ANDRÁS VETIER

PROBABILITY THEORY
WITH SIMULATIONS

2011

This is an introductory textbook to probability theory and statistics with the usual material taught at most universities.
Its special feature, however, is that it contains interactive simulation files. These files are important, because the real life meaning of most of the notions of probability theory and statistics can be experienced only if we make a large number of experiments, not only once, but several times, and not only under a given set of conditions, but under modified conditions, as well.
The simulation files included in this textbook make it possible that the reader could see the results of many experiments, and could repeat them several times, and he or she could modify the parameters of the problem, as well.
Since the simulation files are written in Excel, students themselves can easily construct similar simulation files. Their activity will increase their confidence and interest in the subject.
The book consists of five parts:

1. Probability of events
2. Discrete distributions
3. Continuous distributions in one-dimension
4. Two-dimensional continuous distributions
5. Statistics

The author is devoted to write an exercise-book soon, which will - hopefully - help the students to learn not only the probabilistic and statistic notions but the necessary Excel tricks to construct simulation files according to their own needs.

Key words and phrases: Probability, Random number, Random variable, Discrete distribution, Continuous distribution, Expected value, Statistics, Regression, Confidence interval, Hypothesis test.

Acknowledgement of support:
Prepared within the framework of the project „Scientific training (matemathics and physics) in technical and information science higher education” Grant No. TÁMOP-4.1.2-08/2/A/KMR-2009-0028.

Prepared under the editorship of Budapest University of Technology and Economics, Mathematical Institute.

Professional advisor:
Miklós Ferenczi

Referee:
László Ketskeméty

Prepared for electronic publication by:
Lídia Boglárka Torma

Title page design:
Gergely László Csépány, Norbert Tóth

ISBN: 978-963-279-448-8

Contents

Part - I.

Probability of events

Preface

Usual text-books in probability theory describe the laws of randomness by a text consisting of sentences, formulas, etc., and a collection of examples, problems, figures, etc. which are printed permanently in the book. The reader may read the text, study the examples, the problems, and look at the figures as many times as he or she wants to. That is OK. However, the laws of randomness can be experienced only if many experiments are performed many times. It is important to see the effect of the change of the parameters, as well. In a permanently printed text-book what is printed is printed, and cannot be changed. The reader cannot modify the parameters.

The main purpose of this electronic text-book is to make it possible to simulate the experiments as many times as the reader wishes, and to make it possible to change the parameters according to the wish of the reader. For the simulations, we use the program Excel, it is available in high-schools and universities, and most students know it to a certain level.

I am convinced that having experienced the real life meaning of the notions of probability theory, the mathematical notions and the mathematical proofs become more interesting and attractive for the reader. Since the mathematical proofs are available in many usual textbooks, we give only a few proofs in this textbook.

I am sure you will find mistakes in this text-book. I ask you to let me know them so that I could then correct them. Anyway, I am continuously working on this material, so new, corrected versions (with less or even more mistakes) will occur again and again in the future. Thanks for your cooperation.

A list of suggested textbooks is available from the web-page:

`http://www.math.bme.hu/~vetier/Probability.htm`

Keep in mind that, in the simulation files, whenever you press the F9-key, the computer recalculates all the formulas, among others it gives new values to random numbers, consequently, it generates a new experiment.

Sections marked by *** may be skipped.

Section 1
Introductory problems

Example 1. (Coming home from Salzburg to Vac) My sister-in-law regularly plays the violin in an orchestra in Salzburg almost every Saturday evening, and comes home to Vac on Sundays. (Salzburg, hometown of W. A. Mozart is 600 km west of Budapest, and Vac, a little town next the Danube is 30 km north of Budapest.) Her friend brings her in his car to Nyugati railway station in Budapest, where she takes the train to Vac. The train leaves for Vac every hour. Sometimes she arrives to Nyugati railway station some minutes after the departure of the previous train, and has to wait almost an hour. Other times she arrives to Nyugati railway station some minutes before the departure of the next train, and she has to wait only some minutes. We may be interested in the amount of time passing after the departure of the previous train. Using the following file, you may study a simulation of the amount of time passing after the previous train.

Demonstration file: The amount of time after the departure of the previous train
020-01-00

We may be interested in the amount of time she has to wait until the next train. It is natural to call this amount of time the waiting time until the next train. In the following file, the waiting time is also shown.

Demonstration file: Both the amount of time after the previous train and the waiting time until the next train are shown
020-02-00

As you see the amount of time after the previous train is generated by the command `RAND()`. This command gives a random number between 0 and 1, so the `RAND()*60` command gives a random number between 0 and 60. Rounding is performed by the commands `ROUNDDOWN(_;_)` and `ROUNDUP(_;_)`.

Imagine that you observe the amount of time after the previous train on 10 occasions. You will get 10 real numbers between 0 and 60. In the following file, 10 experiments are simulated.

Demonstration file: 10 experiments for the amount of time after the previous train
020-03-00

As you see, the same command, namely, `RAND()*60` is used in all the 10 cells, but the numerical values returned are different.

If we make 1000 experiments, then - as you see in the next file - the 1000 corresponding dots overcrowd the line.

Demonstration file: 1000 experiments on a line
020-04-00

This is why, for visualization purposes, we give each of the points a different second coordinate, as if the points were moved out from the line into a narrow horizontal strip. In the following file, where only 10 points are shown, you may check how the points jump out of the line.

Demonstration file: 10 experiments on a narrow horizontal strip
020-05-00

When there are 1000 points, the points melt together on the overcrowded line, while the distribution of the points on the narrow horizontal strip is really expressive: whenever you press the F9-key, you may see that the points are uniformly distributed between 0 and 60, they constitute a uniformly distributed point-cloud.

Demonstration file: 1000 experiments on a narrow horizontal strip
020-06-00

For my sister-in-law, it is a rather unpleasant event when she has to wait until the next train for more than 45 minutes. Waiting more than 45 minutes obviously means that the amount of time after the departure of the previous train is less than 15 minutes. In the next file, these points are identified, their number - the so called frequency of the event - is calculated, and then the relative frequency of the event, that is, the frequency divided by the total number of experiments is also calculated.

Demonstration file: Frequency and relative frequency of the unpleasant event
020-07-00

In order keep track of whether `RAND()*60` is less than 15 or not, in the simulation file, we use the `IF(_;_;_)` command. The structure of this command is very simple: the first argument is a condition, the second argument is the value of the command if the condition holds, the third argument is the value of the command if the condition does not hold.

Pressing the F9-key in the previous simulation file, you may be convinced that the relative frequency oscillates around a non-random value, in this problem, around 0.25. This value, around which the relative frequency oscillates, is an important characteristic of the event. We call this number the probability of the event, and we write:

$$\mathbf{P}(\text{amount of time after the previous train} < 15) = 0.25$$

In the next file, the value of the probability is also visualized. You may see that the relative frequency oscillates around it.

Demonstration file: Probability of the unpleasant event
020-08-00

Example 2. (Random numbers) Random numbers generated by computers play an essential role in our simulation files. The basic property of a random number is that, for any $0 \le a \le b \le 1$, the probability of the event that the random number is between a and b is equal to the length of the interval $[a;b]$, which is $b-a$:

$$\mathbf{P}(a \le \mathrm{RND} \le b) = \text{ length of the interval } [a;b] = b-a$$

$$\mathbf{P}(a \le \mathrm{RND} < b) = \text{ length of the interval } [a;b) = b-a$$

$$\mathbf{P}(a < \mathrm{RND} \le b) = \text{ length of the interval } (a;b] = b-a$$

$$\mathbf{P}(a < \mathrm{RND} < b) = \text{ length of the interval } (a;b) = b-a$$

Whether the interval is closed or open, it does make any difference in the value of the probability. In the following file, you may choose the values a and b, the left and right end-points of the interval. You will see that the relative frequency of the event $a \le \mathrm{RND} \le b$ really oscillates around $b-a$.

Demonstration file: Probability of an interval for a random number generated by computer
020-09-00

It is important to remember that, for any fixed number x which is between 0 and 1, we have that

$$\mathbf{P}(\mathrm{RND} \le x) = \text{ length of the interval } [0;x] = x$$

$$\mathbf{P}(\mathrm{RND} < x) = \text{ length of the interval } [0;x) = x$$

Demonstration file: Probability of RND < x
020-10-00

Example 3. (Pairs of random numbers) Another basic property of the `RAND()` command is that using it twice, and putting the two random numbers RND_1 and RND_2 together to define a random point $(\mathrm{RND}_1;\mathrm{RND}_2)$, for this point it holds that, for any set A inside the unite square, it holds that

$$\mathbf{P}\left((\mathrm{RND}_1;\mathrm{RND}_2) \in A\right) = \text{area of } A$$

In order to see this fact, in the following file, A can be any triangle inside the unite square.

Demonstration file: Probability of a triangle
020-11-00

In the following file, the relative frequencies of more complicated events are studied:

Demonstration file: Special triangle combined with a diamond-shaped region - unconditional ...
020-12-00

In the following file not only frequencies and probabilities, but conditional frequencies and probabilities are involved. Playing with the file, you will discover the notion of conditional frequency and conditional probability.

Demonstration file: Special triangle combined with a diamond-shaped region - conditional ...
020-13-00

Example 4. (Non-uniform distributions) Just to see a point-cloud which is **not** uniformly distributed, let us replace the `RAND()` command by the `POWER(RAND();2)` command. The command `POWER(_;2)` stands for taking the square.

Demonstration file: Non-uniformly distributed points using the square of a random number
020-14-00

We get another non-uniformly distributed point-cloud if we apply the square-root function, `POWER(_;1/2)`

Demonstration file: Non-uniformly distributed points using the square-root of a random number
020-15-00

In the next file, relative frequencies related to non-uniform distributions are calculated.

Demonstration file: Relative frequency for non-uniform distribution
020-16-00

In the next file, conditional relative frequencies related to non-uniform distributions are calculated.

Demonstration file: Conditional relative frequency for non-uniform distribution
020-17-00

Example 5. (Waiting time for the bus) My wife goes to work by bus every day. She waits for the bus no more than 10 minutes. The amount of time she waits for the bus is uniformly distributed between 0 and 10. In the next file, we simulate this waiting time, and we study the event that "the waiting time < 4". The probability of this event is obviously 0.4 The relative frequency of the event will clearly oscillate around 0.4.

Demonstration file: Waiting time for the bus
020-18-00

Example 6. (Traveling by bus and metro) My friend goes to work by bus and metro every day. He waits for the bus no more than 10 minutes. The amount of time he waits for the bus is uniformly distributed between 0 and 10. When he changes to the metro, he waits for the metro no more than 5 minutes. No matter how much he waited for the bus, the amount of time he waits for the metro is uniformly distributed between 0 and 5.

This example involves two waiting times. As you will see in the next simulation file, the two waiting times together define a uniformly distributed random point in a rectangle.

Demonstration file: Traveling by bus and metro: uniformly distributed waiting times
020-19-00

Some events are visualized in the following files:

Demonstration file: Waiting time for bus < 4 , using uniform distribution
020-20-00

Demonstration file: Waiting time for metro > 3 , using uniform distribution
020-21-00

Demonstration file: Waiting time for bus < 4 AND waiting time for metro > 3 , using uniform distribution
020-22-00

Demonstration file: Waiting time for bus < waiting time for metro , using uniform distribution
020-23-00

Demonstration file: Total waiting time is more than 4 , using uniform distribution
020-24-00

Demonstration file: Waiting time for bus < waiting time for metro AND total waiting time > 4
020-25-00

Demonstration file: Waiting time for bus < waiting time for metro OR total waiting time > 4 , using uniform distribution
020-26-00

Under certain conditions, the application of uniform distribution for the waiting times is justified, but under other conditions it is not. If the busses and metros follow "strict time-tables" and randomness is involved in the problem only because my friend does not follow a "strict time-table", then the application of uniform distribution for the waiting times gives a good model. However, if the busses and metros arrive to the stations where my friend gets on them, in a "chaotic" way, then - as we learn later - the application of a special non-uniform distribution - called "exponential" distribution - is more correct. You may see in the next file that exponentially distributed waiting times are generated by using the `-LN(RAND())` command, that is, taking the minus of the natural logarithm of a simple random number.

Demonstration file: Traveling by bus and metro, using exponential distribution
020-27-00

The events studied above are visualized with exponentially distributed waiting times in the following files:

Demonstration file: Waiting time for bus < 4 , using exponential distribution
020-28-00

Demonstration file: Waiting time for metro > 3
020-29-00

Demonstration file: Waiting time for bus < 4 AND waiting time for metro > 3 , using exponential distribution
020-30-00

Demonstration file: Waiting time for bus < waiting time for metro, using exponential distribution
020-31-00

Demonstration file: Total waiting time > 4 , using exponential distribution
020-32-00

Demonstration file: Waiting time for bus < waiting time for metro AND total waiting time > 4 , using exponential distribution
020-33-00

Demonstration file: Waiting time for bus < waiting time for metro OR total waiting time > 4 , using exponential distribution
020-34-00

Example 7. (Dice) Toss a fair die and observe the number on top. This random number will be denoted here by X. It is easy to make 10 experiments for X. You may also make 100 experiments. But it would be boring to make 1000 experiments. This is why we will make - in the following file - a simulation of 1000 experiments. We will get 1000 integer numbers. The smallest possible value of X is 1, the largest is 6. We may count how many 1-s, 2-s, ... , 6-s we get. The numbers we get are the frequencies of the possible values. The frequencies divided by the total number of experiments are the relative frequencies.

In the following file, the frequencies are calculated by the `FREQUENCY(_;_)` command, which is a very useful but a little bit complicated command.

Demonstration file: Fair die, 1000 tosses
020-36-00

How to use the `FREQUENCY` command. The first argument of the `FREQUENCY(_;_)` command is the array of the data-set, the second argument is the array containing the list of the possible values. While entering the `FREQUENCY(_;_)` command, one must pay special attention to the following steps:

1. Type the `FREQUENCY(_;_)` command next to the first possible value with the correct arguments. You will get the frequency of the first possible value.
2. Mark - with the mouse - all the cells where the frequencies of the other possible values will be.
3. Press the F2-key.
4. Press the `Ctrl`-key, keep it pressed, and press the `Shift`-key, keep it pressed, too, and press the `Enter`-key. You will get the frequencies of all possible values. (You must not use the copy-paste command instead of the above sequence of commands. That would give false results.) The cells containing the frequencies will be stuck together, which means that later on they can be treated only together as a whole unit.

We see that each relative frequency is oscillating around $1/6$, so the probability of each possible value is $1/6$. This is shown in the next file.

Demonstration file: 1000 tosses with a fair die, relative frequencies and probabilities
020-38-00

In the following file not only frequencies and probabilities, but conditional frequencies and probabilities are involved. Playing with the file, you may study what the notion of conditional frequency and probability mean. Because of the large size of the file, downloading it may take longer time.

Demonstration file: Conditional relative frequency and probability of events related to fair dice
020-39-00

In the following two files, unfair dice are simulated. Because of the large size of the files, downloading them may take longer time.

Demonstration file: Unfair dice (larger values have larger probabilities)
020-40-00

Demonstration file: Unfair dice (smaller values have larger probabilities)
020-41-00

Section 2
Outcomes and events

A **phenomenon** means that, under certain circumstances or conditions, something is happening, or we do something. When the conditions are fulfilled, we say that we perform a **valid experiment**. When the conditions are not fulfilled, we say that this is an invalid experiment. It will be important in our theory that for phenomenon (at least theoretically) the experiments can be repeated as many times as we want. When, related to the phenomenon, we decide or declare what we are interested in, what we observe, we define as an **observation**. The possible results of the observation are called the **outcomes** (or - in some text-books - **elementary events**). The set of all outcomes is the **sample space**. Here are some examples for phenomena and observations.

