-\frac{x-2}{8}=0.9$. This implies $x=2.8$.

Given that the value of the random variable is more than 5, compute the probability that it is less than 8.

$$P(\xi<8\,|\,\xi\ge 5)=\frac{P((\xi<8)\cap(\xi\ge 5))}{P(\xi\ge 5)}=\frac{P(5\le\xi<8)}{P(\xi\ge 5)}=\frac{F(8)-F(5)}{1-F(5)}=\frac{\frac{8-2}{8}-\frac{5-2}{8}}{1-\frac{5-2}{8}}=\frac{3}{5}=0.6.$$

Notice that this conditional probability is proportional to the length of the interval $[5,8)$ if the number is from $[5,10]$.

<u>Theorem</u> If ξ is a uniformly distributed random variable in $[0,1]$, $0<c$ and $d\in R$, then $\eta=c\xi+d$ is uniformly distributed random variable in $[d,c+d]$.

<u>Proof</u> Investigate the cumulative distribution function of η, then take its derivative.

$$F_\eta(x)=P(\eta<x)=P(c\xi+d<x)=P(\xi<\frac{x-d}{c})=F_\xi\left(\frac{x-d}{c}\right).$$

Recalling that $F_\xi(x)=\begin{cases}0 & \text{if } x\le 0\\ x & \text{if } 0<x\le 1\\ 1 & \text{if } 1<x\end{cases}$, $F_\eta(x)=\begin{cases}0 & \text{if } \frac{x-d}{c}\le 0\\ \frac{x-d}{c} & \text{if } 0<\frac{x-d}{c}\le 1\\ 1 & \text{if } 1<\frac{x-d}{c}\end{cases}$.

Summarizing, $F_\eta(x)=\begin{cases}0 & \text{if } x\le d\\ \frac{x-d}{c} & \text{if } d<x\le c+d\\ 1 & \text{if } c+d<x\end{cases}$.

Taking the derivative of $F_\eta(x)$, $f_\eta(x)=\begin{cases}\frac{1}{c} & \text{if } d\le x\le c+d\\ 0 & \text{otherwise}\end{cases}$.

Remarks

- If c is negative, then $\eta=c\xi+d$ is a uniformly distributed random variable in $[c+d,d]$.
- Using the random number generator, we can get a uniformly distributed random variable in $[a,b]$ by multiplying the generated random number by $b-a$ and adding a.
- If ξ is a uniformly distributed random variable in $[0,1]$, then so is $\eta=1-\xi$. To justify it, first take into consideration that all of values of ξ are in $[0,1]$, hence so are the values of $\eta=1-\xi$. Moreover,

$F_\eta(x)=P(\eta<x)=P(1-\xi<x)=P(1-x<\xi)=1-F_\xi(1-x)=1-(1-x)=x$, if $0<x<1$. Therefore $f_\eta(x)=F'_\eta(x)=1$, if $0<x<1$ and zero outside $[0,1]$.

Theorem

Let ξ be a uniformly distributed random variable in [0,1]. Let F be a continuous cumulative distribution function in R. Let $I=\{x\in R: F(x)\ne 0, F(x)\ne 1\}$ and suppose that F is strictly monotone in I. Then $\eta=F^{-1}(\xi)$ is a random variable those cumulative distribution function is F.

Proof

$F^{-1}:(0,1)\to I$, $P(\xi=0)=0$, $P(\xi=1)=0$. $\eta=F^{-1}(\xi)$ is well defined. Take any value $x\in I$, and investigate the cumulative distribution function of η at x. Taking into account that $F_\xi(x)=\begin{cases}0 & \text{if } x\le 0\\ x & \text{if } 0<x\le 1\\ 1 & \text{if } 1<x\end{cases}$.

$F_\eta(x)=P(\eta<x)=P\left(F^{-1}(\xi)<x\right)$.

As F is monotone increasing, $\left\{F^{-1}(\xi)<x\right\}=\left\{F(F^{-1}(\xi))<F(x)\right\}=\left\{\xi<F(x)\right\}$.Consequently, $P(F^{-1}(\xi)<x)=P(\xi<F(x))=F_\xi(x)=F(x)$.

If $x\le\inf I$, then $F(x)=0$ and $F_\eta(x)=P(\eta<x)=P\left(F^{-1}(\xi)<x\right)=0$.

If $\sup I\le x$, then $F(x)=1$ and $F_\eta(x)=P(\eta<x)=P\left(F^{-1}(\xi)<x\right)=1$.

Consequently, the cumulative distribution function of $F^{-1}(\xi)$ is $F(x)$.

Remark

- The previous statement gives us the possibility to generate random variables with cumulative distribution function F.

Example

E2. Generate random variables with cumulative distribution function $F(x)=\begin{cases}0 & \text{if } x\le 1\\ 1-\dfrac{1}{x} & \text{if } 1<x\end{cases}$.

Apply the previous statement. F is a strictly monotone increasing function in the interval $(1,\infty)$, $F^{-1}(y)=\dfrac{1}{1-y}$, $0<y<1$. Consequently, if ξ is uniformly distributed in $[0,1]$, then $F^{-1}(\xi)$ is a random variable with cumulative distribution function F. Consequently, substituting the random number generated by the computer into F^{-1} we get a random variable with cumulative distribution function F. The relative frequencies of the random numbers and the probability density function $f(x)=F'(x)=\dfrac{1}{x^2}$, $1\le x$, can be seen in Fig.g.3.

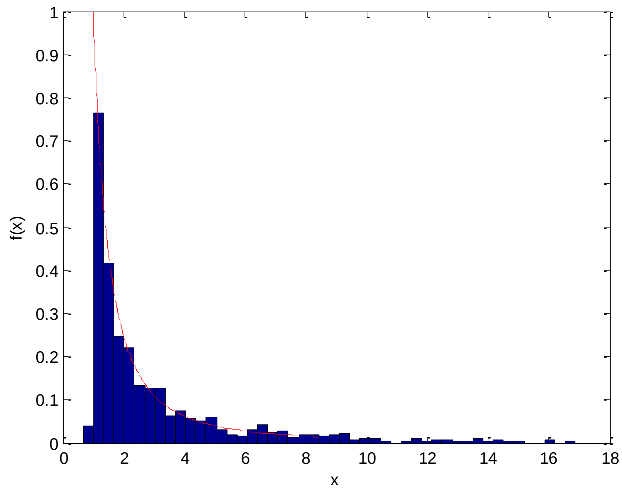


Figure g.3. The relative frequencies of random numbers $F^{-1}(\xi)$ situated in different subintervals and the probability density function

g.2. Exponentially distributed random variables

In this subsection we deal a frequently used continuous distribution, namely the exponential one.

<u>Definition:</u> The random variable ξ is an **exponentially distributed random variable with parameter** $0<\lambda$**,** if its probability density function is $f(x)=\begin{cases}0 & \text{if } x\le 0\\ \lambda e^{-\lambda x} & \text{if } 0<x\end{cases}$.

Remarks

- $0\le f(x)$ is obvious, furthermore,

$\int\limits_{-\infty}^{\infty} f(x)dx=\int\limits_{0}^{\infty}\lambda e^{-\lambda x}dx=\left[\lambda\dfrac{e^{-\lambda x}}{-\lambda}\right]_0^{\infty}=\lim\limits_{x\to\infty}\left(-e^{-\lambda x}\right)-(-e^0)=0+1=1$. These properties imply that $f(x)$ is a probability density function. The graphs of probability density functions of exponentially distributed random variables belonging to different parameters are presented in Fig.g.4.

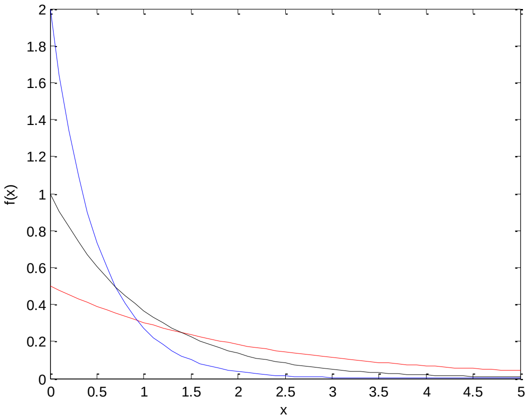


Figure g.4. The probability density functions of exponentially distributed random variables with parameters $\lambda = 1$ (black), $\lambda = 0.5$ (red) and $\lambda = 2$ (blue)

- An exponentially distributed random variable takes its values with large probability around zero, whatever the parameter is. All of its values are nonnegative.

<u>Theorem</u> The cumulative distribution function of an exponentially distributed random variable with parameter $0 < \lambda$ is $F(x) = \begin{cases} 0 & \text{if } x \le 0 \\ 1 - e^{-\lambda x} & \text{if } 0 < x \end{cases}$.

<u>Proof</u>

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} 0dx = 0, \text{ if } x \le 0.$$

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{0}^{x} \lambda e^{-\lambda t}dx = \left[\frac{e^{-\lambda t}}{-1}\right]_0^x = e^{-\lambda x} - (-1) = 1 - e^{-\lambda x}, \text{ if } 0 < x.$$

The graphs of the cumulative distributions function belonging to the previous probability density functions are presented in Fig.g.5.

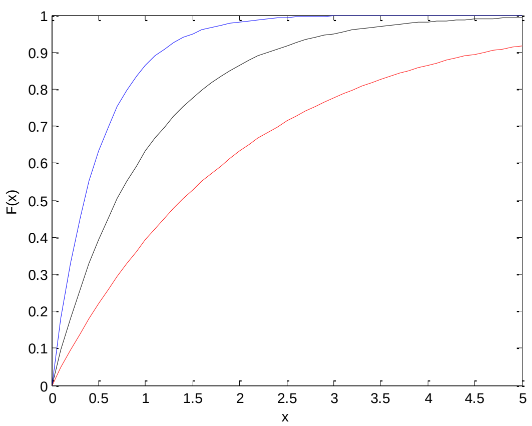


Figure g.5. The cumulative distribution functions of exponentially distributed random variables with parameters $\lambda = 1$ (black), $\lambda = 0.5$ (blue) and $\lambda = 2$ (red)

Remark

- A simple way to generate an exponentially distributed random variable to substitute the uniformly distributed random variable into $F^{-1}(y) = \frac{\ln(1-y)}{-\lambda}$. The relative frequencies of exponentially distributed random variables situated in the interval $[0,5]$ are presented in Fig.g.6. One can notice that the relative frequencies follow the probability density function drawn by red line.

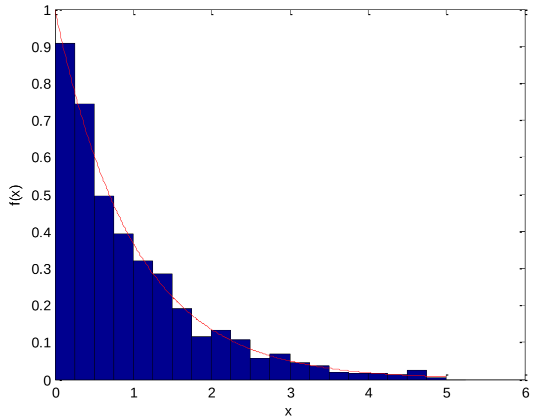


Figure g.6. The relative frequencies of random numbers $-\ln(1-\xi)$ situated in different subintervals and the exponential probability density function with parameter $\lambda = 1$

Numerical characteristics of exponentially distributed random variables:

Expectation

$E(\xi) = \frac{1}{\lambda}$. It follows from

$$E(\xi) = \int_{-\infty}^{\infty} x \cdot f(x)dx = \int_{0}^{\infty} x \cdot \lambda e^{-\lambda x} dx = \left[x \cdot (-e^{-\lambda x})\right]_0^{\infty} - \int_{0}^{\infty} -e^{-\lambda x} dx = 0 - \left[\frac{e^{-\lambda x}}{-\lambda}\right]_0^{\infty} = \frac{1}{\lambda}.$$

Taking the average of random numbers generated previously in the presented way, for $\lambda = 1$, the results are in Table g.1. Differences from the exact expectation 1 are also presented:

N=	1000	10000	100000	1000000	10000000
Average	0.9796	1.0083	1.0015	1.0005	0.9996
Difference	0.0204	0.0083	0.0015	0.0005	0.0004

Table g.1. The average of the values of the random variable $-\ln(1-\xi)$, if ξ is a uniformly distributed random variable in $[0,1]$ in case of different numbers of simulations N

Dispersion

$D(\xi)=\frac{1}{\lambda}$. As a proof, recall that $D^2(\xi)=E\left(\xi^2\right)-\left(E(\xi)\right)^2$. Integrating twice by parts, $E(\xi^2)=\int_{-\infty}^{\infty}x^2 f(x)dx=\frac{2}{\lambda^2}$, $D^2(\xi)=E\left(\xi^2\right)-\left(E(\xi)\right)^2=\frac{2}{\lambda^2}-\frac{1}{\lambda^2}=\frac{1}{\lambda^2}$.

Mode

There is no mode.

Median

$\frac{\ln 0.5}{-\lambda}$. We have to find the value x for which $F(x)=0.5$.In order to do this, we have to solve the following equation $1-e^{-\lambda x}=0.5$. This implies $e^{-\lambda x}=0.5$. Taking the logarithm of both sides, we get $-\lambda x=\ln 0.5$, finally $x=\frac{\ln 0.5}{-\lambda}$.

Example

E1. The lifetime of a bulb is supposed to be an exponentially distributed random variable with expectation 1000 hours. Compute the probability that the bulb breaks down before 500 hours.

Let ξ denote the lifetime of a bulb. As ξ is an exponentially distributed random variable, its cumulative distribution function is $F(x)=1-e^{-\lambda x}$, $x\geq 0$. As $E(\xi)=\frac{1}{\lambda}=1000$, $\lambda=0.001$.

$P(\xi<500)=F(500)=1-e^{-\frac{500}{1000}}=0.393$.

Compute the probability that the bulb goes broke between 1000 and 2000 hours.

$P(1000<\xi<2000)=F(2000)-F(1000)=\left(1-e^{-\frac{2000}{1000}}\right)-\left(1-e^{-\frac{1000}{1000}}\right)=0.233$.

At most how many hours is the lifetime of a bulb with probability 0.98?

x=?, $P(\xi\leq x)=0.98$. $P(\xi\leq x)=F(x)=1-e^{-\frac{x}{1000}}=0.98$, consequently, $e^{-\frac{x}{1000}}=0.02$, and $x=-1000\cdot\ln 0.02=3912$.

At least how many hours is the lifetime of a bulb with probability 0.98?

x=?, $P(\xi\geq x)=0.98$. $P(\xi\geq x)=1-F(x)=1-e^{-\frac{x}{1000}}=0.98$, consequently, $e^{-\frac{x}{1000}}=0.98$, and $x=-1000\cdot\ln 0.98=20.2$.

Compute the probability that, out of 10 bulbs, having independent exponentially distributed lifetimes with expectation 1000 hours, 7 go broke before 1000 hours and 3 operate after 1000 hours.

Let ξ_i denote the lifetime of the ith bulb. They are independent random variables and $P(\xi_i<1000)=F(1000)=1-e^{-\frac{1000}{1000}}=0.632$, $P(\xi_i\geq 1000)=0.368$. If η is the number of bulbs going broke until 1000 hours, η is a binomially distributed random variable with parameters $n=10$ and $p=P(\xi_i<1000)$. Therefore $P(\eta=7)=\binom{10}{7}\cdot 0.632^7\cdot 0.368^3=0.241$.

Now we present the characteristic feature of exponentially distributed random variables.

<u>Theorem</u> If ξ is an exponentially distributed random variable, then for any $0 \le x, 0 \le y$ the following property holds: $P(\xi \ge x+y \mid \xi \ge x) = P(\xi \ge y)$.

<u>Proof</u>

Recall that $P(\xi \ge a) = 1 - F(a) = 1 - (1 - e^{-\lambda a}) = e^{-\lambda a}$.

Moreover,

$$P(\xi \ge x+y \mid \xi \ge x) = \frac{P(\xi \ge x+y \cap \xi \ge x)}{P(\xi \ge x)} = \frac{P(\xi \ge x+y)}{P(\xi \ge x)} = \frac{e^{-\lambda(x+y)}}{e^{-\lambda x}} = e^{-\lambda y} = P(\xi \ge y).$$

<u>Remark</u>

- The previous property can be written in the form $P(\xi < x+y \mid \xi \ge x) = P(\xi < y)$, as well.

Consider that

$P(\xi < x+y \mid \xi \ge x) = 1 - P(\xi \ge x+y \mid \xi \ge x) = 1 - P(\xi \ge y) = P(\xi < y)$.

- As exponentially distributed random variables are continuous random variables, then we do not bother if the strict inequality (>) or ≥ holds. We can also write $P(\xi > x+y \mid \xi > x) = P(\xi > y)$, which coincides with the property stated for geometrically distributed random variables.
- The property can be interpreted as the forever young property. If ξ is the lifetime of an appliance, then ξ is the time point when it goes broke it does not go broke until x, the probability that it will not go broke until further y units of time is the same that it does not go broke until y from the beginning. This is the reason for the name of the property.
- The forever young property is valid only for the exponentially distributed random variable in the set of continuous random variables.

<u>Theorem</u> Let ξ be a continuous random variable with nonnegative values, suppose that its cumulative distribution function is differentiable and $\lim_{x \to 0+} F(x) = \lambda$, $0 < \lambda$. Moreover, for any $0 \le x, y$ $P(\xi \ge x+y \mid \xi \ge x) = P(\xi \ge y)$ holds. Then ξ is an exponentially distributed random variable with parameter λ.

<u>Proof</u> Let $G(x) = 1 - F(x)$. As ξ is nonnegative, $F(0) = 0$, $G(0) = 1$. As the conditional probability exists, $0 < P(\xi \ge x)$, consequently $G(x) < 1$. Let $0 < y = \Delta x$, $P(\xi \ge x+y \mid \xi \ge x) = P(\xi \ge y)$ has the form $P(\xi \ge x+\Delta x \mid \xi \ge x) = P(\xi \ge \Delta x)$. $P(\xi \ge x+\Delta x \mid \xi \ge x) = \dfrac{P(\xi \ge x+\Delta x)}{P(\xi \ge x)} = \dfrac{G(x+\Delta x)}{G(x)} = G(\Delta x)$. This implies the equation $G(x+\Delta x) = G(\Delta x)G(x)$. Subtracting $G(x)$ and using $G(0) = 1$ we get $G(x+\Delta x) - G(x) = G(x)(G(\Delta x) - G(0))$. Dividing by Δx and taking the limit of both sides if $0 < \Delta x \to 0$ we arrive at $G'(x) = G'(0+)G(x)$. $F'(0+) = \lambda$ implies $G'(0+) = -\lambda$, therefore $G'(x) = -\lambda G(x)$. This is an ordinary differential equation which is easy to solve. Dividing by $G(x) \ne 0$, $\dfrac{G'(x)}{G(x)} = -\lambda$, consequently $\ln|G(x)| = -\lambda x + c$. $G(x)$ is nonnegative, hence $\ln G(x) = -\lambda x + c$ and $G(x) = e^{-\lambda x + c}$. $G(0) = e^{-\lambda \cdot 0 + c} = 1$ implies $c = 0$ and $G(x) = e^{-\lambda x}$. Finally, $1 - F(x) = e^{-\lambda x}$, $F(x) = 1 - e^{-\lambda x}$ and $f(x) = F'(x) = \lambda e^{-\lambda x}$.

<u>Remarks</u>

- The assumptions of the previous statement can be slightly relaxed.
- The forever young property can be assumed of the lifetime of appliances when the fault is not caused by age. For example, if ξ is the age of a person, then $P(\xi \geq 100 \mid \xi \geq 90) \neq P(\xi \geq 10)$. In other words, if he survives 90 years, the probability that he survives 10 more years is obviously less than the probability of surviving 10 years from the birth. Examples of exponentially distributed random variables are punctures. Punctures are usually caused by a pin. I we do not run into a pin until x, then the wheel do not remember the previous passage.
- The forever young property of the exponentially and geometrically distributed random variables indicates that the geometrical distributed random variable is the discrete counterpart of the exponentially distributed random variable. This is also supported by the formulas $E(\xi) = \frac{1}{p}$ and $E(\xi) = \frac{1}{\lambda}$, respectively.

Example

E2. The distances between the consecutive punctures are independent exponentially distributed random variables. The probability that there is no puncture until 20000 km equals 0.6. Compute the probability that there is no puncture until 50000 km.

Let ξ_1 be the distance until the first puncture. Because of the forever young property, we can suppose that the distance begins at 0. Actually we do not know the expectation and the value of the parameter, but we know that $P(\xi_1 < 20000) = 0.6$, This is suitable for determining the value of the parameter λ as follows. $P(\xi_1 < 20000) = F(20000) = 1 - e^{-\lambda 20000} = 0.6$. $e^{-\lambda 20000} = 0.4$, which implies $\lambda = \frac{\ln 0.4}{-20000} = 4.58 \cdot 10^{-5}$. Returning to the question, $P(\xi_1 \geq 50000) = 1 - \left(1 - e^{-4.58 \cdot 10^{-5} \cdot 50000}\right) = 0.101$.

Compute the expectation of the distance between consecutive punctures.

$E(\xi_1) = \frac{1}{\lambda} = 21827$.

Given that the first puncture does not happen until 50000 km, compute the probability that it happens within 70000 km.

$P(\xi_1 < 70000 \mid \xi_1 \geq 50000) = P(\xi_1 < 2000) = F(2000) = 1 - e^{\lambda \cdot 2000} = 0.6$.

Given that the first puncture happens within 50000 km, compute the probability that it is until 10000 km.

$$P(\xi_1 < 10000 \mid \xi_1 < 50000) = \frac{P(\xi_1 < 10000 \cap \xi_1 < 50000)}{P(\xi_1 < 50000)} = \frac{P(\xi_1 < 10000)}{P(\xi_1 < 50000)} = \frac{F(10000)}{F(50000)} = 0.408 .$$

Theorem (Relationship between exponentially distributed random variables and Poisson distributed random variables)

Let ξ_i, $i = 1,2,3,...$ be independent exponentially distributed random variables with parameter λ, $0 < T$ fixed and

$$\eta_T = \begin{cases} 0 & \text{if} \quad T \le \xi_1 \\ 1 & \text{if} \quad \xi_1 < T \le \xi_2 \\ 2 & \text{if} \quad \xi_1 + \xi_2 < T \le \xi_1 + \xi_2 + \xi_3 \\ . & \\ . & \\ k & \text{if} \quad \sum_{i=1}^{k} \xi_i < T \le \sum_{i=1}^{k+1} \xi_i \\ . & \\ . & \\ . & \end{cases}.$$

Then, η_T is a Poisson distributed random variable with parameter $\lambda^* = \lambda \cdot T$.

The proof of this statement is omitted as it requires the knowledge of the distribution of the sum of exponentially distributed random variables.

E3. Returning to Example E2, compute the probability that until 100000 km there are at most 2 punctures.

Denote the number of punctures within T (km) by η_T. Applying the previous statement η_{100000} is a Poisson distributed random variable with parameter $\lambda^* = 100000 \cdot \lambda = 100000 \cdot 4.58 \cdot 10^{-5} = 4.58$.

Consequently, $P(\eta_{100000} \le 2) = P(\eta_{100000} = 0) + P(\eta_{100000} = 1) + P(\eta_{100000} = 2) =$

$$\frac{4.58^0}{0!} e^{-4.58} + \frac{4.58^1}{1!} e^{-4.58} + \frac{4.58^2}{2!} e^{-4.58} = 0.165 .$$

How many punctures happen until 200000 km most likely?

η_{200000} is also a Poisson distributed random variable with parameter $\lambda^{**} = 200000 \cdot 4.58 \cdot 10^{-5} = 9.16$. As the parameter λ^{**} is not an integer, there is a unique mode, namely $[\lambda^{**}] = [9.16] = 9$.

Theorem

If ξ is an exponentially distributed random variable with parameter λ, then $\eta = [\xi] + 1$ is geometrically distributed random variable with parameter $p = 1 - e^{-\lambda}$.

Proof As $0 < \xi$, $[\xi]$ takes nonnegative integer values, and η takes positive integer values.

$P(\eta = 1) = P([\xi] + 1 = 1) = P([\xi] = 0) = P(0 \le \xi < 1) = F(1) - F(0) = 1 - e^{-\lambda 1} - 0 = p$.

$P(\eta = 2) = P([\xi] + 1 = 2) = P([\xi] = 1) = P(1 \le \xi < 2) = F(2) - F(1) = \left(1 - e^{-\lambda 2}\right) - \left(1 - e^{-\lambda 1}\right) =$

$e^{-\lambda} - e^{-2\lambda} = e^{-\lambda}(1 - e^{-\lambda}) = p(1-p)$.

In general,

$$P(\eta = k) = P([\xi] + 1 = k) = P([\xi] = k - 1) = P(k - 1 \le \xi < k) = F(k) - F(k-1) = \left(1 - e^{-\lambda k}\right) - \left(1 - e^{-\lambda(k-1)}\right) =$$

$= e^{-\lambda(k-1)}(1 - e^{-\lambda}) = \left(e^{-\lambda}\right)^{k-1}(1 - e^{-\lambda}) = (1-p)^{k-1} \cdot p$, which is the formula to be proved.

Example

E4. Telecommunication companies invoice the fee of calls on the basis of minutes. It means that all minutes which have been begun have to be paid completely. If the duration of a call is an exponentially distributed random variable with expectation 2 minutes, how much is the expectation of its fee if every minute costs 25HUF.

Let ξ denote the duration of a call. The minutes invoiced are $\eta = [\xi] + 1$. The previous statement states that η is a geometrically distributed random variable with parameter $p = 1 - e^{-\lambda} = 1 - e^{-0.5} = 0.393$. Consequently, $E(\eta) = \frac{1}{p} = \frac{1}{0.393} = 2.54$. The expectation of the fee of a call is $E(25 \cdot \eta) = 25 \cdot E(\eta) = 25 \cdot 2.54 = 63.54$.

g.3. Normally distributed random variables

In this subsection we deal with the most important continuous distribution, namely the normal distribution. First of all we investigate the standard normal one.

Definition The continuous random variable ξ is a standard normally distributed random variable, if its probability density function is $f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$, $x \in R$.

Remarks

- The inequality $0 < f(x)$ holds for any value of $x \in R$, and it can be proved that $\int_{-\infty}^{\infty} e^{\frac{-x^2}{2}} dx = \sqrt{2\pi}$. Consequently, $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx = 1$. This means that $f(x)$ is really a probability density function.
- The above function is often called as the Gauss curve and is denoted by $\varphi(x)$.
- The function $\varphi(x)$ is obviously symmetric to the axis x.
- Standard normally distributed random variables can take any value.

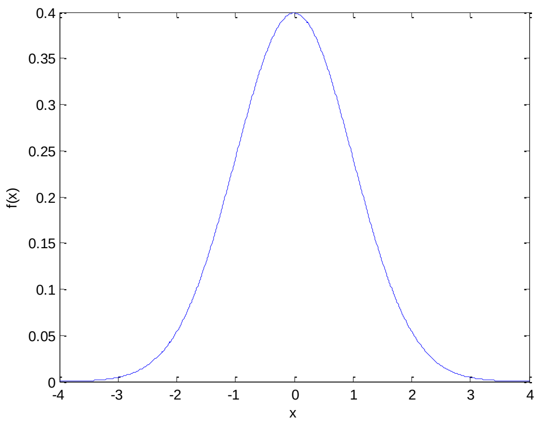


Figure g.7. The probability density function of a standard normally distributed random variable

- The graph of the probability density function can be seen in Fig.g.7.

- The cumulative distribution function of a standard normally distributed random variable is $F(x)=\int_{-\infty}^{x} f(t)dt=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$, which is the area under the Gauss-curve presented in Fig.g.8.

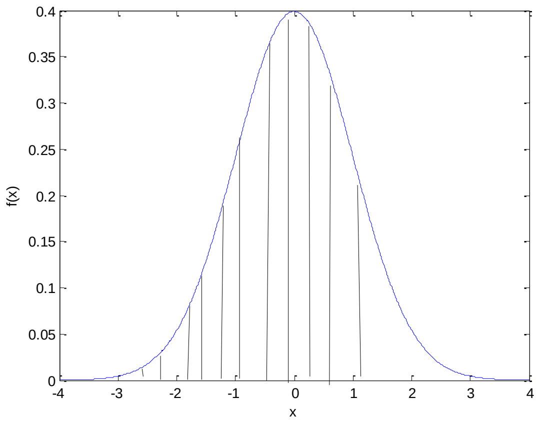


Figure g.8. The value of the cumulative distribution function as the area under the probability density function

- The cumulative distribution function of standard normally distributed random variables is denoted by $\Phi(x)$ (capital F in Greek alphabet). Its graph can be seen in Fig.g.9.

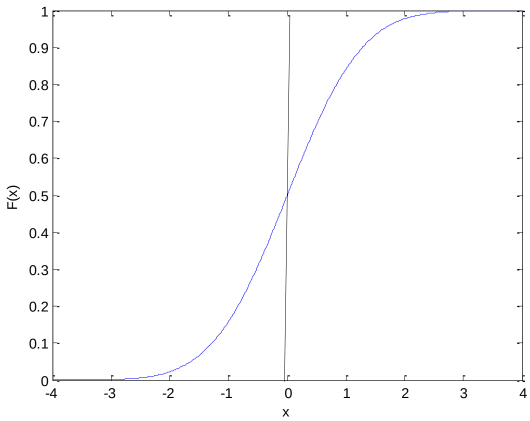


Figure g.9. The cumulative distribution function of standard normally distributed random variables

- We use the following notation: $\xi \sim N(0,1)$. N refers the name “normal”, the numbers 0 and 1 are parameters whose meanings will be explained later.
- The function Φ can not be written in a closed form, its values are computed numerically and are included in a table (see Table 1 at the end of the booklet and Table g.2.)

x	$\Phi(x)$
0	0.5
1	0.8413
2	0.9773
3	0.9986

Table g.2. Some values of the cumulative distribution function of standard normally distributed random variables

Data from this table can be read out as follows: $\Phi(0)=0.5$, $\Phi(1)=0.8413$, $\Phi(2)=0.9773$, $\Phi(3)=0.9986$.

Remarks

- The tables do not contain arguments greater than 3.8. As the cumulative distribution function is monotone increasing and it takes values at most 1, furthermore $\Phi(3.8)=0.99993$, $0.9999<\Phi(x)<1$ holds in case of $3.8<x$. We use $\Phi(x)\approx 1$ for $3.8<x$.
- The tables do not contain arguments less than 0, because the values at negative arguments can be computed as follows.

Theorem
If $0\le x$, then $\Phi(-x)=1-\Phi(x)$.
Proof The proof is based on the symmetry of the probability density function.
$$\Phi(-x)=\int_{-\infty}^{-x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt=1-\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt=1-\Phi(x).$$
Expressively, stripped areas of the Fig.g.9. are equal.