Example 1. (Fair coin) Let the phenomenon mean tossing a fair coin on top of a table. Let an experiment be valid if one of the sides of the coin shows up (that is the coin does not stop on one of its edges). Here are some observations:

1. We observe where the center of the coin stops on a rectangular shaped table. Here the outcomes are the points of the top of the table. The sample space is the surface of the table, that is, a rectangle.

2. We observe how much time the coin rolls on the table before stopping. Here the outcomes are the positive real numbers. The sample space is the positive part of the real line.

3. We observe which side of the coin shows up when it stops. Now the outcomes are *heads* and *tails*. The sample space is the set $\{H,T\}$ consisting of two elements: H stands for *heads*, T stands for *tails*.

Example 2. (Fair die) Let the phenomenon mean rolling a fair die on top of a table. Let an experiment be valid if the die remains on top of the table so that it stands clearly on one of its sides. Here are some observations:

1. We observe where the die stops. Here the outcomes are the points of the top of the table. The sample space is the surface of the table, that is, a rectangle.

2. We observe how much time the die rolls on the table before stopping. Here the outcomes are the positive real numbers. The sample space is the positive part of the real line.

3. We observe which side of the die shows up when it stops. Now the outcomes are 1, 2, 3, 4, 5, 6. The sample space is the set $\{1,2,3,4,5,6\}$.

4. We observe whether we get 6 or we do not get 6. Here there are two outcomes: 6, not 6. The sample space is a set consisting of two elements: $\{6, \text{not } 6\}$.

5. We observe whether we get a number greater than 4 or not greater than 4. Here there are two outcomes again, namely: greater, not greater. The sample space is a set consisting of two elements: $\{\text{greater}, \text{not greater}\}$.

Example 3. (Two fair dice) Let the phenomenon mean rolling two fair dice, a red and a blue, on top of a table. Let an experiment be valid if both dice remain on top of the table so that they stand clearly on one of their sides. Here are some observations:

1. We observe the pair of numbers we get. Let the first number in the pair be taken from the red die, the second from the blue. Here we have 36 outcomes, which can be arranges in a 6 by 6 table. The sample space may be represented as the set of the 36 cells of a 6 by 6 table.

(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)

2. We observe the maximum of the two numbers we toss. Here the outcomes are again the numbers 1, 2, 3, 4, 5, 6. The sample space is the set $\{1,2,3,4,5,6\}$.

3. We observe the sum of the two numbers we toss. Here there are 11 outcomes: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 The sample space is the set $\{1,2,3,4,5,6,7,8,9,10,11,12\}$.

Example 4. (Toss a coin until the first head) Let the phenomenon mean tossing a fair die until the first time the head occurs. Here are some observations:

1. We observe the sequence of heads and tails we get. Now the outcomes are the possible sequences of heads and tails. The sample space is the set of all possible sequences of heads and tails.

2. We observe the number of tosses until the first time the head occurs. Now the outcomes are the positive integers: 1, 2, 3, ... and the symbol ∞. The symbol ∞ means: we never ever get a head. The sample space is the set consisting of all positive integers and the symbol ∞: $\{1,2,3,\ldots,\infty\}$

3. We observe how many tails we get before the first head occurs. Now the outcomes are the non-negative integers: 0, 1, 2, ... and the symbol ∞. The symbol ∞ means: we never get a head, that is why we get an infinite number of tails. The sample space is the set consisting of all non-negative integers and the symbol ∞: $\{0,1,2,\ldots,\infty\}$.

An **event** is a statement related to the phenomenon or to an observation so that whenever an experiment is performed we can decide whether the statement is *true* or *false*. When it is true we say that the event *occurs*, when it is not true, we say that the event *does not occur*. Instead of true and false, the words *yes* and *no* are also often used. We often write the number 1 for the occurrence, and the number 0 for the non-occurrence of an event. An event, that is, a statement related to an observation obviously corresponds to a *subset* of the sample space taken for that observation. The subset consists of those outcomes for which the event occurs. For example, tossing a die and observing the number on the top, the event "greater than 4" corresponds to the subset $\{5,6\}$.

It may happen that two different statements always occur at the same time. In this case we say that the two statements define the *same event*.

Now we list some **operations and relations on events.** We put the corresponding set-theoretical operations and relations into parentheses.

1. The **sure** or **certain** event always occurs. (Whole sample space.)

2. The **impossible** event never occurs. (Empty set.)

3. The **complement** of an event occurs if and only if the event does not occur. (Complementary set.)

4. The **intersection** or **product** of events is the logical *and*-operation, meaning that "each event occurs". (Intersection of sets.)

5. The **union** or **sum** of events is the logical *or*-operation, meaning that "at least one of the events occurs". (Union of sets.)

6. The **difference** of an event and another event means that the first event occurs, but the other event does not occur. (Difference of sets.)

7. Some events are said to be **exclusive** events, and we say that they **exclude** each other if the occurrence of one of them guarantees that the others do not occur. (Disjoint sets.)

8. An event is said to **imply** another event if the occurrence of the first event guarantees the occurrence of the other event. (A set is a subset of the other.)

Drawing a **Venn-diagram** is a possibility to visualize events, operations on events, etc. by sets drawn in the plain.

Section 3
Relative frequency and probability

When we make experiments again and again for a phenomenon or an observation, then we get **sequence of experiments**. Assume now that we make a sequence of experiments for an event. We may take notes at each experiment whether the event occurs or does not occur, and we may count how many times the event occurs. This occurrence number is called the **frequency** of the event. The frequency divided by the number of experiments is the **relative frequency**. Since the occurrence of an event depends on randomness, both the frequency and the relative frequency depend on randomness.

Now the reader may study the following files again, which appeared among the introductory problems in Section 1.

Demonstration file: Waiting time for the bus
020-18-00

Demonstration file: Traveling by bus and metro: uniformly distributed waiting times
020-19-00

Demonstration file: Waiting time for bus < 4
020-20-00

Demonstration file: Waiting time for metro > 3
020-21-00

Demonstration file: Waiting time for bus < 4 AND waiting time for metro > 3
020-22-00

Demonstration file: Waiting time for bus < waiting time for metro
020-23-00

Demonstration file: Total waiting time > 4
020-24-00

Demonstration file: Waiting time for bus is less than waiting time for metro AND total waiting time > 4
020-25-00

Demonstration file: Waiting time for bus < waiting time for metro OR total waiting time > 4
020-26-00

It is an important law, called the law of large numbers, that the relative frequencies of an event in a long sequence of experiments stabilize around a number, which does not depend on randomness, but it is a characteristic of the event itself. This number is called the **probability** of the event. The notion of probability can be interpreted like these:

1. Consider an interval around the probability value. If we make a large number of experiments of a (given) large length, then the great majority of relative frequencies (associated to this large length) will be in this interval.

2. If we could make an infinitely long sequence of experiments, then the sequence of relative frequencies would converge to the probability in the mathematical sense of convergence.

Probability theory deals, among others, with figuring out the probability values without performing any experiments, but using theoretical arguments.

In the following files you may learn how the relative frequencies stabilize around the probability. The first and the second are simpler, the third is a little-bit trickier.

Demonstration file: Event and relative frequency
030-01-00

Demonstration file: Tossing a die - probability
030-02-00

Demonstration file: Relative frequency with balls
030-03-00

Playing with the next file, you may check your ability to guess a probability based on your impression when many experiments are performed. When you open it, choose the option "Don't Update".

Demonstration file: Probability guessed by impression
030-04-00

If you want to change the hidden probability value in the previous file, then save the previous file (File A) and the following file (File B) into a folder, and close both. Then open the second file (File B), press F9 to regenerate a new hidden probability value, and open the first file (File A), and choose the option "Update", and close the second file (File B).

Demonstration file: Auxiliary file to generate a new hidden probability value
030-05-00

Section 4
Random numbers

Most calculators have a special key stroke and most computer programs have a simple command to generate random numbers. Calculators and computer programs are made so that the generated random number, let us denote it by RND, can be considered uniformly distributed between 0 and 1, which means that for any $0 \le a \le b \le 1$, it is true that

$$\mathbf{P}(a < \text{RND} < b) = \text{length of } (a;b) = b - a$$

or, the same way,

$$\mathbf{P}(a \le \text{RND} \le b) = \text{length of } [a;b] = b - a$$

The following file illustrates this fact:

Demonstration file: Probability of an interval for a random number generated by computer
040-01-00

Specifically, for any $0 \le x \le 1$ it is true that

$$\mathbf{P}(\text{RND} < x) = x$$

or, the same way,

$$\mathbf{P}(\text{RND} \le x) = x$$

The following file illustrates this fact:

Demonstration file: Probability of RND $\le x$
020-10-00

The probability that a random number is exactly equal to a given number is equal to 0:

$$\mathbf{P}(\text{RND} = a) = \mathbf{P}(a \le \text{RND} \le a) = \text{length of } [a;a] = a - a = 0 \qquad (\text{for all } a)$$

If two random numbers are generated, say RND_1 and RND_2, then the random point $(\text{RND}_1, \text{RND}_2)$ is uniformly distributed in the unit square Swhich has the vertices $(0,0)$, $(1,0)$, $(1,1)$, $(0,1)$. This means that for any $A \subset S$, it is true that

$$P\left((\text{RND}_1, \text{RND}_2) \in A\right) = \text{area of } A$$

In order to illustrate this fact, in the following file, A can be a triangle inside the unite square.

Demonstration file: Probability of a triangle
020-11-00

In the following file, the relative frequencies of more complicated events are studied.

Demonstration file: Special triangle combined with a diamond-shaped region
020-12-00

If three random numbers are generated, say RND_1, RND_2 and RND_3, then the random point $(\mathrm{RND}_1, \mathrm{RND}_2, \mathrm{RND}_3)$ is uniformly distributed in the unit cube S which has the vertices $(0,0,0)$, $(1,0,0)$, $(1,1,0)$, $(0,1,0)$, $(0,0,1)$, $(1,0,1)$, $(1,1,1)$, $(0,1,1)$. This means that for any $A \subset S$, it is true that

$$P((\mathrm{RND}_1, \mathrm{RND}_2, \mathrm{RND}_3) \in A) = \text{volume of } A \qquad (A \subset S)$$

The following file deals with powers of random numbers.

Demonstration file: Powers of random numbers
040-02-00

Section 5
Classical problems

The simplest way of calculating a probability is when an observation has a finite number of outcomes so that, for some symmetry reasons, each outcome has the same probability. In this case the probability of an event is calculated by the **classical formula:**

$$\text{probability} = \frac{\text{number of favorable outcomes}}{\text{number of all outcomes}}$$

or, briefly:

$$\text{probability} = \frac{\text{favorable}}{\text{all}}$$

In the following files, we simply list all the outcomes, mark those which are favorable for the event in question, and then we use the classical formula to calculate the probability of the event.

Demonstration file: 2 dice, P(Sum = 5)
050-01-00

Demonstration file: 2 dice, P(Sum = k)
050-02-00

Demonstration file: 5 coins, P(Number of heads = k)
050-03-00

Demonstration file: 4 dice, on each dice: 1,2: red, 3,4,5,6: green, P(Number of red = k)
050-04-00

When the number of all outcomes is so large that we are unable to list them, or the problem contains not only numerical values but parameters as well, then combinatorics plays an important role in finding out the number of all outcomes and the number of favorable outcomes. The branch of mathematics dealing with calculating the number of certain cases is called **combinatorics**. It is assumed that the reader is familiar with the basic notions and techniques of elementary combinatorics. Here is only a list of some techniques and formulas we often use in combinatorics:

1. Listing - counting
2. Uniting - adding
3. Leaving off - subtracting
4. Tree-diagram, window technique - multiplication
5. Factorization (considering classes of equal size) - division
6. Permutations without repetition
$$n!$$
7. Permutations with repetition
$$\frac{n!}{k_1!k_2!\dots k_r!}$$
8. Variations without repetition
$$\frac{n!}{(n-k)!}$$
9. Variations with repetition
$$n^k$$
10. Combinations without repetition
$$\binom{n}{k}$$
Remember that the definition of the binomial coefficient $\binom{n}{k}$ is:
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$
When we have to calculate the value of the binomial coefficient $\binom{n}{k}$ without a calculator, it is may be advantageous to use the following form of it:
$$\binom{n}{k} = \frac{n(n-1)(n-2)\dots(n-k+1)}{1\,2\,3\dots k}$$
Notice that in the right side formula, both the numerator and the denominator are a product of k factors. In the numerator, the first factor is n, and the factors are decreasing. In the denominator the first factor is 1, and the factors are increasing. Simplification always reduces the fraction into an integer.
11. Combinations with repetition
$$\binom{n+k-1}{k-1}$$

12. Pascal triangle: if we arrange the binomial coefficients into a triangle-shaped table like this:

$$\begin{array}{ccccccccccccc}
 & & & & & & \binom{0}{0} & & & & & & \\
 & & & & & \binom{1}{0} & & \binom{1}{1} & & & & & \\
 & & & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & & & \\
 & & & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & & & \\
 & & \binom{4}{0} & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & \binom{4}{4} & & \\
 & \binom{5}{0} & & \binom{5}{1} & & \binom{5}{2} & & \binom{5}{3} & & \binom{5}{4} & & \binom{5}{5} & \\
\ldots & & \ldots & & \ldots & & \ldots & & \ldots & & \ldots & & \ldots
\end{array}$$

and calculate the numerical value of each binomial coefficient in this triangle-shaped table, we get the following array:

$$\begin{array}{ccccccccccccc}
 & & & & & & 1 & & & & & & \\
 & & & & & 1 & & 1 & & & & & \\
 & & & & 1 & & 2 & & 1 & & & & \\
 & & & 1 & & 3 & & 3 & & 1 & & & \\
 & & 1 & & 4 & & 6 & & 4 & & 1 & & \\
 & 1 & & 5 & & 10 & & 10 & & 5 & & 1 & \\
\ldots & & \ldots & & \ldots & & \ldots & & \ldots & & \ldots & & \ldots
\end{array}$$

The numbers in this triangle-shaped table satisfy the following two simple rules:

(a) The elements at the edges of each row are equal to 1.

(b) Addition rule: Elements which are not at the edges are equal to the sum of the two numbers which stand above that element.

Based on these rules one can easily construct the table and find out the numerical values of the binomial coefficients. The following file uses this construction.

Demonstration file: Construction of the Pascal triangle using the addition rule
060-01-00

Section 6
Geometrical problems, uniform distributions

Another simple way of calculating a probability is when the outcomes can be identified by an interval S of the (one-dimensional) real line or by a subset S of the (two-dimensional) plane or of the (three-dimensional) space or of an n-dimensional Euclidean space so that the length or area or volume or n-dimensional volume of S is finite but not equal to 0, and the probability of any event, corresponding to some subset A of S, is equal to

$$\mathbf{P}(A) = \frac{\text{length of } A}{\text{length of } S}$$

in the one-dimensional case, or

$$\mathbf{P}(A) = \frac{\text{area of } A}{\text{area of } S}$$

in the two-dimensional case, or

$$\mathbf{P}(A) = \frac{\text{volume of } A}{\text{volume of } S}$$

in the three-dimensional case, or

$$\mathbf{P}(A) = \frac{n\text{-dimensional volume of } A}{n\text{-dimensional volume of } S}$$

in the n-dimensional case. Since the calculation of lengths, areas, volumes, first in the life of most students, is taught in geometry, such problems are called **geometrical problems**.

We also say that a random point is chosen in S according to **uniform distribution** if

$$P\,(\text{the point is in } A) = \frac{\text{length of } A}{\text{length of } S} \qquad (A \subseteq S)$$

in the one-dimensional case, or

$$P\,(\text{the point is in } A) = \frac{\text{area of } A}{\text{area of } S} \qquad (A \subseteq S)$$

in the two-dimensional case, or

$$P(\text{the point is in } A) = \frac{\text{volume of } A}{\text{volume of } S} \qquad (A \subseteq S)$$

in the three-dimensional case, or

$$P(\text{the point is in } A) = \frac{n\text{-dimensional volume of } A}{n\text{-dimensional volume of } S} \qquad (A \subseteq S)$$

in the n-dimensional case.

Now the reader may study the following files again, which appeared among the introductory problems in Section 1.

Demonstration file: Waiting time for the bus
020-18-00

Demonstration file: Traveling by bus and metro: uniformly distributed waiting times
020-19-00

Demonstration file: Waiting time for bus < 4
020-20-00

Demonstration file: Waiting time for metro > 3

Demonstration file: Waiting time for bus < 4 AND waiting time for metro > 3
020-22-00

Demonstration file: Waiting time for bus < waiting time for metro
020-23-00

Demonstration file: Total waiting time > 4
020-24-00

Demonstration file: Waiting time for bus < waiting time for metro AND AND total waiting time > 4
020-25-00

Demonstration file: Waiting time for bus is less than waiting time for metro OR total waiting time is more than 4
020-26-00

The following example may surprise the reader, because the number π appears in the solution.

Example 1. (Buffon's needle problem) Let us draw several long parallel lines onto a big paper so that the distance between adjacent lines is always D. Let us take a needle whose length is L. For simplicity, we assume that $L \leq D$. Let us drop the needle onto the paper "carelessly, in a random way" so that not any direction or position is preferred for the needle the same way. When the needle stops jumping it will either intersect a line (touching without intersection is included) or it will not touch lines at all. We may ask: what is the probability that the needle will intersect a line?

The following two files interpret Buffon's needle problem.

Demonstration file: Buffon's needle problem
070-01-00

Demonstration file: Buffon's needle problem, more experiments
070-02-00

Solution. The line of the needle and the given parallel lines define an acute angle, this is what we denote by X. The center of the needle and the closest line to it define a distance, this is what we denote by Y. Obviously, $0 \leq X \leq \pi/2$ and $0 \leq Y \leq D/2$. The point (X,Y) is obviously a random point inside the rectangle defined by the intervals $(0;\pi/2)$ and $(0;D/2)$. Since X and Y follow uniform distribution and they are independent of each other, the random point (X,Y) follows uniform distribution on the rectangle. The needle intersects a line if and only if $Y \leq L/2\sin(X)$, that is, the points in the rectangle corresponding to intersections constitute the range below the graph of the curve with equation $y = L/2\sin(x)$. Thus, we get that

$$\mathbf{P}(\text{Intersection}) = \frac{\text{Area under the curve}}{\text{Area of the rectangle}} = \frac{\int_0^{\frac{\pi}{2}} \frac{L}{2}\sin(x)\,dx}{\frac{D}{2}\cdot\left(\frac{\pi}{2}\right)} = \frac{2L}{\pi D}$$

Remark. If $2L = D$, that is the distance between the parallel lines is twice the length of the needle, then we get the nice and surprising result:

$$\mathbf{P}(\text{Intersection}) = \frac{1}{\pi}$$

The following sequence of problems may seem a contradiction, because the (seemingly) same questions have different answers in the different solutions.

Example 2. (Bertrand's paradox) Let us consider a circle. For the sake of Bertrand's paradox, a chord of the circle is called long, if it is longer than the length of a side of a regular triangle drawn into the circle. Let us Choose a chord "at random". We may ask: what is the probability that the chord is long? The following files interpret Bertrand's paradox.

Demonstration file: Bertrand paradox, introduction
070-03-00

Demonstration file: Two points on the perimeter
070-04-00

Demonstration file: One point inside
070-05-00

Demonstration file: One point on a radius
070-06-00

Demonstration file: Two points inside
070-07-00

Demonstration file: One point on the perimeter, other point inside
070-08-00

Demonstration file: Point and direction
070-09-00

Demonstration file: Bertrand paradox, comparison
070-10-00

Section 7
Basic properties of probability

The following properties are formulated for probabilities. If we accept some of them as axioms, then the others can be proved. We shall not do so. Instead of such an approach, we emphasize that each of these formulas can be translated into a formula for relative frequencies by replacing the expression "probability of" by the expression "relative frequency of", or replacing the letter "**P**", which is an abbreviation of the expression "probability of", by the expression "relative frequency of". If you make this replacement, you will get properties for relative frequencies which are obviously true.

For example, the first three properties for relative frequencies sound like this:

1. Relative frequency of the **sure event** is 1.
2. Relative frequency of the **impossible event** is 0.
3. **Complement rule for relative frequencies**:

$$\text{relative frequency of } A + \text{relative frequency of } \overline{A} = 1$$

This is why it is easy to accept that the following properties for probabilities hold.

1. The probability of the **sure event** is 1.
2. The probability of the **impossible event** is 0.
3. **Complement rule**:

$$\mathbf{P}(A) + \mathbf{P}(\overline{A}) = 1$$

4. **Addition law of probability for exclusive events**:

   If A, B are exclusive events, then

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$$

   If A, B, C are exclusive events, then

$$\mathbf{P}(A \cup B \cup C) = \mathbf{P}(A) + \mathbf{P}(B) + \mathbf{P}(C)$$

If $A_1, A_2, \ldots, A_n$ are exclusive events, then

$$\mathbf{P}(A_1 \cup A_2 \cup \ldots \cup A_n) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \ldots + \mathbf{P}(A_n)$$

5. **Addition law of probability for arbitrary events**:

   If A, B are arbitrary events, then

$$\mathbf{P}(A_1 \cup A_2) = \mathbf{P}(A_1) + \mathbf{P}(A_2) - \mathbf{P}(A_1 \cap A_2)$$

   If A, B, C are arbitrary events, then

$$\begin{aligned} \mathbf{P}(A_1 \cup A_2 \cup A_3) \quad = \quad & +\mathbf{P}(A_1) + \mathbf{P}(A_2) + \mathbf{P}(A_3) \\ & -\mathbf{P}(A_1 \cap A_2) - \mathbf{P}(A_1 \cap A_3) - \mathbf{P}(A_2 \cap A_3) \\ & +\mathbf{P}(A_1 \cap A_2 \cap A_3) \end{aligned}$$

Remark. Notice that , on the right side
- in the 1st line, there are $\binom{3}{1} = 3$ terms, the probabilities of the individual events with "+" signs,
- in the 2nd line there are $\binom{3}{2} = 3$ terms, the probabilities of the intersections of two events with "−" signs,
- in the 3rd line there is $\binom{3}{3} = 1$ term, the probability of the intersection of all events with a "+" sign.

Poincaré formula: If $A_1, A_2, \ldots, A_n$ are arbitrary events, then

$$\mathbf{P}(A_1 \cup A_2 \cup \ldots \cup A_n) =$$

$$+\mathbf{P}(A_1) + \mathbf{P}(A_2) + \ldots + \mathbf{P}(A_3)$$

$$-\mathbf{P}(A_1 \cap A_2) - \mathbf{P}(A_1 \cap A_3) - \ldots - \mathbf{P}(A_{n-1} \cap A_n)$$

$$+\mathbf{P}(A_1 \cap A_2 \cap A_3) + \mathbf{P}(A_1 \cap A_2 \cap A_4) + \ldots + \mathbf{P}(A_{n-2} \cap A_{n-1} \cap A_n)$$

$$\vdots$$

$$+(-1)^{n+1}\mathbf{P}(A_1 \cap A_2 \cap \ldots \cap A_n)$$

Remark. Notice that, on the right side
- in the 1st line, there are $\binom{n}{1} = n$ terms, the probabilities of the individual events with "+" signs,
- in the 2nd line there are $\binom{n}{2}$ terms, the probabilities of the intersections of two events with "−" signs,
- in the 3rd line there are $\binom{n}{3}$ terms, the probabilities of the intersections of two events, with "+" signs,
- in the nth line there is $\binom{n}{n} = 1$ term, the probability of the intersection of all events with a "+" or "−" sign depending on whether n is odd or even.

6. **Special subtraction rule**: If event A implies event B, then

$$\mathbf{P}(B \backslash A) = P(B) - \mathbf{P}(A)$$

7. **General subtraction rule**: If A and B are arbitrary events, then

$$\mathbf{P}(B \backslash A) = P(B) - \mathbf{P}(A \cap B)$$

Section 8
Conditional relative frequency and conditional probability

Let A and B denote events related to a phenomenon. Imagine that we make N experiments for the phenomenon. Let N_A denote the number of times that A occurs, and let $N_{A\cap B}$ denote the number of times that B occurs together with A. The **conditional relative frequency** is introduced by the fraction:

$$\frac{N_{A\cap B}}{N_A}$$

This fraction shows how often B occurs among the occurrences of A. Dividing both the numerator and the denominator by N, we get that, for large N, if $\mathbf{P}(A) \neq 0$, then

$$\frac{N_{A\cap B}}{N_A} = \frac{\frac{N_{A\cap B}}{N}}{\frac{N_A}{N}} \approx \frac{\mathbf{P}(A\cap B)}{\mathbf{P}(A)}$$

that is, for a large number of experiments, the conditional relative frequency stabilizes around

$$\frac{\mathbf{P}(A\cap B)}{\mathbf{P}(A)}$$

This value will be called the **conditional probability** of B on condition that A occurs, and will be denoted by $\mathbf{P}(B|A)$:

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

This formula is also named as the **division rule for probabilities**.

In the following files, not only frequencies and probabilities, but conditional frequencies and probabilities are involved.

Demonstration file: Special triangle combined with a diamond-shaped region
020-13-00

Demonstration file: Circle and/or hyperbolas
090-01-00

Multiplication rules. Rearranging the division rule, we get the **multiplication rule for two events**:

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

which can be easily extended to the **multiplication rule for arbitrary events**:

$$\begin{array}{lcl}
\mathbf{P}(A_1 \cap A_2) & = & \mathbf{P}(A_1)\ \mathbf{P}(A_2|A_1) \\
\mathbf{P}(A_1 \cap A_2 \cap A_3) & = & \mathbf{P}(A_1)\ \mathbf{P}(A_2|A_1)\ \mathbf{P}(A_3|A_1 \cap A_2) \\
\mathbf{P}(A_1 \cap A_2 \cap A_3 \cap A_4) & = & \mathbf{P}(A_1)\ \mathbf{P}(A_2|A_1)\ \mathbf{P}(A_3|A_1 \cap A_2)\ \mathbf{P}(A_4|A_1 \cap A_2 \cap A_3) \\
 & \vdots &
\end{array}$$

As a special case, we get the **multiplication rule for a decreasing sequence of events**:

If

A_2 is implies A_1, that is, $A_2 \subseteq A_1$, or equivalently, $A_1 \cap A_2 = A_2$,
A_3 is implies A_2, that is, $A_3 \subseteq A_2$, or equivalently, $A_2 \cap A_3 = A_3$,
A_4 is implies A_3, that is, $A_4 \subseteq A_3$, or equivalently, $A_3 \cap A_4 = A_4$,
$\vdots$

then

$$\begin{array}{lcl}
\mathbf{P}(A_2) & = & \mathbf{P}(A_1)\ \mathbf{P}(A_2|A_1) \\
\mathbf{P}(A_3) & = & \mathbf{P}(A_2)\ \mathbf{P}(A_3|A_2) \\
\mathbf{P}(A_4) & = & \mathbf{P}(A_3)\ \mathbf{P}(A_4|A_3) \\
 & \vdots &
\end{array}$$

and, consequently

$$\begin{array}{lcl}
\mathbf{P}(A_2) & = & \mathbf{P}(A_1)\ \mathbf{P}(A_2|A_1) \\
\mathbf{P}(A_3) & = & \mathbf{P}(A_1)\ \mathbf{P}(A_2|A_1)\ \mathbf{P}(A_3|A_2) \\
\mathbf{P}(A_4) & = & \mathbf{P}(A_1)\ \mathbf{P}(A_2|A_1)\ \mathbf{P}(A_3|A_2)\ \mathbf{P}(A_4|A_3) \\
 & \vdots &
\end{array}$$

Example 1. (Birthday paradox) Imagine that in a group of n people, everybody, one after the other, tells which day of the year he or she was born. (For simplicity, loop years are neglected, that is, there are only 365 days in a year.) It may happen that all the n people say different days, but it may happen that there will be one ore more coincidences. The reader, in the future, at parties, may make experiments. Obviously, if n is small, then the probability that at least one coincidence occurs, is small. If n is larger, then this probability is larger. If $n \geq 366$, then the coincidence is sure. The following file simulates the problem:

Demonstration file: Birthday paradox - simulation
090-02-10

We ask two questions:

1. For a given n ($n = 2,3,4,\ldots,366$), how much is the probability that at least one coincidence occurs?
2. Which is the smallest n for which $\mathbf{P}(\text{at least one coincidence occurs}) \geq 0.5$?

Remark. People often argue like this: the half of 365 is $365/2 = 182.5$, so the answer to the second question is 183. We shall see that this answer is very far from the truth. The correct answer is surprisingly small: 23. This means that when 23 people gather together, then the probability that at least one birthday coincidence occurs is more than half, and the probability that no birthday coincidence occurs is less than half. If you do not believe, then make experiments: if you make many experiments with groups consisting of at least 23 people, then the case that at least one birthday coincidence occurs will be more frequent than the case that no birthday coincidence occurs.

Solution. Let us define the event A_k like this:

$$A_k = \text{the first } k \text{ people have different birthdays} \quad (k = 1,2,3,\ldots)$$

The complement of A_k is:

$$\overline{A_k} = \text{at least one coincidence occurs}$$

It is obvious that $\mathbf{P}(A_1) = 1$. The sequence of the events $A_1, A_2, A_3, \ldots$ clearly constitutes a decreasing sequence of events. In order to determine the conditional probability $\mathbf{P}(A_k|A_{k-1})$, let us assume that A_{k-1} occurs, that is, the first $k-1$ people have different birthdays. It is obvious that A_k occurs if and only if the kth person has a birthday different from the previous $k-1$ birthdays, that is, he or she was born on one of the remaining $365-(k-1)$ days. This is why

$$\mathbf{P}(A_k|A_{k-1}) = (365-(k-1))/365 \quad (k \geq 1)$$

that is

$$\begin{aligned}
&\mathbf{P}(A_2|A_1) = 364/365 = 0,9973\\
&\mathbf{P}(A_3|A_2) = 363/365 = 0,9945\\
&\mathbf{P}(A_4|A_3) = 362/365 = 0,9918\\
&\vdots
\end{aligned}$$

Now, using the multiplication rule for our decreasing sequence of events, we get:

$$\begin{array}{lcl}
\mathbf{P}(A_1) & = & 1\\
\mathbf{P}(A_2) & = & \mathbf{P}(A_1)\ \mathbf{P}(A_2|A_1) \quad = \quad 1 \quad 0,9973 \quad = \quad 0,9973\\
\mathbf{P}(A_3) & = & \mathbf{P}(A_2)\ \mathbf{P}(A_3|A_2) \quad = \quad 0,9973 \quad 0,9945 \quad = \quad 0,9918\\
\mathbf{P}(A_4) & = & \mathbf{P}(A_3)\ \mathbf{P}(A_4|A_3) \quad = \quad 0,9918 \quad 0,9918 \quad = \quad 0,9836\\
\vdots & &
\end{array}$$

Since the events A_n mean no coincidences, in order to to get the probabilities of the birthday coincidences we need to find the probabilities of their complements :

$$\begin{aligned}
&\mathbf{P}\left(\overline{A_1}\right) = 1-\mathbf{P}(A_1) = 1-1 &&= 0\\
&\mathbf{P}\left(\overline{A_2}\right) = 1-\mathbf{P}(A_2) = 1-0,9973 &&= 0,0027\\
&\mathbf{P}\left(\overline{A_3}\right) = 1-\mathbf{P}(A_3) = 1-0,9918 &&= 0,0082\\
&\mathbf{P}\left(\overline{A_4}\right) = 1-\mathbf{P}(A_4) = 1-0,9836 &&= 0,0164\\
&\vdots
\end{aligned}$$

The following file shows how such a table can be easily constructed and extended up to $n = 366$ in Excel:

Demonstration file: Birthday paradox - calculation
090-02-00

In this Excel table, we find the answer to our first question: the probability that at least one coincidence occurs is calculated for all $n = 1,2,\ldots,366$. In order to get the answer to the second question, we must find where the first time the probability of the coincidence is larger than half in the table. Wee see that

$$\mathbf{P}\left(\overline{A_{22}}\right) = 0,4757$$

$$\mathbf{P}\left(\overline{A_{23}}\right) = 0,5073$$

which means that 23 is the smallest n for which the probability that at least one coincidence occurs is greater than half.