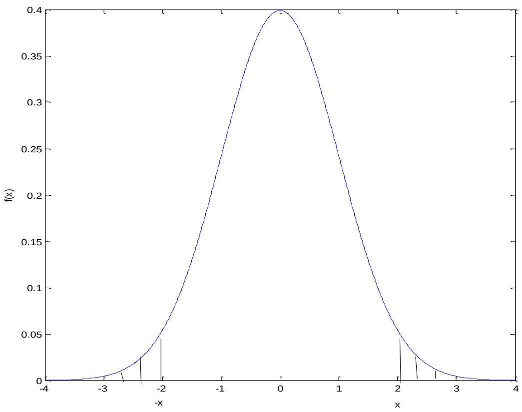


Figure g.9. Equal areas under the standard normal probability density function due to its symmetry

Obviously, $\Phi(-x)=1-\Phi(x)$ holds for any value of x.

Theorem
If $\xi\sim N(0,1)$, then $-\xi\sim N(0,1)$ holds, as well.
Proof Let $\eta=-\xi$.

$F_\eta(x) = P(\eta < x) = P(-\xi < x) = P(0 < \xi + x) = P(-x < \xi) = 1 - \Phi(-x) = 1 - (1 - \Phi(x)) = \Phi(x).$

Now $f_\eta(x) = F_\eta'(x) = \Phi'(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$, which proves the statement.

Numerical characteristics of standard normally distributed random variables:

Expectation

$E(\xi) = 0$. It follows from the fact that $\int x \cdot e^{-\frac{x^2}{2}} dx = -e^{-\frac{x^2}{2}}$ and

$$E(\xi) = \int_{-\infty}^{\infty} x \cdot f(x)dx = \int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \lim_{x\to\infty} -\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} - \lim_{x\to-\infty} -\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = 0.$$

Dispersion

$D(\xi) = 1$. As a proof, recall that $D^2(\xi) = E(\xi^2) - (E(\xi))^2$. Applying partially integration

$$E(\xi^2) = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_{-\infty}^{\infty} x^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \int_{-\infty}^{\infty} x \cdot x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \left[x \frac{-1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

.

Recalling L' Hopital's rule we get

$\lim_{x\to\infty} x \frac{-1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = \lim_{x\to-\infty} x \frac{-1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = 0$. Moreover, $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 1$, as $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ is a probability density function. Consequently, $D^2(\xi) = E(\xi^2) - (E(\xi))^2 = 1 - 0^2 = 1$, which proves the statement.

Mode

Local maximum of φ is at $x = 0$, consequently the mode is zero.

Median

$me = 0$. We have to find the value x for which $\Phi(x) = 0.5$. Using the table of cumulative distribution function of standard normal distribution, we get $x = 0$.

Example

E1. Let ξ be a standard normally distributed random variable. Compute the probability that ξ is less than 2.5.

$P(\xi < 2.5) = \Phi(2.5) = 0.9938$.

Compute the probability that ξ is greater than -1.2.

$P(-1.2 < \xi) = 1 - \Phi(-1.2) = 1 - (1 - \Phi(1.2)) = \Phi(1.2) = 0.8849$.

Compute the probability that ξ is between -0.5 and 0.5.

$P(-0.5 < \xi < 0.5) = \Phi(0.5) - \Phi(-0.5) = \Phi(0.5) - (1 - \Phi(0.5)) = 2\Phi(0.5) - 1 = 2 \cdot 0.6915 - 1 = 0.3830$.

At most how much is ξ with probability 0.9?

x=? $P(\xi \le x) = 0.9$. $P(\xi \le x) = \Phi(x) = 0.9$. We have to find the value 0.9 in the columns of Φ, as the value of the function equals 0.9. Therefore, $x = 1.28$.

At least how much is ξ with probability 0.95?

x=? $P(\xi \geq x) = 0.95$. $1 - \Phi(x) = 0.95 \Rightarrow \Phi(x) = 0.05$. As $\Phi(x) < 0.5$ and Φ is monotone increasing function, $x < 0$. If we denote $x = -a$, $0 < a$ and $\Phi(x) = \Phi(-a) = 1 - \Phi(a) = 0.05$. This implies $\Phi(a) = 0.95$ and $a = 1.645$ .Finally, we end in $x = -1.645$.

Determine an interval which is symmetric to 0 and in which the values of ξ are situated with probability 0.99.

x=? $P(-x < \xi < x) = 0.99$. $P(-x < \xi < x) = \Phi(x) - \Phi(-x) = 2\Phi(x) - 1 = 0.99$. This implies $\Phi(x) = 0.995$ and $x = 2.58$. The interval is $(-2.58, 2.58)$

Now we turn to the general form of normal distribution.

Definition Let ξ be a standard normally distributed random variable, $m \in R$ and $0 < \sigma$. The random variable $\eta = \sigma\xi + m$ is called a **normally distributed random variable with parameters** m **and** σ**.** We use the notation $\eta \sim N(m, \sigma)$.

Remarks

- With $m = 0$ and $\sigma = 1$, $\eta = \sigma\xi + m = \xi$ is a standard normally distributed random variable. It fits with the notation $\xi \sim N(0,1)$.
- η is a linear transformation of a standard normally distributed random variable.
- If $a < 0$ and $m \in R$, then $\eta = a\xi + m = (-a)(-\xi) + m$. Recall that $-\xi \sim N(0,1)$ holds as well, furthermore $0 < -a$, consequently η is a normally distributed random variable with parameters m and $-a$.

Theorem Let ξ be a standard normally distributed random variable, $m \in R$ and $0 < \sigma$. The cumulative distribution function of the random variable $\eta = \sigma\xi + m$ is $F_\eta(x) = \Phi(\frac{x-m}{\sigma})$ and the probability density function of η is $f_\eta(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$.

Proof $F_\eta(x) = P(\sigma\xi + m < x) = P(\xi < \frac{x-m}{\sigma}) = \Phi(\frac{x-m}{\sigma})$.

$$f_\eta(x) = F_\eta{}'(x) = \left(\Phi(\frac{x-m}{\sigma})\right)' \cdot \frac{1}{\sigma} = \varphi(\frac{x-m}{\sigma}) \cdot \frac{1}{\sigma} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

The graph of the cumulative distribution functions can be seen in Fig.g.10. In all cases $m = 0$, the red line is for $\sigma = 1$, the yellow line is for $\sigma = 2$, the blue line is for $\sigma = 4$ and the green line is for $\sigma = 0.5$.

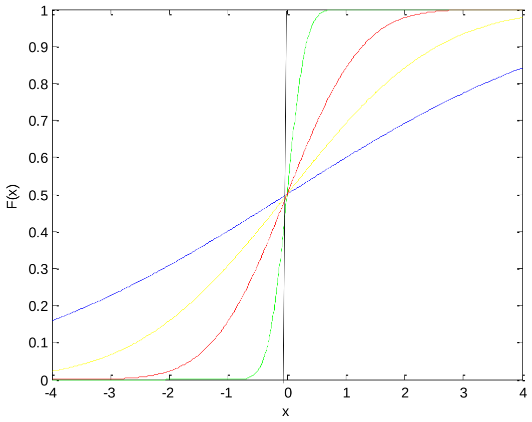


Figure g.10. The cumulative distribution functions for normally distributed random variables for different values of σ

The graph of the probability density functions be seen in Fig.g.11. In all cases $m=0$, the red line is for $\sigma=1$, the yellow line is for $\sigma=2$, the blue line is for $\sigma=4$ and the green line is for $\sigma=0.5$.

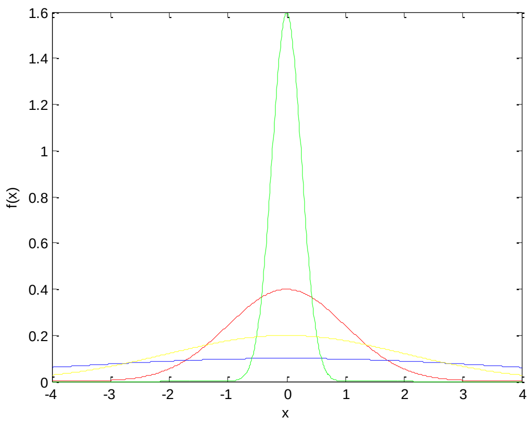


Figure g.10. The probability density functions for normally distributed random variables for different values of σ

One can notice that if the value of σ is large, then the curve is flat, if the value of σ is small, then the curve is peaky. It is the obvious consequence of the fact that the peak is at height of $\frac{1}{\sqrt{2\pi}\sigma}$.

If we want to present the role of the parameter m, then we can notice that the probability density function is symmetric to m. In Fig.g.11., the parameter σ equals 1, red line is for $m=0$, blue line is for $m=1$ and green line is for $m=-1$.

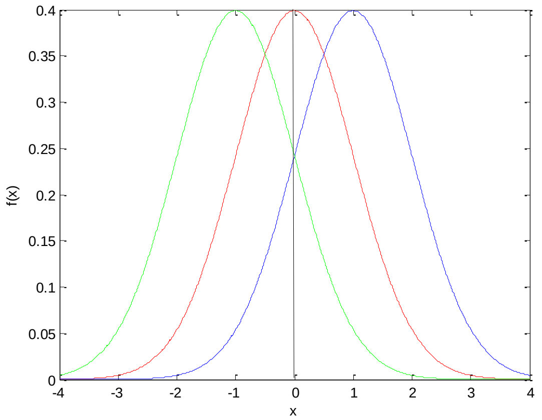


Figure g.11. The probability density functions for normally distributed random variables for different values of m

Numerical characteristics of normally distributed random variables:

Expectation

If $\eta \sim N(m,\sigma)$, then $E(\eta)=m$. It follows from the fact that $E(\eta)=E(\sigma\cdot\xi+m)=\sigma E(\xi)+m=\sigma\cdot 0+m=m$.

Dispersion

If $\eta \sim N(m,\sigma)$, then $D(\eta)=\sigma$. To prove it, take into consideration that $D(\eta)=D(\sigma\cdot\xi+m)=\sigma D(\xi)=\sigma\cdot 1=\sigma$.

Summarizing, the first parameter is the expectation, the second one is the dispersion.

Mode

Local maximum of $f_\eta(x)$ is at $x=m$, consequently the mode is m.

Median

$me=m$. We have to find the value x for which $F_\eta(x)=0.5$.This means $\Phi(\frac{x-m}{\sigma})=0.5$. It implies $\frac{x-m}{\sigma}=0 \Rightarrow x=m$.

Example

E2. Let $\eta \sim N(5,2)$. Compute the probability that η is less than 0.

$P(\eta<0)=F_\eta(x)=\Phi(\frac{0-5}{2})=\Phi(-2.5)=1-\Phi(2.5)=1-0.9938=0.0062$.

Compute the probability that the value of η is between 0 and 6.

$P(0<\eta<6)=F_\eta(6)-F_\eta(0)=\Phi(\frac{6-5}{2})-\Phi(\frac{0-5}{2})=\Phi(0.5)-\Phi(-2.5)=0.6915-0.0062=0.6853$.

Compute the probability that the value of η is greater than 6.

$P(6<\eta)=1-F_\eta(6)=1-\Phi(\frac{6-5}{2})=1-\Phi(0.5)=1-0.6915=0.3085$.

At most how much is the value of η with probability 0.8?

x=? $P(\eta\le x)=0.8$. $\Phi(\frac{x-5}{2})=0.8$. Since $\Phi(0.84)\approx 0.8$, therefore $\frac{x-5}{2}=0.84$. This implies $x=5+0.84\cdot 2=6.68$.

At least how much is the value of η with probability 0.98?

x=? $P(x\le\eta)=0.98$. $P(x\le\eta)=1-\Phi(\frac{x-5}{2})=0.98$. $\Phi(\frac{x-5}{2})=0.02$. If we introduce the new variable $y=\frac{x-5}{2}$, we reduce our task to determine the solution of $\Phi(y)=0.02$. This type of problem was previously solved. We can first realize that y is negative and if $y=-a$, then $\Phi(a)=0.98$. Consequently, $a=2.06$, $y=-2.06$, that is $\frac{x-5}{2}=-2.06$. Finally, arranging the equation we get $x=5-2.06\cdot 2=0.88$.

Compute the value of the probability density function at 6.

$$f_\eta(6)=\frac{1}{\sqrt{2\pi}}e^{\frac{-(6-5)^2}{2\cdot 2^2}}=0.352 \ .$$

<u>Theorem</u> (k times σ rule) If $\eta\sim N(m,\sigma)$, then $P(m-k\sigma<\eta<m+k\sigma)=2\Phi(k)-1$.

<u>Proof</u> The proof is very simple, just compute the probability.

$P(m-k\sigma<\eta<m+k\sigma)=F_\eta(m+k\sigma)-F_\eta(m-k\sigma)=\Phi(\frac{m+k\sigma-m}{\sigma})\ \ \Phi(\frac{m-k\sigma-m}{\sigma})=$

$\Phi(k)-\Phi(-k)=\Phi(k)-(1-\Phi(k))=2\Phi(k)-1$.

<u>Remarks</u>

- Substituting the values $k=0,1,2,3$ into the previous formula, we get

$P(m-\sigma<\eta<m+\sigma)=2\Phi(1)-1=2\cdot 0.8413-1=0.6826$,

$P(m-2\sigma<\eta<m+2\sigma)=2\Phi(2)-1=2\cdot 0.9772-1=0.9544$,

$P(m-3\sigma<\eta<m+3\sigma)=2\Phi(3)-1=2\cdot 0.9987-1=0.9974$.

- The last equality states that a normally distributed random variable takes its values in the interval which is symmetric to the expectation and has radius 3 times dispersion with probability almost 1.
- The probability density function with parameters $m=1$ and $\sigma=1$, for k=1,2 present the k times σ rule (see Fig.g.12.).

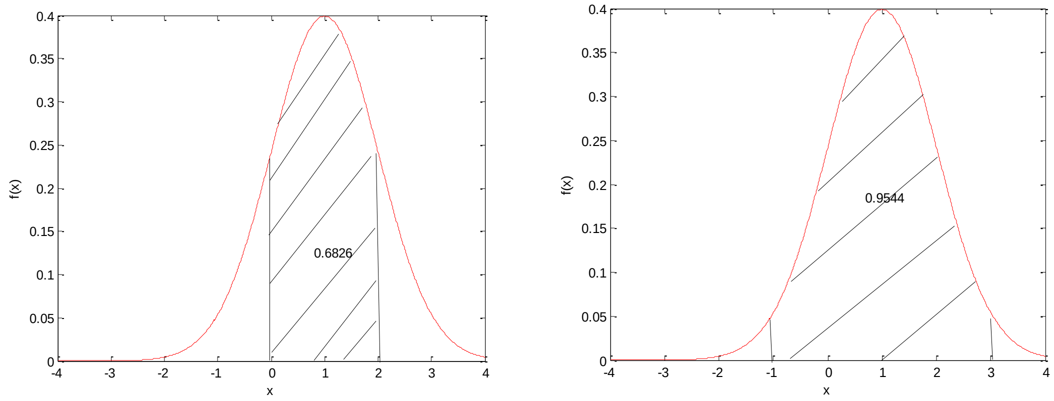


Figure g.12.The areas under the probability density function

Example

E3. Let $\eta \sim N(3,12)$. Determine an interval, symmetric to 3, in which the values of η are situated with probability 0.99!

Apply the "k times σ rule". As the required probability equals 0.99, consequently, $2\Phi(k)-1=0.99$. This implies $\Phi(k)=0.995$, and as a consequence, $k=2.58$. Therefore the interval has the form $(m-k\sigma, m+k\sigma)=(3-12\cdot 2.58,\quad 3+12\cdot 2.58)=(-27.96,\quad 33.96)$. It is also presented in Fig.g.13.

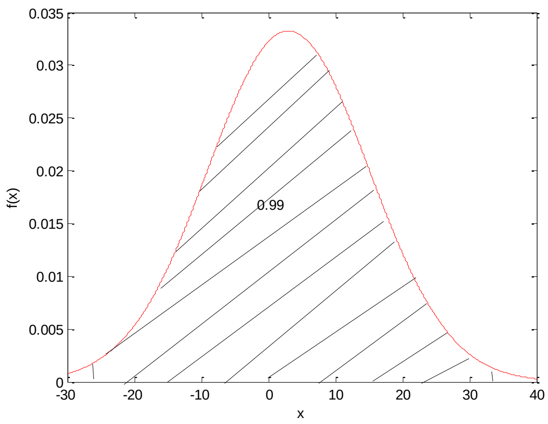


Figure g.13. The area 0.99 under the probability density function

Theorem If η is a normally distributed random variable, then so is its linear transformation. Namely, if $\eta \sim N(m,\sigma)$, $a \neq 0$, then $\theta = a\eta + b \sim N(a\cdot m + b, |a|\cdot\sigma)$.

Proof

Recall the definition of the normally distributed random variable, $\eta = \sigma\cdot\xi + m$ with $\xi \sim N(0,1)$. $\theta = a\eta + b = a(\sigma\cdot\xi + m) + b = a\sigma\xi + am + b$. If $0 < a$, then $\theta \sim N(am+b, a\sigma)$, if $a < 0$, then $\theta \sim N(am+b, -a\sigma)$. Summarizing these formulas we get the statement to be proved.

Theorem If $\eta_1 \sim N(m_1, \sigma_1)$, $\eta_2 \sim N(m_2, \sigma_2)$ furthermore η_1 and η_2 are independent, then $\eta_1 + \eta_2 \sim N(m_1 + m_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.

Remarks

- Although we can not prove the previous statement, notice, that the parameters are calculated according to the properties of expectation and variance. The first parameter is the expectation. Expectation of the sum is the sum of expectations. The second parameter is the dispersion. Dispersions can not be given, but variances can. $D^2(\eta_1 + \eta_2) = D^2(\eta_1) + D^2(\eta_2)$, therefore $D(\eta_1 + \eta_2) = \sqrt{\sigma_1^2 + \sigma_2^2}$.
- As a consequence of the previous statement we emphasize the following: If ξ_i $i = 1,2,3,...,n$ are independent identically distributed random variables, $\xi_i \sim N(m, \sigma)$, then $\boxed{\sum_{i=1}^{n} \xi_i \sim N(n \cdot m, \sigma \cdot \sqrt{n})}$.
- If ξ_i $i = 1,2,3,...,n$ are independent identically distributed random variables, $\xi_i \sim N(m, \sigma)$, then $\boxed{\frac{\sum_{i=1}^{n} \xi_i}{n} \sim N\left(m, \frac{\sigma}{\sqrt{n}}\right)}$.

Example

E4. The weights of adults are normally distributed random variables with expectation 75 kg and dispersion 10 kg. The weights of 5 year old children are also normally distributed random variables with expectation 18 kg and dispersion 3 kg. Compute the probability that the average weight of 20 adults is less than 70 kg.

$\xi_a \sim N(75,10)$, $\xi_c \sim N(18,2)$. $\frac{\sum_{i=1}^{20} \xi_{a,i}}{20} \sim N(75, \frac{10}{\sqrt{20}})$,

$$P(\frac{\sum_{i=1}^{20} \xi_{a,i}}{20} < 70) = F_{\frac{\sum_{i=1}^{20} \xi_{a,i}}{20}}(70) = \Phi(\frac{70 - 75}{2.236}) = \Phi(-2.236) = 1 - \Phi(2.236) = 1 - 0.9873 = 0.0127 .$$

Give an interval symmetric to 75 kg in which the average weight of 10 adults is with probability 0.9.

$\frac{\sum_{i=1}^{20} \xi_{a,i}}{10} \sim N(75, \frac{10}{\sqrt{10}})$. To answer the question apply the "k times σ rule" with expectation 75 and dispersion $10/\sqrt{10}$. $2\Phi(k) - 1 = 0.9$ implies $k = 1.645$, therefore the required interval has the form $(75 - 1.645 \cdot 3.16,\quad 75 + 1.645 \cdot 3.16) = (69.8,\quad 80.2)$.

At most how much is the total weight of 6 adults in the elevator with probability 0.98?

x=? $P(\sum_{i=1}^{6} \xi_{a,i} < x) = 0.98$. $\sum_{i=1}^{6} \xi_{a,i} \sim N(6 \cdot 75, \sqrt{6} \cdot 10)$. It means that $\Phi_{\sum_{i=1}^{6} \xi_{a,i}}(x) = 0.98$.

$\Phi(\frac{x - 450}{24.495}) = 0.98$. Consequently, $\frac{x - 450}{24.495} = 2.06$, finally

$x = 450 + 24.495 \cdot 2.06 = 500.46\text{kg} \approx 500\,\text{kg}$.

Compute the probability that the total weight of an adult and a 5 year old child is more than 100 kg, if their weights are independent.

$\xi_a + \xi_c \sim N(75 + 18, \sqrt{10^2 + 3^2})$,

$P(100 < \xi_a + \xi_c) = 1 - F_{\xi_a + \xi_c}(100) = 1 - \Phi(\frac{100 - 93}{10.44}) = 1 - 0.7486 = 0.2514$.

E5. The daily return of a shop is a normally distributed random variable with expectation 1 million HUF and dispersion 0.2 million HUF. Suppose that the returns belonging to different days are independent random variables. Compute the probability that there is at most 0.1 million HUF difference between the returns of two different days.

Let ξ_1 denote the return of the first day, ξ_2 denote the return of the second day. $\xi_1 \sim N(1,0.2)$, $\xi_2 \sim N(1,0.2)$. The question is $P(|\xi_1 - \xi_2| < 0.1)$.

$P(|\xi_1 - \xi_2| < 0.1) = P(-0.1 < \xi_1 - \xi_2 < 0.1) = F_{\xi_1 - \xi_2}(0.1) - F_{\xi_1 - \xi_2}(-0.1)$.

If we knew the cumulative distribution function of $\xi_1 - \xi_2$, then we could substitute 0.1 and -0.1 into it.

As $\xi_1 - \xi_2 = \xi_1 + (-\xi_2)$, furthermore $-\xi_2 \sim N(-1, 0.2)$, $\xi_1 - \xi_2 \sim N(1 - 1, \sqrt{0.2^2 + 0.2^2})$.

Consequently, $\xi_1 - \xi_2 \sim N(0, 0.283)$. This implies $F_{\xi_1 - \xi_2}(x) = \Phi(\frac{x - 0}{0.283})$.

Finally, $P(|\xi_1 - \xi_2| < 0.1) = \Phi(\frac{0.1}{0.283}) - \Phi(\frac{-0.1}{0.283}) = 2\Phi(\frac{0.1}{0.283}) - 1 = 2 \cdot 0.6381 - 1 = 0.2762$.

Compute the probability that the return of a fixed day is less than the 80% of the return of another day.

$P(\xi_1 < 0.8 \cdot \xi_2) = ?$ $P(\xi_1 < 0.8 \cdot \xi_2) = P(\xi_1 - 0.8 \cdot \xi_2 < 0) = F_{\xi_1 - 0.8\xi_2}(0)$.

If we knew the cumulative distribution function of $\xi_1 - 0.8\xi_2$, then we could substitute 0 into it.

$\xi_2 \sim N(0.8 \cdot 1, 0.8 \cdot 0.2)$, $-\xi_2 \sim N(-0.8 \cdot 1, 0.8 \cdot 0.2)$.

$\xi_1 - 0.8 \cdot \xi_2 \sim N(1 - 0.8, \sqrt{0.2^2 + (0.8 \cdot 0.2)^2})$. Consequently, $\xi_1 - 0.8 \cdot \xi_2 \sim N(0.2, 0.256)$.

Now we can finish the computations as follows:

$P(\xi_1 < 0.8 \cdot \xi_2) = F_{\xi_1 < 0.8 \cdot \xi_2}(0) = \Phi(\frac{0 - 0.2}{0.256}) = \Phi(-0.78) = 0.2173$.

g.4. Further random variables derived from normally distributed ones

In statistics, there are many other distributions which originate from normal ones. Actually we investigate the chi-square and Student's t distributions. We will use them in chapter j, as well.

Definition Let $\xi \sim N(0,1)$. Then $\theta = \xi^2$ is called a **chi-squared distributed random variable with degree of freedom 1** and it is denoted by $\theta \sim \chi_1^2$

Theorem The cumulative distribution function of $\theta = \xi^2$ is $F_\theta(x) = \begin{cases} 0 & \text{if } x \le 0 \\ 2\Phi(\sqrt{x}) - 1 & \text{if } 0 < x \end{cases}$.

The probability density function of η is $f_{\theta}(x)=\begin{cases}0 & \text{if } x<0\\ \frac{1}{\sqrt{2\pi}}e^{\frac{-x}{2}}\cdot\frac{1}{\sqrt{x}} & 0<x\end{cases}$.

<u>Proof</u>

All of the values of χ_1^2 are nonnegative, consequently, $F_{\chi_1^2}(x)=0$, if $x\leq 0$. For positive x values,

$$F_{\theta}(x)=P(\theta<x)=P(\xi^2<x)=P(-\sqrt{x}<\xi<\sqrt{x})=F_{\xi}(\sqrt{x})-F_{\xi}(-\sqrt{x})=\Phi(\sqrt{x})-\Phi(-\sqrt{x})$$

$$=2\Phi(\sqrt{x})-1.$$

$$f_{\theta}(x)=(F_{\theta})'(x)=\begin{cases}0 \text{ if } x<0\\ 2\cdot\Phi'(\sqrt{x})\cdot\left(\sqrt{x}\right)'=2\cdot\frac{1}{\sqrt{2\pi}}e^{\frac{-(\sqrt{x})^2}{2}}\cdot\frac{1}{2\sqrt{x}}=\frac{1}{\sqrt{2\pi}\cdot\sqrt{x}}e^{-\frac{x}{2}} & \text{if } 0<x\end{cases}.$$

The graph of the above cumulative distribution function and the probability density function can be seen in Fig. g.14.

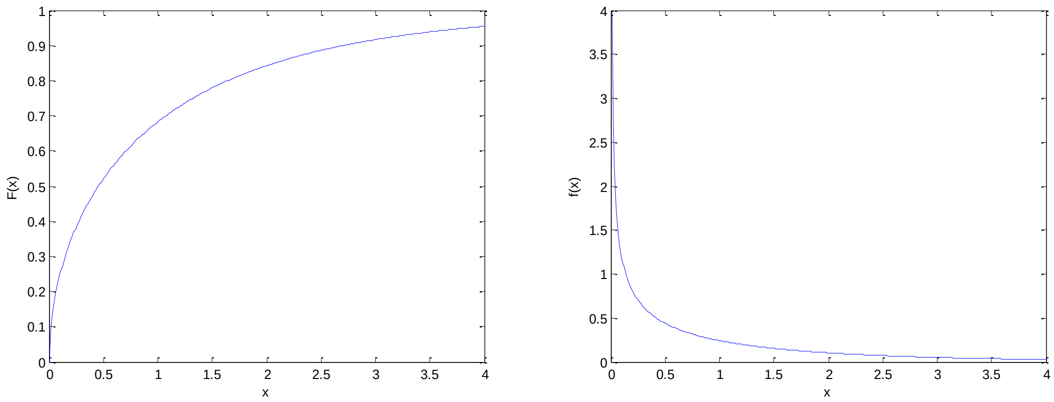


Figure g.14. The graphs of the cumulative distribution function and the probability density function of χ_1^2 distributed random variables

<u>Numerical characteristics of chi-squared distributed random variables with degree of freedom 1:</u>

<u>Expectation</u>

$E(\theta)=1$, which is a straightforward consequence of $E(\xi^2)=D^2(\xi)+(E(\xi))^2=1+0=1$.

<u>Dispersion</u>

$D(\theta)=\sqrt{2}$, which can be computed by partial integration.

<u>Mode</u>

There is no local maximum for the probability density function.

<u>Median</u>

$me=0.675$. We have to solve the equation $2\Phi(\sqrt{x})-1=0.5$, that is $\Phi(\sqrt{x})=0.75$. It is satisfied by $\sqrt{x}=0.675$, x=0.456.

Definition Let $\xi_i \sim N(0,1), i = 1,2,3,.., n$, and let ξ_i be independent. Then $\theta = \sum_{i=1}^{n} \xi_i^2$ is called a **chi-squared distributed random variable with degree of freedom n** and is denoted by $\theta \sim \chi_n^2$

Theorem

The probability density function of a χ_n^2 distributed random variable is

$$f_\theta(x) = \begin{cases} \dfrac{x^{\frac{n}{2}-1} e^{\frac{-x}{2}}}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} & \text{if} \quad 0 < x \\ 0 & \text{otherwise} \end{cases}.$$

The function Γ is the generalization of the factorial for non-integer values. $\Gamma(0.5) = \sqrt{\pi}$, furthermore $\Gamma(x+1) = x \cdot \Gamma(x)$.

The graph of the probability density function of a χ_n^2 distributed random variable with degree of freedom $n = 5$ can be seen in Fig.g.15.

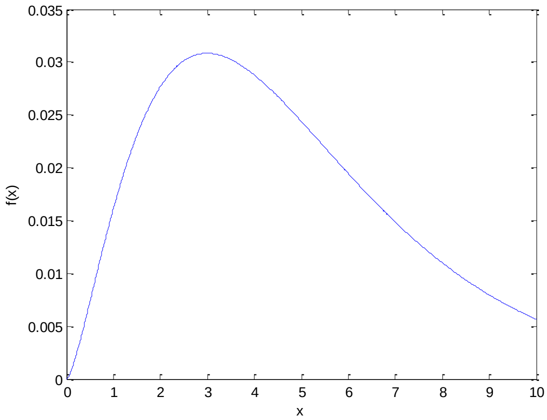


Figure g.15. The graph of the probability density function of a χ_5^2 distributed random variables

Remarks

- For n=2, the probability density function coincides with that of an exponentially distributed random variable with parameter $\lambda = 0.5$.
- For general values of n, the explicit form of the cumulative distribution function of χ_n^2 is quite complicated, it is not used usually. The values for which the cumulative distribution function reaches certain levels are included in tables used in statistics. These tables are used in chapter j, as well. For example, if we seek the value x for which $P(\chi_5^2 < x) = 0.95$ holds, we get $x = 11.07$ (see Table 3 at the end of the booklet).

Usually, the real number x for which $P(\theta < x) = \alpha$ holds, can be found in tables and is denoted by $\chi_{n,\alpha}^2$ (see Table 2 at the end of the booklet).

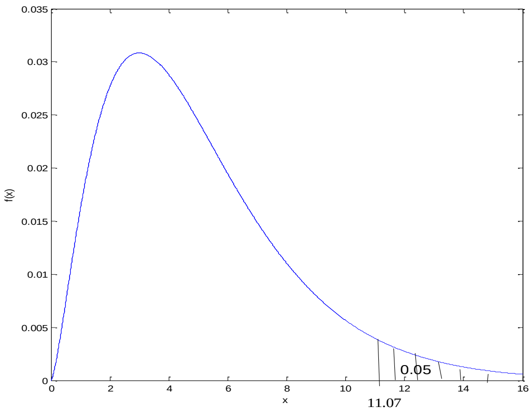


Figure g.15. The value exceeded with probability 0.05 in case of χ_5^2

Numerical characteristics of chi-squared distributed random variables:

Expectation

$E(\theta)=n$, which is a straightforward consequence of $E(\sum_{i=1}^{n}\xi_i^2)=\sum_{i=1}^{n}E(\xi_i^2)=n$.

Dispersion

$D(\theta)=\sqrt{2n}$, which follows from $D^2(\sum_{i=1}^{n}\xi_i^2)=\sum_{i=1}^{n}D^2(\xi_i^2)=2\cdot n$.

Mode

There is no mode if $n\le 2$, and it is $n-2$, if $2<n$.

Median

It can not be expressed explicitly, it is about $n(1-\frac{2}{9n})^3$

Definition Let $\xi_1,\xi_2,\ldots,\xi_n$ and η be independent standard normally distributed random variables. The random variable $\theta=\frac{\eta}{\sum_{i=1}^{n}\xi_i^2}$ is called a **Student's t distributed random variable with degree of freedom n** and is denoted by $\theta\sim\tau_n$.

Theorem

The probability density function of a Student's t distributed random variable with degree of freedom n is $f_n(x)=\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\cdot\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$.

Remarks

- If n is even, then the normalising factor is $\frac{1}{2\sqrt{n}}\frac{(n-1)(n-3)...5\cdot 3}{(n-2)(n-4)...4\cdot 2}$, if n is odd, then it is $\frac{1}{\pi\sqrt{n}}\frac{(n-1)(n-3)...4\cdot 2}{(n-2)(n-4)...5\cdot 3}$.
- If n=1, then $f_1(x)=\frac{1}{\pi}\cdot\frac{1}{1+x^2}$. The random variable with this probability density function is called a Cauchy distributed random variable.
- If $n\to\infty$, then $\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}\to e^{\frac{-x^2}{2}}$, consequently $f_n(x)\to\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}=\varphi(x)$ for any values of x.
- The probability density functions of τ_n distributed random variable can be seen in Fig.g.16.

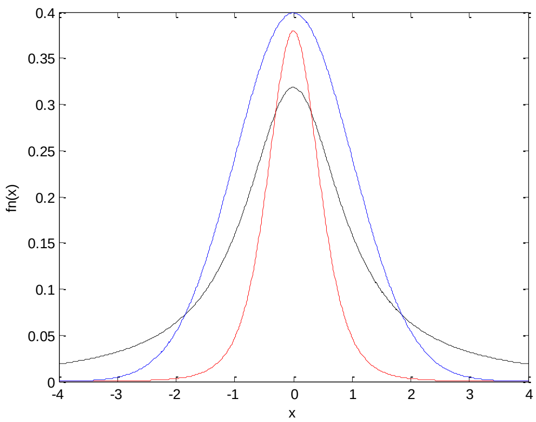


Figure g.16. The probability density functions of τ_n distributed random variable for $n=1$ (black), $n=5$ (red) and $n=100$ (blue)

- The closed form of the cumulative distribution functions do not exist. The values for which the cumulative distribution function reach different levels are included in tables used in statistics (see Table 2 at the end of the booklet). These tables are used in chapter j, as well. Supposing $\theta\sim\tau_n$, the value, for which $P(-x\le\theta\le x)=1-\alpha$ and $P(x<|\theta|)=\alpha$ is usually denoted by $t_{n,\alpha}$. For example, if $\alpha=0.2$ and $n=5$, $t_{5,0.2}=1.476$. It is also presented in Fig. g.17.

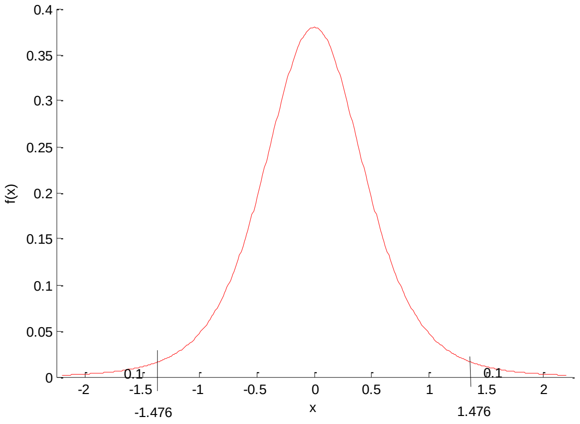


Figure g.17. The bounds for τ_5 distributed random variables with probability 0.8

<u>Numerical characteristics of Student's t distributed random variable:</u>

<u>Expectation</u>

If $\theta \sim \tau_n$, then $E(\theta) = 0$, if $1 < n$. It is the straightforward consequence of the symmetry of the probability density function. If $n = 1$, expectation does not exist.

<u>Dispersion</u>

$D(\theta) = \dfrac{\sqrt{n-2}}{\sqrt{n}}$, if $2 < n$, otherwise it does not exist. It can be computed by partial integration.

<u>Mode</u>

It is always zero.

<u>Median</u>

It is always zero, due to the symmetry of the probability density function.

h. The law of large numbers

The aim of this chapter

In this chapter we present asymptotical theorems which characterize the behaviour of the average of many independent identically distributed random variables. We return to the relative frequency, as well, and we prove that it is about the probability of the event. These theorems are the theoretical basis of the polls and computer simulations.

Preliminary knowledge

Expectation, dispersion and their properties. Binomially distributed random variables.

Content

h.1. Markov's and Chebisev's inequalities.

h.2. The law of large numbers.

h.3. Bernoulli's theorem.

h.1. Markov and Chebisev's inequalities

First we provide estimations for certain probabilities. Although these estimations are quite rough, they are appropriate to be applied for proving asymptotical statements. Their main advantage is that they do not require the knowledge of the distribution of the random variable, they use only the expectation and dispersion.

Theorem (Markov's inequality)
Let ξ be a random variable all of whose values are nonnegative and $E(\xi)$ exists. Then, for any $0<\varepsilon$ the following inequality holds: $P(\xi \geq \varepsilon) \leq \frac{E(\xi)}{\varepsilon}$.

Proof
The proof is based on the following: $\varepsilon \cdot \mathbf{1}_{\xi\geq\varepsilon} \leq \xi$. Recall that $\mathbf{1}_A = \begin{cases} 1 & \text{if A holds} \\ 0 & \text{if A does not hold} \end{cases}$.

This implies $\mathbf{1}_{\xi\geq\varepsilon} = \begin{cases} 1 & \text{if } \xi \geq \varepsilon \text{ holds} \\ 0 & \text{if } \xi < \varepsilon \text{ holds} \end{cases}$.

Multiplying by ε we get $\varepsilon \cdot \mathbf{1}_{\xi\geq\varepsilon} = \begin{cases} \varepsilon & \text{if } \xi \geq \varepsilon \text{ holds} \\ 0 & \text{if } \xi < \varepsilon \text{ holds} \end{cases}$. Taking into account the non-negativity of ξ, this means that $\varepsilon \cdot \mathbf{1}_{\xi\geq\varepsilon} \leq \xi$. Applying the property of expectation that if $\eta_1 \leq \eta_2$ then $E(\eta_1) \leq E(\eta_2)$, we can see that $E(\varepsilon \cdot \mathbf{1}_{\xi\geq\varepsilon}) = \varepsilon \cdot E(\mathbf{1}_{\xi\geq\varepsilon}) \leq E(\xi)$. Recalling that $E(\mathbf{1}_A) = P(A)$ and dividing both sides by $0<\varepsilon$ the inequality becomes $P(\xi \geq \varepsilon) \leq \frac{E(\xi)}{\varepsilon}$. This is the statement to be proved.

Theorem (Chebisev's inequality)
Let η be a random variable those dispersion exists. Then for any $0<\lambda$, the following inequality is satisfied: $P(|\eta - E(\eta)| \geq \lambda) \leq \frac{D^2(\eta)}{\lambda^2}$.

Proof Note that $|\eta - E(\eta)| \geq \lambda$ holds if and only if $(\eta - E(\eta))^2 \geq \lambda^2$. Consequently, $P(|\eta - E(\eta)| \geq \lambda) = P((\eta - E(\eta))^2 \geq \lambda^2)$. Apply Markov's inequality with $\xi = (\eta - E(\eta))^2$ and $\varepsilon = \lambda^2$. The non-negativity obviously holds, and $E(\xi) = E((\eta - E(\eta))^2) = D^2(\eta)$. Therefore, $P(|\eta - E(\eta))| \geq \lambda) = P(\xi \geq \varepsilon) \leq \frac{E(\xi)}{\varepsilon} = \frac{D^2(\eta)}{\lambda^2}$, and this is the statement to be proved.

Remark

- Chebisev's inequality can be also written in the following form: $P(|\eta - E(\eta)| < \lambda) \geq 1 - \frac{D^2(\eta)}{\lambda^2}$. $\{|\eta - E(\eta)| < \lambda)\}$ is the complement of the event $\{|\eta - E(\eta)| \geq \lambda\}$. If $P(A) \leq x$, then $P(\overline{A}) = 1 - P(A) \geq 1 - x$, which implies the statement.

- Chebisev's inequality can be also written as follows: $P(|\eta - E(\eta)| \geq kD(\eta)) \leq \frac{1}{k^2}$

and $P(|\eta - E(\eta)| < kD(\eta)) \geq 1 - \frac{1}{k^2}$. Substitute $\lambda = kD(\eta)$. This can be done with $k = \frac{\lambda}{D(\eta)}$, supposing $D(\eta) \neq 0$. If $D(\eta) = 0$, then on the basis of the property of dispersion, $P(\eta = E(\eta)) = 1$, therefore $P(|\eta - E(\eta)|) \geq kD(\eta)) = 0$ which is less than $\frac{1}{k^2}$ for any value of k.

- The inequality $P(|\eta - E(\eta)| \geq kD(\eta)) \leq \frac{1}{k^2}$ expresses that the random variable η takes its values outside the neighbourhood with radius $kD(\eta)$ of its expectation with probability not larger than $\frac{1}{k^2}$. Large deviation is with small probability.
- The inequality $P(|\eta - E(\eta)| < kD(\eta)) \geq 1 - \frac{1}{k^2}$ states that a random variable η takes its values in the neighbourhood with radius $kD(\eta)$ of its expectation with probability no smaller than $1 - \frac{1}{k^2}$. Small deviation is with large probability.
- The proofs do not use the distribution of the random variable.
- If we know the distribution of η, the probabilities $P(|\eta - E(\eta)| \geq kD(\eta))$ and $P(|\eta - E(\eta)| < kD(\eta))$ can be computed explicitly.

<u>Example</u>

E1. Let η be a Poisson distributed random variable with parameter $\lambda = 2$. Compute the probability that the values of η are in the neighbourhood with radius $D(\eta)$ of its expectation.

$E(\eta) = \lambda = 2$, $D(\eta) = \sqrt{\lambda} = \sqrt{2} = 1.41$. $|\eta - E(\eta)| < D(\eta)$ means that $E(\eta) - D(\eta) < \eta < E(\eta) + D(\eta)$. Explicitly, $2 - \sqrt{2} < \eta < 2 + \sqrt{2}$, that is $0.59 < \eta < 3.41$.

Now $P(0.59 < \eta < 3.41) = P(\eta = 1) + P(\eta = 2) + P(\eta = 3) = \frac{2^1}{1!}e^{-2} + \frac{2^2}{2!}e^{-2} + \frac{2^3}{3!}e^{-2} = 0.722$.

E2. Let η be a uniformly distributed random variable in $[-1,2] = [a, b]$. Compute the probability that η takes its value in the in the neighbourhood with radius $1.5 \cdot D(\eta)$ of its expectation.

$E(\eta) = \frac{a+b}{2} = \frac{-1+2}{2} = 0.5$. $D(\eta) = \frac{b-a}{\sqrt{12}} = \frac{2-(-1)}{\sqrt{12}} = \frac{3}{2\sqrt{3}} = 0.866$. $1.5 \cdot D(\eta) = 1.299$.

The interval is $(0.5 - 1.2999,\ 0.5 + 1.299) = (-0.799,\ \ 1.799)$. The question can be written as $P(\eta \in (-0.799, 1.799)) = P(-0.799 < \eta < 1.799) = F(1.799) - F(-0.799)$. Recalling that

$$F(x) = \begin{cases} 0 & \text{if } x \leq a = -1 \\ \frac{x-a}{b-a} = \frac{x+1}{3} & \text{if } -1 = a < x \leq b = 2, \\ 1 & \text{if } b = 2 \leq x \end{cases}$$

we get $P(-0.799 < \eta < 1.799) = \frac{1.7999+1}{3} - \frac{-0.799+1}{3} = 0.866$.
We note that one can check that the result ends in the same probability independently of the endpoints of the interval $[a, b]$.

E3. Let η be an exponentially distributed random variable. Determine the interval symmetric to expectation of η in which the values of η are situated with probability 0.99.

Let the radius of the interval be $kD(\eta)$. Since $E(\eta) = \frac{1}{\lambda} = D(\eta)$, the interval looks like $\left(\frac{1}{\lambda} - k \cdot \frac{1}{\lambda}, \frac{1}{\lambda} + k \cdot \frac{1}{\lambda}\right)$.

$P(\eta \in \left(\frac{1}{\lambda} - k \cdot \frac{1}{\lambda}, \frac{1}{\lambda} + k \cdot \frac{1}{\lambda}\right)) = P\left(\frac{1}{\lambda} - k \cdot \frac{1}{\lambda} < \eta < \frac{1}{\lambda} + k \cdot \frac{1}{\lambda}\right) = F(\frac{1}{\lambda} + k \cdot \frac{1}{\lambda}) - F(\frac{1}{\lambda} - k \cdot \frac{1}{\lambda})$.

Recalling that $F(x) = \begin{cases} 0 & \text{if } x \le 0 \\ 1 - e^{-\lambda x} & \text{if } 0 < x \end{cases}$, $F(\frac{1}{\lambda} + k \cdot \frac{1}{\lambda}) = 1 - e^{-\lambda(\frac{1}{\lambda} + k \cdot \frac{1}{\lambda})} = 1 - e^{-(1+k)}$.

The value of $F(\frac{1}{\lambda} - k \cdot \frac{1}{\lambda})$ depends on the sign of its argument. One can notice that $\frac{1}{\lambda} - k \cdot \frac{1}{\lambda} < 0$, if $1 < k$ and $0 < \frac{1}{\lambda} - k \cdot \frac{1}{\lambda}$ if $k < 1$. If $k = 1$, then $P(\eta \in (\frac{1}{\lambda} - 1 \cdot \frac{1}{\lambda}, \frac{1}{\lambda} + 1 \cdot \frac{1}{\lambda})) = P(0 < \eta < \frac{2}{\lambda}) = F(\frac{2}{\lambda}) - 0 = 1 - e^{-2} = 0.865 < 0.99$. This implies $1 < k$. Therefore, $F(\frac{1}{\lambda} - k \cdot \frac{1}{\lambda}) = 0$. Consequently,

$P(\eta \in \left(\frac{1}{\lambda} - k \cdot \frac{1}{\lambda}, \frac{1}{\lambda} + k \cdot \frac{1}{\lambda}\right)) = 1 - e^{-(1+k)} = 0.99$, $e^{-(1+k)} = 0.01$, $1 + k = -\ln 0.01 = 4.605$,

$k = 3.605$. As a check, $P(\frac{1}{\lambda} - 3.605 \frac{1}{\lambda} < \eta < \frac{1}{\lambda} + 3.605 \frac{1}{\lambda}) = 1 - e^{-4.605} - 0 = 1 - 0.01 = 0.99$, which was the requirement.
We note that the value of k is independent of the value of the parameter λ .

E4. We do not know the distribution of a random variable η, but we know its expectation and dispersion, $E(\eta) = 200$ and $D(\eta) = 10$. Construct an interval in which the values of η are situated with probability at least 0.95!

According to the Chebisev's inequality. If $1 - \frac{1}{k^2} = 0.95$, then $k = 4.472$, and the interval looks like $(200 - 10 \cdot 4.472, \quad 200 + 10 \cdot 4.472) = (155.28, 244.72)$.

E5. Let η be a binomially distributed random variable with expectation 200 and dispersion 10. Compute the probability that values of η are situated in the neighbourhood of its expectation with radius $4.472 \cdot D(\eta)$.

As η is binomially distributed with parameters n and p, $E(\eta) = n \cdot p = 200$, $D(\eta) = \sqrt{np(1-p)} = 10$, consequently $1 - p = \frac{100}{200} = 0.5$, which implies $p = 0.5$ and $n = 400$. The question is $P(200 - 4.472 \cdot 10 < \eta < 200 + 4.472 \cdot 10)$ $= P(155.28 < \eta < 244.72)$. η takes only nonnegative integer values, hence $P(155.28 < \eta < 244.72) = P(\eta = 156) + P(\eta = 157) + ... + P(\eta = 244)$. As η is a binomially distributed random variable, $P(\eta = k) = \binom{n}{k} p^k \cdot (1-p)^{n-k} = \binom{400}{k} 0.5^k \cdot 0.5^{400-k}$.

$$P(155.28 < \eta < 244.72) = \binom{400}{156} 0.5^{156} \cdot 0.5^{400-156} + \binom{400}{157} 0.5^{157} \cdot 0.5^{400-157} + ... + \binom{400}{244} 0.5^{244} \cdot 0.5^{400-244}$$
$= 0.99999$.

E6. Let η be a random variable with expectation 200 and dispersion 10. Determine the probability that the values of η are situated in the interval $(175, 225)$.
As we do not know the distribution of η, we can not determine the required probability exactly, but we can give an estimation for it. The interval $(175, 225)$ is symmetric to the expectation 200, it can be written as
$(200 - 2.5 \cdot 10, \quad 200 + 2.5 \cdot 10) = (E(\eta) - k \cdot D(\eta), \quad E(\eta) + k \cdot D(\eta))$ with $k = 2.5$.
$P(|\eta - E(\eta)| < kD(\eta)) \geq 1 - \frac{1}{k^2}$ implies $P(175 < \eta < 225) \geq 1 - \frac{1}{2.5^2} = 0.84$.

E7. Let η be binomially distributed random variable with expectation 200 and dispersion 10. Compute the probability that the values of η are situated in the interval $(175, 225)$.
$P(175 < \eta < 225) = P(\eta = 176) + P(\eta = 177) + ... + P(\eta = 224) =$
$\binom{400}{176} 0.5^{176} 0.5^{224} + \binom{400}{177} 0.5^{177} 0.5^{223} + ... + \binom{400}{224} 0.5^{224} 0.5^{176} = 0.9858$, which is much more than the estimation 0.84 given by Chebisev's inequality. We point out that actually we know the distribution of the random variable, and it is an extra information to E6.

E8. Let η be a normally distributed random variable with expectation 200 and dispersion 10. Compute the probability that the values of η are situated in the interval $(175, 225)$.

Now, $\eta \sim N(200, 10)$, and $F(x) = \Phi(\frac{x - 200}{10})$. Now $P(175 < \eta < 225) = F(225) - F(175) =$ $= \Phi(\frac{225 - 200}{10}) - \Phi(\frac{175 - 200}{10}) = \Phi(2.5) - \Phi(-2.5) = 2 \cdot \Phi(2.5) - 1 = 0.9876$. We note that this probability is also much more than the estimation given by Chebisev's inequality due to the extra information of distribution. Furthermore it is close to the probability computed in the previous example. The reason of this latter phenomenon will be given in the next section i.

h.2. The law of large numbers

In this subsection we provide a form the law of large numbers which is easy to prove and which can give estimations for the probability of large deviations. This statement is the basis of computer simulations. One can state stronger forms of the law of large numbers and one also can give statements under weaker assumptions, as well.

<u>Theorem</u> Let $\xi_1, \xi_2, \dots, \xi_n, \dots$ be independent identically distributed random variables with $E(\xi_i) = m$ and $D(\xi_i) = \sigma$. Then , for any $0 < \varepsilon$,

$$P\left(\left| \frac{\sum_{i=1}^{n} \xi_i}{n} - m \right| < \varepsilon \right) \to 1, \text{ if } n \to \infty,$$

and

$$P\left(\left| \frac{\sum_{i=1}^{n} \xi_i}{n} - m \right| \geq \varepsilon \right) \to 0 \text{ if } n \to \infty.$$

<u>Proof</u>

Let $\eta_n = \frac{\sum_{i=1}^{n} \xi_i}{n}$. Now $E\left(\frac{\sum_{i=1}^{n} \xi_i}{n}\right) = m$ and $D\left(\frac{\sum_{i=1}^{n} \xi_i}{n}\right) = \frac{\sigma}{\sqrt{n}}$. Apply Chevisev' inequality for η_n. This gives us $P(|\eta_n - m| > \varepsilon) \leq \frac{D^2(\eta_n)}{\varepsilon^2}$, which implies $P\left(\left| \frac{\sum_{i=1}^{n} \xi_i}{n} - m \right| > \varepsilon \right) \leq \frac{\sigma^2}{n\varepsilon^2}$. As ε and σ are fixed, $\frac{\sigma^2}{n\varepsilon^2} \to 0$, if $n \to \infty$, which coincides with the second part of the statement. The formula $P\left(\left| \frac{\sum_{i=1}^{n} \xi_i}{n} - m \right| \leq \varepsilon \right) \geq 1 - \frac{\sigma^2}{n\varepsilon^2} \to 1 - 0$ is the first part of the statement.

Probability theory and mathematical statistics– Law of large numbers

Example

E1. Let $\xi_1, \ldots \xi_{1000}$ be independent uniformly distributed random variable in $[0,1]$. Give an estimation for the probability $P\left(\left|\frac{\sum_{i=1}^{1000}\xi_i}{1000}-0.5\right|>0.05\right)$.

Apply the above inequality $P\left(\left|\frac{\sum_{i=1}^{n}\xi_i}{n}-m\right|\geq\varepsilon\right)\leq\frac{\sigma^2}{n\varepsilon^2}$.

Now $E(\xi_i)=0.5=m$, $D(\xi_i)=\frac{1}{\sqrt{12}}=0.2887=\sigma$. Substitute $\varepsilon=0.05$,

$\frac{\sigma^2}{n\varepsilon^2}=\frac{1}{12\cdot 1000\cdot 0.05^2}=0.033$.

Consequently, $P\left(\left|\frac{\sum_{i=1}^{1000}\xi_i}{1000}-0.5\right|\geq 0.05\right)\leq 0.033$.

At most how much is the difference between the average and 0.5 with probability 0.95?

The question is the value of ε, for which $P\left(\left|\frac{\sum_{i=1}^{n}\xi_i}{n}-m\right|<\varepsilon\right)=0.95$. As we do not know the exact distribution of $\frac{\sum_{i=1}^{n}\xi_i}{n}$, we can not compute the exact probability, but we are able to estimate the probability. $P\left(\left|\frac{\sum_{i=1}^{n}\xi_i}{n}-m\right|<\varepsilon\right)\geq 1-\frac{\sigma^2}{n\varepsilon^2}$, if $1-\frac{\sigma^2}{n\varepsilon^2}=0.95$, then

$P\left(\left|\frac{\sum_{i=1}^{n}\xi_i}{n}-m\right|<\varepsilon\right)\geq 0.95$ holds. $1-\frac{\sigma^2}{n\varepsilon^2}=0.95$ implies $\frac{1}{12\cdot 1000\cdot 0.05}=\varepsilon^2$, consequently

$\varepsilon^2=1.6667\times 10^{-3}$, $\varepsilon=0.041$.

How many random variables have to be averaged in order to assure that the difference between the average and 0.5 should be at most 0.01 with probability 0.98?

The question is the value of n for which $P\left(\left|\frac{\sum_{i=1}^{n}\xi_i}{n}-m\right|<0.01\right)=0.98$. Applying the formula

$P\left(\left|\frac{\sum_{i=1}^{n}\xi_i}{n}-m\right|<\varepsilon\right)\geq 1-\frac{\sigma^2}{n\varepsilon^2}$ again, substitute $1-\frac{\sigma^2}{n\varepsilon^2}=0.98$ and $\varepsilon=0.01$.

$$\frac{\sigma^2}{\varepsilon^2\cdot 0.01}=\frac{1}{12\cdot 0.01^2\cdot 0.02}=n,\ n=41667\ .$$

How many random variables have to be averaged in order to assure that the difference between the average and 0.5 should be at most 0.005 with probability 0.98?

If $\varepsilon=0.005$, then , $n=1.6667\times 10^5$, which is four times larger than the previous number of experiments. If we want to increase the accuracy to the half, we need 2^2 times more experiments.

<u>Remark</u>

- If we fix the accuracy ε, and the value of n, then $P\left(\left|\frac{\sum_{i=1}^{n}\xi_i}{n}-m\right|<\varepsilon\right)\geq 1-\frac{\sigma^2}{n\varepsilon^2}$

gives us an estimation for the probability that the maximal difference between the average and the expectation exceeds the accuracy.

Probability theory and mathematical statistics– Law of large numbers

- If we fix the probability $1-\alpha$ (reliability) and the value of n, then $1-\frac{\sigma^2}{n\varepsilon^2}\geq 1-\alpha$

implies $\varepsilon \leq \sqrt{\frac{\sigma^2}{n\cdot\alpha}}$. Consequently, the accuracy is proportional to the square root of the reciprocal of the number of experiments.

- If we fix the probability $1-\alpha$ (reliability) and the accuracy ε, then $1-\frac{\sigma^2}{n\varepsilon^2}\leq 1-\alpha$

implies $\frac{\sigma^2}{\varepsilon^2\alpha}\leq n$. This means that the number of experiments is proportional to the square of reciprocal of the accuracy.

- As an illustration of the law of large numbers, we present Table h.1. The random variables were uniformly distributed in $[0,1]$, the reliability level was fixed as $1-\alpha=0.95$ and $1-\alpha=0.99$. The table shows that the difference between the average and the expectation is getting smaller and smaller as the number of simulations was increased. The total requested time was less than 1 minute. The theoretical accuracy $\varepsilon=\sqrt{\frac{\sigma^2}{n\cdot 0.05}}$ and $\varepsilon=\sqrt{\frac{\sigma^2}{n\cdot 0.01}}$ were computed for the reliability levels 0.95 and 0.99, respectively.

n	$\frac{\sum_{i=1}^{n}\xi_i}{n}$	$\left\lvert\frac{\sum_{i=1}^{n}\xi_i}{n}-0.5\right\rvert$	$\sqrt{\frac{\sigma^2}{n\cdot 0.05}}$	$\sqrt{\frac{\sigma^2}{n\cdot 0.01}}$
10	0.432756065694353	0.067243934305647	0.11785	0.2635
100	0.530898496906201	0.030898496906201	0.03 7268	0.0833
1000	0.506786612848606	0.006786612848606	0.011785	0.02635
10000	0.496156685345852	0.003843314654148	0.003 7268	0.00833
100000	0.500349684591498	0.000349684591498	0.0011785	0.002635
1000000	0.500158856526807	0.000158856526807	0.0003 7268	0.000833
10000000	0.499726933610529	0.000273066389471	0.00011785	0.0002635
100000000	0.499951340487525	0.000048659512475	0.000037268	0.0000833
1000000000	0.499985939301628	0.000014060698372	0.000011785	0.00002635

Table h.1. The averages and their differences from the expectation in case of uniformly distributed random numbers

Secondly, the random variables were exponentially distributed with expectation 0.1 and 10. Table h.2. shows that the difference between the average and the expectation depends on the value of the parameter. The parameter is the reciprocal of the dispersion, consequently, the larger the dispersion, the larger the difference.

	$\lambda = 0.1$	$\lambda = 0.1$	$\lambda = 10$	$\lambda = 10$
N	$\frac{\sum_{i=1}^{n} \xi_i}{n}$	$\left\lvert \frac{\sum_{i=1}^{n} \xi_i}{n} - 10 \right\rvert$	$\frac{\sum_{i=1}^{n} \xi_i}{n}$	$\left\lvert \frac{\sum_{i=1}^{n} \xi_i}{n} - 0.1 \right\rvert$
10	6.2277618964331	3.7722381035668	0.09447373893621	0.055276
100	11.756814668520	1.7568146685202	0.10392000570707	0.00392
1000	9.5670585169631	0.4329414830368	0.09696619091756	0.00304
10000	9.9932193771582	0.0067806228417	0.100150679660307	0.00015
100000	9.9708942677258	0.0291057322741	0.100629035751288	0.00063
1000000	9.9943200370807	0.0056799629192	0.100039656754390	0.00004
10000000	10.003113268035	0.0031132680354	0.099950954820648	0.00004
100000000	9.9994289522126	0.00057104778736	0.100000507690485	0.00000005
100000000	10.000097147933	0.00009714793369	0.100000729791939	0.00000007

Table h.2. The averages and their differences from the expectation in case of exponentially distributed random numbers

- The law of large numbers is expressed by the sentence that the expectation is **about** the average of many values of random variable. Not exactly the same, but it is not far from it.
- As the expectation is an integral, the law of large numbers provides the possibility to compute integrals numerically as follows: Let $g: H \to R$, $H \subset R$, $[a,b] \subset H$, suppose that g is continuous in $[a,b]$. Taking into account the properties of the expectation, $I = \int_a^b g(x)dx = (b-a)\int_a^b g(x) \cdot \frac{1}{b-a} dx = (b-a) \cdot E(g(\eta))$, where η is a uniformly distributed random variable in $[a,b]$. $E(g(\eta))$ is about the average of many values of $g(\eta)$. η can be constructed as a linear transformation of a uniformly distributed random variable in [0,1]. Consequently, the algorithm of computing the approximate value of the integral $\int_a^b g(x)dx$ is the following: generate a random number, multiply it by $b-a$ and add "a", then

substitute this value into the function g. Substitution can be made as the all the values we get are in the domain of g. Repeat the process n times and take the average of the values. Multiplying the average by $b-a$ we get the approximate value of the integral. The necessary number of simulations can be determined as follows:

$$P\left(\left|\int_a^b g(x)dx-(b-a)\cdot\frac{\sum_{i=1}^n g(\eta_i)}{n}\right|<\varepsilon\right)=P\left(\left|\frac{\int_a^b g(x)dx}{(b-a)}-\frac{\sum_{i=1}^n g(\eta_i)}{n}\right|<\frac{\varepsilon}{b-a}\right)\geq$$

$$\geq 1-(b-a)^2\frac{D^2(g(\eta_i))}{n\cdot\varepsilon^2}=1-\alpha\,.$$

As η_i is in [a,b], $D(g(\eta_i))\leq\frac{\max_{a\leq x\leq b} g(x)-\min_{a\leq x\leq b} g(x)}{2}$.

$$1-\alpha=P\left(\left|\int_a^b g(x)dx-(b-a)\cdot\frac{\sum_{i=1}^n g(\eta_i)}{n}\right|<\varepsilon\right)\geq 1-(b-a)^2\frac{\left(\max_{a\leq x\leq b} g(x)-\min_{a\leq x\leq b} g(x)\right)^2}{4\cdot n\cdot\varepsilon^2},$$ which

implies $(b-a)^2\cdot\frac{\left(\max_{a\leq x\leq b} g(x)-\min_{a\leq x\leq b} g(x)\right)^2}{4\alpha\cdot\varepsilon^2}\leq n$.

Example

E2. Compute $\int_0^1\frac{1}{1+x}dx$ by random simulation.