We say that the events A_1, A_2, ... constitute a **total system** if they are exclusive, and their union is the sure event.

Total probability formula. If the events A_1, A_2, ... have a probability different from zero, and they constitute a total system, then

$$\mathbf{P}(B) = \sum_i \mathbf{P}(A_i)\mathbf{P}(B|A_i)$$

The following example illustrates how the total probability formula may be used.

Example 2. (Is it defective?) There are three workshops in a factory: A_1, A_2 A_3. Assume that
- workshop A_1 makes 30 percent,
- workshop A_2 makes 40 percent,
- workshop A_3 makes 30 percent of all production.

We assume that

- the probability that an item made in workshop A_1 is defective is 0,05,
- the probability that an item made in workshop A_2 is defective is 0.04,
- the probability that an item made in workshop A_3 is defective is 0.07.

Now taking an item made in the factory, what is the probability that it is defective?

Solution. The following file - using the total probability formula - gives the answer:

Demonstration file: Application of the total probability formula
090-02-50

The Bayes formula expresses a conditional probability in terms of other conditional and unconditional probabilities.

Bayes formula. If the events A_1, A_2, ... have a probability different from zero, and they constitute a total system, then

$$\mathbf{P}(A_k|B) = \frac{\mathbf{P}(A_k)\mathbf{P}(B|A_k)}{\mathbf{P}(B)} = \frac{\mathbf{P}(A_k)\mathbf{P}(B|A_k)}{\sum_i \mathbf{P}(A_i)\mathbf{P}(B|A_i)}$$

The following files illustrate and use the Bayes formula.

Example 3. (Which workshop made the defective item?) Assuming that an item made in the factory in the previous problem is defective, we may ask: Which workshop made it? Obviously, any of them may make defective items. So, the good question consists of 3 questions, which may sound like this:
- What is the probability that the defective item was made in workshop A_1?
- What is the probability that the defective item was made in workshop A_2?
- What is the probability that the defective item was made in workshop A_3?

Solution. The following file - using the Bayes formula - gives the answer to these questions:

Demonstration file: Application of the Bayes formula
090-03-00

Example 4. (Is he sick or healthy?) Assume that 0.001 part of people are infected by a certain bad illness, 0.999 part of people are healthy. Assume also that if a person is infected by the illness, then he or she will be correctly diagnosed sick with a probability 0.9, and he or she will be mistakenly diagnosed healthy with a probability 0.1. Moreover, if a person is healthy, then he or she will be correctly diagnosed healthy with a probability 0.8. and he or she will be mistakenly diagnosed sick with a probability 0.2, Now imagine that a person is examined, and the test says the person is sick. Knowing this fact what is the probability that this person is really sick?

Solution. The answer is surprising. Using the Bayes formula, it is given in the following file.

Demonstration file: Sick or healthy?
090-04-00

Section 9
Independence of events

Independence of two events. The event B and its complement $\bar{B}$ are called to be **independent of** the event A and its complement $\bar{A}$ if

$$\mathbf{P}(B|A) = \mathbf{P}(B|\bar{A}) = \mathbf{P}(B)$$

$$\mathbf{P}(\bar{B}|A) = \mathbf{P}(\bar{B}|\bar{A}) = \mathbf{P}(\bar{B})$$

It is easy to see that in order for these four equalities to hold it is enough that one of them holds, because the other three equalities are consequences of the chosen one. This is why many textbooks introduce the notion of independence so that the event B is called to be **independent of** the event A if

$$\mathbf{P}(B|A) = \mathbf{P}(B)$$

On the left side of this equality, replacing $\mathbf{P}(B|A)$ by $\frac{\mathbf{P}(A\cap B)}{\mathbf{P}(A)}$, we get that independence means that

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

or, equivalently,

$$\mathbf{P}(A\cap B) = \mathbf{P}(A)\mathbf{P}(B)$$

Now dividing by $\mathbf{P}(B)$, we get that

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

that is

$$\mathbf{P}(A|B) = \mathbf{P}(A)$$

which means that event A is independent of the event B. Thus, we see that independence is a symmetrical relation, and we can simply say, that events A and B are independent of each other, or more generally the pair $A,\bar{A}$ and the pair $B,\bar{B}$ are **independent of each other**.

Independence of three events. The notion of independence of three events is introduced in the following way. The sequence of events A, B, C is called independent if

$$\mathbf{P}(B|A) = \mathbf{P}(B|\bar{A}) = \mathbf{P}(B)$$

$$\mathbf{P}(\bar{B}|A) = \mathbf{P}(\bar{B}|\bar{A}) = \mathbf{P}(\bar{B})$$
$$\mathbf{P}(C|A\cap B) = \mathbf{P}(C|A\cap\bar{B}) = \mathbf{P}(C|\bar{A}\cap B) = \mathbf{P}(C|\bar{A}\cap\bar{B}) = \mathbf{P}(C)$$
$$\mathbf{P}(\bar{C}|A\cap B) = \mathbf{P}(|A\cap\bar{B}) = \mathbf{P}(\bar{C}|\bar{A}\cap B) = \mathbf{P}(\bar{C}|\bar{A}\cap\bar{B}) = \mathbf{P}(\bar{C})$$

Pairwise and total independence. It can be shown (we omit the proof) that these equalities hold if and only if the following $2^3 = 8$ **multiplication rules** hold:

$$\begin{aligned}
\mathbf{P}(A\cap B\cap C) &= \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C)\\
\mathbf{P}(A\cap B\cap\bar{C}) &= \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(\bar{C})\\
\mathbf{P}(A\cap\bar{B}\cap C) &= \mathbf{P}(A)\mathbf{P}(\bar{B})\mathbf{P}(C)\\
\mathbf{P}(A\cap\bar{B}\cap\bar{C}) &= \mathbf{P}(A)\mathbf{P}(\bar{B})\mathbf{P}(\bar{C})\\
\mathbf{P}(\bar{A}\cap B\cap C) &= \mathbf{P}(\bar{A})\mathbf{P}(B)\mathbf{P}(C)\\
\mathbf{P}(\bar{A}\cap B\cap\bar{C}) &= \mathbf{P}(\bar{A})\mathbf{P}(B)\mathbf{P}(\bar{C})\\
\mathbf{P}(\bar{A}\cap\bar{B}\cap C) &= \mathbf{P}(\bar{A})\mathbf{P}(\bar{B})\mathbf{P}(C)\\
\mathbf{P}(\bar{A}\cap\bar{B}\cap\bar{C}) &= \mathbf{P}(\bar{A})\mathbf{P}(\bar{B})\mathbf{P}(\bar{C})
\end{aligned}$$

The multiplication rules are symmetrical with respect to any permutation of the events A, B, C, which means that in the terminology we do not have to take into account the order of the events A, B, C, and we can just say that the events A, B, C are independent of each other.

It is important to keep in mind that it may happen that any two of the three events A, B, C are independent of each other, that is,

1. A and B are independent of each other,
2. A and C are independent of each other,
3. B and C are independent of each other,
4. but the three events A, B, C are not independent of each other.

If this is the case, then we say that the events A, B, C are **pairwise independent**, but they are not **totally independent**. So, pairwise independence does not imply total independence.

Independence of more events. The independence of n events can be introduced similarly to the independence of three events. It can be shown that the independence of n events can also be characterized by 2^n **multiplication rules**:

$$\begin{aligned}
\mathbf{P}(A_1\cap A_2\cap\ldots\cap A_n) &= \mathbf{P}(A_1)\mathbf{P}(A_2)\ldots\mathbf{P}(A_n)\\
\mathbf{P}(A_1\cap A_2\cap\ldots\cap\bar{A}_n) &= \mathbf{P}(A_1)\mathbf{P}(A_2)\ldots\mathbf{P}(\bar{A}_n)\\
&\vdots\\
\mathbf{P}(\bar{A}_1\cap\bar{A}_2\cap\ldots\cap\bar{A}_n) &= \mathbf{P}(\bar{A}_1)\mathbf{P}(\bar{A}_2)\ldots\mathbf{P}(\bar{A}_n)
\end{aligned}$$

The following files illustrate and use the multiplication rules for independent events.

Demonstration file: Multiplication rules for independent events
100-01-00

Demonstration file: How many events occur?
100-02-00

Playing with the following file, you may check your ability to decide - on the basis of performed experiments - whether two events are dependent or independent.

Demonstration file: Colors dependent or independent
100-03-00

Section 10
*** Infinite sequences of events

The following rule is a generalization of the addition law of the probability for a finite number of exclusive events, which was described among the basic properties of probability.

Addition law of probability for an infinite number of exclusive events: If $A_1, A_2, \ldots, A_n$ are exclusive events, then

$$\mathbf{P}(A_1 \cup A_2 \cup \ldots) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \ldots$$

Example 1. (Odd or even?) My friend I play with a fair coin. We toss it until the first time a head occurs. We agree that I win if the number of tosses is an odd number, that is, 1 or 3 or 5 ..., and my friend wins if the number of tosses is an even number, that is, 2 or 4 or 6 .... What is the probability that I win? What is the probability that my friend wins?

Remark. You may think that odd end even numbers "balance" each other, so the asked probabilities are equal. However, if you play with the following simulation file for a couple of times, or you read the theoretical solution, then you will experience that this is not true:

Demonstration file: Odd or even?
100-03-50

Solution.

$$\mathbf{P}(\text{I win}) =$$

$$\mathbf{P}(\text{First head occurs at the 1st toss or 3rd toss or 5th toss or ... }) =$$

$$\mathbf{P}(1\text{st}) + \mathbf{P}(3\text{rd}) + \mathbf{P}(5\text{th}) + \ldots =$$

$$0.5+0.5^3+0.5^5+\ldots=\frac{0.5}{1-0.5^2}=\frac{0.5}{1-0.25}=\frac{0.5}{0.75}=\frac{2}{3}$$

$$\mathbf{P}(\text{My friend wins})=$$

$$\mathbf{P}(\text{First head occurs at the 2nd toss or 4th toss or 6th toss or ... })=$$

$$\mathbf{P}(\text{2nd})+\mathbf{P}(\text{4th})+\mathbf{P}(\text{6th})+\ldots=$$

$$0.5^2+0.5^4+0.5^6+\ldots=\frac{0.5^2}{1-0.5^2}=\frac{0.25}{1-0.25}=\frac{0.25}{0.75}=\frac{1}{3}$$

The following two properties are closely related to the addition law of probability for an infinite number of exclusive events.

Limit for an increasing sequence of events. Let $A_1,A_2,\ldots$ be an **increasing sequence of events**, that is, A_k implies A_{k+1} for al k. Let A be the union of the events $A_1,A_2,\ldots$. A clearly means that one of the infinitely many events $A_1,A_2,\ldots$ occurs. The following limit relation holds.

$$\mathbf{P}(A)=\lim_{n\to\infty}\mathbf{P}(A_n)$$

Limit for a decreasing sequence of events. Let $A_1,A_2,\ldots$ be a **decreasing sequence of events**, that is, A_k is implied by A_{k+1} for all k. Let A be the intersection of the events $A_1,A_2,\ldots$. A clearly means the event that all the infinitely many events $A_1,A_2,\ldots$ occur. The following limit relation holds:

$$\mathbf{P}(A)=\lim_{n\to\infty}\mathbf{P}(A_n)$$

The following example gives us an important message: if we consider an event which has a positive probability, and we make an unlimited number of independent experiments, then regardless of how small that probability of the event is, the event, sooner or later, will occur for sure.

Example 2. (Unlimited number of exams) Let us assume that a student passes each of his exams, independently of the previous exams, with a positive probability p, and fails with a probability $q=1-p$. We will show that if p is positive and our student has an unlimited number of possibilities to take the exam in a course, then it is sure that the student, sooner ar later, passes the course.

Solution. Let the event A_n mean that our student fails the first n exams. Because of the independence of the exams, the probability of A_n is:

$$\mathbf{P}(A_n) = q^n$$

Let the event A mean that the student fails all the infinite number of exams. Obviously A implies A_n for all n, so

$$\mathbf{P}(A) \leq \mathbf{P}(A_n) \quad \text{for all } n$$

Since $\mathbf{P}(A_n) \to 0$, when $n \to 0$, the value of the probability $\mathbf{P}(A)$ cannot be positive. Thus, it is 0, which means that its complement has a probability 1, that is, the student, sooner ar later, passes the course for sure.

In order to simulate the above problem see the following file. Whenever you press the F9 key, 10000 experiments are performed. Pressing the F9 key again and again, you will see that, regardless how small the probability of the success is, sooner or later success will occur.

Demonstration file: Many experiments for an event which has a small probability
100-04-00

The purpose of the following numerical example is to show that, if our student's knowledge is strongly declining, then, in spite of the fact that he or she has an infinite number of possibilities, the probability $\mathbf{P}(A)$ may be positive, that is, it is not sure at all that the student ever passes the course.

Example 3. (Student's knowledge strongly declining) Let us assume that our student fails the first exam with a probability

$$\mathbf{P}(A_1) = 0.6 + 0.4/2 = 0.8000$$

and if our student fails the first n exams, then the probability of failing the next exam is:

$$\mathbf{P}(A_{n+1}|A_n) = \frac{0.6 + 0.4/(n+1)}{0.6 + 0.4/n} \quad \text{for all } n$$

meaning that

$$\mathbf{P}(A_2|A_1) = \frac{0.6 + 0.4/3}{0.6 + 0.4/2} = 0.9167$$

$$\mathbf{P}(A_3|A_2) = \frac{0.6 + 0.4/4}{0.6 + 0.4/3} = 0.9545$$

$$\mathbf{P}(A_4|A_3) = \frac{0.6 + 0.4/5}{0.6 + 0.4/4} = 0.9714$$

$$\vdots$$

Using the multiplication rule for a decreasing sequence of events we get, for example, that the value of $\mathbf{P}(A_4)$ is:

$$\mathbf{P}(A_4) = \mathbf{P}(A_1)\ \mathbf{P}(A_2|A_1)\ \mathbf{P}(A_3|A_2)\ \mathbf{P}(A_4|A_3) =$$

$$= \frac{0.6+0.4/2}{1}\ \frac{0.6+0.4/3}{0.6+0.4/2}\ \frac{0.6+0.4/4}{0.6+0.4/3}\ \frac{0.6+0.4/5}{0.6+0.4/4}$$

$$= 0.6+0.4/5$$

In a similar way, it can be shown that the value of $\mathbf{P}(A_n)$ is:

$$\mathbf{P}(A_n) = 0.6+0.4/(n+1)$$

Since the events $A_1, A_2, \ldots$ constitute a decreasing sequence of events, we get:

$$\mathbf{P}(A) = \lim_{n\to\infty} \mathbf{P}(A_n) = \lim_{n\to\infty} 0.6+0.4/(n+1) = 0,6.$$

which means that the student fails all the infinite number of exams with a probability 0.6.