Notice that $\int_0^1\frac{1}{1+x}dx=E(\frac{1}{1+\xi})$ where ξ is a uniformly distributed random variable in [0,1]. Consequently, generate random numbers with the computer, add 1, and take the reciprocal. This process has to be repeated many times. Take the average of the numbers you get, and this average is the approximate value of the integral. As $\xi\in([0,1]$, $\frac{1}{\xi+1}\in[0.5,1]$, $D^2(\frac{1}{1+\xi})\leq\frac{0.5^2}{4}=0.0625$. If we fix the reliability level $1-\alpha=0.99$, the necessary number of simulation is $\frac{0.0625}{\varepsilon^2\cdot 0.01}\leq n$. If we would like to compute the integral with difference less than 0.01, then we have to make $0.0625\cdot 10^6=62500\leq n$ simulations. As $\int_0^1\frac{1}{1+x}dx=[\ln(1+x)]_{x=0}^{x=1}=\ln 2-\ln 1=\ln 2$, we can follow the difference between the exact value and the approximate value of the integral in Table h.3.

ε is computed as $\sqrt{\frac{\left(\frac{\max_{a\leq x\leq b} g(x)-\min_{a\leq x\leq b} g(x)}{2}\right)^2}{n\cdot\alpha}}=\sqrt{\frac{0.0625}{0.01\cdot n}}=\varepsilon$.

N	average	Difference	ε
62	0.702627791231423	0.009480610671478	0.3175
625	0.694214696993436	0.001067516433491	0.1
6250	0.695502819777260	0.002355639217315	0.03175
62500	0.693417064411419	0.000269883851474	0.01
625000	0.693095119363388	0.000052061196558	0.003175
6250000	0.693134534818101	0.000012645741844	0.001
62500000	0.693167969772721	0.000020789212776	0.0003175
625000000	0.693142704368027	0.000004476191918	0.0001

Table h.3. The averages and their differences from the expectation in case of transformed random variables

For all the simulations, the elapsed time was 42.9 seconds.

E3. Compute the value of the integral $\int_1^3 \sin\frac{1}{x}dx$ with accuracy 0.01.

Note, that $\int_1^3 \sin\frac{1}{x}dx = 2\cdot\int_1^3 \sin\frac{1}{x}\cdot\frac{1}{2}dx = 2\cdot E(\eta)$, where $\eta=\sin(\frac{1}{\xi})$ and ξ is uniformly distributed random variable in $[1,3]$. $-1\le\sin\frac{1}{x}\le 1$, $D^2(\sin\frac{1}{\eta})\le\frac{(1-(-1))^2}{4}=1$,

$$P\left(\left|2\cdot\frac{\sum_{i=1}^{n}\frac{1}{1+\eta_i}}{n}-\int_1^3\sin\frac{1}{x}dx\right|\le\varepsilon\right)\ge 1-4\cdot\frac{1}{n\varepsilon^2}.$$ $1-4\cdot\frac{1}{n\varepsilon^2}=0.99$ and $\varepsilon=0.01$ implies

$n=4000000$ .We can follow the average and the theoretical accuracy as the function of the number of simulations in Table g.4. Elapsed time, together for all simulations, was 36.82 seconds.

n	average	ε
40	4.044413814196310	3.162
400	3.124480498240279	1
4000	3.266154820794264	0.3162
40000	3.241221397791890	0.1
400000	3.252187207202902	0.03162
4000000	3.251025444611742	0.01
40000000	3.251126290354754	0. 003162
400000000	3.250561315440294	0.001

Table g.4. Averages of random variables given by $\eta=\sin(\frac{1}{\xi})$ and the theoretical accuracy

We note that better estimations for the variance can be also given, we used $-1\le\sin y\le 1$ for the sake of simplicity.

E4. Compute $\int_{-100}^{100} e^{\frac{-x^2}{2}}dx$ by random simulation.

Note that $\int_{-100}^{100} e^{\frac{-x^2}{2}} dx = 200 \cdot E(e^{\frac{-\xi^2}{2}})$ where ξ is a uniformly distributed random variable in $[-100, 100]$. As $0 \le e^{\frac{-\eta^2}{2}} \le 1$ $D^2(e^{\frac{-\xi^2}{2}}) \le \frac{1}{4}$,

$P(\left| 200 \cdot \frac{\sum_{i=1}^{n} e^{\frac{-\eta_i^2}{2}}}{n} - \int_{-100}^{100} e^{\frac{-x^2}{2}} dx \right| < \varepsilon) \ge 1 - 200^2 \frac{1}{4n \cdot \varepsilon^2}$. $1 - 10000 \frac{1}{4n \cdot \varepsilon^2} \ge 0.99$ implies $n \ge 25000000$. Since from the standard normal probability density function we know that $\int_{-100}^{100} e^{\frac{-x^2}{2}} dx \approx \int_{-\infty}^{\infty} e^{\frac{-x^2}{2}} dx \approx \sqrt{2\pi}$, comparing the average to $\sqrt{2\pi}$ we get Table h.5.:

n	average	Difference	ε
25	8.323342326487701	5.816714051856701	10
250	3.015562934762770	0.508934660131769	3.16227
2500	2.264787314861209	0.241840959769791	1
25000	2.441972159407621	0.064656115223379	0.316227
250000	2.451752388622218	0.054875886008782	0.1
2500000	2.511696184700974	0.005067910069974	0.0316227
25000000	2.508097777785709	0.001469503154709	0.01
250000000	2.504753761626246	0.001874513004754	0.00316227

Table h.5. Averages of the transformed random variable and their differences from $\sqrt{2\pi}$ in case of different numbers of simulations

We can see that the actual difference is always smaller than the theoretical accuracy.

h.3. Bernoulli's theorem

In this subsection we apply the law of large numbers to characteristically distributed random variables and we get a statement for relative frequencies. This statement tells us that the relative frequency of an event A is close to the probability of A.

Theorem (Bernoulli's theorem) Let A be an event, and $k_A(n)$ be the frequency of the event performing n independent experiments. Then, for any $0 < \varepsilon$, $P(\left|\frac{k_A(n)}{n} - P(A)\right| \ge \varepsilon) \to 0$ if $n \to \infty$ and $P(\left|\frac{k_A(n)}{n} - P(A)\right| < \varepsilon) \to 1$ supposing $n \to \infty$.

Proof Recall that $k_A(n)$ is a binomially distributed random variable with parameters n and $p = P(A)$, and $k_A(n)$ can be written as a sum of n independent characteristically distributed random variables $\mathbf{1}_A^i$ with parameter p. $E(\mathbf{1}_A^i) = p = P(A)$, $D^2(\mathbf{1}_A^i) = \sqrt{p(1-p)}$, consequently, $P\left(\left|\frac{k_A(n)}{n} - P(A)\right| < \varepsilon\right) \geq 1 - \frac{p(1-p)}{n\varepsilon^2} \to 1 - 0$ supposing $n \to \infty$ and $P\left(\left|\frac{k_A(n)}{n} - P(A)\right| \geq \varepsilon\right) \leq \frac{p(1-p)}{n\varepsilon^2} \to 0$ supposing $n \to \infty$.

Remarks

- The above statement tells us that large difference between the relative frequency and the probability occurs with small probability, small difference occurs with large probability.
- Roughly spoken, the relative frequency is about the probability, if the number of simulations is large. This is the theoretical background of computer simulations and pools.
- $0 \leq p(1-p) \leq \frac{1}{4}$, consequently $P\left(\left|\frac{k_A(n)}{n} - P(A)\right| < \varepsilon\right) \geq 1 - \frac{1}{4n\varepsilon^2}$. This inequality provides the possibility to estimate the necessary number of simulations.
- If we fix the number of simulations and the accuracy (ε), we can estimate the probability that the difference between the relative frequency and the probability exceeds the accuracy ε.
- If we fix the number of simulations and the reliability ($1-\alpha$), we can compute the accuracy ε by $1 - \frac{1}{4n\varepsilon^2} \geq 1 - \alpha$, $\varepsilon \leq \frac{1}{\sqrt{4n\alpha}}$.
- If we fix the reliability ($1-\alpha$) and the accuracy ε, we can determine the necessary number of simulations by $\frac{1}{4\alpha\varepsilon^2} \leq n$.

Examples

E1. To illustrate the above statement we present the following simulation example: flip a fair coin four times and determine the probability that there are both heads and tails among the results.

Of course our computer can not flip a coin but it can generate a random number uniformly distributed on $[0,1]$. Imagine that if the result (random number) is less than 0.5, then we get a head, in the opposite case we get a tail. Repeat it four times and decide whether the results of flips are the same in all cases or there are at least one head and at least one tail. Repeat the composite experiment n times and compute how many times you get both heads and tails. The relative frequency is about the probability. If we would like to approximate the probability of the event "you get both heads and tails " with accuracy $\varepsilon = 0.01$ with probability 0.99, we need $\frac{1}{4\alpha\varepsilon^2} = \frac{1}{4 \cdot 0.01 \cdot 0.01^2} = 250000 \leq n$ experiments. The relative frequencies arising from simulations and their differences from the exact probability $\frac{14}{16}$ can be seen in Table h.6. One can notice that the real difference is much smaller than the accuracy showing that the estimation is not sharp. We can see better estimations in the next chapter.

n	Relative frequency	Difference	ε
25	0.960000000000000	0.085000000000000	1
250	0.892000000000000	0.017000000000000	0.3162
2500	0.874800000000000	0.000200000000000	0.1
25000	0.873840000000000	0.001160000000000	0.03162
250000	0.875000000000000	0	0.01
2500000	0.875041600000000	0.000041600000000	0.003162
25000000	0.875081200000000	0.000081200000000	0.001
250000000	0.874980140000000	0.000019860000000	0.003162

Table h.6. Relative frequencies and their differences from the exact probability

The computer program is very simple and the elapsed time is small. The program for simulation was written in MatLab and it can be seen as follows:

```
function szim16
format long
tic
er=zeros(8,1)
for j=1:1:8
  jo=0;
  for i=1:1:(2.5*10^j);
   head=0;
     for k=1:1:4
      vel=rand(1);
      if vel<0.5
      head=head+1;
      end
     end
     if 0<head & head<4
      jo=jo+1;
     end
   end
 szim=jo/(2.5*10^j);
 er(j,1)=szim;
end
toc
er
kul=abs(er-14/16)
```

The relative frequencies and their differences from the exact probability are plotted in Fig.h.1. and Fig.h.2. with $n = 2.5 \cdot 10^k$.

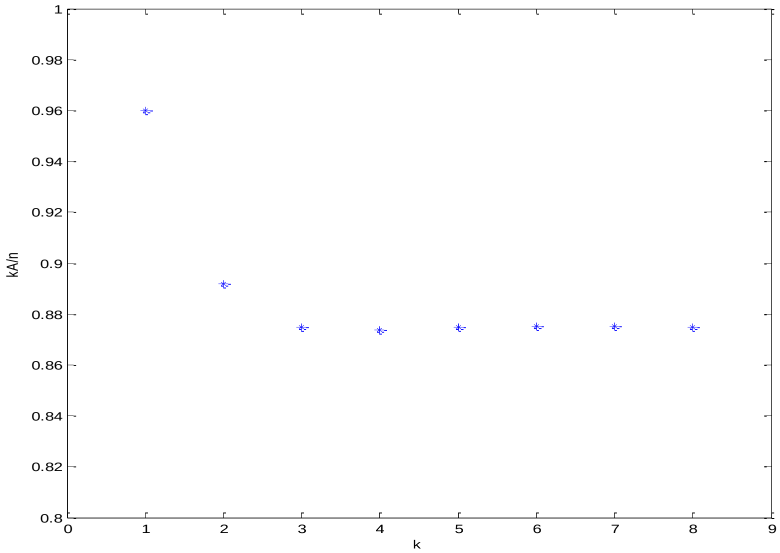


Figure h.1. Relative frequencies as a function of the number of simulations on logarithmic scale

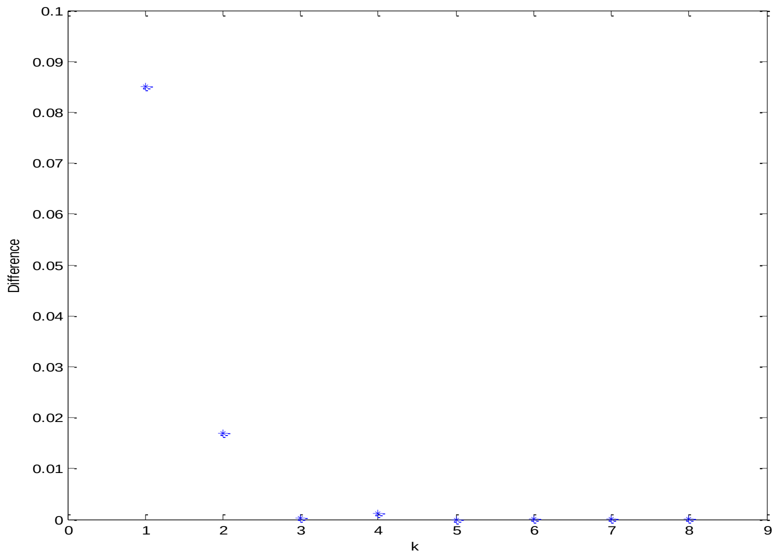


Figure h.2. Differences of the relative frequencies and the probability as a function of the number of simulations on logarithmic scale

Of course, it is easy to find events whose probability is complicated to compute but a computer program for the simulation is easy to elaborate. In those cases the approximation of the probability by relative frequencies is a useful tool for people who are able to apply informatics.

i. Central limit theorem

The aim of this chapter

In this chapter we present asymptotical theorems about the distribution of the sum and the average of many independent identically distributed random variables. We will approximate the cumulative distribution functions and probability density functions with the help of those of normal distributions.

Preliminary knowledge

Convergence of functions. Cumulative distribution function, normal distribution, properties of expectation, dispersion.

Content

i.1. Central limit theorem for the sum of independent identically distributed random variables.

i.2. Moivre-Laplace formula.

i.3. Central limit theorem for the average of independent identically distributed random variables.

i.4. Central limit theorem for relative frequency.

i.1. Central limit theorem for the sum of independent identically distributed random variables

In the previous section we have dealt with the difference of the average of many independent identically distributed random variables and their expectation. We have proved that the difference is small with large probability, if the number of random variables is large. In this chapter we deal with the distribution of the sum and the average of many independent random variables. We state that they are approximately normally distributed. We use this theorem for computations, as well.

Theorem (Central limit theorem) Let $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ be independent identically distributed random variables with expectation $E(\xi_i) = m$ and dispersion $D(\xi_i) = \sigma$, $i = 1,2,\ldots$. Then,

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^{n}\xi_i - nm}{\sigma\sqrt{n}} < x\right) = \Phi(x) \text{ for any } x \in R.$$

The proof of the theorem requires additional tools in probability theory and analysis, consequently we omit it.

Remarks

- $P\left(\frac{\sum_{i=1}^{n}\xi_i - nm}{\sigma\sqrt{n}} < x\right)$ is the value of the cumulative distribution function of the random variable $\frac{\sum_{i=1}^{n}\xi_i - nm}{\sigma\sqrt{n}}$ at the point x.
- $$E\left(\frac{\sum_{i=1}^{n}\xi_i - nm}{\sigma\sqrt{n}}\right) = \frac{1}{\sigma\sqrt{n}} E\left(\sum_{i=1}^{n}\xi_i - nm\right) = \frac{1}{\sigma\sqrt{n}}\left(E\left(\sum_{i=1}^{n}\xi_i\right) - nm\right) = \frac{1}{\sigma\sqrt{n}}(nm - nm) = 0.$$
  $$D\left(\frac{\sum_{i=1}^{n}\xi_i - nm}{\sigma\sqrt{n}}\right) = \frac{1}{\sigma\sqrt{n}} D\left(\sum_{i=1}^{n}\xi_i - nm\right) = \frac{1}{\sigma\sqrt{n}} D\left(\sum_{i=1}^{n}\xi_i\right) = \frac{\sigma\sqrt{n}}{\sigma\sqrt{n}} = 1.$$
- The random variable $\frac{\sum_{i=1}^{n}\xi_i - nm}{\sigma\sqrt{n}}$ is usually called as the standardized sum.
- The central limit theorem states that the limit of the cumulative distribution function of the random variables $\frac{\sum_{i=1}^{n}\xi_i - nm}{\sigma\sqrt{n}}$ equals the cumulative distribution function of standard normally distributed random variables. Consequently, for large values of n, the cumulative

distribution function of the standardized sum is approximately the function Φ. It can be written in the form $F_{\frac{\sum_{i=1}^{n}\xi_i - nm}{\sigma\sqrt{n}}}(x) \approx \Phi(x)$.

- The distribution of ξ_i can be arbitrary. In practice, the approximation is good for $100 \le n$, and many times for $30 \le n$.
- The relative frequencies of the standardized sums can be seen in the following Figs.i.1, i.2. and i.3., if we sum up n=1, n=2, n=5, n=10, $n = 30$, n=100 independent random variables. The random variables were uniformly distributed in $[0,1]$. The red line is the graph of the probability density function of standard normal distribution. One can see that the shape of histogram follows more and more the shape of the Gauss curve.

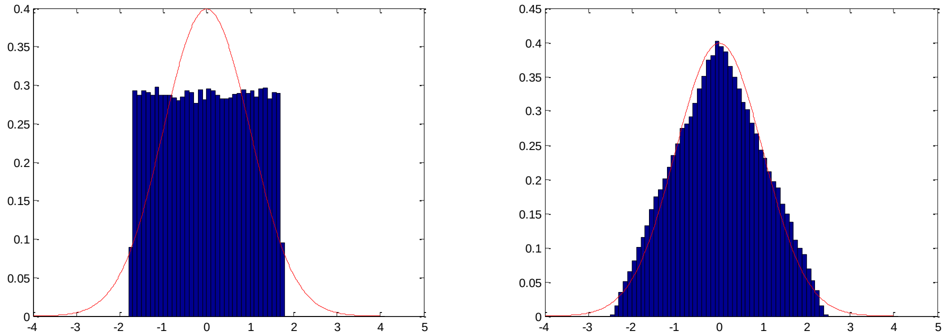


Figure i.1. The relative frequencies of the values of the standardized sums if we sum up n=1 and n=2 random variables

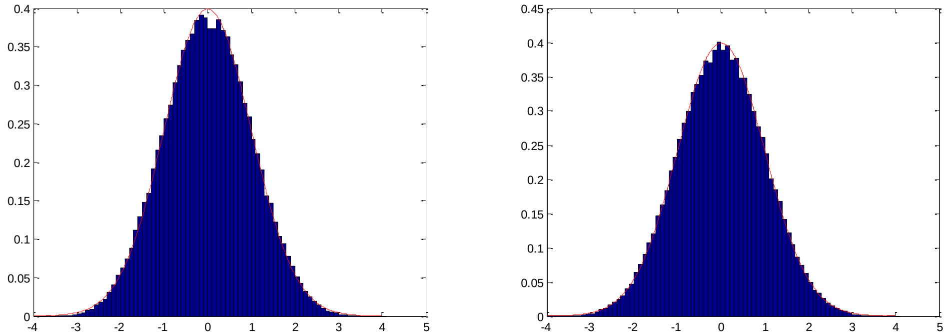


Figure i.2. The relative frequencies of the values of the standardized sums if we sum up n=5 and n=10 random variables

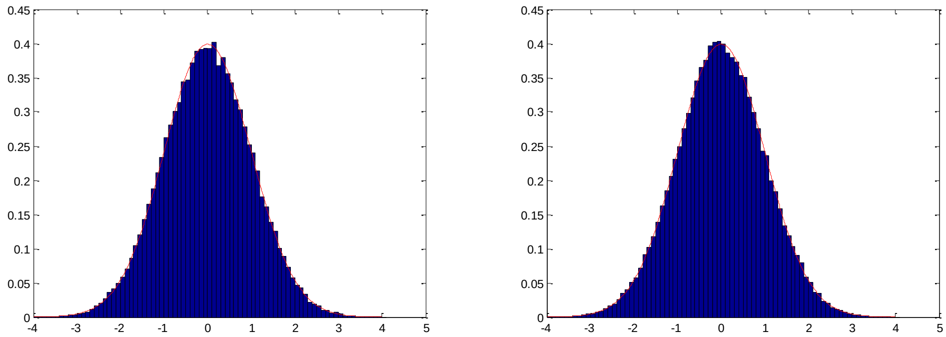


Figure i.3. The relative frequencies of the values of the standardized sums if we sum up n=30 and n=100 random variables

- The distribution of ξ_i can be arbitrary. In Figs. i.4., i.5. and i.6. the relative frequencies of the standardized sum of n exponentially distributed random variables with expectation $E(\xi_i) = 1 = \frac{1}{\lambda}$ $(n = 1, 2, 5, 10, 30, 100)$ are presented. One can realize that the shape of the Gauss curve appears for larger values of n than previously, due to the asymmetry of the exponential probability density function.

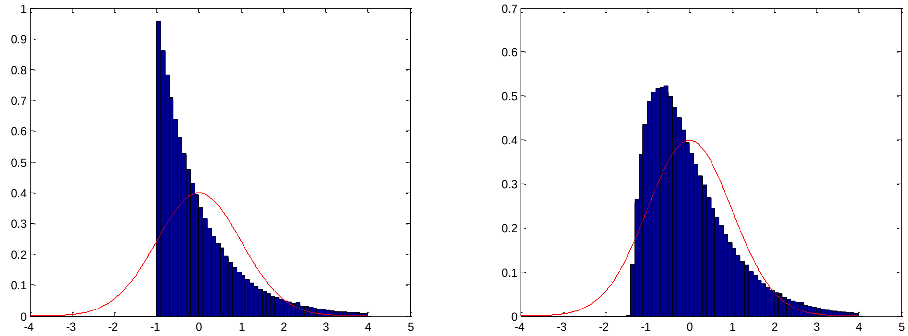


Figure i.4. The relative frequencies of the values of the standardized sums of exponentially distributed random variables, if we sum up n=1 and n=2 random variables

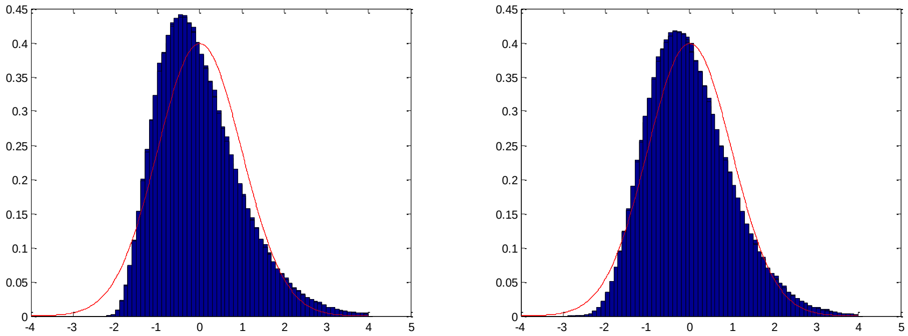


Figure i.5. The relative frequencies of the values of the standardized sums of exponentially distributed random variables, if we sum up n=5 and n=10 random variables

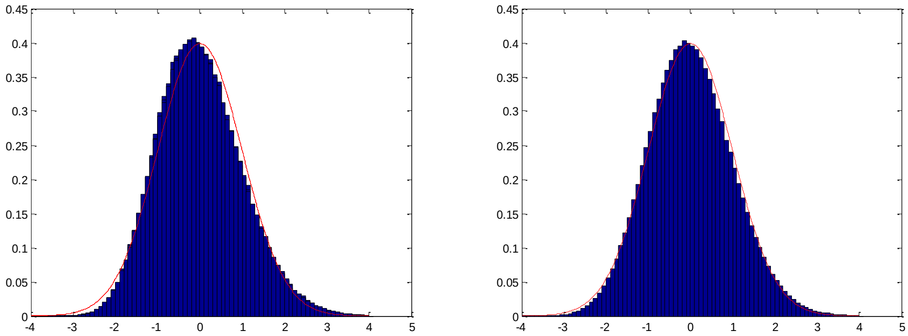


Figure i.6. The relative frequencies of the values of the standardized sums of exponentially distributed random variables, if we sum up n=30 and n=100 random variables

- Finally we illustrate the central limit theorem in the case when $\theta_i \sim N(0,1)$, and $\xi_i = \theta_i^2$, that is $\sum_{i=1}^{n} \xi_i \sim \chi_n^2$. The standardized sums are approximately normally distributed random variables. We note that many program languages have a random number generator which provides normally distributed random variables, as well.

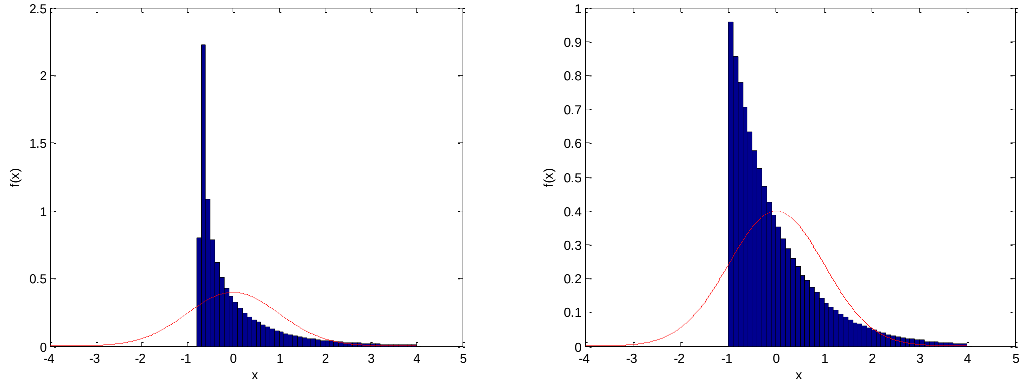


Figure i.7. The relative frequencies of the values of chi-squared distributed random variables with degree of freedom n=1 and n=2

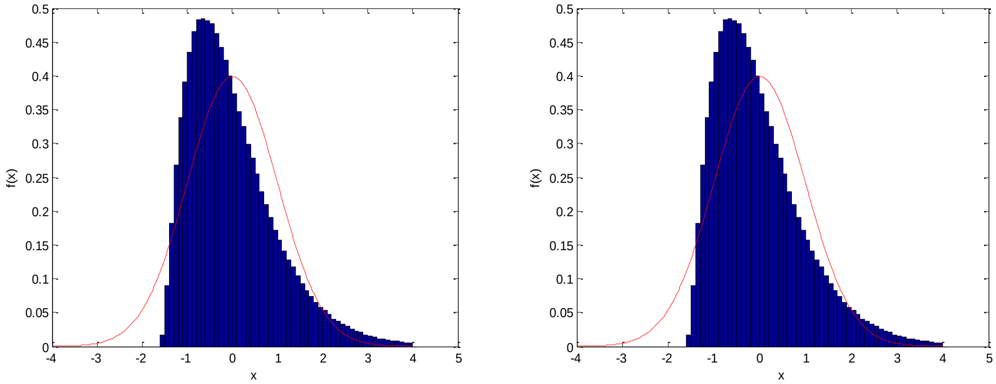


Figure i.8. The relative frequencies of the values of chi-squared distributed random variables with degree of freedom n=5 and n=10

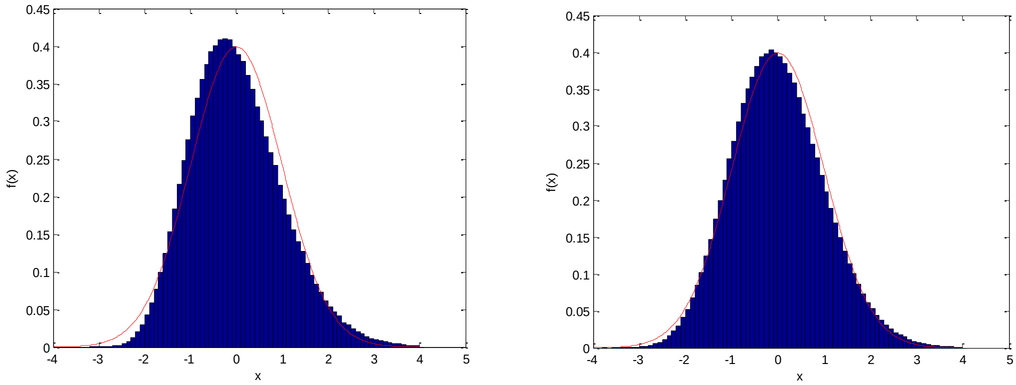


Figure i.9. The relative frequencies of the values chi-squared distributed random variables with degree of freedom n=30 and n=100

After the illustrations we consider what can be stated about the distribution function of the sums, without standardization.

Remark

- The cumulative distribution function of the sum $\sum_{i=1}^{n}\xi_i$ is about that of the normal distribution function with expectation $n\cdot m$ and dispersion $\sigma\sqrt{n}$, that is $F_{\sum_{i=1}^{n}\xi_i}(x)\approx\Phi(\frac{x-nm}{\sigma\sqrt{n}})$. This can be proved as follows:

$F_{\sum_{i=1}^{n}\xi_i}(y)=P(\sum_{i=1}^{n}\xi_i<y)=P(\frac{\sum_{i=1}^{n}\xi_i-nm}{\sigma\sqrt{n}}<\frac{y-nm}{\sigma\sqrt{n}})\approx\Phi(\frac{y-nm}{\sigma\sqrt{n}})$, which coincides with the cumulative distribution function of $\eta\sim N(nm,\sigma\sqrt{n})$. We emphasize that $E(\sum_{i=1}^{n}\xi_i)=nm$ and $D(\sum_{i=1}^{n}\xi_i)=\sigma\sqrt{n}$.

Examples

E1. Flip a fair coin. If the result is a head, then you gain 10 HUF, if the result is a tail, you pay 8 HUF. Applying the central limit theorem, compute the probability that after 100 games you are in loss. Determine the same probability by computer simulation.

Let ξ_i be the gain during the ith game. $\xi_i\sim\begin{pmatrix}-8 & 10\\ 0.5 & 0.5\end{pmatrix}$, $i=1,2,...,100$. ξ_i are independent, identically distributed random variables. Moreover, $E(\xi_i)=-8\cdot\frac{1}{2}+10\cdot\frac{1}{2}=1$, $D(\xi_i)=\sqrt{(-8)^2\cdot\frac{1}{2}+10^2\cdot\frac{1}{2}-1^2}=9$. The question is the probability $P(\sum_{i=1}^{100}\xi_i<0)$. Recall that $P(\sum_{i=1}^{100}\xi_i<0)=F_{\sum_{i=1}^{100}\xi_i}(0)$. According to the central limit theorem, $F_{\sum_{i=1}^{100}\xi_i}(x)\approx\Phi(\frac{x-100\cdot 1}{9\cdot\sqrt{100}})$, consequently,

$F_{\sum_{i=1}^{100}\xi_i}(0)\approx\Phi(\frac{0-100\cdot 1}{9\cdot\sqrt{100}})=\Phi(-1.111)=1-\Phi(1.111)=0.1336$.

In order to approximate the probability by relative frequency with accuracy 0.001, according to the previous section, we need 25000000 simulations. After making the required number of simulations, we get $\frac{k_A(n)}{n}=0.13568732$ which is quite close to the approximate value obtained by the central limit theorem.

E2. Supposing the previous game, how many games have to be played in order not to be in negative with probability 0.99?

Our question is the value of n for which $P(\sum_{i=1}^{n}\xi_i \geq 0) = 0.99$. This question can be expressed with the cumulative distribution function of the sum as follows: n=? $1 - F_{\sum_{i=1}^{n}\xi_i}(0) = 0.99$. As $F_{\sum_{i=1}^{n}\xi_i}(x) \approx \Phi(\frac{x - n\cdot 1}{9\cdot\sqrt{n}})$, we have to solve the equation $\Phi(\frac{0-n}{9\cdot\sqrt{n}}) = 0.01$. This was detailed in the subsection of normally distributed random variables in subsection g.3. $\Phi(y) = 0.01$ implies $y = -2.3263$, therefore $\frac{0 - n\cdot 1}{9\cdot\sqrt{n}} = -2.3263$, $n = 438.35$, that is $n = 439$. As a control, performing the simulation 25000000 times, the relative frequency was 0.98914.

E3. The accounts in the shops are rounded to 0 or 5. If the finial digit of the account equals 0, 1, 2, 8, or 9 than the money to be paid ends in 0. If the finial digit of the account equals 3, 4, 5, 6, or 7, then the money to be paid ends in 5. Suppose that all final digits are equally probable and they are independent during different payments. Applying the central limit theorem, determine the probability that the loss of the shop due to 300 payments is at least -30 and less than 30!

Let the ξ_i $i = 1,2,3,...,300$ be the loss of the shop during the ith payment. $\xi_i \sim \begin{pmatrix} -2 & -1 & 0 & 1 & 2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{pmatrix}$, which are independent identically distributed random variables. The total loss during 300 payments equals $\sum_{i=1}^{300}\xi_i$. The question is $P(-30 \leq \sum_{i=1}^{300}\xi_i < 30)$ which can be expressed with the cumulative distribution function of $\sum_{i=1}^{100}\xi_i$ as follows: $P(-30 \leq \sum_{i=1}^{300}\xi_i < 30) = F_{\sum_{i=1}^{300}\xi_i}(30) - F_{\sum_{i=1}^{300}\xi_i}(-30)$. According to the central limit theorem, $F_{\sum_{i=1}^{300}\xi_i}(x) \approx \Phi(\frac{x - 300\cdot m}{\sigma\sqrt{300}})$, where

$m = E(\xi_i) = -2\cdot 0.2 - 0.1\cdot 0.2 + 0\cdot 0.2 + 1\cdot 0.2 + 2\cdot 0.2 = 0$ and

$\sigma = D(\xi_i) = \sqrt{(-2)^2\cdot 0.2 + (-1)^2\cdot 0.2 + 0^2\cdot 0.2 + 2^2\cdot 0.2 + 1^2\cdot 0.2 - 0^2} = \sqrt{2}$.

Consequently, $F_{\sum_{i=1}^{300}\xi_i}(30) \approx \Phi(\frac{30-0}{\sqrt{2}\sqrt{300}}) = 0.88966$,

$F_{\sum_{i=1}^{300}\xi_i}(-30) \approx \Phi(\frac{-30-0}{\sqrt{2}\sqrt{300}}) = 1 - 0.88966 = 0.11034$ and

$P(-30 \leq \sum_{i=1}^{300}\xi_i < 30) = F_{\sum_{i=1}^{300}\xi_i}(30) - F_{\sum_{i=1}^{300}\xi_i}(-30) \approx 0.88966 - 0.11034 = 0.77932 \approx 0.8$.

Give an interval in which the loss is situated with probability 0.99.

The interval in which a normally distributed random variable with parameters m= 0 and $\sigma = \sqrt{600}$ takes its values with probability 0.99 is $(-63.1, 63.1)$. Therefore the loss is between -63.1 and 63.1 with probability 0.99. Notice that the loss may be -300, it is in a loose interval with large probability. This fact is appropriate for checking based on random phenomenon.

E4. Throw a fair die 1000 times. At least how much is the sum of the results with probability 0.95?

Let the result of the ith throw be denoted by ξ_i, i=1,2,…,1000. Now $\xi_i \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}$, which are independent identically distributed random variables with expectation $E(\xi_i) = 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + 3\cdot\frac{1}{6} + 4\cdot\frac{1}{6} + 5\cdot\frac{1}{6} + 6\cdot\frac{1}{6} = 3.5 = m$ and dispersion $D(\xi_i) = \sqrt{1^2\cdot\frac{1}{6} + 2^2\cdot\frac{1}{6} + 3^2\cdot\frac{1}{6} + 4^2\cdot\frac{1}{6} + 5^2\cdot\frac{1}{6} + 6^2\cdot\frac{1}{6} - 3.5^2} = 1.7078 = \sigma$.

The central limit theorem states that $F_{\sum_{i=1}^{n}\xi_i}(x) \approx \Phi(\frac{x - n\cdot 3.5}{1.7078\sqrt{n}})$. The question is the value of x for which $P(\sum_{i=1}^{1000}\xi_i \geq x) = 0.95$, that is $1 - F_{\sum_{i=1}^{n}\xi_i}(x) = 0.95$. Solving the equation $1 - \Phi(\frac{x - 1000\cdot 3.5}{1.7078\cdot\sqrt{1000}}) = 0.95$, $x = 3411.2$. Summarizing, the sum of 1000 throws is at least 3412 with probability 0.95. Although we do not know what happens during one experiment, the sum of 1000 experiments can be well predicted.

i.2. Moivre-Laplace formula

The Moivre-Laplace formula is a special form of the central limit theorem, the form applied to the cumulative distribution function of binomially distributed random variables.

<u>Theorem</u> (Moivre-Laplace formula) Let $k_A(n)$ be the frequency of the event A ($P(A) = p,\quad 0 < p < 1$) during $2 \leq n$ independent experiments, that is $k_A(n)$ is binomially distributed random variable with parameters n and p. Then, for any $x \in R$, $\lim_{n\to\infty} P(\frac{k_A(n) - np}{\sqrt{np(1-p)}} < x) = \Phi(x)$.

<u>Proof</u> Recall that $\mathbf{k_A}(n) = \sum_{i=1}^{n}\mathbf{1}_A^i$ with

$$\mathbf{1}_A^i = \begin{cases} 1 & \text{if A occurs during the ith experiments} \\ 0 & \text{if A does not occur during the ith experiments} \end{cases}.$$

$\mathbf{1}_A^i$ $i = 1,2,...$ are independent, characteristically distributed random variables with parameter p, $E(\mathbf{1}_A^i) = p$, $D(\mathbf{1}_A^i) = \sqrt{p(1-p)}$. Applying the central limit theorem we get the statement to be proved.

<u>Remarks</u>

- $P(\eta < x)$ equals the cumulative distribution function of $k_A(n)$ at point x.
- $E(k_A(n)) = np$, $D(k_A(n)) = \sqrt{np(1-p)}$.
- The Moivre-Laplace formula states that $F_{\frac{k_A(n)-np}{\sqrt{np(1-p)}}}(x) \approx \Phi(x)$.

- $F_{k_A(n)}(x) \approx \Phi(\frac{x-np}{\sqrt{np(1-p}})$, which can be proved as follows:

$$F_{k_A(n)}(y) = P(k_A(n) < y) = P(\frac{k_A(n)-np}{\sqrt{np(1-p)}} < \frac{y-np}{\sqrt{np(1-p)}}) \approx \Phi(\frac{y-np}{\sqrt{np(1-p)}}).$$

- For any $a < b$,

$$P(a \le k_A(n) < b) = F_{k_A(n)}(b) - F_{k_A(n)}(a) \approx \Phi(\frac{b-np}{\sqrt{np(1-p)}}) - \Phi(\frac{a-np}{\sqrt{np(1-p)}}).$$

- The approximation is good if $100 \le n$ and $10 \le np$.
- $P(k_A(n) = k) = P(k \le k_A(n) < k+1) = F_{k_A(n)}(k+1) - F_{k_A(n)}(k) \approx$

$$\approx \Phi(\frac{(k+1)-np}{\sqrt{np(1-p)}}) - \Phi(\frac{k-np}{\sqrt{np(1-p)}}).$$

Consequently, $P(k_A(n) = k) = \binom{n}{k} p^k (1-p)^{n-k}$ can be approximated with the help of the cumulative distribution function of a normally distributed random variable. The differences between the exact and the approximate values can be seen in Fig.i.10. The values of parameters are $n = 100$ and $p = 0.1$. Largest difference between the exact and the approximate values is less then 0.01.

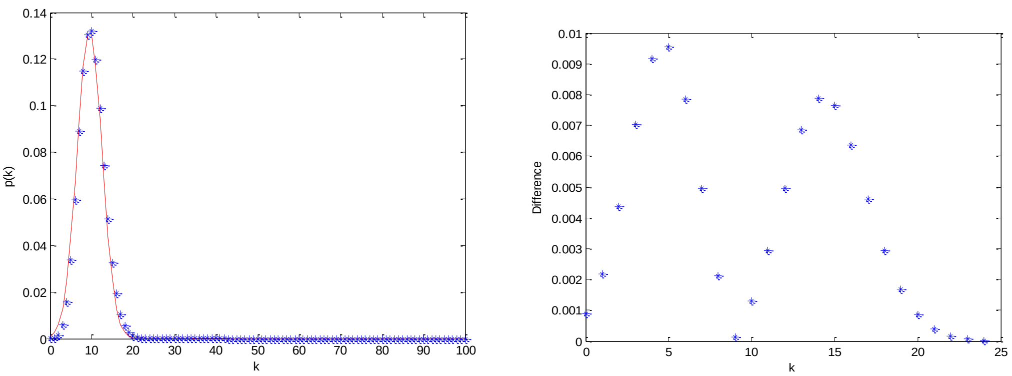


Figure i.10. The exact and the approximate probabilities and their differences in case of binomial distribution

- $P(k_A(n) = k) = \binom{n}{k} p^k (1-p)^{n-k}$ can be also approximated with the help of

the probability density function of normally distributed random variables. From analysis one can recall that if the function G is continuously differentiable in $[a,b]$, then $G(b) - G(a) = G'(c)(b-a)$, for some $c \in (a,b)$. Applying this theorem for $a = k$ and $b = k+1$ we get $P(k \le k_A(n) < k+1) = F_{k_A(n)}(k+1) - F_{k_A(n)}(k) \approx$

$$\Phi(\frac{k+1-np}{\sqrt{np(1-p)}}) - \Phi(\frac{k-np}{\sqrt{np(1-p)}}) = \Phi'(\frac{c-np}{\sqrt{np(1-p)}}) \cdot \frac{1}{\sqrt{np(1-p)}}(k+1-k).$$

As $\Phi'(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$,

$$\Phi'\left(\frac{c-np}{\sqrt{np(1-p)}}\right)\cdot\frac{1}{\sqrt{np(1-p)}}=\frac{1}{\sqrt{2\pi}}e^{-\frac{\left(\frac{c-np}{\sqrt{np(1-p)}}\right)^2}{2}}\cdot\frac{1}{\sqrt{np(1-p)}},$$

which coincides with the probability density function of a normally distributed random variable with expectation $m=np$ and dispersion $\sigma=\frac{1}{\sqrt{np(1-p)}}$ at some point $c\in(k,k+1)$.

If we choose the middle of he interval, that is $c=k+0.5$ we get $P(k_A(n)=k)\approx \frac{1}{\sqrt{np(1-p)}}\varphi(\frac{k+0.5-np}{\sqrt{np(1-p)}})$. The exact and the approximate probabilities and their differences are plotted in Fig.i.11. One can see that the largest difference between the approximate and exact probability is less than 0.01.

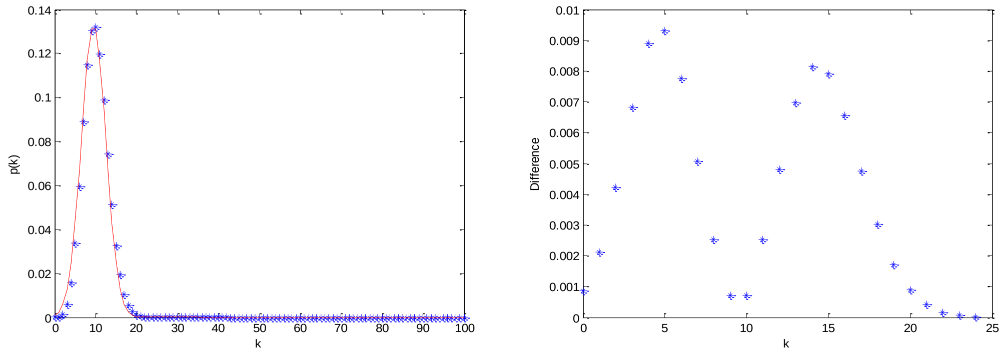


Figure i.11. The exact and the approximate probabilities and their differences in case of binomial distribution

<u>Example</u>

E1. In an airport, the number of tickets sold for a flight is 500. Suppose that all of the ticket holders come to the check in with probability 0.95 independently of each other. Compute the probability that the number of people come to the check in is at least 490.

Let η denote the number of people coming to the check in. η is a binomially distributed random variable with parameters $n=500$ and $p=0.95$. The question is $P(\eta\geq 490)$. Now

$$P(\eta\geq 490)=P(\eta=490)+P(\eta=491)+P(\eta=492)+P(\eta=493)+...+P(\eta=500)=$$

$$\binom{500}{490}0.95^{490}\cdot 0.05^{10}+\binom{500}{491}0.95^{491}\cdot 0.05^{9}+...+\binom{500}{500}0.95^{500}\cdot 0.05^{0}=0.00046 .$$

If one applies the Moivre-Laplace formula,

$P(\eta\geq 490)=1-F_\eta(490)\approx 1-\Phi(\frac{490-500\cdot 0.95}{\sqrt{500\cdot 0.95\cdot 0.05}})=1-0.99896=0.00104$. The difference between the exact and approximate probabilities is less than 0.001. One can conclude that that the probability of having at least 490 passengers on the flight is very small. More than 500 tickets may be sold, if the number of places is 500 and we would like to have less than 0.01 probability for overfilling.

E2. How many tickets may be sold in order to assure that at least 500 passengers come to the check in?

Let η_n be the number of passengers coming to the check in in case of n tickets sold. The question is the value of n for which $P(\eta_n\leq 500)=0.99$. We require $F_{\eta_n}(501)=0.99$.

Applying the Moivre-Laplace formula, $F_{\eta_n}(x) \approx \Phi(\frac{x - n \cdot 0.95}{\sqrt{n \cdot 0.95 \cdot 0.05}})$. Solving the equation $\Phi(\frac{501 - n \cdot 0.95}{\sqrt{n \cdot 0.95 \cdot 0.05}}) = 0.99$ we get $\frac{501 - n \cdot 0.95}{\sqrt{n \cdot 0.95 \cdot 0.05}} = 2.3263$, which is a quadratic equation for n. Solving it, we obtain n=515. As a control,

$$P(\eta_{515} \le 500) = \sum_{i=0}^{500} P(\eta_{515} = i) = 1 - \sum_{i=501}^{515} P(\eta_{515} = i) = 1 - \sum_{i=501}^{515} \binom{515}{i} 0.95^i \cdot 0.05^{515-i} = 0.9926 > 0.99.$$

E3. How many passengers come to the check in most likely? Compute/approximate the probability belonging to the mode in case of n=515 tickets sold.
The mode of a binomially distributed random variable is
$[(n+1) \cdot p] = [516 \cdot 0.95] = [490.2] = 490$, as $(n+1) \cdot p$
is not an integer. $P(\eta_{515} = 490) = \binom{515}{490} 0.95^{490} \cdot 0.05^{25} = 8.0585 \cdot 10^{-2}$.

Approximating this value by the normal cumulative distribution function, we get $P(\eta_{515} = 490) = P(490 \le \eta_{515} < 491) = F_{\eta_{515}}(491) - F(490) \approx$

$$\approx \Phi(\frac{491 - 515 \cdot 0.95}{\sqrt{515 \cdot 0.95 \cdot 0.05}}) - \Phi(\frac{490 - 515 \cdot 0.95}{\sqrt{515 \cdot 0.95 \cdot 0.05}}) = 0.63826 - 0.56026 = 0.078 .$$

If we apply approximation by probability density function, we get

$$P(\eta_{515} = 490) \approx \frac{1}{\sqrt{2\pi}\sqrt{515 \cdot 0.95 \cdot 0.05}} \varphi(\frac{490 - 515 \cdot 0.95}{\sqrt{515 \cdot 0.95 \cdot 0.05}}) = 7.8125 \times 10^{-2},$$

which is almost the same as the previous approximation.

E4. Flip a fair coin 400 times. Determine approximately the probability that the number of heads is at least 480 and less than 520.
Let η_{1000} be the frequency of heads in case of 1000 flips. η_{1000} is a binomially distributed random variable with parameters $n = 1000$ and $p = 0.5$. The question is $P(480 \le \eta_{1000} < 520)$, which can be expressed with the cumulative distribution function of η_{1000} in the following way: $P(480 \le \eta_{1000} < 520) = F_{\eta_{1000}}(520) - F_{\eta_{1000}}(480)$. Applying the Moivre-Laplace formula, $F_{\eta_{1000}}(x) \approx \Phi(\frac{x - 1000 \cdot 0.5}{\sqrt{1000 \cdot 0.5 \cdot 0.5}})$, and

$$P(480 \le \eta_{1000} < 520) \approx \Phi(\frac{520 - 500}{\sqrt{250}}) - \Phi(\frac{480 - 500}{\sqrt{250}}) =$$

$$\Phi(\frac{520 - 500}{\sqrt{250}}) - \Phi(\frac{480 - 500}{\sqrt{250}}) = 2\Phi(\frac{20}{\sqrt{250}}) - 1 = 0.7941 .$$

Give an interval symmetric to 500 in which the number of heads is situated with probability 0.99.
If $\theta \sim N(500, \sqrt{250})$, then $P(500 - 2.5758 \cdot \sqrt{250} < \theta < 500 + 2.5758 \cdot \sqrt{250}) = 0.99$. That means $P(459 < \eta_{1000} < 541) = 0.99$.
What do you think if you count 455 heads in case of 1000 flips?
If we realize that the frequency of heads is less than 459, then there are two possibilities . The first one is that an event with very small probability occurs. The second one is that the coin is not fair. People tend to believe the second one. This is the basic thinking of mathematical statistics.

At the end of this subsection we present the approximation of Poisson distribution by normal distribution. The possibility of that is not surprising: Poisson distribution is the limit of binomial distribution.

<u>Theorem</u> Let η_n be a sequence of Poisson distributed random variables with parameters $\lambda_n = n$ .Then $\lim_{n\to\infty} P(\frac{\eta_n - n}{\sqrt{n}} < x) = \Phi(x)$.

<u>Proof</u> η_n can be written as the sum of n independent Poisson distributed random variables with parameter $\lambda = 1$, consequently the central limit theorem provides the formula presented above.

<u>Remarks</u>

- The condition $\lambda_n = n$ is not crucial. If η is a Poisson distributed random variable with parameter λ and $10 \le \lambda$, then $P(\eta < x) \approx \Phi(\frac{x-\lambda}{\sqrt{\lambda}})$.
- The expectation of η is $E(\eta) = \lambda$, the dispersion of η is $D(\eta) = \sqrt{\lambda}$. Roughly spoken, the expectations of the approximated and the approximate distributions are the same values. The same can be stated about the dispersions.
- Similarly to the binomially distributed random variable, $P(\eta = k) = \frac{\lambda^k}{k!} e^{-\lambda} = P(k \le \eta < k+1) \approx \Phi(\frac{k+1-\lambda}{\sqrt{\lambda}}) - \Phi(\frac{k-\lambda}{\sqrt{\lambda}})$ .The goodness of the approximation can be seen in Fig.i.12. in case of $\lambda = 10$ and in Fig.i.13. in case of $\lambda = 50$.

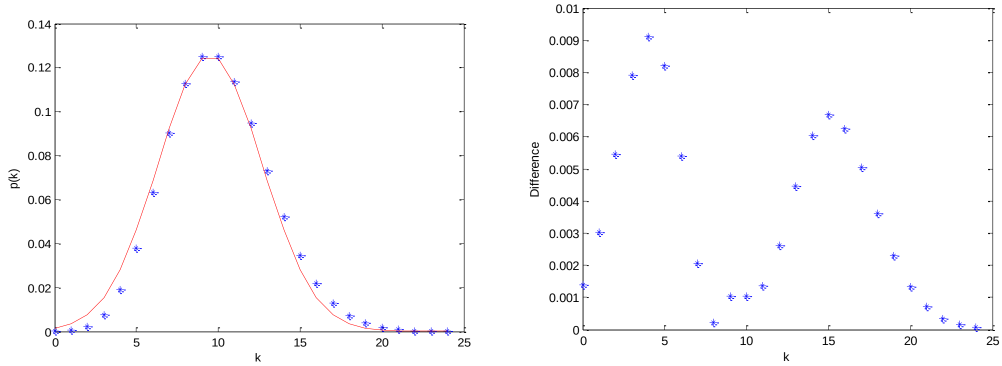


Figure i.12. The exact and the approximate probabilities and their differences in case of Poisson distribution with parameter $\lambda = 10$

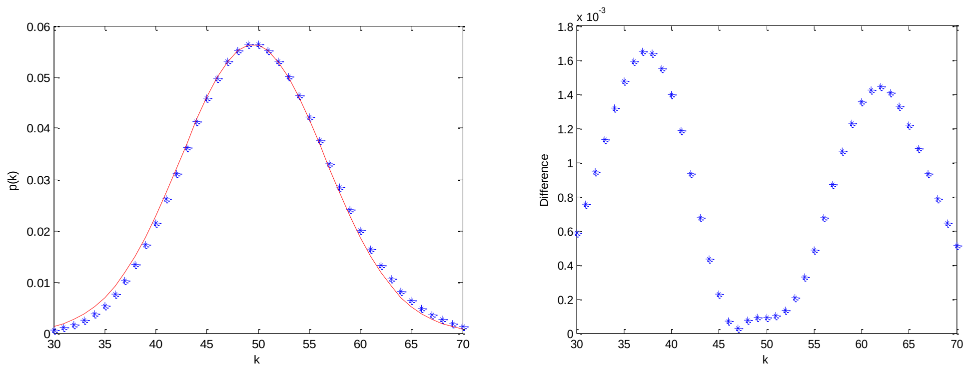


Figure i.13. The exact and the approximate probabilities and their differences in case of Poisson distribution with parameter $\lambda = 50$

Example

E5. The working times of a certain part of a machine between consecutive failings are supposed to be independent exponentially distributed random variables with expectation 24 hours. If a part goes wrong, it is changed immediately. How many spare parts should we have in order to have enough for a time period 90 days with probability 0.99.
Let η_T be denote the failings from time $t = 0$ to T. Recall that η_T is a Poisson distributed random variable with parameter $\lambda_T = \lambda \cdot T$, where λ is the parameter of the exponential distribution. Actually, if the time unit is a day, then $\lambda = \frac{1}{E(\xi_i)} = \frac{1}{1} = 1$, where ξ_i $i = 1,2,3,...$ denotes the time between the (i-1)th and ith failings. Consequently, η_{90} is a Poisson distributed random variable with parameter $\lambda_{90} = 90$. The question is the value of x for which $P(\eta_{90} < x) = 0.99$. $P(\eta_{90} < x) = F_{\eta_{90}}(x) \approx \Phi(\frac{x-90}{\sqrt{90}})$. Solving the equation $\Phi(\frac{x-90}{\sqrt{90}}) = 0.99$ we get $\frac{x-90}{\sqrt{90}} = 2.3263$, which implies $x = 90 + 2.3263 \cdot \sqrt{90} = 112.07$.
Consequently, we should have 113 spare parts in order not to run out them with probability 0.99. As a control, $P(\eta_{90} \le 113) = \sum_{i=0}^{113} \frac{90^i}{i!} e^{-90} = 0.99172$, but $P(\eta_{90} \le 112) = \sum_{i=0}^{112} \frac{90^i}{i!} e^{-90} = 0.98924$. This also supports the goodness of the presented method.

i.3. Central limit theorem for the average of independent identically distributed random variables

The central limit theorem was presented for the sum of many independent random variables. The average can be computed as a product of the sum and a constant value, consequently, the central limit theorem can be written for the average, as well.

Theorem Let $\xi_1, \xi_2, ..., \xi_n, ...$ be independent identically distributed random variables with expectation $E(\xi_i) = m$ and dispersion $D(\xi_i) = \sigma$, $i = 1,2,...$. Then,

$$\lim_{n\to\infty} P\left(\frac{\frac{\sum_{i=1}^{n} \xi_i}{n} - m}{\frac{\sigma}{\sqrt{n}}} < x \right) = \Phi(x) \text{ for any } x \in R .$$

Proof Notice that

$$P\left(\frac{\frac{\sum_{i=1}^{n}\xi_i}{n}-m}{\frac{\sigma}{\sqrt{n}}}<x\right)=P\left(\frac{\frac{\sum_{i=1}^{n}\xi_i-nm}{n}}{\frac{\sigma}{\sqrt{n}}}<x\right)=P\left(\frac{\frac{\sum_{i=1}^{n}\xi_i-nm}{n}}{\frac{\sqrt{n}\cdot\sigma}{n}}<x\right)=P\left(\frac{\sum_{i=1}^{n}\xi_i-nm}{\sigma\cdot\sqrt{n}}<x\right).$$

Therefore the statement is the straightforward consequence of the central limit theorem for sums.

<u>Remarks</u>

- $E\left(\frac{\sum_{i=1}^{n}\xi_i}{n}\right)=m$, $D\left(\frac{\sum_{i=1}^{n}\xi_i}{n}\right)=\frac{\sigma}{\sqrt{n}}$.
- $P\left(\frac{\frac{\sum_{i=1}^{n}\xi_i}{n}-m}{\frac{\sigma}{\sqrt{n}}}<x\right)$ is the cumulative distribution function of $\frac{\frac{\sum_{i=1}^{n}\xi_i}{n}-m}{\frac{\sigma}{\sqrt{n}}}$, that is the standardized average.
- $F_{\frac{\sum_{i=1}^{n}\xi_i}{n}}(x)\approx\Phi\left(\frac{x-m}{\frac{\sigma}{\sqrt{n}}}\right)$. This can be proved as follows:

$$F_{\frac{\sum_{i=1}^{n}\xi_i}{n}}(y)=P\left(\frac{\sum_{i=1}^{n}\xi_i}{n}<y\right)=P\left(\frac{\frac{\sum_{i=1}^{n}\xi_i}{n}-m}{\frac{\sigma}{\sqrt{n}}}<\frac{y-m}{\frac{\sigma}{\sqrt{n}}}\right)\approx\Phi\left(\frac{y-m}{\frac{\sigma}{\sqrt{n}}}\right).$$

- The cumulative distribution function of the average can be approximated by the cumulative distribution function of a normally distributed random variable. The expectations of the approximated and the approximate distributions are the same and so are their dispersions.
- The distribution of the averaged random variables can be arbitrary.
- The approximation can be applied if the number of random variables is at least 100.
- The fact that the average is approximately a normally distributed random variable and data are frequently averaged in statistics, is the reason of the leading role of normal distribution in statistics.

<u>Example</u>

E1. Let us suppose that the lifetime of bulbs are independent exponentially distributed random variables with expectation 1000 hours. Give and interval symmetric to 1000 in which the lifetime of one bulb is situated with probability 0.8.

$E(\xi_i)=\frac{1}{\lambda}=1000$. As $P(\xi_i<2000)=1-e^{-\frac{2000}{1000}}=0.865$, consequently, the interval looks like $(1000-x, 1000+x)$ with $x<1000$.

$$P(1000-x<\xi_i<1000+x)=F_{\xi_i}(1000+x)-F_{\xi_i}(1000-x)=1-e^{-\frac{1000+x}{1000}}-\left(1-e^{-\frac{1000-x}{1000}}\right)$$

$=e^{-\frac{1000-x}{1000}}-e^{-\frac{1000+x}{1000}}$. Solving the equation $e^{-\frac{1000-x}{1000}}-e^{-\frac{1000+x}{1000}}=0.8$, we get $e^{\frac{x}{1000}}-e^{-\frac{x}{1000}}=0.8\cdot e=2.1746$. Defining the new variable $y=e^{\frac{x}{1000}}$ we get $y-\frac{1}{y}=2.1746$. This is a quadratic equation for the variable y. Solving it we end up with

$y=-0.38993$ and $y=2.5645$. $y=e^{\frac{x}{1000}}$ can not be negative, therefore $y=2.5645$. This implies $x=1000\cdot\ln(2.5645)=941.76$.

The interval looks like $(1000-941.76,\quad 1000+941.76)=(58.24,\quad 1941.76)$ .We note that the interval is quite large, almost 1900 hours its length is.

As a control,

$$P(58.24<\xi_i<1941.76)=F_{\xi_i}(1941.76)-F_{\xi_i}(58.24)=1-e^{\frac{-1941.76}{1000}}-(1-e^{-\frac{58.24}{1000}})=0.8.$$

Give and interval symmetric to 1000 in which the average lifetime of 200 bulbs is situated with probability 0.8.

Turning to the average,

$$P\left(1000-y<\frac{\sum_{i=1}^{200}\xi_i}{n}<1000+y\right)=F_{\frac{\sum_{i=1}^{200}\xi_i}{n}}(1000+y)-F_{\frac{\sum_{i=1}^{200}\xi_i}{n}}(1000-y)\ .$$

Taking into account that $F_{\frac{\sum_{i=1}^{200}\xi_i}{n}}(x)\approx\Phi\left(\frac{x-1000}{\frac{1000}{\sqrt{200}}}\right)$, we should determine the value y for which $\Phi\left(\frac{1000+y-1000}{\frac{1000}{\sqrt{200}}}\right)-\Phi\left(\frac{1000-y-1000}{\frac{1000}{\sqrt{200}}}\right)=0.8$ holds. This implies

$2\cdot\Phi\left(\frac{y}{\frac{1000}{\sqrt{200}}}\right)-1=0.8$, that is $\frac{y\sqrt{200}}{1000}=1.2816$, that is $y=90.623$.

The interval in which the average is situated with probability 0.8 is $(1000-90.623,\ 1000+90.623)=(909,1091)$. Notice that its length is about 182 hours, which is much less than it was in the case of exponential distribution.

i.4. Central limit theorem for relative frequency

At the end of this chapter, we present the central limit theorem for relative frequency. As the relative frequency is the average of independent characteristically distributed random variable with parameter p, this form of the central limit theorem is a special case of that concerning average.

Theorem Let $k_A(n)$ be the frequency of the event A for which $P(A)=p, \quad 0<p<1$, during $2\le n$ independent experiments. Then, for any $x\in R$, $\lim_{n\to\infty} P\left(\frac{\frac{k_A(n)}{n}-p}{\sqrt{\frac{p(1-p)}{n}}}<x\right)=\Phi(x)$.

Remarks

- $E(\frac{k_A(n)}{n})=p$, $D(\frac{k_A(n)}{n})=\sqrt{\frac{p(1-p)}{n}}$.
- $P(\frac{\frac{k_A(n)}{n}-p}{\sqrt{\frac{p(1-p)}{n}}}<x)$ is the value of the cumulative distribution function of the standardized relative frequency.
- Returning to the relative frequency, $F_{\frac{k_A(n)}{n}}(x)\approx\Phi\left(\frac{x-p}{\sqrt{\frac{p(1-p)}{n}}}\right)$. This can be argued by $P(\frac{k_A(n)}{n}<y)=P\left(\frac{\frac{k_A(n)}{n}-p}{\sqrt{\frac{p(1-p)}{n}}}<\frac{y-p}{\sqrt{\frac{p(1-p)}{n}}}\right)\approx\Phi\left(\frac{y-p}{\sqrt{\frac{p(1-p)}{n}}}\right)$.
- $P(\left|\frac{k_A(n)}{n}-p\right|<\varepsilon)=P(p-\varepsilon<\frac{k_A(n)}{n}<p+\varepsilon)\approx\Phi\left(\frac{p+\varepsilon-p}{\sqrt{\frac{p(1-p)}{n}}}\right)-\Phi\left(\frac{p-\varepsilon-p}{\sqrt{\frac{p(1-p)}{n}}}\right)=$

$=2\Phi(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}})-1$.