Remark. Notice that in this numerical example the probability that the student fails all the infinite number of exams has a probability not only positive but greater than half. So, in spite of the fact that the student has an infinite number of possibilities, failure forever is more likely than a success ever.

Remark. Let us choose the positive numbers a and b so that $a+b=1$. If in the above example, we replace the value 0.6 by a and the value 0.4 by b , then obviously

$$\mathbf{P}(A_n) = a+b/(n+1)$$

and the probability that the student fails all the infinite number of exams is:

$$\mathbf{P}(A) = \lim_{n\to\infty} \mathbf{P}(A_n) = \lim_{n\to\infty} a+b/(n+1) = a.$$

The following file illustrates this more general case.

Demonstration file: Student's knowledge strongly declining
100-05-00

Remark. We may also calculate how much is the probability that the student fails the first $n-1$ exams, but passes the nth exam:

$$\mathbf{P}(\text{fails the first } n-1 \text{ exams, but passes the } n\text{th exam}) =$$

$$\mathbf{P}(\text{fails the first } n-1 \text{ exams}) - \mathbf{P}(\text{fails the first } n \text{ exams}) =$$

$$\left(a+\frac{b}{n}\right) - \left(a+\frac{b}{n+1}\right) =$$

$$\frac{b}{n(n+1)}$$

The following file show these probabilities.

Demonstration file: When does the student pass the exam?
100-06-00

Section 11
*** Drawing with or without replacement. Permutations

Drawing with replacement. Assume that a box contains 5 tickets with different letters on them, say A, B, C, D, E. Let us draw a ticket, write down the letter written on it, and let us put back the ticket into the box. Then, let us draw again, write down the letter on this ticket, and let us put back this ticket into the box, too. What we did is called drawing twice **with replacement**. Obviously, we may draw several times with replacement, as well. The following files illustrate what drawing with replacement means.

Demonstration file: Drawing with replacement from 10 elements
110-01-00

Demonstration file: Drawing with replacement from 4 red balls and 6 blue balls
110-02-00

Drawing without replacement. Now let us draw a ticket, write down the letter on it, and let us put aside the ticket. Then, let us draw another ticket from the box, write down the letter on this ticket, and let us put aside this ticket, too. What we did is called drawing twice **without replacement**. Obviously, if there are n tickets in the box, we may draw at most n times without replacement. The following files illustrate what drawing without replacement means.

Demonstration file: Drawing without replacement from 10 elements
110-03-00

Demonstration file: Drawing without replacement from 4 red balls and 6 blue balls
110-04-00

Permutations. If there are n tickets in the box, and we draw n times without replacement, then we get a permutation of the n tickets. Obviously, all possible permutations have the same probability. The following file gives a random permutation of the 10 given elements.

Demonstration file: Permutations of 10 elements
110-03-05

The following problem entitled "Catching the Queen" may seem an artificial problem which is far from real life. But as you will see the solution of this problem will help us to find the optimal strategy of a typical real life problem which will be presented later under the title "Sinbad and the 100 beautiful girls".

Example 1. (Catching the Queen) First, let us choose and fix a number c between 0 and 9. Take, for example, $c = 4$. Then let us consider a permutation of the numbers $1, 2, \ldots, 10$, for example, $6, 5, 7, 4, 1, 2, 8, 10, 9, 3$. The largest number, that is, the 10 is called "the Queen", and the largest number before the Queen is called the Servant. In the above example, the Servant is the number 7. Let us denote the position of the Queen by X, and let us denote the position of the Servant by Y. In the above example, the position of the Queen is 8, that is, $X = 8$. and the position of the Servant is 3, that is, $Y = 3$. We are interested in the probability of the event that the position of the Queen is larger than c and the position of the Servant is smaller than or equal to c, that is, $X > c$ and $Y \leq c$. This probability obviously depends on c. We will express this probability in terms of c.,

Solution.

$$\mathbf{P}(\ X > c \ \text{ and } \ Y \leq c\ ) =$$

$$\sum_{k=c+1}^{10} \mathbf{P}(\ X = k \ \text{ and } \ Y \leq c\ )) =$$

$$\sum_{k=c+1}^{10} \mathbf{P}(\ X = k\ )\ \mathbf{P}(\ Y \leq c \mid X = k\ ) =$$

$$\sum_{k=c+1}^{10} \frac{1}{10}\ \frac{c}{k-1} =$$

$$\frac{c}{10}\ \sum_{k=c+1}^{10} \frac{1}{k-1}$$

Remark. Let us consider a permutation of the numbers $1, 2, \ldots, 100$. The largest number, that is, the 100 is called "the Queen", and the largest number before the Queen is called the Servant. Let us choose a number c between 0 and 99. The probability of the event that the position of the Queen is larger than c and the position of the Servant is smaller than or equal to c, that is, $X > c$ and $Y \leq c$ can be calculated the same way as in the previous example:

$$\mathbf{P}(\ X > c \ \text{ and } \ Y \leq c\ ) =$$

$$\sum_{k=c+1}^{100} \mathbf{P}(\ X=k \ \text{ and } \ Y \le c\ )) =$$

$$\sum_{k=c+1}^{100} \mathbf{P}(\ X=k\ )\ \mathbf{P}(\ Y \le c \mid X=k\ ) =$$

$$\sum_{k=c+1}^{100} \frac{1}{100}\ \frac{c}{k-1} =$$

$$\frac{c}{100}\ \sum_{k=c+1}^{100} \frac{1}{k-1}$$

For each number c between 0 and 99, the value of this probability can be calculated by Excel:

Demonstration file: Catching the Queen
110-03-08

We see that the maximal probability occurs when $c = 37$, and the value of the maximal probability rounded to 6 decimal places is 0.371043, or rounded to 2 decimal places is 0.37. This fact will be used in the following example.

Example 2. (Sinbad and the 100 beautiful girls) Imagine that the sultan offers Sinbad to choose one of the 100 beautiful girls in his harem. Sinbad has never seen the girls before. The method how Sinbad is allowed to make his choice is very strict: the girls show up for Sinbad separately, one after the other, in a random order, and Sinbad has the right to say "this is the girl I choose" only once. This means that when a girls shows up and then disappears because Sinbad does not choose her, then Sinbad has a vary small chance to meet this girl again. The purpose of Sinbad is to catch the most beautiful girl, so when he makes his choice, he will ask whether he could catch the most beautiful girl. If he realizes that has caught the most beautiful girl, then he will be happy, otherwise he will be not. (We assume that "beauty" is well defined for each girls so that the girls could be arranged into a well defined order according to their beauties.)

You probably think that Sinbad has no much chance to become happy: 100 unknown girls are too many for Sinbad being able to catch one specific one, namely, the most beautiful one! If Sinbad picks a girl at random, then the probability to catch the most beautiful one is 0.01. You will see soon that, applying a tricky strategy, Sinbad can catch the most beautiful girl with 0.37 probability, which is much larger than 0.01.

Solution. What Sinbad can actually do is to apply a strategy like this. First he observes, say c, girls, without choosing any of them, but keeping in mind the maximal beauty of them. Then he compares the beauty of each later girl to this maximal beauty. If one of the later girls is more beautiful than this maximal beauty, then he chooses that girl. If none of the later girls is more beautiful than this maximal beauty, then he will choose the last girl. Let us think out what is the probability that he can catch the most beautiful girl. Using the terminology and notation of the previous example, we may write:

$$\mathbf{P}(\text{ Sinbad catches the Queen }) =$$

$$\mathbf{P}(\ X > c \ \text{ and } \ Y \le c\ ) =$$

$$\sum_{k=c+1}^{100} \mathbf{P}(\ X = k \ \text{ and } \ Y \le c\ )) =$$

$$\sum_{k=c+1}^{100} \mathbf{P}(\ X = k\ )\ \mathbf{P}(\ Y \le c \mid X = k\ ) =$$

$$\sum_{k=c+1}^{100} \frac{1}{100}\ \frac{c}{k-1} =$$

$$\frac{c}{100}\ \sum_{k=c+1}^{100} \frac{1}{k-1}$$

The maximal value of this expression is calculated in the following file:

Demonstration file: Catching the Queen
110-03-08

So we see that if $c = 37$, then $\mathbf{P}(\text{ Sinbad catches the Queen }) = 0.37$. The best strategy for Sinbad is to observe 37 girls, remembering the maximal beauty of them, and then waiting for a girl who is more beautiful than this maximal beauty, and choosing this girl.

Remark. A very good approximation to the optimal strategy for Sinbad can be derived by the following analytical method. Using the approximation

$$\sum_{k=c+1}^{100} \frac{1}{k-1} \approx \ln(100) - \ln(c) = -\ln(\frac{c}{100})$$

we may write

$$\mathbf{P}(\text{ Sinbad catches the Queen }) =$$

$$\frac{c}{100}\ \sum_{k=c+1}^{100} \frac{1}{k-1} \approx -\frac{c}{100}\ \ln(\frac{c}{100}) = -x\ \ln(x)$$

where $x = -\frac{c}{100}\ \ln(\frac{c}{100})$. We know that the function

$$f(x) = -x\ \ln(x)$$

takes its maximum at $x = \frac{1}{e}$, and the maximal value of this function is $\frac{1}{e}$. This is how we get that the best strategy for Sinbad is to use approximately a c so that

$$\frac{c}{100} \approx \frac{1}{e} \approx 0.37$$

that is

$$c \approx \frac{100}{e} \approx 37$$

and then

$$\mathbf{P}(\quad \text{Sinbad catches the Queen} \quad) \approx \frac{1}{e} \quad \approx 0.37$$

which is really an excellent approximation to the exact solution.

If the number of girls is not 100 but n, then this analytical method obviously gives the approximation

$$c \approx \frac{n}{e} \approx 0.37\ \ n$$

and

$$\mathbf{P}(\quad \text{Sinbad catches the Queen} \quad) \approx \frac{1}{e} \quad \approx 0.37$$

for the optimal solution. The best strategy for Sinbad is to observe approximately 37 percent of the girls, remembering the maximal beauty of them, and then waiting for a girl who is more beautiful than this maximal beauty, and choosing this girl.

Part - II.

Discrete distributions

Section 12
Discrete random variables and distributions

When there are a finite or countably infinite number of outcomes, and each is assigned a (probability) value so that each value is non-negative and their sum is equal to 1, we say that a **discrete distribution** is given on the set of these outcomes. If x is an outcome, and the probability value assigned to it is denoted by $p(x)$, then the function p is called the **weight function** or **probability function** of the distribution. We emphasize that

$$p(x) \geq 0 \quad \text{for all} \quad x$$

$$\sum_x p(x) = 1$$

The second property is called the **normalization property**.

The reader must have learned about the abstract notion of a **point-mass** in mechanics: certain amount of mass located in a certain point. This notion is really an abstract notion, because in reality, any positive amount of mass has a positive diameter, while the diameter of a point is zero. Although, point-masses in reality do not exist, our fantasy helps us to imagine them. It is advantageous to interpret the term $p(x)$ of a discrete distribution not only as the probability of the possible value x, but also as if an amount of mass $p(x)$ were located in the point x. Thus, a discrete distribution can be interpreted as a **point-mass distribution**.

The following file offers several ways to visualize a discrete distribution.

Demonstration file: Methods of visualization of discrete distributions
120-01-00

When the possible results of an observation are real numbers, we say that we work with a **random variable**. Thus, to define a random variable, it means to refer to a (random) numerical value or ask a question so that the answer to the question means some (random) number. It is useful to abbreviate the declaration or question by a symbol (most often used are capital letters like X, Y, ..., or Greek letters like α, β, ...) so that later we can refer to the random variable by its symbol. Here are some examples for random variables:

1. Tossing a fair die, let X denote the random variable defined by

   $$X = \text{the number which shows up on the top of the die}$$

   or, equivalently,

   $$X = \text{which number shows up on the top of the die?}$$

   Then, the equality $X = 6$ means the event that *we toss a six with the die*, the inequality $X < 3$ means the event that *we toss a number which is less than 3*.

2. Tossing two coins, let Y denote the random variable defined by

   $$Y = \text{the number of heads we get with the two coins}$$

When a random variable is thought of, we may figure out its **possible values**. The possible values of the random variable X defined above are $1,2,3,4,5,6$. The possible values of the random variable Y defined above are $\{0,1,2\}$. When there are a finite number or countably infinite number of outcomes, then we say that the distribution and the random variable are **discrete**. When we figure out the probability of each outcome of a discrete random variable, we say, that we figure out the **distribution of the random variable**. The distribution of the random variable X is quite simple: each of the numbers $1,2,3,4,5,6$ has the same probability, namely, $\frac{1}{6}$. This can be described, among others, by a table:

x	1	2	3	4	5	6
$p(x)$	1/6	1/6	1/6	1/6	1/6	1/6

The distribution of the random variable Y is:

x	0	1	2
$p(x)$	0.25	0.50	0.25

Distributions are also described by formulas, as in the following chapters, where the most important discrete distributions will be listed.

The following file visualizes the most important discrete distributions.

Demonstration file: Most important discrete distributions
120-10-00

Calculating a probability by summation. If an event corresponds to a subset A of the sample space, then the probability of the event can be calculated by the sum:

$$\mathbf{P}(A) = \sum_{x:x\in A} p(x)$$

In this sum, we summarize the probabilities of those outcomes which belong to the set A corresponding to the event.

Example 1. (Will everybody play?) Imagine that there is group of 10 people who like to play the card game called "bridge". Each evening, those of them who are free that evening come together in the house of one of them. As you probably know 4 persons are needed for this game. When all the 10 people come together, then 8 people can play, and 2 just stay there and watch the others playing. When only 9 people come together, then 8 people can play, and 1 just stays there and watches the others playing. When only 8 people come together, then all the 8 people can play. When only 7 people come together, then 4 people can play, and 3 just stay there and watch the others playing. And so on. Assume that the probability that exactly x people come together is $p(x)$, where $p(x)$ is given by the following table:

x	1	2	3	4	5	6	7	8	9	10
$p(x)$	0.01	0.04	0.06	0.09	0.10	0.15	0.25	0.20	0.07	0.03

The question is: what is the probability that all the gathering people can play, that is, nobody has to only stay and watch the others playing? In other words: what is the probability that 4 or 8 gather together?

Solution. The solution is obvious: in order to get the answer, we have to add $p(4)$ and $p(8)$: $p(4)+p(8)=0.09+0.20=0.29$.

Using Excel. The above summation was very easy. However, when there are many terms to add, then it may be convenient to perform the addition in Excel using simple summation, or using the `SUM`-command (in Hungarian: `SZUM`), or using the `SUMIF`-command (in Hungarian: `SZUMHA`), or using the `SUMPRODUCT`-command (in Hungarian: `SZORZATÖSSZEG`).

The following file shows these possibilities to perform the summation in order to calculate probabilities:

Demonstration file: Calculating probabilities by summation, using Excel - Case 1
020-35-00

Example 2. (Is the number of people greater than 6 and less than 9 ?) Assume that 5 men and 5 women go for a hike so that the probability that the number of men is x and the number of women is y equals $p(x,y)$, where the value of $p(x,y)$ is given by the table in the following file. In the following file, the probability that the number of people is greater than 6 and less than 9 is calculated by the same 4 ways as in the previous example.

Demonstration file: Calculating probabilities by summation, using Excel - Case 2
020-35-50

Section 13
Uniform distribution (discrete)

A fair die has 6 possible values so that their probabilities are equal.

The following file simulates 1000 tosses with a fair die:

Demonstration file: Fair die, 1000 tosses
020-36-00

If the possible values of a random variable constitute an interval $[A,B]$ of integer numbers, and the probabilities of these values are equal, then the weight function is constant on this interval of integer numbers:

Weight function (probability function): It is a constant function. The value of the constant is the reciprocal of the number of integers in $[A,A+1,\ldots,B]$, which is $B-A+1$.

$$p(x) = \frac{1}{B-A+1} \quad \text{if } x = A,\ldots,B$$

Here A and B are parameters:
A = left endpoint of the interval
B = right endpoint of the interval

The following file shows the uniform distribution:

Demonstration file: Uniform distribution (discrete)
120-00-05

Proof of the normalization property.

$$\sum_x p(x) = \sum_{x=A}^{B} \frac{1}{B-A+1} = (B-A+1)\frac{1}{B-A+1} = 1$$

Section 14
Hyper-geometrical distribution

Application: Phenomenon: Red and green balls are in a box. A given number of draws are made *without replacement*.

Definition of the random variable:

$$X = \text{ the number of times we draw red}$$

Parameters:

$$A = \text{ number of red balls in the box}$$

$$B = \text{ number of green balls in the box}$$

$$n = \text{ number of draws}$$

The following file simulates a hyper-geometrical random variable.

Demonstration file: Drawing without replacement from 4 red balls and 6 blue balls
110-04-00

Weight function (probability function):

$$p(x) = \frac{\binom{A}{x}\binom{B}{n-x}}{\binom{A+B}{n}} \quad \text{if } \max(0, n-B) \leq x \leq \min(n, A)$$

The following file shows the hyper-geometrical distribution:

Demonstration file: Hyper-geometrical distribution
120-10-10

Proof of the formula of the weight function. First let us realize that the possible values of X must satisfy the inequalities: $x \geq 0$, $x \leq n$, $x \leq A$, $n - x \leq B$. The last of these four inequalities means that $x \geq n - B$. Obviously, $x \geq 0$ and $x \geq n - b$ together mean that $x \geq \max(0, n - B)$, and $x \leq n$ and $x \leq A$ together mean that $x \leq \min(n, A)$. This is how we get that $\max(0, n - B) \leq x \leq \min(n, A)$. In order to find an expression for $p(x)$, we need to study the event "X=x" which means that "there are exactly x red balls among the n balls" which we have drawn". In order to use the classical formula, first we realize that there are $\binom{A+B}{n}$ equally probable possible combinations. The favorable combinations are characterized by the property that x of the chosen balls are red, and $n - x$ are green. The number of such combinations is $\binom{A}{x}\binom{B}{n-x}$. The ratio of the two expressions gives the formula of the weight function.

Proof of the normalization property.

$$\sum_x p(x) = \sum_{x=\max(0,n-B)}^{\min(n,A)} \frac{\binom{A}{x}\binom{B}{n-x}}{\binom{A+B}{n}} = 1$$

We used the fact that

$$\sum_{x=\max(0,n-B)}^{\min(n,A)} \binom{A}{x}\binom{B}{n-x} = \binom{A+B}{n}$$

which can be derived by the following combinatorial argument: Assume that there are A red and B blue balls in a box. If n balls are drawn without replacement, then the number of combinations in which there are exactly x red balls is $\binom{A}{x}\binom{B}{n-x}$, so the total number of combinations is

$$\sum_{x=\max(0,n-B)}^{\min(n,A)} \binom{A}{x}\binom{B}{n-x}$$

On the on the hand, the number of all possible combinations is obviously $\binom{A+B}{n}$.

Remark. Some textbooks define the same distribution by a different parameter setting, so that $N = A + B$, $K = A$ and n are considered parameters. Then the probability of x looks like as this:

$$\frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \quad \text{if } \max(0, n - N + A) \leq x \leq \min(n, A)$$

In this approach, the parameters:

$$K = \text{ number of red balls in the box}$$

$$N = \text{ number of all balls in the box}$$
$$n = \text{ number of draws}$$

Using Excel. In Excel, the function `HYPGEOMDIST` (in Hungarian: `HIPERGEOM.ELOSZLÁS`) is associated to this second approach:

$$\frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} = \texttt{HYPGEOMDIST(x;n;K;N)}$$

In case of the first approach, the Excel-function `HYPGEOMDIST` should be used with the following parameter-setting:

$$\frac{\binom{A}{x}\binom{B}{n-x}}{\binom{A+B}{n}} = \texttt{HYPGEOMDIST(x;n;A;A+B)}$$

Here is an example for the application of the hyper-geometrical distribution:

Example 1. (Lottery) There are two popular lotteries in Hungary. One is called 5-lottery, the other is called 6-lottery. On a 5-lottery ticket, the player has to fill in 5 numbers out of 90, and on Sunday evening 5 numbers are chosen at random out of the 90 numbers. If the chosen numbers are the same as the filled numbers, then the player has 5 hits, and wins a huge amount of money. If 4 of the chosen numbers are among the filled numbers, then the player has 4 hits, and wins a good amount of money. 3 hits mean a not negligible amount of money. 2 hits mean a some small amount of money. For 1 hit or 0 hits, the player does not get any money. If you play this lottery, you will be interested in knowing how much is the probability of each of the possible hits. On a 6-lottery ticket, the player has to fill in 6 numbers out 45 of , and on Saturday evening 6 numbers are chosen at random out of the 45 numbers. We give the probability of each of the possible hits for the 6-lottery, as well.

Solution. The answers are given by the formula of the hyper-geometrical distribution.

5-lottery:

$$\mathbf{P}(5 \text{ hits}) = \frac{\binom{5}{5}\binom{85}{0}}{\binom{90}{5}} = 0,00000002 = 2 \cdot 10^{-8}$$

$$\mathbf{P}(4 \text{ hits}) = \frac{\binom{5}{4}\binom{85}{1}}{\binom{90}{5}} = 0,00000967 = 1 \cdot 10^{-5}$$

$$\mathbf{P}(3 \text{ hits}) = \frac{\binom{5}{3}\binom{85}{2}}{\binom{90}{5}} = 0,00081230 = 2 \cdot 10^{-4}$$

$$\mathbf{P}(2 \text{ hits}) = \frac{\binom{5}{2}\binom{85}{3}}{\binom{90}{5}} = 0,02247364 = 2 \cdot 10^{-2}$$

$$\mathbf{P}(1 \text{ hit }) = \frac{\binom{5}{1}\binom{85}{4}}{\binom{90}{5}} = 0,23035480 = 2 \cdot 10^{-1}$$

$$\mathbf{P}(0 \text{ hits}) = \frac{\binom{5}{0}\binom{85}{5}}{\binom{90}{5}} = 0,74634956 = 7 \cdot 10^{-1}$$

6-lottery:

$$\mathbf{P}(6 \text{ hits}) = \frac{\binom{6}{6}\binom{39}{0}}{\binom{45}{6}} = 0,00000012 = 1 \cdot 10^{-7}$$

$$\mathbf{P}(5 \text{ hits}) = \frac{\binom{6}{5}\binom{39}{1}}{\binom{45}{6}} = 0,00002873 = 3 \cdot 10^{-5}$$

$$\mathbf{P}(4\text{ hits}) = \frac{\binom{6}{4}\binom{39}{2}}{\binom{45}{6}} = 0,00136463 = 1 \cdot 10^{-3}$$

$$\mathbf{P}(3\text{ hits}) = \frac{\binom{6}{3}\binom{39}{3}}{\binom{45}{6}} = 0,02244060 = 2 \cdot 10^{-2}$$

$$\mathbf{P}(2\text{ hits}) = \frac{\binom{6}{2}\binom{39}{4}}{\binom{45}{6}} = 0,15147402 = 2 \cdot 10^{-1}$$

$$\mathbf{P}(1\text{ hit }) = \frac{\binom{6}{1}\binom{39}{5}}{\binom{45}{6}} = 0,42412726 = 4 \cdot 10^{-1}$$

$$\mathbf{P}(0\text{ hits}) = \frac{\binom{6}{0}\binom{39}{6}}{\binom{45}{6}} = 0,40056464 = 4 \cdot 10^{-1}$$

For the calculation, made by Excel, see the following file.

Demonstration file: Lottery probabilities
120-01-50

Section 15
Binomial distribution

Applications: 1. Phenomenon: Red and green balls are in a box. A given number of draws are made *with replacement*.

Definition of the random variable:

$$X = \text{the number of times we draw red}$$

Parameters:

$$n = \text{number of draws}$$

$$p = \text{probability of drawing red at one draw} = \frac{\text{number of red}}{\text{number of all}}$$

2. Phenomenon: We make a given number of experiments for an event.

Definition of the random variable:

$$X = \text{number of times the event occurs}$$

Parameters:

$$n = \text{number of experiments}$$

$$p = \text{probability of the event}$$

3. Phenomenon: A given number of independent events which have the same probability are observed.

Definition of the random variable:

$$X = \text{how many of the events occur}$$

Parameters:

$$n = \text{number of events}$$

$$p = \text{common probability value of the events}$$

The following file simulates a binomial random variable.

Demonstration file: Binomial random variable: simulation with bulbs
120-03-00

Weight function (probability function):

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{if } x = 0,1,2,\ldots,n$$

The following file shows the binomial distribution:

Demonstration file: Binomial distribution
120-10-20

Proof of the formula of the weight function. In order to find an expression for $p(x)$, we need to study the event "X=x" which means that "the number times we draw red is x", which automatically includes that the number of times we draw green is $n-x$. If the variation of the colors were prescribed, for example, we would prescribe that the 1st, the 2nd, and so on the xth should be red, and the $(x+1)$th, the $(x+2)$th, and so on the nth should be green, then the probability of each of these variations would be $p^n(1-p)^{n-x}$. Since there are $\frac{n!}{x!\,(n-x)!} = \binom{n}{x}$ variations, the the product of $\binom{n}{x}$ and $p^n(1-p)^{n-x}$ really yields the formula of the weight function.

Proof of the normalization property.

$$\sum_x p(x) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = (p+(1-p))^n = 1^n = 1$$

We used the binomial formula

$$\sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x} = (a+b)^n$$

known from algebra, with $a = p$, $b = 1-p$.

Approximation with hyper-geometrical distribution: Let n and p be given numbers. If A and B are large, and $\frac{A}{A+B}$ is close to p, then the terms of the hyper-geometrical distribution with parameters A, B, n approximate the terms of the binomial distribution with parameters n and p:

$$\frac{\binom{A}{x}\binom{B}{n-x}}{\binom{A+B}{n}} \approx \binom{n}{x} p^x (1-p)^{n-x}$$

More precisely: for any fixed n, p and x, $x = 0,1,2,\ldots$, it is true that if $A \to \infty$, $B \to \infty$, $\frac{A}{A+B} \to p$, then

$$\frac{\binom{A}{x}\binom{B}{n-x}}{\binom{A+B}{n}} \to \binom{n}{x} p^x (1-p)^{n-x}$$

Using the following file, you may compare the binomial- and hyper-geometrical distributions:

Demonstration file: Comparison of the binomial- and hyper-geometrical distributions 120-10-05

Proof of the approximation is left for the interested reader as a limit-calculation exercise.

Remark. If we think of the real-life application of the hyper-geometrical and binomial distributions, then the statement becomes quite natural: if we draw a given (small) number of times from a box which contains a large number of red and blue balls, then the fact whether we draw without replacement (which would imply hyper-geometrical distribution) or with replacement (which would imply binomial distribution) has only a negligible effect.

Using Excel. In Excel, the function `BINOMDIST` (in Hungarian: `BINOM.ELOSZLÁS`) is associated to this distribution. If the last parameter is `FALSE`, we get the weight function of the binomial distribution:

$$\binom{n}{x} p^x (1-p)^{n-x} = \texttt{BINOMDIST(x;n;p;FALSE)}$$

If the last parameter is `TRUE`, then we get the so called accumulated probabilities for the binomial distribution:

$$\sum_{x=0}^{k} \binom{n}{x} p^x (1-p)^{n-x} = \texttt{BINOMDIST(k;n;p;TRUE)}$$

Example 1. (Air-plane tickets) Assume that there are 200 seats on an air-plane, and 202 tickets are sold for a flight on that air-plane. If some passengers - for different causes - miss the flight, then there remain empty seats on the air-plane. This is why some air-lines sell more tickets than the number of seats in the air-plain. Clearly, if 202 tickets are sold, then it may happen that more people arrive at the gate of the flight at the air-port than 200, which is a bad situation for the air-line. Let us assume that each passenger may miss the flight independently of the others with a probability $p = 0.03$. If $n = 202$ tickets are sold, then how much is the probability that there are more than 200 people?

Solution. The number of occurring follows, obviously, binomial distribution with parameters $n = 202$ and $p = 0.03$.

$$\mathbf{P}(\text{More people occur than 200}) =$$

$$\mathbf{P}(\text{ 0 or 1 persons miss the flight}) =$$

$$\texttt{BINOMDIST(1; 202; 0,03; TRUE)} \approx 0,015$$

We see that under the assumptions, the bad situation for the air-line will take place only in 1-2 % of the cases.

Using Excel. It is important for an air-line which uses this strategy to know how the bad situation depends on the parameters n and p. The answer is easily given by an Excel formula:

$$\texttt{BINOMDIST(n-201; n; p; TRUE)}$$

Using this formula, it is easy to construct a table in Excel which expresses the numerical values of the probability of the bad situation in terms of n and p:

Demonstration file: Air-plane tickets: probability of the bad event
120-03-90

Example 2. (How many chairs?) Let us assume that each of the 400 students at a university attends a lecture independently of the others with a probability 0.6. First, let us assume that there are, say, only 230 chairs in the lecture-room. If more than 230 students attend, then some of the attending students will not have a chair. If 230 or less students attend, then all attending students will have a chair. The probability that the second case holds:

$$\mathbf{P}(\text{All attending students will have a chair}) =$$

$$\mathbf{P}(\text{230 or less students attend}) =$$

$$\texttt{BINOMDIST(230; 400; 0,6; TRUE)} \approx 0,17$$

Now, let us assume that there are 250 chairs. If more than 250 students attend, then some of the students will not have a chair. Now:

$$\mathbf{P}(\text{All attending students will have a chair}) =$$

$$\mathbf{P}(\text{250 or less students attend}) =$$

$$\texttt{BINOMDIST(250; 400; 0,6; TRUE)} \approx 0,86$$

We may want to know: how many chairs are needed to guarantee that

$$\mathbf{P}(\text{All attending students will have a chair}) \geq 0,99$$

Remark. The following wrong argument is quite popular among people who have not learnt probability theory. Clearly, if there are 400 chairs, then :

$$\mathbf{P}(\text{all attending students will have a chair}) = 1$$

So, they think, taking the 99 % of 400, the answer is 396. We will see that much less chairs are enough, so 396 chairs would be a big waste here.