It provides possibility to compute

1. the reliability $1-\alpha$ as the function of ε and n,
2. ε (accuracy) as the function of reliability $1-\alpha$ and n
3. number of necessary experiments (n) as the function of ε and $1-\alpha$.

- This formula can be directly applied if p is known.

Example

E1. Throw a fair die 500 times. Compute the probability that the relative frequency of “six” is at least 0.15 and less than 0.18.
Let A be the event that the result is “six” performing one throw. The question is $P(0.15\le\frac{k_A(500)}{500}<0.18)$. Recall that

$$P(0.15 \le \frac{k_A(500)}{500} < 0.18) = F_{\frac{k_A(500)}{500}}(0.18) - F_{\frac{k_A(500)}{500}}(0.15) \approx \Phi\left(\frac{0.18 - \frac{1}{6}}{\sqrt{\frac{\frac{1}{6}\cdot\frac{5}{6}}{500}}}\right) - \Phi\left(\frac{0.15 - \frac{1}{6}}{\sqrt{\frac{\frac{1}{6}\cdot\frac{5}{6}}{500}}}\right) =$$

$= \Phi(0.8) - \Phi(-1) = 0.78814 - 0.15866 = 0.62948 \approx 0.63$.

Making computer simulations, applying 10^6 repetitions, we get an approximate value for $P(0.15 \le \frac{k_A(500)}{500} < 0.18)$. This means that $500 \cdot 10^6 = 5 \cdot 10^8$ random experiments were performed, which required 0.31 sec. Computer simulation resulted in 0.627480.

E2. Throw a fair die 500 times. At most how much is the difference between the exact probability and the relative frequency with reliability 0.9?

Applying $P(\left|\frac{k_A(n)}{n} - p\right| < \varepsilon) \approx 2\Phi(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}) - 1$, $2\Phi(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}) - 1 = 0.90$ implies

$\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}} = 1.645$. Substituting $n = 500$ and $p = \frac{1}{6}$ we get $\varepsilon = \frac{1.654 \cdot \sqrt{\frac{1}{6}\cdot\frac{5}{6}}}{\sqrt{500}} = 0.0274$. It means that $P(\frac{1}{6} - 0.0274 < \frac{k_A(500)}{500} < \frac{1}{6} + 0.0274) = P(0.1393 < \frac{k_A(500)}{500} < 0.1941) \approx 0.90$. Computer simulation resulted in 0.907078. If we would like to increase the reliability, for example, $1 - \alpha = 0.99$, then $2\Phi(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}) - 1 = 0.99$, $\frac{\varepsilon\sqrt{500}}{\sqrt{\frac{1}{6}\cdot\frac{5}{6}}} = 2.5758$, $\varepsilon = 0.04293$.

Consequently, the interval is $(\frac{1}{6} - 0.04293, \frac{1}{6} + 0.04293) = (0.12374,\ 0.20960)$. We can realize that the greater the reliability, the larger the interval.

E3. Throw a fair die 500 times repeatedly. How many throws should be done, if the relative frequency of "six" is closer to the exact probability than 0.01 with reliability 0.99?

Apply the formula $P(\left|\frac{k_A(n)}{n} - p\right| < \varepsilon) \approx 2\Phi(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}) - 1$ again with $\varepsilon = 0.01$ and $1 - \alpha = 0.99$. $2\Phi(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}) - 1 = 0.99$ implies $\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}) = 2.5758$, that is $\sqrt{n} = \frac{2.5758}{0.01}\sqrt{\frac{1}{6}\cdot\frac{5}{6}}$, $n = \left(\frac{2.5758}{0.01}\sqrt{\frac{1}{6}\cdot\frac{5}{6}}\right)^2 = 9215$ instead of 500 experiments. As $2\Phi(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}) - 1$ is a monotone increasing function of n, if we increase the value of n, we increase the reliability, as well. If we apply the estimation $P(\left|\frac{k_A(n)}{n} - p\right| < \varepsilon) \ge 1 - \frac{p(1-p)}{n\varepsilon^2}$ presented in the previous chapter, substituting $\varepsilon = 0.01$, $p = \frac{1}{6}$ and $1 - \frac{p(1-p)}{n\varepsilon^2} = 0.99$ we

get $n = \dfrac{\frac{1}{6}\cdot\frac{5}{6}}{0.01\cdot 0.01^2} \approx 13890$ which is about the 1.5 times larger than the previously determined simulation number. It means that it is rather worth computing by central limit theorem, than by the law of large numbers.
Note that if we would like to have accuracy $\varepsilon = 0.001$, then the number of simulation has to be $10^2 = 100$ times larger than in the case of $\varepsilon = 0.01$.

We would like to emphasize that in the previous examples the probability of the event A was known. But in many cases it is unknown and we would like to approximate the unknown probability by the relative frequency. In those cases we can apply upper estimation for the probability $2\Phi(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}})-1$.

<u>Theorem</u> For any value of $0 < p < 1$, $2\Phi(2\varepsilon\sqrt{n})-1 \le 2\Phi(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}})-1$.

<u>Proof</u> If $0 < p < 1$, then $p(1-p) \le \frac{1}{4}$, therefore $\sqrt{p(1-p)} \le \frac{1}{2}$. This implies $\frac{\varepsilon\sqrt{n}}{\frac{1}{2}} \le \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}$, that is $2\cdot\varepsilon\sqrt{n} \le \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}$. Since Φ is a monotone increasing function, so is $2\Phi-1$, therefore $2\cdot\varepsilon\sqrt{n} \le \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}$ implies $2\Phi(2\cdot\varepsilon\sqrt{n})-1 \le 2\Phi(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}})-1$, which is the statement to be proved.

<u>Remarks</u>

- The formula $2\Phi(2\varepsilon\sqrt{n})-1$ does not contain the unknown value of p, therefore the inequality $2\Phi(2\varepsilon\sqrt{n})-1 \le P(\left|\frac{k_A}{n}-p\right| < \varepsilon)$ is suitable for estimating the accuracy, the reliability and the necessary number of simulations in the case of unknown p value.

For the sake of applications, we determine the reliability as the function of n and ε, the accuracy ε as the function of n and reliability $1-\alpha$, and the necessary number of simulations as the function of ε and $1-\alpha$.

1. If n and ε are fixed then $2\Phi(2\varepsilon\sqrt{n})-1 \le P(\left|\frac{k_A}{n}-p\right| < \varepsilon)$ supplies a direct lower bound for the reliability.
2. If n and the reliability $1-\alpha$ are fixed, with the choice $2\Phi(2\varepsilon\sqrt{n})-1 = 1-\alpha$, then $2\varepsilon\sqrt{n} = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$ and $\varepsilon = \dfrac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}{2\sqrt{n}}$. Notice that the accuracy ε is proportional to the

reciprocal of the square root of the number of simulations. We note that $\Phi^{-1}\left(1-\frac{\alpha}{2}\right)=y$ means that $\Phi(y)=1-\frac{\alpha}{2}$. Summarizing, if $\varepsilon=\frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}{2\sqrt{n}}$, then $1-\alpha\le P(\left|\frac{k_A}{n}-p\right|<\varepsilon)$.

- If the accuracy ε and the reliability $1-\alpha$ are fixed, then $2\Phi(2\varepsilon\sqrt{n})-1=1-\alpha$ serves again for the formula $2\varepsilon\sqrt{n}=\Phi^{-1}\left(1-\frac{\alpha}{2}\right)$ and, $n=\left(\frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}{2\varepsilon}\right)^2$. If n increases, then the reliability increases supposing ε is fixed. If the reliability is fixed and n increases, then ε decreases. Note that the required number of simulations is proportional to the square of the reciprocal of the accuracy. Summarizing, if $\left(\frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}{2\varepsilon}\right)^2\le n$, then $1-\alpha\le P(\left|\frac{k_A}{n}-p\right|<\varepsilon)$.

Examples

E1. At a survey, $n=1000$ people are asked about a yes/no question. The relative frequency of the answer "yes" is 0.35. Estimate the probability that the relative frequency is closer to the probability of the answer "yes" (p) than 0.05, that is $P(0.3<p<0.4)$.
Let A be the event that the answer is yes, $P(A)=p$ is unknown. Recalling $2\Phi(2\varepsilon\sqrt{n})-1\le P(\left|\frac{k_A}{n}-p\right|<\varepsilon)$ and substituting $n=1000$ and $\varepsilon=0.05$, $2\Phi(2\varepsilon\sqrt{n})-1=2\Phi(2\cdot 0.05\cdot\sqrt{1000})-1=2\cdot 0.9992\ -1=0.9984$. Therefore, $0.9984\le P(\left|k_A-p\right|<0.05)$.

E2. At a survey, $n=1000$ people are asked about a yes/no question. How much is the largest difference between the relative frequency and the exact probability p with reliability 0.95?

We have a formula for the accuracy, namely $\varepsilon=\frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}{2\sqrt{n}}$. Now, $1-\alpha=0.95$, $1-\frac{\alpha}{2}=0.975$, $\Phi^{-1}\left(1-\frac{\alpha}{2}\right)=1.96$ and $\frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}{2\sqrt{n}}=\frac{1.96}{2\sqrt{1000}}=0.031$. That means $0.95\le P(0.35-0.031<\frac{k_A}{1000}<0.35+0.031)$. This is the reason why surveys publish the results with $\pm 3\%$ in case of 1000 people.

E3. At a survey some people are asked about a yes/no question. If we need accuracy $\varepsilon = 0.01$ with reliability 0.95, how many people should be asked to be able to do this?

Apply $\left(\frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}{2\varepsilon} \right)^2 \le n$ with $\varepsilon = 0.01$, $1-\alpha = 0.95$.

$$\left(\frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}{2\varepsilon} \right)^2 = \left(\frac{1.96}{2 \cdot 0.01} \right)^2 = 98^2 = 9604 .$$

This is the reason why 10000 people are asked to have accuracy 0.01 with reliability 0.95.

Summarizing our result, in case of $9604 \le n$, $0.95 \le P(\left|\frac{k_A}{n} - p\right| < 0.01) = P(0.34 < p < 0.35)$.

Of course, the above questions should have been asked for computer simulation as well. The main difference between survey and computer simulation is that the number of simulation can be easily increased but the increment of number of people asked at a survey requires lots of money.

Finally we present Tables i.1.and i.2., which contain the required number of simulations for given accuracy, in case of reliability levels $1-\alpha = 0.95$ and $1-\alpha = 0.99$. These reliability levels are often used in practice. In Tables i.3. and i.4., we present accuracy at given numbers of simulation.

$1-\alpha = 0.95$

n	ε
10	0.3099
100	0.098
500	0.043827
1000	0.03099
5000	0.013859
10000	0.0098
50000	0.0043827
100000	0.003099
500000	0.0013859
1000000	0.00098
500000	0.00043827
1000000	0.0003099
5000000	0.00013859
10000000	0.000098
50000000	0.000043827
100000000	0.00003099

Table i.1.The accuracy in the function of number of simulations in case of reliability level 0.95

$1-\alpha = 0.99$

n	ε
10	0.40727
100	0.12879
500	0.05 7597

1000	0.040727
5000	0.018214
10000	0.012879
50000	0.005 7597
100000	0.0040727
500000	0.0018214
1000000	0.0012879
500000	0.0005 7597
1000000	0.00040727
5000000	0.00018214
10000000	0.00012879
50000000	0.00005 7597
100000000	0.000040727

Table i.2.The accuracy in the function of number of simulations in case of reliability level 0.95

$1-\alpha=0.95$

ε	n
0.1	97
0.05	385
0.025	1537
0.01	9604
0.005	38416
0.0025	153660
0.001	960400
0.0005	3841600
0.00025	15366000
0.0001	96040000

Table i.3.Necessary number of simulations to a given accuracy in case of reliability level 0.95

$1-\alpha=0.99$

ε	n
0.1	166
0.05	664
0.025	2654
0.01	16587
0.005	66347
0.0025	265390
0.001	1658700
0.0005	6634700
0.00025	26539000
0.0001	165870000

Table i.4.Necessary number of simulations to a given accuracy in case of reliability level 0.99

j. Basic concepts of mathematical statistics

The aim of this chapter

In this chapter we present the basic concepts of mathematical statistics and we sketch some branches of it. We introduce the empirical cumulative distribution function, the empirical density function, estimations of expectations and dispersions. We also present how to test hypothesis in some cases.

Preliminary knowledge

Properties of average. Normal distribution. Student’s t distribution. Chi-squared distribution.

Content

j.1. Empirical cumulative distribution functions and histogram.

j.2. Estimation of probability, expectation and variance.

j.3. Testing hypothesis.

j.1. Empirical cumulative distribution function and histogram

In the previous chapters we have dealt with probabilities. In this last section we present how to draw conclusions from data on the basis of probabilistic arguments. As the cumulative distribution function contains all information about the random variable, our primary aim is to approximate it on the basis of data. Data have dual nature, before performing the sampling they are random variables, after performing the sampling they are real numbers as the results of observations of a random phenomenon. The statistical methods are executed on the numbers, but they are elaborated for the random variables.

First, clarify the concept of sample.

Definition A **sample** is a series of independent observations concerning a random variable $\underline{\xi}$. More precisely, a sample is $\underline{\xi} = (\xi_1, \xi_2, ..., \xi_n)$, where ξ_i $i = 1,2,...,n$ are independent identically distributed random variable with common distribution function F. The **number of elements** of the sample equals n.

Definition Let the values of the sample be $x_1, x_2, ..., x_n$, $x_i \in R$, i=1,2,…,n. The e**mpirical cumulative distribution function** belonging to the values of the sample $\underline{x} = (x_1, x_2, ..., x_n)$ is defined as $F_{(x_1, x_2, ... x_n)} : R \to R$ $F_{(x_1, x_2, ... x_n)}(z) = F_e(z) = \dfrac{\sum_{i=1}^{n} 1_{\{x_i < z\}}}{n}$.

Remarks

- The argument of the function is denoted by z because the letter x is related to the sample.
- $F_{(x_1, x_2, ... x_n)}(z)$ is briefly denoted by $F_e(z)$.
- The cumulative distribution function is the relative frequency of the event $\{\xi < z\}$ if

we perform independent experiments for this event. $F_e(z) = \dfrac{\sum_{i=1}^{n} \mathbf{1}_{\{x_i < z\}}}{n}$ is a step function which has jumps at $z = x_i$. It is constant zero previous to the smallest element of the sample, and it is constant 1 following the greatest one.

- The elements of the sample x_i and x_j may be equal.
- The function $F_e(z)$ has all the properties of cumulative distribution function.

Namely,

1. $\sum_{i=1}^{n} \mathbf{1}_{\{x_i < z\}} \le \sum_{i=1}^{n} \mathbf{1}_{\{x_i < y\}}$ for any values of $z < y$, which implies the monotone increasing property.
2. Its limit is zero at $-\infty$ and 1 at ∞.
3. It is left hand-side continuous. Consequently, it is really a cumulative distribution function.

- The random variable $\theta \sim \begin{pmatrix} x_1 & x_2 & . & . & x_n \\ \frac{1}{n} & \frac{1}{n} & & & \frac{1}{n} \end{pmatrix}$ has the same cumulative distribution function if all the values x_i are different i=1,2,…,n. If some values x_i are repeatedly in the sample, then the probability belonging to this value is the relative frequency of this element in the sample.

Example

E1. Let the elements of the sample be $x_1 = 12$, $x_2 = 10$, $x_3 = 15$, $x_4 = 12$, $x_5 = 13$. Draw the empirical cumulative distribution function belonging to these sample elements.

$$F_e(z) \frac{= \sum_{i=1}^{5} \mathbf{1}_{\{x_i < z\}}}{5} = \begin{cases} 0 & \text{if } z \le 10 \\ \frac{1}{5} & \text{if } 10 \le z < 12 \\ \frac{3}{5} & \text{if } 12 \le z < 13 \\ \frac{4}{5} & \text{if } 13 \le z < 15 \\ 1 & \text{if } 15 \le z \end{cases}.$$

This function can be seen in Fig.j.1.

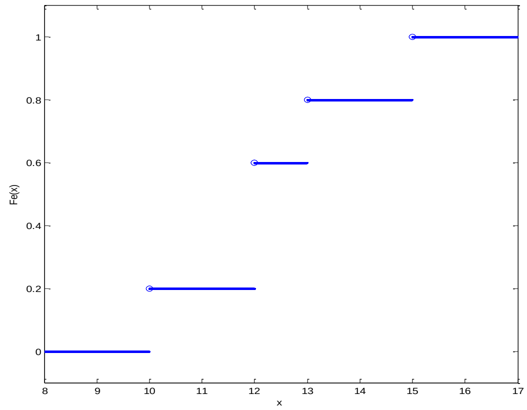


Figure j.1. The empirical distribution function belonging to the sample elements in E.1.

Theorem If $F_e(z)$ is the empirical cumulative distribution function belonging to the sample elements $(x_1,....,x_n)$ and $F(z)$ is the cumulative distribution function of ξ_i, $i = 1,2,3,...$, then for any value of $x \in R$ and $0 < \varepsilon$, $P(|F_e(z) - F(z)| < \varepsilon) \to 1$ if $n \to \infty$.

Proof Let A be the event that the random variable $\xi *$ is less than z, that is $A = \{\xi * < z\}$. Now $F_e(z)$ is the relative frequency of A during n independent trials. Moreover, $F(z) = P(A)$. The law of large numbers states that the relative frequency of an event and the probability of that are close to each other, that is $P(|F_e(z) - F(z)| < \varepsilon) \ge 1 - \frac{F(z)(1 - F(z))}{n\varepsilon^2} \to 1 - 0$, supposing $n \to \infty$.

Remarks

- The above theorem is the consequence of the law of large numbers.
- The theorem states that the values of the cumulative distribution function can be approximated by the empirical cumulative distribution function. The necessary number of simulations to a given accuracy can be determined by applying the central limit theorem presented in the previous section. For example, if $\varepsilon = 0.01$, then $n = 9604$, if the reliability level is 0.95.

Example

E1. Let ξ^* be an exponentially distributed random variable with parameter $\lambda = 1$. Take a sample of n elements independently with respect to ξ^*. Draw the empirical cumulative distribution function of the sample if n=10 and n=100 and $n = 1000$.
The empirical cumulative distribution functions together with the exact one can be seen in Figs.j.2. and j.3.

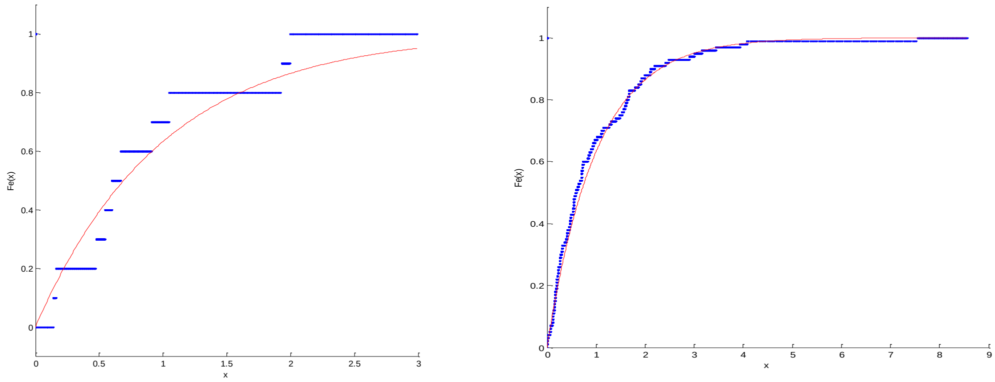


Figure j.2. The empirical distribution function belonging to an exponentially distributed sample of 10 and 100 elements

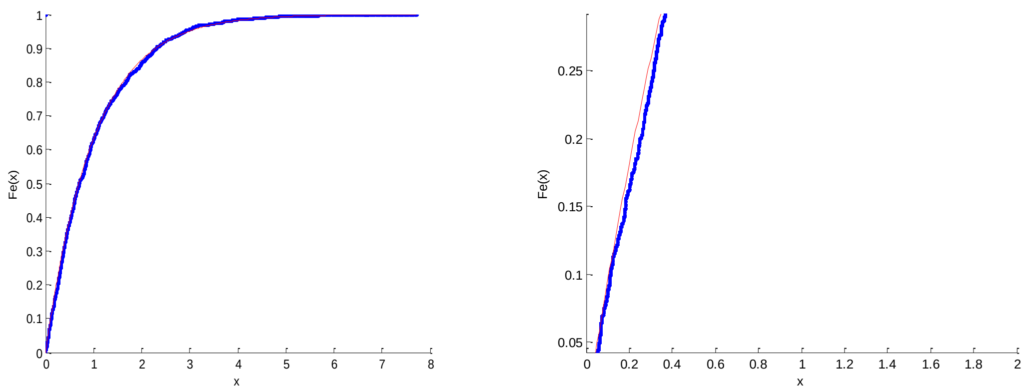


Figure j.3. Empirical distribution function belonging to an exponentially distributed sample of 1000 elements and a segment of the function

One can realize that there is hardly any difference between the exact cumulative distribution function and the empirical one if the number of sample elements is large.

E2. The exact cumulative distribution function and the empirical one is presented in Fig. j.4. in case of $\xi \sim \begin{pmatrix} 0 & 1 \\ 0.5 & 0.5 \end{pmatrix}$. The number of sample elements was n=10 and n=100.

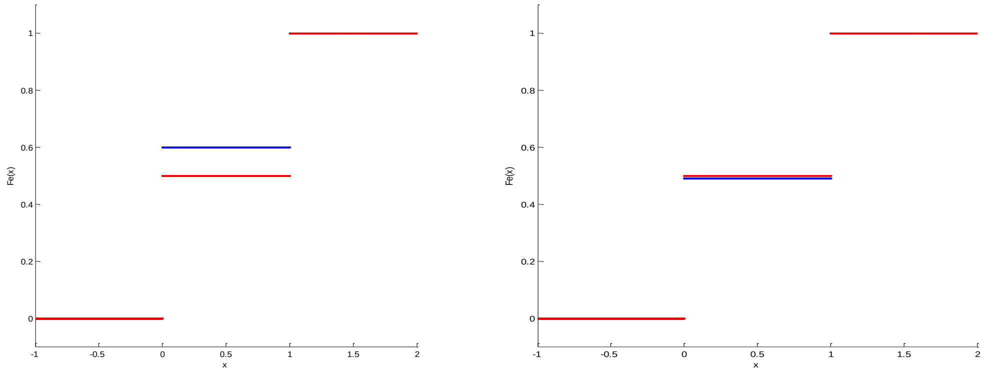


Fig.j.4. The empirical cumulative distribution function (blue) and the exact one (red) in case of 10 and 100 sample elements, respectively

One can see that if the number of elements is large, then they are close to each other.

The following statement is a stronger one than the previously proved statement. We present it without proof.

Theorem (Glivenko)
If $F_e(z)$ is the empirical cumulative distribution function belonging to the sample elements $(x_1,...., x_n)$ and $F(z)$ is the cumulative distribution function of ξ^* and ξ_i, $i = 1,2,3,...$. Then $\sup_{z \in R} |F_e(z) - F(z)| \to 0$ if $n \to \infty$ with probability 1.

Remarks

- Glivenko's theorem is often used as the fundamental theorem of mathematical statistics.
- Its philosophical interpretation is that the world is knowable.
- The main differences of Glivenko's theorem and the theorem presented at the beginning of the section are that this later states uniform convergence (not for every z separately) and states probability 1 (strong law of large numbers).
- A test for distribution function can be given on the basis of maximal difference. It is called Kolmogorov-Smirnov's test, and will be presented in the last subsection.

Now we turn to the approximation of the probability density function by histogram. Histograms are used for presentation of relative frequencies. We usually compared them to the probability density function. We mention that relative frequency and frequency differs in a constant multiplier, therefore the shape the figures are very similar.

Definition Let $x_1, x_2, ..., x_n$ be the values of the sample. Let $a = \min_{i=1,2,..,n} x_i$, $b = \max_{i=1,2,..,n} x_i$ and $1 \le m$ fixed. Then consider the points $y_0 = a - \frac{b-a}{2m}$, $y_i = y_{i-1} + i\frac{b-a}{m}$, $i = 1,2,...,m+1$.

Let $k_i(n,m) = \sum_{j=1}^{n} \mathbf{1}_{\{x_j \in [y_{i-1}, y_i)\}}$, $i = 1,2,...,m$ and

$$f_e(z) = \begin{cases} \frac{k_i(n,m)}{n} \cdot \frac{1}{\frac{b-a}{m}} & \text{if} \quad z \in [y_{i-1}, y_i) \quad i = 1,2,...,m \\ 0 \quad \text{otherwise} \end{cases}.$$

The function $f_e(z)$ is called the **histogram** with m equal length subintervals belonging to the sample elements $x_1, x_2, ..., x_n$.

Remarks

- The histogram strongly depends on the value of m. If m is too small or too large as compared to n the shape of the graph of the histogram will not be appropriate. To see this, we present Fig.j.5. The number of sample elements was n=100 in all cases. The sample was uniformly distributed, m=4, m=10, m=50 and m=100. The sample elements were the same in case of all histograms.

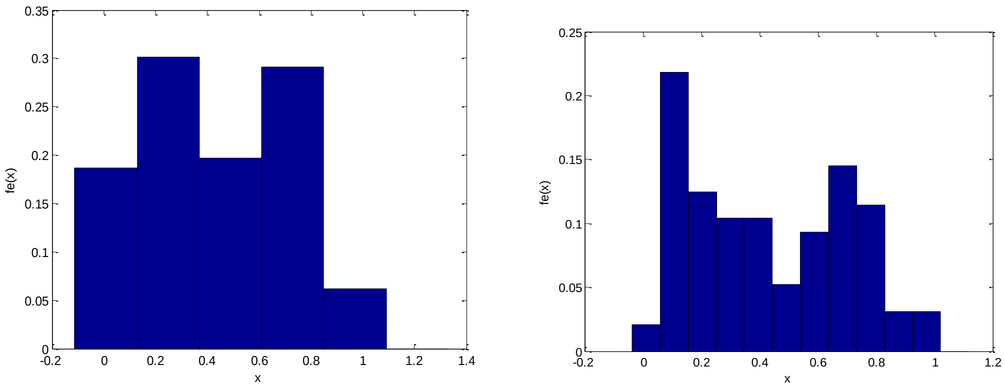


Figure j.5. Histograms of a sample of 100 elements in case of 5 and 11 subintervals

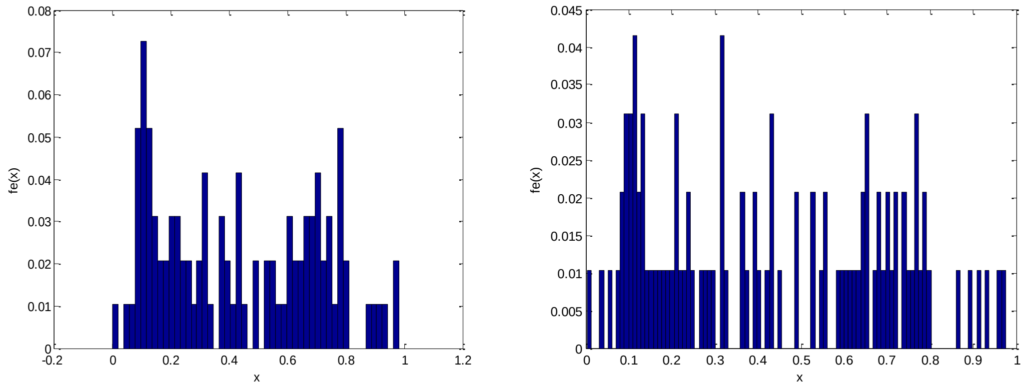
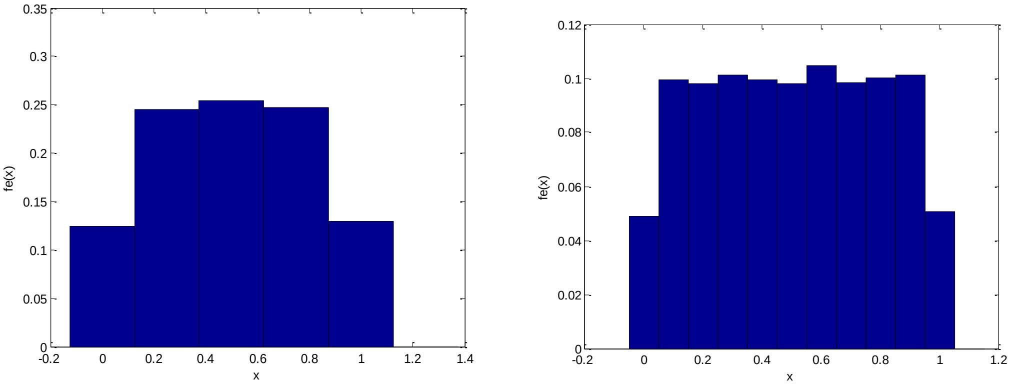
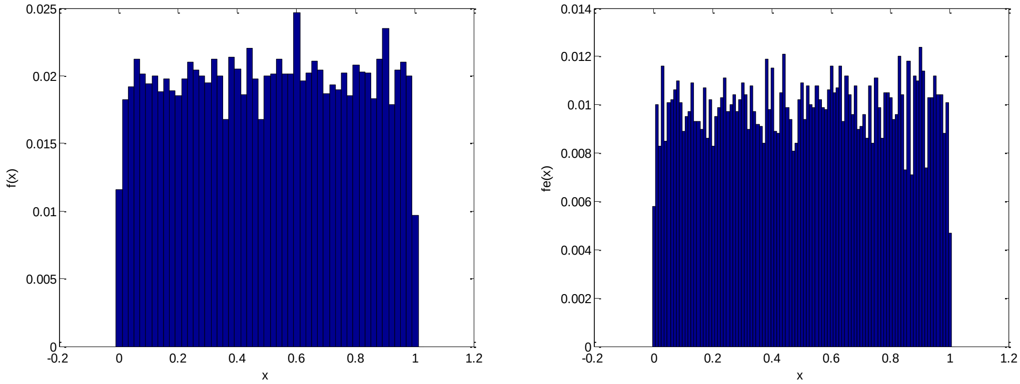


Figure j.5. Histograms of a sample of 100 elements in case of m=50 and m=100

If the number of sample elements is 10000 and they are uniformly distributed in [0,1], then the histograms for $m = 4$, $m = 10$, $m = 50$ and $m = 100$ look as follows:



j.5. Histograms of sample of 10000 elements in case of 5 and 11 subintervals



j.6. Histograms of sample of 10000 elements in case of 51 and 101 subintervals

The histograms belonging to $m=4$ and $m=10$ seem to be better approximations of the probability density function of uniformly distributed random variable. The height of the first and last rectangle is the half of the others because the smallest value of the sample is about zero, the first subinterval is $[-0.05, 0.05]$, and $P(-0.05 \le \xi < 0.05) = P(0 \le \xi < 0.05) = 0.05$, while $P(0.05 \le \xi < 0.15) = 0.1$. The last subinterval is $(0.95, 1.05]$, $P(0.95 \le \xi < 1.05) = 0.05$.