Solution. To give the answer we have to find c so that

$$\mathbf{P}(\text{All attending students will have a chair}) =$$

$$\mathbf{P}(c \text{ or less students attend}) =$$

$$\texttt{BINOMDIST(c; 400; 0,6; TRUE)} \geq 0,99$$

Using Excel, we construct a table for `BINOMDIST(c; 400; 0,6; TRUE)`

Demonstration file: How many chairs?
120-03-95

A part of the table is printed here:

c	**P**(all attending students will have a chair)
260	0,9824
261	0,9864
262	0,9897
263	0,9922
264	0,9942
265	0,9957

We see that if $c < 263$, then

$$\mathbf{P}(\text{All attending students will have a chair}) < 0,99$$

if $c \geq 263$, then

$$\mathbf{P}(\text{All attending students will have a chair}) \geq 0,99$$

Thus, we may conclude that 263 chairs are enough.

Remark. The way how we found the value of c was the following: we went down in the second column on the table, and when we first found a number greater than or equal to 0.99, we took the c value standing there in the first column.

Using Excel. In Excel, there is a special command to find the value c in such problems: `CRITBINOM(n;p;y)` (in Hungarian: `KRITBINOM(n;p;y)`) gives the smallest c value for which `BINOMDIST(c; n; p; TRUE)` $\geq y$. Specifically, as you may be convinced

$$\texttt{CRITBINOM(400;0,6;0,99)} = 263$$

Using the `CRITBINOM(n;p;y)`-command, we get that

y	CRITBINOM(400 ; 0,6 ; y)
0,9	253
0,99	263
0,999	270
0,9999	276
0,99999	281
0,999999	285

which shows, among others, that with 285 chairs:

$$\mathbf{P}(\text{all attending students will have a chair}) \geq 0,999999$$

Putting only 285 chairs instead of 400 into the lecture-room, we may save 125 chairs on the price of a risk which has a probability less than 0,0000001. Such facts are important when the size or capacity of an object is planned.

Example 3. (Computers and viruses) There are 12 computers in an office. Each of them, independent of the others, has a virus with a probability 0.6. Each computer which has a virus still works with a probability 0.7, independent of the others. The number of computers having a virus is a random variable V. It is obvious that V has a binomial distribution with parameters 12 and 0.6. The number of computers having a virus, but still working is another random variable, which we denote by W. It is obvious that if $V = i$, then W has a binomial distribution with parameters i and 0.7. It is not difficult to see that W has a binomial distribution with parameters 12 and $(0.6)(0.7) = 0.42$. In the following file, we simulate V and W, and first calculate the following probabilities:

$$P(V = 4)$$

$$P(W = 3|V = 4)$$

$$P(V = 4 \text{ and } W = 3)$$

Then we calculate the more general probabilities:

$$P(V = i)$$

$$P(W = j|V = i)$$

$$P(V = i \text{ and } W = j)$$

Finally, we calculate the probability

$$P(W = j)$$

in two ways: first from the probabilities

$$P(V = i \text{ and } W = j)$$

by summation, and then using the BINOMDIST Excel function with parameters 12 and 0.42. You can see that we get the same numerical values for

$$P(W = j)$$

in both ways.

Demonstration file: Computers and viruses
120-04-00

Example 4. (Analyzing the behavior of the relative frequency) If we make 10 experiments for an event which has a probability 0.6, then the possible values of the frequency of the event (the number of times it occurs) are the numbers $0, 1, 2, \ldots, 10$, and the associated probabilities come from the binomial distribution with parameters 10 and 0.6. The possible values of the relative frequency of the event (the number of times it occurs divided by 10) are the numbers $0.0, 0.1, 0.2, \ldots, 1.0$, and the associated probabilities are the same: they come from the binomial distribution with parameters 10 and 0.6. We may call this distribution as a **compressed binomial distribution**. In the following first file, take n, the number of experiments to be 10, 100, 1000, and observe how the theoretical distribution of the relative frequency, the compressed binomial distribution gets closer and closer to the value 0.6, and - by pressing the F9 key - observe also how the simulated relative frequency is oscillating around 0.6 with less and less oscillation as n increases. The second file offers also ten-thousand experiments, but the size of file is 10 times larger, so downloading it may take much time.

Demonstration file: Analyzing the behavior of the relative frequency (maximum thousand experiments)
120-04-50

Demonstration file: Analyzing the behavior of the relative frequency (maximum ten-thousand experiments)
120-04-51

The special case of of the binomial distribution when $n = 1$ has a special name:

Indicator distribution with parameter p

Application: Phenomenon: An event is considered. We perform an experiment for the event.
Definition of the random variable:

$$X = \begin{cases} 0 & \text{if the event does not occur} \\ 1 & \text{if the event occurs} \end{cases}$$

Parameter:

$$p = \text{probability of the event}$$

Weight function (probability function):

$$p(x) = \begin{cases} 1-p & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases}$$

Section 16
Geometrical distribution (pessimistic)

Remark. The adjective *pessimistic* will become clear when we introduce and explain the meaning of an *optimistic* geometrical distribution as well.

Applications: 1. Phenomenon: There are red and green balls in a box. We make draws *with replacement* until we draw the first red ball.

Definition of the random variable:

$$X = \text{how many green balls are drawn before the first red}$$

Parameter:

$$p = \text{the probability of red at each draw}$$

2. Phenomenon: We make experiments for an event until the first occurrence of the event (until the first "success").

Definition of the random variable:

$$X = \text{the number of experiments needed before the first occurrence}$$

or, with the other terminology,

$$X = \text{the number failures before the first success}$$

Parameter:

$$p = \text{probability of the event}$$

3. Phenomenon: An infinite sequence of independent events which have the same probability is considered.

Definition of the random variable:

$$X = \text{the number of non-occurrences before the first occurrence}$$

Parameter:

$$p = \text{common probability value of the events}$$

The following file simulates a "pessimistic" geometrical random variable.

Demonstration file: Geometrical random variable, pessimistic: simulation with bulbs
120-07-00

Weight function (probability function):

$$p(x) = (1-p)^x p \quad \text{if } x = 0,1,2,\dots$$

The following file shows the pessimistic geometrical distribution:

Demonstration file: Geometrical distribution, pessimistic
120-10-35

Proof of the formula of the weight function. In order to find an expression for $p(x)$, we need to study the event "X=x" which means that "the before the first red ball there are x green balls", which means that the 1st draw is green, and the 2nd draw is green, and the 3rd draw is green, nd so on the xth draw is green, and the $(x+1)$th draw is red. The probability of this is equal to $(1-p)^x p$, which is the formula of the weight function.

Proof of the normalization property.

$$\sum_x p(x) = \sum_{x=0}^{\infty} (1-p)^x p = \frac{p}{1-(1-p)} = \frac{p}{p} = 1$$

We used the summation formula for infinite geometrical series verbalized as "First term divided by one minus the quotient":

$$\sum_{x=0}^{n} q^x a = \frac{a}{1-q} = \frac{p}{p} = 1$$

Using Excel. In Excel, there is no a special function associated to this distribution. However, using the power function `POWER` (in Hungarian: `HATVÁNY` and multiplication, it is easy to construct a formula for this distribution:

$$(1-p)^x p = \texttt{POWER(1-p;x)*p}$$

Section 17
Geometrical distribution (optimistic)

Applications: 1. Phenomenon: Red and green balls are in a box. We make draws with replacement until we draw the first red ball

Definition of the random variable:

$$X = \text{how many draws are needed until the first red}$$

Parameter:

$$p = \text{the probability of red at each draw}$$

2. Phenomenon: We make experiments for an event until the first occurrence of the event.

Definition of the random variable:

$$X = \text{the number of experiments needed until the first occurrence}$$

Parameter:

$$p = \text{probability of the event}$$

3. Phenomenon: An infinite sequence of independent events which have the same probability is considered.

Definition of the random variable:

$$X = \text{the rank of the first occurring event in the sequence}$$

Parameter:

$$p = \text{common probability value of the events}$$

The following file simulates an 'optimistic' geometrical random variables:

Demonstration file: Geometrical random variable, optimistic: simulation with bulbs 120-06-00

Weight function (probability function):

$$p(x) = (1-p)^{x-1}p \quad \text{if } x = 1,2,\dots$$

Demonstration file: Geometrical distribution, pessimistic, optimistic
120-10-40

Proof of the formula of the weight function. In order to find an expression for $p(x)$, we need to study the event "X=x" which means that "the first red ball occurs at the xth draw", which means that the 1st draw is green, and the 2nd draw is green, and the 3rd draw is green, nd so on the $(x-1)$th draw is green, and the xth draw is red. The probability of this is equal to $(1-p)^{x-1}p$, which is the formula of the weight function.

Proof of the normalization property.

$$\sum_x p(x) = \sum_{x=1}^{\infty}(1-p)^{x-1}p = \frac{p}{1-(1-p)} = \frac{p}{p} = 1$$

We used the summation formula for infinite geometrical series verbalized as "First term divided by one minus the quotient":

$$\sum_{x=0}^{n} q^x a = \frac{a}{1-q} = \frac{p}{p} = 1$$

The following file shows the geometrical distribution:

The following file shows the geometrical distributions:

Demonstration file: Geometrical distributions, pessimistic, optimistic
120-10-40

Using Excel. In Excel, there is no a special function associated to this distribution. However, using the power function `POWER` (in Hungarian: `HATVÁNY` and multiplication, it is easy to construct a formula for this distribution:

$$(1-p)^{x-1}p = \texttt{POWER(1-p;x-1)*p}$$

Remark. The terms **pessimistic** and **optimistic** are justified by the attitude that drawing a red ball at any draw may be interpreted as a success, drawing a green ball at any draw may be interpreted as a failure. Now, a person interested in the number of draws until *the first success* can be regarded as an optimistic person compared to someone else who is interested in the *number of failures* before the first success.

Section 18
*** Negative binomial distribution (pessimistic)

Applications: 1. Phenomenon: Red and green balls are in a box. We make draws with replacement until we draw the rth red.

Definition of the random variable:

$$X = \text{how many green balls are drawn before the } r\text{th red ball}$$

Parameters:

$$r = \text{the number of times we want to pick red}$$

$$p = \text{the probability of drawing red at each draw}$$

2. Phenomenon: We make experiments for an event until the rth occurrence of the event (until the rth "success").

Definition of the random variable:

$$X = \text{the number of non-occurrences before the } r\text{th occurrence}$$

or, with the other terminology,

$$X = \text{the number of failures before the } r\text{th success}$$

Parameters:

$$r = \text{the number of times we want to pick red}$$

$$p = \text{the probability of the event}$$

3. Phenomenon: An infinite sequence of independent events which have the same probability is considered.

Definition of the random variable:

$$X = \text{the number of non-occurrences before the } r\text{th occurrence}$$

or, with the other terminology,

$$X = \text{the number of failures before the } r\text{th success}$$

Parameters:

$$r = \text{the number of times we want occurrence}$$

$$p = \text{common probability value of the events}$$

Demonstration file: Negative binomial random variable, pessimistic: simulation with bulbs 120-09-00

Weight function (probability function):

$$p(x) = \binom{x+r-1}{x} p^r (1-p)^x \quad \text{if } x = 0,1,2,\ldots \quad \text{(combinatorial form)}$$

$$p(x) = \binom{-r}{x} p^r \left(-(1-p)\right)^x \quad \text{if } x = 0,1,2,\ldots \quad \text{(analytical form)}$$

The following file shows the pessimistic negative binomial distribution:

Demonstration file: Negative binomial distribution, pessimistic 120-10-45

Proof of the combinatorial form. In order to find an expression for $p(x)$, we need to study the event "X=x" which means that "before the rth red ball, we draw exactly x green balls". This means that among the first $x+r-1$ draws there are exactly x green balls, and the $(x+r)$th draw is a red. The probability that among the first $x+r-1$ draws there are exactly x green balls is equal to

$$\binom{x+r-1}{x} (1-p)^x p^{r-1}$$

and probability that the $(x+r)$th draw is a red is equal to p. The product of these two expressions yields the combinatorial form of the weight function.

Proof of the analytical form. We derive it from the combinatorial form. We expand the binomial coefficient into a fraction of products, and we get that

$$\binom{x+r-1}{x} = \frac{(x+r-1)\ (x+r-2)\ \ldots\ (r+1)\ (r)}{x!} =$$

$$\frac{(-(x+r-1))\ (-(x+r-2))\ \ldots\ (-(r+1))\ (-r)}{x!}\ (-1)^x =$$

$$\frac{(-r)\ (-(r+1))\ \ldots\ (-(x+r-3))\ (-(x+r-1))}{x!}\ (-1)^x =$$

$$\frac{(-r)\ \ (-r-1))\ \ \ldots\ \ (-r-(x-2))\ \ (-r-(x-1))}{x!}\ (-1)^x = \binom{-r}{x}(-1)^x$$

If both the leftmost and the rightmost side of this equality are multiplied by $p^r(1-p)^x$, we get the combinatorial form on the left side and the analytical form on the right side.

Remark. Since the analytical form of the weight function contains the negative number $-r$ in the upper position of the binomial coefficient, the name "negative binomial with parameter r" is used for this distribution.

Proof of the normalization property.

$$\sum_x p(x) = \sum_{x=0}^{\infty} == 1$$

We used the summation formula

The following file simulates a 'pessimistic' negative binomial random variable.

Using Excel. In Excel, the function `NEGBINOMDIST` (in Hungarian: `NEGBINOM.ELOSZL`) is gives the individual terms of this distribution:

$$\binom{x+r-1}{r-1} p^r(1-p)^x = \texttt{NEGBINOMDIST(x;r;p)}$$

This Excel function does not offer a `TRUE`-option to calculate the summarized probabilities. The summarized probability

$$\sum_{i=0}^{k} \binom{i+r-1}{r-1} p^r(1-p)^i$$

can be calculated, obviously, by summation. However, using a trivial relation between negative binomial and binomial distribution, the summarized probability can be directly calculated by the Excel formula `1-BINOMDIST(r-1;k;p;TRUE)`

Section 19
*** Negative binomial distribution (optimistic)

Applications: 1. Phenomenon: Red and green balls are in a box. We make draws with replacement until we draw the rth red.

Definition of the random variable:

$$X = \text{how many draws are needed until the } r\text{th red}$$

Parameters:

$$r = \text{the number of times we want to pick red}$$

$$p = \text{the probability of drawing red at each draw}$$

2. Phenomenon: We make experiments for an event until the rth occurrence of the event.

Definition of the random variable:

$$X = \text{the number of experiments needed until the } r\text{th occurrence}$$

Parameters:

$$r = \text{the number of times we want occurrence}$$

$$p = \text{probability of the event}$$

3. Phenomenon: An infinite sequence of independent events which have the same probability is considered.

Definition of the random variable:

$$X = \text{the rank of the } r\text{th occurring event in the sequence}$$

Parameters:

$$r = \text{the number of times we want occurrence}$$

$$p = \text{common probability value of the events}$$

The following file simulates an "optimistic" negative binomial random variables:

Demonstration file: Negative binomial random variable, optimistic: simulated with bulbs
120-08-00

Weight function (probability function):

$$p(x) = \binom{x-1}{x-r} p^r (1-p)^{x-r} \quad \text{if } x = r, r+1, r+2, \ldots$$

The following file shows the negative binomial distributions, both pessimistic and optimistic:

Demonstration file: Negative binomial distributions: pessimistic, optimistic
120-10-50

Using Excel. In Excel, the function `NEGBINOMDIST` (in Hungarian: `NEGBINOM.ELOSZL`) can be used for this distribution. However, keep in mind that the function `NEGBINOMDIST` is directly associated to the **pessimistic** negative binomial distribution, so for the **optimistic** negative binomial distribution we have to use the function `NEGBINOMDIST` with the following parameter-setting:

$$\binom{x-1}{r-1} p^r (1-p)^{x-r} = \texttt{NEGBINOMDIST(x-r;r;p)}$$

Section 20
Poisson-distribution

Applications: 1. Phenomenon: We make a large number of experiments for an event which has a small probability.