- Although there are many theorems concerning the relationship of the empirical cumulative distribution function and the real cumulative distribution function, it is difficult to give a limit theorem concerning the histogram and the probability density function. Roughly speaking, for appropriate fixed m values, the histogram is close the real probability density function, if n is large. Examples were presented in section g.

j.2. Estimation of probability, expectation and variance

After approximating the cumulative distribution function and the probability density function, we estimate the probability of an event, furthermore the expectation and the variance of a random variable. This will be done by a function of the sample.

Definition Let $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$ be a sample and $g: H \subset R^n \to R$ a real valued function with $\operatorname{Im}\xi \subset H$. Then $g \circ \xi = g(\xi)$ is called **statistics**.

Remarks

- Statistics are the function of the sample. The question in which cases which function should be applied is an important question of mathematical statistics.
- The function $g \circ \eta : \Omega \to R$ is a random variable, and $g(x_1, x_2, ... x_n)$ is a real number. The dual property appears in this case, as well.

Estimation of probability

Let $\underline{\xi} = (\xi_1, \xi_2, ..., \xi_n)$ be a sample, $\xi_i = 1_A{}^i = \begin{cases} 1 & \text{if A occurs at the ith experiment} \\ 0 & \text{if } \overline{A} \text{ occurs at the ith experiment} \end{cases}$ are characteristically distributed random variables with parameter $0 < p < 1$. Let $g : R^n \to R$ be $g(y_1, y_2, ..., y_n) = \frac{\sum_{i=1}^{n} y_i}{n}$. Then $g \circ \underline{\xi} = g(\underline{\xi}) = \frac{\sum_{i=1}^{n} \xi_i}{n}$ is the sample average. It can be considered as the relative frequency of an event A with $P(A) = p$. Now, $E\left(g(\underline{\xi})\right) = E\left(\frac{\sum_{i=1}^{n} \xi_i}{n}\right) = \frac{np}{n} = p$, $D(g(\xi)) = \sqrt{\frac{p(1-p)}{n}} \to 0$. Consequently, if we estimate the probability $p = P(A)$ by $\hat{p} = \frac{\sum_{i=1}^{n} \xi_i}{n} = \frac{k_A(n)}{n}$, then the expectation of the estimation equals the exact probability p and the dispersion of the estimation tends to zero if $n \to \infty$. These two properties imply the consistency of the estimation, which means that the estimate value fluctuates around the value to be estimated and the fluctuation tends zero if the number of sample elements tends to infinity.

Moreover, applying the central limit theorem, for $100 \le n$, $10 \le np$, we can write that $P(p - u_\alpha \sqrt{\frac{p(1-p)}{n}} \le \frac{k_A}{n} \le p + u_\alpha \sqrt{\frac{p(1-p)}{n}}) = 1 - \alpha$, with $\Phi(u_\alpha) = 1 - \frac{\alpha}{2}$. Arranging the sides of both inequalities, we end up with (approximately)

$$P\left(\frac{k_A}{n} - u_\alpha \sqrt{\frac{\frac{k_A}{n}(1-\frac{k_A}{n})}{n}} \le p \le \frac{k_A}{n} + u_\alpha \sqrt{\frac{\frac{k_A}{n}(1-\frac{k_A}{n})}{n}}\right) = 1 - \alpha \text{ with } \Phi(u_\alpha) = 1 - \frac{\alpha}{2}.$$

Summarizing, the interval

$$\boxed{\left[\frac{k_A}{n} - u_\alpha \sqrt{\frac{\frac{k_A}{n}(1-\frac{k_A}{n})}{n}}, \frac{k_A}{n} + u_\alpha \sqrt{\frac{\frac{k_A}{n}(1-\frac{k_A}{n})}{n}}\right]}$$

contains the exact probability p with probability (reliability level) $1 - \alpha$. This interval is usually called the confidence interval for the probability belonging to the reliability level $1 - \alpha$.

Remarks

- We list the values u_α for some frequently used reliability levels $1-\alpha$, and give confidence intervals for the probability in case of relative frequency $\frac{\sum_{i=1}^{n}\xi_i}{n}=0.450$ and n=500 in Table j.1.

$1-\alpha$	u_α	Confidence interval
0.9	1.645	$[0.413,\ 0.487]$
0.95	1.960	$[0.406,\ 0.493]$
0.98	2.326	$[0.398,\ 0.502]$
0.99	2.575	$[0.393, 0.507]$

Table j.1. Values u_α and confidence intervals for the probability belonging to reliability level $1-\alpha$

- The larger the reliability, the wider the interval.

Estimation of the expectation in case of known value of dispersion

Let $\underline{\xi}=(\xi_1,\xi_2,...,\xi_n)$ be a sample, ξ_i are random variables with expectation m and dispersion σ. Let $g:R^n\to R$ $g(y_1,y_2,...,y_n)=\frac{\sum_{i=1}^{n}y_i}{n}$. Then $g(\xi)=\frac{\sum_{i=1}^{n}\xi_i}{n}=\overline{\xi}$ is the sample average. Now, $E(g(\xi))=E\left(\frac{\sum_{i=1}^{n}\xi_i}{n}\right)=\frac{nm}{n}=m$, $D(g(\eta))=\frac{\sigma}{\sqrt{n}}$. Consequently, if we estimate the expectation by the sample average, then, with notation $\hat{m}=\frac{\sum_{i=1}^{n}\xi_i}{n}$, $E(\hat{m})=m$, and $D\left(\hat{m}\right)\to 0$. This means that the sample average is a consistent estimation for the expectation. Note that the sample average is the expectation belonging to the empirical cumulative distribution function. Moreover, if $\xi_i\sim N(m,\sigma)$, or $100\le n$, then $\frac{\sum_{i=1}^{n}\xi_i}{n}\sim N(m,\frac{\sigma}{\sqrt{n}})$, or this holds approximately. Applying the $k\cdot\sigma$ rule with notation

$k = u_\alpha$, we get $P(m - u_\alpha \frac{\sigma}{\sqrt{n}} < \frac{\sum_{i=1}^{n} \xi_i}{n} < m + u_\alpha \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$. Arranging both sides of the inequalities we end up with $P(\frac{\sum_{i=1}^{n} \xi_i}{n} - u_\alpha \frac{\sigma}{\sqrt{n}} < m < \frac{\sum_{i=1}^{n} \xi_i}{n} + u_\alpha \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$.

$$\left[\frac{\sum_{i=1}^{n} \xi_i}{n} - u_\alpha \frac{\sigma}{\sqrt{n}}, \frac{\sum_{i=1}^{n} \xi_i}{n} + u_\alpha \frac{\sigma}{\sqrt{n}} \right]$$

is an interval in which the expectation m is situated with probability $1-\alpha$, It is called the confidence interval of the expectation belonging to the reliability level $1-\alpha$.

Remarks

- The above formula can be applied in the case when the dispersion is given.
- If we have the sample elements $(x_1, x_2, ..., x_n)$, we have to substitute these values into the formula $\left[\frac{\sum_{i=1}^{n} \xi_i}{n} - u_\alpha \frac{\sigma}{\sqrt{n}}, \frac{\sum_{i=1}^{n} \xi_i}{n} + u_\alpha \frac{\sigma}{\sqrt{n}} \right]$ to get the confidence interval for the expectation belonging to the reliability level $1-\alpha$ in case of σ=0.2 . For example, if $x_1 = 1.5$, $x_2 = 1.7$, $x_3 = 1.4$, $x_4 = 1.9$, $x_5 = 1.7$ then

$\frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{1.5 + 1.7 + 1.4 + 1.9 + 1.7}{5} = 1.64$. The confidence interval belonging to the reliability levels 0.9, 0.95, 0.98 and 0.99 are contained in the Table j.2.

$1-\alpha$	u_α	$\left[1.64 - u_\alpha \cdot \frac{0.2}{\sqrt{5}}, 1.64 + u_\alpha \cdot \frac{0.2}{\sqrt{5}}\right]$
0.9	1.645	$[1.493, 1.787]$
0.95	1.960	$[1.465, 1.815]$
0.98	2.326	$[1.432, 1.848]$
0.99	2.575	$[1.409, 1.871]$

Table j.2. Confidence intervals for the expectation in case of reliability level $1-\alpha$

- If the reliability level is increased, then the length of the interval increases, as well.
- If the number of sample elements tends to infinity, the length of the confidence interval tends to zero.
- If the accuracy is given, we can compute the necessary number of sample elements to a given reliability level. For example, if we would like to have a confidence interval to

the reliability level 0.99 with length 0.1, then $u_\alpha \cdot \frac{\sigma}{\sqrt{n}} \le \frac{0.1}{2}$, $\left(\frac{u_\alpha \cdot \sigma}{0.05}\right)^2 \le n$, that is . $\left(\frac{2.576}{0.05} \cdot 0.2\right)^2 = 107 \le n$. The number of the necessary elements is proportional to the variance and to the square of the reciprocal of the accuracy.

If the dispersion of the random variable is not known then we have to estimate it on the basis of the sample.

Estimation of the variance and the dispersion

As the sample average is the expectation belonging to the empirical distribution function, it is a natural idea to estimate the variance σ^2 by the variance belonging to the empirical distribution function.

Let $s^2 : R^n \to R$, $s^2(y_1, y_2, ..., y_n) = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n}$. Then

$s^2 \circ \xi = s^2(\xi_1, \xi_2, ..., \xi_n) = \frac{\sum_{i=1}^{n} (\xi_i - \overline{\xi})^2}{n}$, where $\overline{\xi} = \frac{\sum_{i=1}^{n} \xi_i}{n}$. $s^2(\xi_1, \xi_2, ..., \xi_n)$ is a random variable.

$$E(s^2(\xi_1, \xi_2, ..., \xi_n)) = E\left(\frac{\sum_{i=1}^{n} (\xi_i - \overline{\xi})^2}{n}\right) =$$

$$E\left(\frac{\sum_{i=1}^{n} (\xi_i - m + m - \overline{\xi})^2}{n}\right) = \frac{1}{n} E\left(\sum_{i=1}^{n} (\xi_i - m - (\overline{\xi} - m))^2\right) =$$

$$\frac{1}{n} E\left(\sum_{i=1}^{n} (\xi_i - m)^2 - 2\sum_{i=1}^{n} (\xi_i - m)(\overline{\xi} - m) + n(\overline{\xi} - m))^2\right) = \frac{1}{n} E\left(\sum_{i=1}^{n} (\xi_i - m)^2\right) - E((\overline{\xi} - m)^2).$$

$$E((\overline{\xi} - m)^2) = E\left(\left(\frac{\sum_{i=1}^{n} \xi_i}{n} - m\right)^2\right) = \frac{1}{n^2} E\left(\sum_{i=1}^{n} (\xi_i - m)^2 - 2\sum_{i<j} (\xi_i - m)(\xi_j - m)\right)$$

$= \frac{1}{n^2} \sum_{i=1}^{n} E\left((\xi_i - m)^2\right) = \frac{1}{n}\sigma^2$. This implies $E\left(\frac{\sum_{i=1}^{n} (\xi_i - \overline{\xi})^2}{n}\right) = \sigma^2 - \frac{1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2$.

Consequently, $E\left(\frac{\sum_{i=1}^{n}\left(\xi_i-\overline{\xi}\right)^2}{n}\right)\neq\sigma^2$. Let

$$s^{*2}(y_1,y_2,...,y_n)=\frac{\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2}{n-1}=\frac{n}{n-1}s^2(y_1,y_2,...,y_n).$$

$$E(s^{*2}(\xi_1,\xi_2,...,\xi_n))=E\left(\frac{\sum_{i=1}^{n}\left(\xi_i-\overline{\xi}\right)^2}{n-1}\right)=\frac{n}{n-1}\cdot E\left(\frac{\sum_{i=1}^{n}\left(\xi_i-\overline{\xi}\right)^2}{n}\right)=\frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2=\sigma^2.$$

$s^{*2}(\xi_1,\xi_2,...,\xi_n)$ is briefly denoted by s^{*2}. It can be proved that if $E(\xi_i{}^4)$ exists, then $D^2(s^{*2}(\xi_1,\xi_2,...,\xi_n))\to 0$, if $n\to\infty$. Summarizing, s^{*2} is a consistent estimation of the variance. Now it is worth estimating the dispersion by the statistics

$$s^*(\xi_1,\xi_2,...\xi_n)=\sqrt{s^{*2}(\xi_1,\xi_2,...\xi_n)}=\sqrt{\frac{\sum_{i=1}^{n}\left(\xi_i-\overline{\xi}\right)^2}{n-1}}.$$

<u>Definition</u> The statistics $s^*(\xi_1,\xi_2,...\xi_n)=\sqrt{\frac{\sum_{i=1}^{n}\left(\xi_i-\overline{\xi}\right)^2}{n-1}}$ is called the **corrected empirical dispersion**.

To construct confidence interval for the variance and the dispersion we state the following theorem without proof (Fisher-Cochran's theorem)

<u>Theorem</u> If $\xi_i\sim N(m,\sigma)$, then $(n-1)\frac{s^{*2}(\xi_1,\xi_2,...\xi_n)}{\sigma^2}\sim\chi^2_{n-1}$, furthermore $\overline{\xi}$ and $s^{*2}(\xi_1,\xi_2,...\xi_n)$ are independent random variables. By definition of Student's t distribution (see chapter g), this also implies that $\frac{\overline{\xi}-m}{s^*(\xi_1,\xi_2,...\xi_n)}\sqrt{n}\sim\tau_{n-1}$.

<u>Remarks</u>

- χ^2_n distributed random variables were presented in Chapter g. The explicit forms of their cumulative distribution functions are not usually used. There are tables (see Table 3.) which contain the real values $\chi^2_{n,\alpha}$ for which $P(\chi^2_{n,\alpha}\leq\theta)=\alpha$ supposing $\theta\sim\chi^2_n$. This means that $P(\theta<\chi^2_{n,\alpha})=1-\alpha$. These values $\chi^2_{n,\alpha}$ are called the critical values belonging to the reliability level $1-\alpha$.
- By the help of the critical values belonging to $1-\frac{\alpha}{2}$ and $\frac{\alpha}{2}$ one can construct an interval in which the values of a χ^2_n distributed random variable are situated with probability $1-\alpha$. Namely, $P(\chi^2_{n,1-\alpha/2}\leq\theta\leq\chi^2_{n,\alpha/2})$. These intervals will be used to

construct such intervals in which the variance and the dispersion are situated with probability $1-\alpha$.

If $\xi_i \sim N(m,\sigma)$, then $(n-1)\dfrac{s^{*2}(\xi_1,\xi_2,...,\xi_n)}{\sigma^2}=\dfrac{\sum_{i=1}^{n}\left(\xi_i-\bar{\xi}\right)^2}{\sigma^2}\sim\chi^2_{n-1}$, consequently,

$P(\chi^2_{n,1-\alpha/2}\le(n-1)\dfrac{s^{*2}}{\sigma^2}\le\chi^2_{n,\alpha/2})=1-\alpha$. Arranging the sides of the inequalities we end up with $P((n-1)\dfrac{s^{*2}}{\chi^2_{n,\alpha/2}}\le\sigma^2\le(n-1)\dfrac{s^{*2}}{\chi^2_{n,1-\alpha/2}})=1-\alpha$. As a straightforward consequence, $P(\sqrt{(n-1)\dfrac{s^{*2}}{\chi^2_{n,\alpha/2}}}\le\sigma\le\sqrt{(n-1)\dfrac{s^{*2}}{\chi^2_{n,1-\alpha/2}}})=1-\alpha$. Summarizing, supposing normally distributed samples or large number of elements, the confidence interval for the variance belonging to the reliability level $1-\alpha$ looks like

$$\boxed{\left[(n-1)\frac{s^{*2}}{\chi^2_{n,\alpha/2}},(n-1)\frac{s^{*2}}{\chi^2_{n,1-\alpha/2}}\right]}$$

and that for the dispersion it is

$$\boxed{\left[\sqrt{(n-1)\frac{s^{*2}}{\chi^2_{n,\alpha/2}}},\sqrt{(n-1)\frac{s^{*2}}{\chi^2_{n,1-\alpha/2}}}\right]}.$$

Remarks

- Due to the central limit theorem, the assumption of normally distributed sample can be omitted if n is large.
- If we have the value of the sample, we can construct the confidence intervals for the variance and the dispersion by the following steps: compute the value of s^{*2}, find the critical value belonging to the reliability levels $\dfrac{\alpha}{2}$ and $1-\dfrac{\alpha}{2}$, then substitute them into the formulae in the boxes.
- For example, assuming normally distributed sample, if $x_1=1.5$, $x_2=1.7$, $x_3=1.4$, $x_4=1.9$, $x_5=1.7$ then $\bar{x}=1.64$ and $s^{*2}=\dfrac{\sum_{i=1}^{5}\left(x_i-\bar{x}\right)^2}{4}=$

$$=\frac{(1.5-1.64)^2+(1.7-1.64)^2+(1.4-1.64)^2+(1.9-1.64)^2+(1.7-1.64)^2}{4}==0.038\,.$$

Confidence intervals belonging to the reliability levels 0.9, 0.95, 0.98 and 0.99 are included in Table j.3.

$1-\alpha$	$\chi^2_{4,1-\alpha/2}$	$\chi^2_{4,\alpha/2}$	$\left[4\cdot\frac{s*^2}{\chi^2_{4,\alpha/2}}, 4\cdot\frac{s*^2}{\chi^2_{4-\alpha/2}}\right]$	$\left[\sqrt{(n-1)\frac{s*^2}{\chi^2_{4,\alpha/2}}}, \sqrt{(n-1)\frac{s*^2}{\chi^2_{4,1-\alpha/2}}}\right]$
0.9	0.711	9. 488	$[0.016, \ 0.214]$	$[0.127, \ 0.462,]$
0.95	0.484	8. 496	$[0.018, 0.314]$	$[0.134, \ 0.560]$
0.98	0.297	13.277	$[0.011, 0.512]$	$[0.107, \ 0.715]$
0.99	0.207	14. 86	$[0.010, 0.734]$	$[0.101, \ 0.857]$

Table j.3. Critical values and confidence intervals for the variance and dispersion in case of reliability levels $1-\alpha$

- The greater the reliability, the larger the interval is.

Finally let us return to the estimation of the expectation in case of unknown dispersion.

Estimation of the expectation in case of unknown dispersion

Taking the sample average does not require the knowledge of the dispersion. Furthermore, estimating the expectation by the sample average, $E(\hat{m}) = E(\frac{\sum_{i=1}^{n}\xi_i}{n}) = m$, and $D\left(\frac{\sum_{i=1}^{n}\xi_i}{n}\right) \to 0$ holds in the case of unknown value of σ, as well.

Turning to the confidence interval for the expectation, apply Fisher-Cochran' theorem and the formula $\frac{\bar{\xi}-m}{s*}\sqrt{n} \sim \tau_{n-1}$ in case of normally distributed samples.

There are tables of Student's t distribution, in which one can find the real numbers $t_{n,\alpha}$, for which $P(-t_{n,\alpha} \le t_n \le t_{n,\alpha}) = 1-\alpha$. The value $t_{n,\alpha}$ is called the critical value belonging to the reliability level $1-\alpha$. Now, $P(-t_{n-1,\alpha} \le \frac{\bar{\xi}-m}{s*}\sqrt{n} \le t_{n-1,\alpha}) = 1-\alpha$. Arranging both sides of the inequalities we end up with $P(\bar{\xi} - \frac{t_{n-1,\alpha}\cdot s*}{\sqrt{n}} \le m \le \bar{\xi} + \frac{t_{n-1,\alpha}\cdot s*}{\sqrt{n}}) = 1-\alpha$.

Summarizing, the confidence interval for the expectation belonging to the reliability level $1-\alpha$ is

$$\boxed{\left[\bar{\xi} - \frac{t_{n-1,\alpha}\cdot s*}{\sqrt{n}}, \bar{\xi} + \frac{t_{n-1,\alpha}\cdot s*}{\sqrt{n}}\right]}.$$

Remarks

- Note that the confidence intervals for the expectation are very similar in the cases of known and unknown dispersion. In case of unknown dispersion, σ is replaced by its estimation, s^*, and the critical value is $t_{n-1,\alpha}$ instead of u_α.
- The larger the reliability level, the larger the interval.
- The larger the number of elements, the smaller the critical value.
- The limit of the critical values $t_{n,\alpha}$ is u_α, that is $\lim_{n\to\infty} t_{n,\alpha} = u_\alpha$.This is due to the statement that the cumulative function of a standard normally distributed random variable is the limit of the cumulative distribution functions of Student's t distributed random variables.
- The confidence intervals belonging to a given reliability level can be constructed after executing the following steps: compute s^* on the basis of the sample, find the critical value and substitute into the above formula. In case of a normally distributed sample and $x_1 = 1.5$, $x_2 = 1.7$, $x_3 = 1.4$, $x_4 = 1.9$, $x_5 = 1.7$, $\overline{x} = 1.64$ and $s^* = \sqrt{0.038}$. The confidence intervals belonging to the reliability levels 0.9, 0.95, 0.975 and 0.99 are presented in Table j.5.

$1-\alpha$	$t_{4,\alpha}$	$\left[\overline{\xi} - \frac{t_{4,\alpha} \cdot s^*}{\sqrt{5}}, \overline{\xi} + \frac{t_{4,\alpha} \cdot s^*}{\sqrt{5}}\right]$
0.9	2.132	$[1.454, 1.826]$
0.95	2.776	$[1.398, 1.882]$
0.975	3.495	$[1.335, 1.945]$
0.99	4.604	$[1.239, 2.041]$

Table j.5. Critical values and confidence intervals for the expectation in case of unknown value of dispersion

j.3. Testing hypothesis

An important branch of mathematical statistics is testing hypothesis. Hypothesis is an idea about the value of probability, expectation, dispersion, a parameter or about the cumulative distribution function itself. We check whether the hypothesis can be true or not, more exactly, the data contradict the hypothesis or not. The main idea of testing hypothesis is the following: if the hypothesis holds, then a certain function of the sample has a known distribution. This implies that one can determine an interval in which the function of the sample is situated with a given reliability $1-\alpha$. If the hypothesis does hold, the values of the function (test function) are outside that interval with probability α. The mentioned interval is called the acceptation region; its compliment is the critical region. Then, check whether the test function is really in the acceptation region. If it is, then the data do not contradict the hypotheses. If it is not, there are two reasons for which this may happen: the hypothesis does not hold or the hypothesis holds and an event with small probability α occurs. Statisticians vote for the first one, hence we do not accept the hypothesis, because we rather trust in the alternative than in the occurrence of rare event. Of course, the decision may be wrong.

The name of the basic idea is null hypotheses (H_0), the name of the opposite is alternative hypothesis (H_1). They have to be mutually exclusive but they may not cover all the possibilities concerning the parameter. For example, H_0 is that the probability of an event is 0.4, the alternative hypothesis is that the probability of the event is smaller than 0.4.

The decision, whether we accept (fail to reject) H_0 or reject it, may be right or wrong. The following four cases can be distinguished:

	H_0 is accepted	H_0 is rejected
H_0 is true	Right decision	Wrong decision
H_0 is not true	Wrong decision	Right decision

Table j.6. Possibilities concerning the decisions in testing a hypothesis

Decision that H_0 is true, although it is rejected is called as error of the first kind (type I. error), its probability is α. The probability of the first kind error is usually called as the level of significance.
Decision that H_0 is not true, although it is not failed to reject is called as error of the second kind (type II. error). Its probability depends on the value of the tested parameter, for example. Consequences of the different kind of errors are of various severities.

Remarks

- Usually applied significance levels are $\alpha = 0.05$ and $\alpha = 0.01$.
- Some test functions are connected with the statistics presented in the previous subsection.
- The elaborated tests can be executed as a recipe in the kitchen. Their steps are the following:

State H_0 and H_1, fix the level of significance.

Determine the critical region and the acceptance region.

Compute the actual value of the test function by substituting the values of the sample elements into the test function.

Check whether the actual value of the test function is in the critical region or in the acceptance region.

Make your decision: if the actual value of the test function is in the critical region, reject H_0, if it is in the acceptance region, accept H_0.

- If H_0 is accepted, then H_0 may be untrue but the data do not contradict to this assumption. If you doubt in H_0 you should take a sample of more elements.

In the latest part of this subsection we present tests for the probability, expectation, variance and cumulative distribution function. We explain the task, present the test function, critical and acceptance region and decision itself in all cases, separately.

Test for the probability

During this problem we have to decide about the probability of an event, whether it can be a fixed number or not.

Let $\underline{\xi} = (\xi_1, \xi_2, ..., \xi_n)$ be the sample, $\xi_i = 1_A{}^i = \begin{cases} 1 & \text{if } A \text{ occurs at the } i\text{th experiment} \\ 0 & \text{if } \overline{A} \text{ occurs at the } i\text{th experiment} \end{cases}$.

Now, $\sum_{i=1}^{n} \xi_i = k_A(n)$, the frequency of A, and $\dfrac{\sum_{i=1}^{n} \xi_i}{n} = \dfrac{k_A(n)}{n}$ is its relative frequency.

Let $H_0 : P(A) = p_0$, $H_1 : P(A) \neq p_0$, where p_0 is the idea about the probability of the event. If $100 \leq n$, $10 \leq np_0$ is satisfied, then by the central limit theorem we can state, that $\dfrac{\frac{k_A(n)}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$ supposing that H_0 holds. Consequently, let the test function be

$u = \dfrac{\frac{k_A(n)}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$. If H_0 holds, then $P(-u_\alpha \leq \dfrac{\frac{k_A(n)}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leq u_\alpha) = 1-\alpha$, where

$\Phi(u_\alpha) = 1 - \dfrac{\alpha}{2}$, coinciding with the previous subsection. The critical region is $(-\infty, -u_\alpha) \cup (u_\alpha, \infty)$ and the acceptance region is $[-u_\alpha, u_\alpha]$. The critical value u_α and its opposite are the bounds of the critical region. If the actual value of $\dfrac{\frac{k_A(n)}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ is in the interval $[-u_\alpha, u_\alpha]$, then H_0 is accepted, in the opposite case H_0 is rejected and H_1 is accepted. The level of significance equals α.

Let $H_0 : P(A) = p_0$ and $H_1 : P(A) < p_0$ a one sided alternative hypothesis. Then, if H_0 holds, then $\dfrac{\frac{k_A(n)}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$, and $P(-u_{2\alpha} < \dfrac{\frac{k_A(n)}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}) = 1-\alpha$ supposing $100 \leq n$, $10 \leq np_0$. The critical region is $(-\infty, -u_{2\alpha})$, the acceptance region is $[-u_{2\alpha}, \infty)$. If the actual value of the test function $\dfrac{\frac{k_A(n)}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ is at least $-u_{2\alpha}$, then we accept H_0, if it is less than $-u_{2\alpha}$ we reject H_0 and we accept H_1. Then the data rather support that $P(A) < p_0$ and they contradict to $P(A) = p_0$.

<u>Remarks</u>

- The alternative hypothesis $H_1 : p_0 < p$ can be similarly handled.
- The smaller the significance level, the larger the acceptance region.
- The larger the number of sample elements, the smaller the value of $\sqrt{\dfrac{p_0(1-p_0)}{n}}$

and the larger is its reciprocal. Consequently, greater difference can be accepted between the relative frequency and the real probability in case of a small number of sample elements. Same difference between the relative frequency and the real probability may result in acceptance of H_0 for a small number of elements of the sample and in rejection of H_0 in case of large number of elements of the sample.

- Same difference between the relative frequency and the real probability may result in acceptance of H_0 for small number of elements of sample and in rejection of H_0 in case of large number of elements of the sample.
- Acceptance of in case of two sided alternative hypothesis and rejection of in case of one sided alternative hypothesis may happen at the same significance level. Example will be presented later.

Example

E1. Let the relative frequency of an event A during n independent experiments be 0.35. Test the hypothesis $H_0: P(A) = 0.4$ and $H_1: P(A) \neq 0.4$ in the case of significance levels $\alpha = 0.1$, $\alpha = 0.05$, $\alpha = 0.01$ and number of sample elements $n = 100$, $n = 300$, $n = 600$. Results are included in Table j.7.

α, n	u_α	Critical region	Actual value of the test function	Decision
$\alpha = 0.1$, n=100	1.645	$(-\infty, -1.645) \cup (1.645, \infty)$	-1. 0206	H_0 is accepted, H_1 is rejected
$\alpha = 0.1$, n=300	1.645	$(-\infty, -1.645) \cup (1.645, \infty)$	-1. 7678	H_0 is rejected, H_1 is accepted
$\alpha = 0.1$, n=600	1.645	$(-\infty, -1.645) \cup (1.645, \infty)$	-2. 5	H_0 is rejected, H_1 is accepted
$\alpha = 0.05$, n=100	1.96	$(-\infty, -1.96) \cup (1.96, \infty)$	-1. 0206	H_0 is accepted, H_1 is rejected
$\alpha = 0.05$, n=300	1.96	$(-\infty, -1.96) \cup (1.96, \infty)$	-1. 7678	H_0 is accepted, H_1 is rejected
$\alpha = 0.05$, n=600	1.96	$(-\infty, -1.96) \cup (1.96, \infty)$	-2. 5	H_0 is rejected, H_1 is accepted
$\alpha = 0.01$, n=100	2.576	$(-\infty, -2.576) \cup (2.576, \infty)$	-1. 0206	H_0 is accepted, H_1 is rejected
$\alpha = 0.01$, n=300	2.576	$(-\infty, -2.576) \cup (2.576, \infty)$	-1. 7678	H_0 is accepted, H_1 is rejected
$\alpha = 0.01$, n=600	2.576	$(-\infty, -2.576) \cup (2.576, \infty)$	-2. 5	H_0 is accepted, H_1 is rejected

Table j.7. Testing hypothesis $p = 0.4$ with two sided alternative hypothesis

E2. Let the relative frequency of an event A during n independent experiments be 0.35. Test the hypothesis $H_0: P(A) = 0.4$ and $H_1: P(A) < 0.4$ in the case of significance levels $\alpha = 0.1$, $\alpha = 0.05$, $\alpha = 0.01$ and number of elements of the samples $n = 100$, $n = 300$, $n = 600$. Results are included in Table j.8.

α, n	$u_{2\alpha}$	Critical region	Actual value of the test function	Decision
$\alpha = 0.1$, n=100	1.282	$(-\infty, -1.282)$	-1. 0206	H_0 is accepted, H_1 is rejected

$\alpha = 0.1$, n=300	1.282	$(-\infty,-1.282)$	-1. 7678	H_0 is rejected, H_1 is accepted
$\alpha = 0.1$, n=600	1.282	$(-\infty,-1.282)$	-2. 5	H_0 is rejected, H_1 is accepted
$\alpha = 0.05$, n=100	1.645	$(-\infty,-1.645)$	-1. 0206	H_0 is accepted, H_1 is rejected
$\alpha = 0.05$, n=600	1.645	$(-\infty,-1.645)$	-1. 7678	H_0 is rejected, H_1 is accepted
$\alpha = 0.05$, n=600	1.645	$(-\infty,-1.645)$	-2. 5	H_0 is rejected, H_1 is accepted
$\alpha = 0.01$, n=100	2.326	$(-\infty,-2.326)$	-1. 0206	H_0 is accepted, H_1 is rejected
$\alpha = 0.01$, n=300	2.326	$(-\infty,-2.326)$	-1. 7678	H_0 is accepted, H_1 is rejected
$\alpha = 0.01$, n=600	2.326	$(-\infty,-2.326)$	-2. 5	H_0 is rejected, H_1 is accepted

Table j.8. Testing hypothesis $p = 0.4$ with one sided alternative hypothesis

Test for the expectation in case of known value of dispersion

Let $\eta = (\xi_1, \xi_2, ..., \xi_n)$ be a sample, ξ_i are random variables with expectation m and with known dispersion σ. We would like to check weather $H_0 : m = m_0$ holds or conversely, $H_1 : m \neq m_0$. If $\xi_i \sim N(m, \sigma)$ or $100 \le n$, then $\dfrac{\frac{\sum_{i=1}^{n} \xi_i}{n} - m}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$. Consequently, if H_0 holds, then

$$P\left(-u_\alpha < \frac{\frac{\sum_{i=1}^{n} \xi_i}{n} - m_0}{\frac{\sigma}{\sqrt{n}}} < u_\alpha \right) = 1 - \alpha.$$

The critical region is $(-\infty, -u_\alpha) \cup (u_\alpha, \infty)$, the acceptance region is $[-u_\alpha, u_\alpha]$. Using the test function $u = \dfrac{\frac{\sum_{i=1}^{n} \xi_i}{n} - m_0}{\frac{\sigma}{\sqrt{n}}}$, if the actual value of the test function is in the critical region then H_0 is rejected, if it is in the acceptance region then H_0 is accepted.

If the alternative hypothesis is $H_1 : m < m_0$, then the critical region is $(-\infty, -u_{2\alpha})$, the acceptance region is $(-u_{2\alpha}, \infty)$. If the actual value of the test function $u = \dfrac{\frac{\sum_{i=1}^{n} \xi_i}{n} - m_0}{\frac{\sigma}{\sqrt{n}}}$ is in the acceptance region then H_0 is accepted, if it is in the critical region then H_0 is rejected and H_1 is accepted.

Remarks

- The alternative hypothesis $H_1 : m_0 < m$ can be similarly handled.
- The smaller the significance level, the larger the acceptance region.
- The larger the number of elements of the sample, the smaller difference between the average and the real expectation can be allowed if H_0 is accepted.
- The necessary number of elements of the sample to detect difference ε between the real and the hypothetical expectation is $\left(\frac{u_\alpha \sigma}{\varepsilon}\right)^2 \le n$. It is proportional to the variance and the square of the reciprocal of the difference to be detected.
- The case when H_0 is rejected applying two sided alternative hypothesis and H_0 is accepted applying one sided alternative hypothesis may occur.
- The test function requires the knowledge of the dispersion.

Example

E3. Let $\xi_i \sim N(m, \sigma)$. Let us assume that the dispersion of the random variable investigated equals 1.2. The computed sample average is supposed to be 100.5. Test the hypothesis that $H_0 : m = 100$ and $H_1 : m \ne 100$ if the level is significance is $\alpha = 0.1$, $\alpha = 0.05$, $\alpha = 0.01$ and the number of sample elements are $n = 10$, $n = 30$, $n = 50$.
Results are included in Table j.9.

α, n	u_α	Critical region	Actual value of the test function	Decision
$\alpha = 0.1$, n=10	1.645	$(-\infty, -1.645) \cup (1.645, \infty)$	1. 3176	H_0 is accepted
$\alpha = 0.1$, n=30	1.645	$(-\infty, -1.645) \cup (1.645, \infty)$	2. 2822	H_0 is rejected, H_1 is accepted
$\alpha = 0.1$, n=50	1.645	$(-\infty, -1.645) \cup (1.645, \infty)$	2. 9463	H_0 is rejected, H_1 is accepted
$\alpha = 0.05$, n=10	1.96	$(-\infty, -1.96) \cup (1.96, \infty)$	1. 3176	H_0 is accepted
$\alpha = 0.05$, n=30	1.96	$(-\infty, -1.96) \cup (1.96, \infty)$	2. 2822	H_0 is rejected, H_1 is accepted
$\alpha = 0.05$, n=50	1.96	$(-\infty, -1.96) \cup (1.96, \infty)$	2. 9463	H_0 is rejected, H_1 is accepted
$\alpha = 0.01$,	2.576	$(-\infty, -2.576) \cup (2.576, \infty)$	1. 3176	H_0 is accepted

n=10				
$\alpha=0.01$, n=30	2.576	$(-\infty,-2.576)\cup(2.576,\infty)$	2. 2822	H_0 is accepted
$\alpha=0.01$, n=50	2.576	$(-\infty,-2.576)\cup(2.576,\infty)$	2. 9463	H_0 is rejected, H_1 is accepted

Table j.9. Testing hypothesis $m=100$ with two sided alternative hypothesis

E4. Let $\xi_i \sim N(m,\sigma)$. Let us assume that the dispersion of the random variable investigated equals 1.2. The computed sample average is supposed to be 100.5. Test the hypothesis that $H_0: m=100$ and $H_1: 100<m$, if the level is significance is $\alpha=0.1$, $\alpha=0.05$, $\alpha=0.01$ and the number of sample elements are $n=10$, $n=30$, $n=50$.
Results are included in Table j.10.

α, n	$u_{2\alpha}$	Critical region	Actual value of the test function	Decision
$\alpha=0.1$, n=10	1.282	$(1.282,\infty)$	1. 3176	H_0 is rejected, H_1 is accepted
$\alpha=0.1$, n=30	1.282	$(1.282,\infty)$	2. 2822	H_0 is rejected, H_1 is accepted
$\alpha=0.1$, n=50	1.282	$(1.282,\infty)$	2. 9463	H_0 is rejected, H_1 is accepted
$\alpha=0.05$, n=10	1.645	$(1.645,\infty)$	1. 3176	H_0 is accepted, H_1 is rejected
$\alpha=0.05$, n=30	1.645	$(1.645,\infty)$	2. 2822	H_0 is rejected, H_1 is accepted
$\alpha=0.05$, n=50	1.645	$(1.645,\infty)$	2. 9463	H_0 is rejected, H_1 is accepted
$\alpha=0.01$, n=10	2.326	$(2.326,\infty)$	1. 3176	H_0 is accepted, H_1 is rejected
$\alpha=0.01$, n=30	2.326	$(2.326,\infty)$	2. 2822	H_0 is accepted, H_1 is rejected
$\alpha=0.01$, n=50	2.326	$(2.326,\infty)$	2. 9463	H_0 is rejected, H_1 is accepted

Table j.10. Testing hypothesis $m=100$ with one sided alternative hypothesis

Test for the expectation in case of unknown value of dispersion

Let $\eta=(\xi_1,\xi_2,...,\xi_n)$ be the sample, ξ_i are random variables with expectation m and dispersion σ but the value of the dispersion is unknown. Let us assume that $\xi \sim N(m,\sigma)$ or the number of the elements of the sample is large. We would like to check whether $H_0: m=m_0$ holds or conversely, $H_1: m\neq m_0$. If $\xi_i \sim N(m,\sigma)$ or $100\leq n$, then

$\dfrac{\dfrac{\sum_{i=1}^{n}\xi_i}{n}-m}{\dfrac{\sigma}{\sqrt{n}}} \sim N(0,1)$. As we do not know the value of σ, we can not compute the actual value of the above statistics. If we use s^* instead of σ, then $\dfrac{\dfrac{\sum_{i=1}^{n}\xi_i}{n}-m_0}{\dfrac{s^*}{\sqrt{n}}} \sim \tau_{n-1}$ supposing H_0 holds. Consequently,

$$P\left(-t_{n-1,\alpha} < \frac{\dfrac{\sum_{i=1}^{n}\xi_i}{n}-m_0}{\dfrac{s^*}{\sqrt{n}}} < t_{n-1,\alpha}\right) = 1-\alpha .$$

The critical region is $(-\infty,-t_{n-1,\alpha}) \cup (t_{n-1,\alpha},\infty)$, the acceptance region is $[-t_{n-1,\alpha}, t_{n-1,\alpha}]$.

Using the test function $t = \dfrac{\dfrac{\sum_{i=1}^{n}\xi_i}{n}-m_0}{\dfrac{s^*}{\sqrt{n}}}$, if the actual value of the test function is in the critical region then H_0 is rejected, if it is in the acceptance region then H_0 is accepted.

If the alternative hypothesis is $H_1: m < m_0$, then the critical region is $(-\infty,-t_{2\alpha})$, the acceptance region is $[-t_{2\alpha},\infty)$. If the actual value of the test function, that is $\dfrac{\dfrac{\sum_{i=1}^{n}x_i}{n}-m_0}{\dfrac{s^*}{\sqrt{n}}}$, is in the acceptance region, then H_0 is accepted, if it is in the critical region then H_0 is rejected and H_1 is accepted.

If $H_0: m = m_0$ and $H_1: m < m_0$, then the critical region is $(-\infty,-t_{n,2\alpha})$ and acceptance region is $[-t_{2\alpha},\infty)$. If the actual value of the test function is in the acceptance region then H_0 is accepted, if it is in the critical region then H_0 is rejected and H_1 is accepted.

<u>Remarks</u>

- The alternative hypothesis $H_1: m_0 < m$ can be similarly handled.
- The smaller the significance level, the larger the acceptance region.
- The larger the number of elements of the sample, the smaller difference between the average and the real expectation can be allowed if H_0 is expected.

- The case when H_0 is rejected applying two sided alternative hypothesis and H_0 is accepted applying one sided alternative hypothesis may occur.
- Note that test functions in case of known and unknown dispersion are very similar.

Example

E5. Let $\xi_i \sim N(m,\sigma)$. Let us assume that the corrected empirical dispersion computed from the sample equals 1.2. The sample average is supposed to be 100.5. Test the hypothesis that $H_0: m = 100$ and $H_1: m \neq 100$, if the level is significance are $\alpha = 0.1$, $\alpha = 0.05$, $\alpha = 0.01$ and the number of sample elements are $n = 10$, $n = 30$, $n = 50$.
The results can be seen in Table j.11.

α, n	t_α	Critical region	Actual value of the test function	Decision
$\alpha = 0.1$, n=10	1.833	$(-\infty,-1.833)\cup(1.833,\infty)$	1. 3176	H_0 is accepted
$\alpha = 0.1$, n=30	1.697	$(-\infty,-1.697)\cup(1.697,\infty)$	2. 2822	H_0 is rejected, H_1 is accepted
$\alpha = 0.1$, n=50	1.676	$(-\infty,-1.676)\cup(1.676,\infty)$	2. 9463	H_0 is rejected, H_1 is accepted
$\alpha = 0.05$, n=10	2.262	$(-\infty,-2.262)\cup(2.262,\infty)$	1. 3176	H_0 is accepted, H_1 is rejected
$\alpha = 0.05$, n=30	2.042	$(-\infty,-2.042)\cup(2.042,\infty)$	2. 2822	H_0 is rejected, H_1 is accepted
$\alpha = 0.05$, n=50	2.009	$(-\infty,-2.009)\cup(2.009,\infty)$	2. 9463	H_0 is rejected, H_1 is accepted
$\alpha = 0.01$, n=10	3.250	$(-\infty,-3.250)\cup(3.250,\infty)$	1. 3176	H_0 is accepted, H_1 is rejected
$\alpha = 0.01$, n=30	2.750	$(-\infty,-2.750)\cup(2.750,\infty)$	2. 2822	H_0 is accepted, H_1 is rejected
$\alpha = 0.01$, n=50	2.678	$(-\infty,-2.678)\cup(2.678,\infty)$	2. 9463	H_0 is rejected, H_1 is accepted

Table j.11. Testing hypothesis $m = 100$ in case of unknown dispersion with two sided alternative hypothesis

E6. Let $\xi_i \sim N(m,\sigma)$. Let us assume that corrected empirical dispersion computed by the sample equals 1.2. The sample average is supposed to be 100.5. Test the hypothesis that $H_0: m=100$ and $H_1: 100<m$ if the level is significance are $\alpha=0.1$, $\alpha=0.05$, $\alpha=0.01$ and the number of sample elements are $n=10$, $n=30$, $n=50$.
Results can be followed in Table j.12.

α, n	$t_{2\alpha}$	Critical region	Actual value of the test function	Decision
$\alpha=0.1$, n=10	1.383	$(1.383,\infty)$	1. 3176	H_0 is accepted, H_1 is rejected
$\alpha=0.1$, n=30	1.310	$(1.310,\infty)$	2. 2822	H_0 is rejected, H_1 is accepted
$\alpha=0.1$, n=50	1.299	$(1.299,\infty)$	2. 9463	H_0 is rejected, H_1 is accepted
$\alpha=0.05$, n=10	1.833	$(1.833,\infty)$	1. 3176	H_0 is accepted, H_1 is rejected
$\alpha=0.05$, n=30	1.697	$(1.697,\infty)$	2. 2822	H_0 is rejected, H_1 is accepted
$\alpha=0.05$, n=50	1.676	$(1.676,\infty)$	2. 9463	H_0 is rejected, H_1 is accepted
$\alpha=0.01$, n=10	2.821	$(2.821,\infty)$	1. 3176	H_0 is accepted, H_1 is rejected
$\alpha=0.01$, n=30	2.462	$(2.462,\infty)$	2. 2822	H_0 is accepted, H_1 is rejected
$\alpha=0.01$, n=50	2.405	$(2.405,\infty)$	2. 9463	H_0 is rejected, H_1 is accepted

Table j.12. Testing hypothesis $m=100$ in case of unknown dispersion with one sided alternative hypothesis

Test for the value of variance

Let $\underline{\xi}=(\xi_1,\xi_2,...,\xi_n)$ be a sample, ξ_i are random variables with expectation m and dispersion σ. We would like to check weather $H_0:\sigma^2=\sigma_0^2$ holds or conversely, $H_1:\sigma^2\neq\sigma_0^2$. Recall that if $\xi_i \sim N(m,\sigma)$ or n is large, then $\frac{(n-1)s^{*2}}{\sigma_0^2}\sim\chi^2_{n-1}$ supposing

H_0 holds. Consequently, $P(\chi^2_{n-1,1-\alpha/2} \le \frac{(n-1)s^{*2}}{\sigma_0^2} \le \chi^2_{n-1,\alpha/2}) = 1-\alpha$. The test function is $\chi^2 = \frac{(n-1)s^{*2}}{\sigma_0^2}$. The critical region is $\left(0,\chi^2_{n-1,1-\alpha/2}\right)\cup(\chi^2_{n-1,\alpha/2},\infty)$, the acceptance region is $\left[\chi^2_{n-1,1-\alpha/2},\chi^2_{n-1,\alpha/2}\right]$. If the actual value of the test function is in the acceptance region then H_0 is accepted, if it is in the critical region then H_0 is rejected and H_1 is accepted.

If the alternative hypothesis is $H_1:\sigma^2<\sigma_0^2$, then $P(\frac{(n-1)s^{*2}}{\sigma_0^2}<\chi^2_{n-1,1-\alpha})=\alpha$. Now, acceptance region is $(\chi^2_{n-1,\alpha},\infty)$, critical region is $[0,\chi^2_{n-1,1-\alpha}]$. If the actual value of the test function is in the acceptance region then H_0 is accepted, if it is in the critical region then H_0 is rejected and H_1 is accepted.

Finally, if the alternative hypothesis is $H_1:\sigma_0^2<\sigma^2$, then apply $P(\frac{(n-1)s^{*2}}{\sigma_0^2}\le\chi^2_{n-1,\alpha})=1-\alpha$. Now, acceptance region is $\left[0,\chi^2_{n-1,\alpha}\right)$, critical region is $\left[\chi^2_{n-1,\alpha},\infty\right)$. If the actual value of the test function is in the acceptance region then H_0 is accepted, if it is in the critical region then H_0 is rejected and H_1 is accepted.

E7. Let $\xi_i \sim N(m,\sigma)$. Let us assume that corrected empirical dispersion computed by the sample equals s^*=1.3. Test the hypothesis that $H_0:\sigma=1.1$ and $H_1:\sigma\ne1.1$ if the level is significance are $\alpha=0.1$, $\alpha=0.05$, $\alpha=0.01$ and the number of sample elements are $n=10$, $n=30$, $n=50$.

α, n	$\chi_{1-\alpha/2}$	$\chi_{\alpha/2}$	Critical region	Actual value of the test statistics	Decision
$\alpha=0.1$, n=10	16. 919	3. 325	$\left[0, 3.325\right)\cup\left(16.919, \infty\right)$	12. 57	H_0 is accepted, H_1 is rejected
$\alpha=0.1$, n=30	42. 557	17. 708	$\left[0, 17.708\right)\cup\left(42.557, \infty\right)$	40. 504	H_0 is accepted, H1 is rejected
$\alpha=0.1$, n=50	66. 339	33. 93	$\left[0, 33.93\right)\cup\left(66.339, \infty\right)$	68. 438	H_0 is rejected, H_1 is accepted
$\alpha=0.05$, n=10	19. 023	2. 7004	$\left[0,2.7004\right)\cup\left(19.023, \infty\right)$	12. 57	H_0 is accepted, H_1 is rejected
$\alpha=0.05$,	45. 722	16. 047	$\left[0,16.047\right)\cup\left(45.722, \infty\right)$	40. 504	H_0 is

n=30					accepted, H_1 is rejected
$\alpha = 0.05$, n=50	70. 222	31. 555	$[0, 31.555) \cup (70.222, \infty)$	68. 438	H_0 is accepted, H_1 is rejected
$\alpha = 0.01$, n=10	23. 589	1. 7349	$[0, 1.7349) \cup (23.589, \infty)$	12. 57	H_0 is accepted, H_1 is rejected
$\alpha = 0.01$, n=30	52. 336	13. 121	$[0, 13.121) \cup (52.336, \infty)$	40. 504	H_0 accepted, H_1 is rejected
$\alpha = 0.01$, n=50	78. 231	23. 983	$[0, 23.983) \cup (78.231, \infty)$	68. 438	H_0 is accepted, H_1 is rejected

Table j.13. Testing hypothesis $\sigma = 1.1$ with two sided alternative hypothesis

E8. Let $\xi_i \sim N(m, \sigma)$. Let us assume that the corrected empirical dispersion computed from the sample equals $s^* = 1.3$. Test the hypothesis that $H_0 : \sigma = 1.1$ and $H_1 : 1.1 < \sigma$ if the level is significance are $\alpha = 0.1$, $\alpha = 0.05$, $\alpha = 0.01$ and the number of sample elements are $n = 10$, $n = 30$, $n = 50$.

α, n	χ_α	Critical region	Actual value of the test statistics	Decision
$\alpha = 0.1$, n=10	14. 684	$(14.684, \infty)$	12. 57	H_0 is accepted, H_1 is rejected
$\alpha = 0.1$, n=30	39. 087	$(39.087, \infty)$	40. 504	H_0 is rejected, H_1 is accepted
$\alpha = 0.1$, n=50	62. 038	$(62.038, \infty)$	68. 438	H_0 is rejected, H_1 is accepted
$\alpha = 0.05$, n=10	16. 919	$(16.919, \infty)$	12. 57	H_0 accepted, H_1 is rejected
$\alpha = 0.05$, n=30	42. 557	$(42.557, \infty)$	40. 504	H_0 is accepted, H_1 is rejected
$\alpha = 0.05$, n=50	66. 339	$(66.339, \infty)$	68. 438	H_0 is rejected, H_1 is accepted
$\alpha = 0.01$, n=10	21. 666	$(21.666, \infty)$	12. 57	H_0 is accepted, H_1 is rejected
$\alpha = 0.01$, n=30	49. 588	$(49.588, \infty)$	40. 504	H_0 is accepted, H_1 is rejected

$\alpha = 0.01$, n=50	74. 919	$(74.919, \infty)$	68. 438	H_0 is accepted, H_1 is rejected

Table j.14. Testing hypothesis $\sigma = 1.2$ with one sided alternative hypothesis

Kolmogorov-Smirnov' test for the cumulative distribution function

Finally, we present the Komogorov-Smirnov' test to test the distribution of the sample. Namely, the hypothesis is that the cumulative distribution function is a given function or, alternatively data contradict to that. To do that we use the maximum difference between the empirical distribution function constructed from the sample and the hypothetical distribution function.

Let $\eta = (\xi_1, \xi_2, ..., \xi_n)$ be the sample, its values are $x_1, x_2, ..., x_n$. Let $F_e(z)$ be the empirical distribution function constructed on the basis of the sample. Let the null hypothesis be $H_0 : F = F_0$ and let $H_1 : F \neq F_0$. If H_0 holds, then $K(y) = P(\lim_{n\to\infty} \sqrt{n} \sup_{z\in R} |F_e(z) - F(z)| < y)$ can be given for any value of y. The values of this function are included in Table 4.

Therefore, if H_0 holds then fixing the value $1-\alpha$ one can find the value k_α for which $P(\lim_{n\to\infty} \sqrt{n} \sup_{z\in R} |F_e(z) - F(z)| \le k_\alpha) = 1-\alpha$. The critical region is (k_α, ∞), the acceptance region is $[0, k_\alpha]$. Test function is $\sqrt{n} \sup_{z\in R} |F_e(z) - F(z)|$. If the actual value of the test function is in the critical region then H_0 is rejected, if it is in the acceptance region then H_0 is accepted. Referring to the shape of the empirical distribution function, the supremum can be computed as the maximal difference of the cumulative distribution function and the empirical distribution function and its right hand side limit at the points of the values of the sample. Consequently, it is enough to compute the values of the hypothetical distribution function at the points of the sample values, the right hand side limit of that at the same points, furthermore the values of the empirical distribution function and their limits at these points. Taking the differences, and their maximum we get the actual value of the test function.

<u>Example</u>

E9. Let the elements of the sample be $x_1 = 2$, $x_2 = 0.5$, $x_3 = 0.1$, $x_4 = 0.7$, $x_5 = 0.2$. Test that $H_0 : F(z) = 1 - e^{-z}$ or $H_1 : F(z) \neq 1 - e^{-z}$ holds.

First note that the basis of Kolmogorov's test is an asymptotic theorem, hence it is not recommended using it for a sample of 5 elements. Nevertheless, for the sake of simplicity we do that.

The empirical cumulative distribution function is $F_e(z) = \begin{cases} 0 \text{ if } z \le 0.1 \\ 0.2 \text{ if } 0.1 < z \le 0.2 \\ 0.4 \text{ if } 0.2 < z \le 0.5 \\ 0.6 \text{ if } 0.5 < z \le 0.7 \\ 0.8 \text{ if } 0.7 < z \le 2 \\ 1 \text{ if } 2 < z \end{cases}$.

x_i	$F_e(x_i)$	$\lim\limits_{z \to x_i+} F_e(z)$	$F_0(x_i)$	$\lvert F_e(x_i) - F_0(x_i) \rvert$	$\left\lvert \lim\limits_{z \to x_i+} F_e(z) - F_0(x_i) \right\rvert$
0.1	0	0.2	0.095	0.095	0.105
0.2	0.2	0.4	0.181	0.019	0.219
0.5	0.4	0.6	0.393	0.007	0.207
0.7	0.6	0.8	0.503	0.097	0.297
2	0.8	1	0.865	0.065	0.135

Table j.13. Testing hypothesis $F(z) = 1 - e^{-z}$

One can see that $\max \lvert F_e(x_i) - F_0(x_i) \rvert = 0.097$, $\max \left\lvert \lim\limits_{z \to x_i+} F_e(z) - F_0(x_i) \right\rvert = 0.297$, therefore $\max\limits_{x \in R} \lvert F_e(x) - F_0(x) \rvert = 0.297$. Thus the actual value of the test function is $\sqrt{5} \cdot 0.297 = 0.664$.

The critical values for $\alpha = 0.1$, $\alpha = 0.05$, $\alpha = 0.01$ are 1.23, 1.36 and 1.63, respectively, (see Table of Kolmogorov's function), consequently H_0 is accepted in all cases of level of significance. One can check that the hypothesis $H_0 : F(z) = 1 - e^{-1.1z}$ is also excepted on the basis of this data. This means that the conclusion „H_0 is accepted" means that data do not contradict to the hypothesis.

Of course, many other tests exist for testing hypothesis, but their presentation is out of the frame of this booklet.

References

- Csernyák László: Valószínűségszámítás, Nemzeti Tankönyvkiadó, Budapest, 1990.
- Gnedenko B.V.: The theory of probability, Mir Publisher, Moszkva, 1976.
- Halmos Pál: Measure Theory, Springer-Verlag, Berlin, 1974.
- http://wiki.stat.ucla.edu/socr/index.php/AP_Statistics_Curriculum_2007_Prob_Simul#Binomial_Coin_Toss_Experiment
- Prékopa András: Valószínűségelmélet műszaki alkalmazásokkal, Műszaki Könyvkiadó, Budapest, 1974.
- Reimann József –Tóth Júlianna: Valószínűségszámítás és matematikai statisztika, Tankönyvkiadó, Budapest, 1991.
- Rényi Alfréd: Probability theory, Tankönyvkiadó, Budapest, 1981.
- Snell: Introduction to Probability with computing. Random House/Birkhauser Math. Series, 1988.
- Solt György: Valószínűségszámítás, Műszaki Könyvkiadó, Budapest, Budapest, 1991.
- Walpole, Myers, Ye: Probability and Statistics for Engineers and Scientists, Pearson, 2011.

Cumulative distribution function of standard normally distributed random variables

$\Phi(x) = P(\xi < x)$

$\xi \sim N(0,1)$

x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.7	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.8	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999

Table 1. Cumulative distribution function of standard normally distributed random variables

Critical values of Student's t distributed random variables

$$P(t_{n,\alpha} < |\xi|) = \alpha$$
$$\xi \sim \tau_n$$

n\α	0.2	0.1	0.05	0.025	0.01	0.001
1	3.078	6.314	12.706	25.452	63.657	636.621
2	1.886	2.920	4.303	6.205	9.925	31.599
3	1.638	2.353	3.182	4.177	5.841	12.924
4	1.533	2.132	2.776	3.495	4.604	8.610
5	1.476	2.015	2.571	3.163	4.032	6.869
6	1.440	1.943	2.447	2.969	3.707	5.959
7	1.415	1.895	2.365	2.841	3.499	5.408
8	1.397	1.860	2.306	2.752	3.355	5.041
9	1.383	1.833	2.262	2.685	3.250	4.781
10	1.372	1.812	2.228	2.634	3.169	4.587
11	1.363	1.796	2.201	2.593	3.106	4.437
12	1.356	1.782	2.179	2.560	3.055	4.318
13	1.350	1.771	2.160	2.533	3.012	4.221
14	1.345	1.761	2.145	2.510	2.977	4.140
15	1.341	1.753	2.131	2.490	2.947	4.073
16	1.337	1.746	2.120	2.473	2.921	4.015
17	1.333	1.740	2.110	2.458	2.898	3.965
18	1.330	1.734	2.101	2.445	2.878	3.922
19	1.328	1.729	2.093	2.433	2.861	3.883
20	1.325	1.725	2.086	2.423	2.845	3.850
25	1.316	1.708	2.060	2.385	2.787	3.725
30	1.310	1.697	2.042	2.360	2.750	3.646
35	1.306	1.690	2.030	2.342	2.724	3.591
40	1.303	1.684	2.021	2.329	2.704	3.551
50	1.299	1.676	2.009	2.311	2.678	3.496
60	1.296	1.671	2.000	2.299	2.660	3.460
70	1.294	1.667	1.994	2.291	2.648	3.435
80	1.292	1.664	1.990	2.284	2.639	3.416
90	1.291	1.662	1.987	2.280	2.632	3.402
100	1.290	1.660	1.984	2.276	2.626	3.390
∞	1.282	1.645	1.960	2.241	2.576	3.291

Table 2. Critical values of Student's t distributed random variables

Critical values of χ^2 distributed random variables

$$P(\chi^2_{n,\alpha} < \xi) = \alpha$$
$$\xi \sim \chi^2_n$$

n\ α	0.999	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.001
1	.00	.00	.00	.00	.02	2.71	3.84	5.02	6.63	10.83
2	.00	.02	.05	.10	.21	4.61	5.99	7.38	9.21	13.82
3	.02	.11	.22	.35	.58	6.25	7.81	9.35	11.34	16.27
4	.09	.30	.48	.71	1.06	7.78	9.49	11.14	13.28	18.47
5	.21	.55	.83	1.15	1.61	9.24	11.07	12.83	15.09	20.52
6	.38	.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	22.46
7	.60	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	24.32
8	.86	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	26.12
9	1.15	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	27.88
10	1.48	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	29.59
11	1.83	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	31.26
12	2.21	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	32.91
13	2.62	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	34.53
14	3.04	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	36.12
15	3.48	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	37.70
16	3.94	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	39.25
17	4.42	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	40.79
18	4.90	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	42.31
19	5.41	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	43.82
20	5.92	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	45.31
25	8.65	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	52.62
30	11.59	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	59.70
35	14.69	18.51	20.57	22.47	24.80	46.06	49.80	53.20	57.34	66.62
40	17.92	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	73.40
50	24.67	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	86.66
60	31.74	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	99.61
70	39.04	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43	112.32
80	46.52	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33	124.84
90	54.16	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12	137.21
100	61.92	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81	149.45

Table 3.Critical values of χ^2 distributed random variables

Kolmogorov's function

$$K(y) = P(\lim_{n\to\infty} \sqrt{n} \sup_{z\in R} |F_e(z) - F(z)| < y)$$

y	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.4	.003	.004	.005	.007	.010	.013	.016	.020	.025
0.5	.036	.043	.050	.059	.068	.077	.088	.099	.110
0.6	.136	.149	.163	.178	.193	.208	.224	.240	.256
0.7	.289	.305	.322	.339	.356	.373	.390	.406	.423
0.8	.456	.472	.488	.504	.519	.535	.550	.565	.579
0.9	.607	.621	.634	.647	.660	.673	.685	.696	.708
1.0	.730	.741	.751	.761	.770	.780	.789	.798	.806
1.1	.822	.830	.837	.845	.851	.858	.864	.871	.877
1.2	.888	.893	.898	.903	.908	.912	.916	.921	.925
1.3	.932	.935	.939	.942	.945	.948	.951	.953	.956
1.4	.960	.962	.965	.967	.968	.970	.972	.973	.975
1.5	.978	.979	.980	.981	.983	.984	.985	.986	.986
1.6	.988	.989	.989	.990	.991	.991	.992	.992	.993
1.7	.994	.994	.995	.995	.995	.996	.996	.996	.996
1.8	.997	.997	.997	.998	.998	.998	.998	.998	.998
1.9	.999	.999	.999	.999	.999	.999	.999	.999	.999

Table 4. Kolmogorov's function