Definition of the random variable:

$$X = \text{number of times the event occurs}$$

Parameter:

$$\lambda = \text{the theoretical average of the number of the times the event occurs}$$

Remark. If the probability of the event is p, and we make n experiments, then

$$\lambda = np$$

Phenomenon: Many, independent events which have small probabilities are observed.

Definition of the random variable:

$$X = \text{how many of them occur}$$

Parameter:

$$\lambda = \text{the theoretical average of the number of the occurring events}$$

2.

Remark. If the number of events is n, and each event has the same probability p, then

$$\lambda = np$$

The following file simulates a Poisson random variable.

Demonstration file: Poisson random variable: simulation with bulbs
120-05-00

Weight function (probability function):

$$p(x) = \frac{\lambda^x}{x!}e^{-\lambda} \quad \text{if } x = 0,1,2,\ldots$$

The following file shows the Poisson distribution:

Demonstration file: Poisson-distribution
120-10-30

Approximation with binomial distribution: if n is large and p is small, then the terms of the binomial distribution with parameters n and p approximate the terms of the Poisson distribution with parameter $\lambda = np$:

$$\binom{n}{x} p^x (1-p)^{n-x} \approx \frac{\lambda^x}{x!}e^{-\lambda}$$

More precisely: for any fixed λ and x so that $\lambda > 0$, $x = 0,1,2,\ldots$, it is true that if $n \to \infty$, $p \to 0$ so that $np \to \lambda$, then

$$\binom{n}{x} p^x (1-p)^{n-x} \to \frac{\lambda^x}{x!}e^{-\lambda}$$

Using the following file, you may compare the binomial- and Poisson distributions:

Demonstration file: Comparison of the binomial- and Poisson distributions
120-10-06

Proof of the approximation.

$$\binom{n}{x} p^x (1-p)^{n-x} = \frac{n(n-1)(n-2)\ldots(n-(x-1))}{x!} p^x (1-p)^{n-x} =$$

$$\frac{\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\ldots\left(\frac{n-(x-1)}{n}\right)}{x!} (np)^x \frac{(1-p)^n}{(1-p)^n} \to$$

$$\frac{(1)\,(1)\,(1)\ldots(1)}{x!}\;(\lambda)^x\,\frac{e^{-\lambda}}{1} \;=\; \frac{\lambda^x}{x!}e^{-\lambda}$$

We used the fact that

$$(1-p)^n \to e^{-\lambda}$$

which follows from the well-known calculus rule stating that if $u \to 1$ and $v \to \infty$, then $\lim u^v = e^{\lim uv}$.

Proof of the normalization property.

$$\sum_x p(x) = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}e^{-\lambda} = e^{-\lambda}\;\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} = e^{-\lambda}\;e^{\lambda} \;=\; 1$$

We used the fact known from the theory of Taylor-series that

$$\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} = e^{\lambda}$$

Using Excel. In Excel, the function `POISSON` (in Hungarian: `POISSON`, too) is associated to this distribution. If the last parameter is `FALSE`, we get the weight function of the Poisson-distribution:

$$\frac{\lambda^x}{x!}e^{-\lambda} = \texttt{POISSON(x;}\lambda\texttt{;FALSE)}$$

If the last parameter is `TRUE`, then we get the so called accumulated probabilities for the Poisson-distribution:

$$\sum_{x=0}^{k}\frac{\lambda^x}{x!}e^{-\lambda} = \texttt{POISSON(k;}\lambda\texttt{;TRUE)}$$

Example 1. (How many fish?) Some years ago I met an old fisherman. He was fishing in a big lake, in which many small fish were swimming regardless of each other. He raised his net from time to time, and collected the fish if there were any. He told me that out of 100 cases the net is empty only 6 times or so, and then he added: "If you can guess the number of fish in the net when I raise it out of the water the next time, I will give you a big sack of money." I am sure he would not have said such a promise if he knew that his visitor was a well educated person in probability theory! I was thinking a little bit, then I made some calculation, and then a said a number. Which number did I say?

Solution. I started to think like this: The number of fish in the net is a random variable. Since the number of fish in the lake is large, and for each fish the probability of being caught is approximately equal to the area of the net compared to the area of the lake, which is a small value, and the fish swim independently, this random variable follows a Poisson distribution.

The information that "out of 100 cases the net is empty only 6 times or so" means that the probability of 0 is $6/100$. The formula for the Poisson distribution at 0 is $\frac{\lambda^0}{0!}e^{-\lambda} = e^{-\lambda}$, so $e^{-\lambda} = 6/100$, from which we get $\lambda = \ln(100/6) \approx 2.8$. This means that the number of fish in the net follows the Poisson distribution with parameter 2.8. Using the file

Demonstration file: How many fish?
120-01-60

The numerical values of this distribution are calculated:

x	0	1	2	3	4	5	6	7	8	9
$p()$	0,06	0,17	0,24	0,22	0,16	0,09	0,04	0,02	0,01,	0,00

We see that the most probable value is the number 2. So, I said "2".

Weight function (probability function):

$$p(x) = \frac{\binom{x-1}{r-1}\binom{n-x}{s-r}}{\binom{n}{s}} \quad \text{if } r \le x \le n-s+r$$

Section 21
Higher dimensional discrete random variables and distributions

When we observe not only one but two random variables, then we may put them together as coordinates of a two-dimensional random variable: if X_1, X_2 are random variables, then (X_1, X_2) is a two-dimensional random variable. If the random variables X_1, X_2 have a finite or countably infinite number of possible values, then the vector (X_1, X_2) has a finite or countably infinite number of possible values, as well, so (X_1, X_2) has a discrete distribution on the plane. Such a distribution can be defined by a formula or by a table of the numerical values of the probabilities, as in the following examples.

Example 1. (Smallest and largest lottery numbers) No direct practical use of studying what the smallest and largest lottery numbers are, nevertheless we shall now consider the following random variables:

$$X_1 = \text{smallest lottery number}$$

$$X_2 = \text{largest lottery number}$$

For simplicity, let us consider a simpler lottery, when 3 numbers are drawn out of 10 (instead of 5 out of 90, or 6 out of 45, as it is in Hungary). Let us first figure out the probability $\mathbf{P}(X_1 = 2, X_2 = 8)$. In order to use the classical formula, we divide the number of the favorable combinations by the number of all possible combinations. Since there are 3 favorable outcomes, namely $(2,3,6), (2,4,6), (2,5,6)$, among the $\binom{10}{3}$ combinations, the probability is

$$\mathbf{P}(X_1 = 2, X_2 = 6) = \frac{3}{\binom{10}{3}} = 0.025$$

In a similar way, whenever $1 \le k_1 < k_2 \le 10$, we get that

$$\mathbf{P}(X_1 = k_1, X_2 = k_2) = \frac{k_2 - k_1 - 1}{\binom{10}{3}}$$

In the following Excel file, the distribution of the vector (X_1, X_2) is given so that these probabilities are arranged into a table:

Demonstration file: Lottery when 3 numbers are drawn out of 10
X_1 = smallest lottery number
X_2 = largest lottery number
Distribution of (X_1, X_2)
120-10-55

In a similar way, for the 90 lottery in Hungary, when 5 numbers are drawn from 90, we get, in a similar way, that

$$\mathbf{P}(X_1 = k_1, X_2 = k_2) = \frac{\binom{k_2 - k_1 - 1}{3}}{\binom{90}{5}} \quad \text{if } 1 \le k_1 < k_2 \le 89$$

In the following Excel file, the distribution of the vector (X_1, X_2) is given so that these probabilities are arranged into a table:

Demonstration file: 90-lottery:
X_1 = smallest lottery number
X_2 = largest lottery number
Distribution of (X_1, X_2)
120-10-56

We may also study the random vector with coordinates

$$X_1 = \text{second smallest lottery number}$$

$$X_2 = \text{second largest lottery number}$$

The distribution of this random vector is given by the formula

$$\mathbf{P}(X_1 = k_1, X_2 = k_2) = \frac{(k_1 - 1)(k_2 - k_1 - 1)(90 - k_2)}{\binom{90}{5}} \quad \text{if } k_1 \ge 2, k_2 \ge k_1 + 2, k_2 \le 90$$

In the following Excel file, the distribution of the vector (X_1, X_2) is given so that these probabilities are arranged into a table:

Demonstration file: 90-lottery:
X_1 = second smallest lottery number
X_2 = second largest lottery number
Distribution of (X_1, X_2)
120-10-57

Example 2. (Drawing until both a red and a blue is drawn) Let us consider a box which contains a certain number of red, blue and white balls. If the probability of drawing a red is denoted by p_1, the probability of drawing a blue is denoted by p_2, then the probability of drawing a white is $1-p_1-p_2$. Let us draw from the box with replacement until we draw both a red and a blue ball. The random variables X_1 and X_2 are defined like this:

$$X_1 = \text{the number of draws until the first red}$$

$$X_2 = \text{the number of draws until the first red}$$

The random variable X_1 obviously has a geometrical distribution with parameter p_1, the random variable X_2 obviously has a geometrical distribution with parameter p_2, so

$$\mathbf{P}(X_1 = k_1) = (1-p_1)^{k_1-1}\ p_1 \quad \text{if } k_1 \geq 1$$

$$\mathbf{P}(X_2 = k_2) = (1-p_2)^{k_2-1}\ p_2 \quad \text{if } k_2 \geq 1,;\ k_2 \geq 1$$

If the draws for X_1 and X_2 are made from different boxes, then - because of the independence - we have that

$$\mathbf{P}(X_1 = k_1, X_2 = k_2) = \mathbf{P}(X_1 = k_1)\mathbf{P}(X_2 = k_2) =$$
$$(1-p_1)^{k_1-1}\ p_1\ (1-p_2)^{k_2-1}\ p_2 \quad \text{if } k_1 \geq 1$$

In the following Excel file, the distribution of the vector (X_1, X_2) is given so that a finite number of these probabilities are arranged into a table:

Demonstration file: Drawing from a box:
X_1 = number of draws until the first red is drawn
X_2 = number of draws until the first blue is drawn
(X_1 and X_2 are independent)
Distribution of (X_1, X_2)
120-10-59

Now imagine that we use only one box, and X_1 and X_2 are related to the same draws. In the following Excel file, a simulation is given for this case:

Demonstration file: Drawing from a box:
X_1 = number of draws until the first red is drawn
X_2 = number of draws until the first blue is drawn
X_1 and X_2 are dependent
Simulation for (X_1, X_2)
120-10-60

It is obvious that X_1 and X_2 cannot be equal to each other. In order to determine the probability $\mathbf{P}(X_1 = k_1, X_2 = k_2)$, first let us assume that $1 \leq k_1 < k_2$. Using the multiplication rule, we get that

$$\mathbf{P}(X_1 = k_1, X_2 = k_2) =$$

$$\mathbf{P}(\text{we draw } k_1-1 \text{ white, then a red, then } k_2-k_1-1 \text{ white or red, then a blue }) =$$
$$(1-p_1-p_2)^{k_1-1}\ p_1\ (1-p_2)^{k_2-k_1-1}\ p_2$$

If $1 \le k_2 < k_1$, then by exchanging the indices, we get that

$$\mathbf{P}(X_1 = k_1, X_2 = k_2) =$$

$$\mathbf{P}(\text{we draw } k_2-1 \text{ white, then a blue, then } k_1-k_2-1 \text{ white or blue, then a red }) =$$
$$(1-p_1-p_2)^{k_2-1}\ p_2\ (1-p_1)^{k_1-k_2-1}\ p_1$$

In the following Excel file, the distribution of the vector (X_1, X_2) is given so that a finite number of these probabilities are arranged into a table:

Demonstration file: Drawing from a box:
X_1 = number of draws until the first red is drawn
X_2 = number of draws until the first blue is drawn
X_1 and X_2 are dependent
Simulation for (X_1, X_2)
120-10-61

When we observe not only one but several random variables, then we may put them together as coordinates of a higher dimensional random variable: if $X_1, \dots; X_n$ are random variables, then $(X_1, \dots; X_n)$ is an n-dimensional random variable. If all the random variables $X_1, \dots; X_n$ have a finite number of possible values, then the vectors $(X_1, \dots; X_n)$ has a finite number of possible values, as well, so $(X_1, \dots; X_n)$ have a discrete distribution.

In the following chapters, some important higher dimensional discrete distributions are described.

Section 22
*** Poly-hyper-geometrical distribution

Application: Phenomenon: Let us consider r different colors:

"1st color"
$\vdots$
"rth color"

Let us put balls into a box so that

A_1 of them are of the "1st color"
$\vdots$
A_r of them are of the "rth color"

The total number of balls in the box is $A = A_1 + \ldots + A_r$. If we draw a ball from the box, then - obviously -

$$\begin{array}{rcl} p_1 & = & \text{probability of drawing a ball of the "1st color"} = \frac{A_1}{A} \\ & \vdots & \\ p_r & = & \text{probability of drawing a ball of the "}r\text{th color"} = \frac{A_r}{A} \end{array}$$

Now let us make a given number of draws from the box *without replacement*. Definition of the coordinates of the random variable X:

$$\begin{array}{rcl} X_1 & = & \text{the number of times we draw balls of the "1st color"} \\ & \vdots & \\ X_r & = & \text{the number of times we draw balls of the "}r\text{th color"} \end{array}$$

Now X is the r-dimensional random variable defined by these coordinates:

$$X_r = (X_1, \ldots, X_r)$$

Parameters:

$$\begin{array}{lcl} n & = & \text{the number of times we draw balls from the box} \\ A_1 & = & \text{number of balls of the "1st color" in the box} \\ & \vdots & \\ A_r & = & \text{number of balls of the "}r\text{th color" in the box} \end{array}$$

Weight function (probability function):

$$p(x_1,\ldots,x_r) = \frac{\binom{A_1}{x_1}\ldots\binom{A_r}{x_r}}{\binom{A_1+\ldots+A_r}{n}}$$

$$\text{if } x_1,\ldots,x_r \text{ are integers, } x_1+\ldots+x_r = n,$$
$$0 \le x_1 \le A_1, \ \ldots \ , 0 \le x_r \le A_r$$

Using Excel. In Excel, the function `COMBIN` (in Hungarian: `KOMBINÁCIÓK`) may be used for this distribution, since

$$\text{COMBIN}(A;x) = \binom{A}{x}$$

Thus, the mathematical formula

$$\frac{\binom{A_1}{x_1}\ldots\binom{A_1}{x_r}}{\binom{A_1+\ldots+A_r}{n}}$$

for the poly-hyper-geometrical distribution can be composed in Excel like this:

$$\frac{\text{COMBIN}(A_1;x_1)\ldots\text{COMBIN}(A_r;x_r)}{\text{COMBIN}(A_1+\ldots+A_r;n)}$$

In Hungarian:

$$\frac{\text{KOMBINÁCIÓK}(A_1;x_1)\ldots\text{KOMBINÁCIÓK}(A_r;x_r)}{\text{KOMBINÁCIÓK}(A_1+\ldots+A_r;n)}$$

Section 23
*** Polynomial distribution

Applications: 1. Phenomenon: Let us consider r different colors:

"1st color"
$\vdots$
"rth color"

Let us put balls into a box so that

A_1 of them are of the "1st color"
$\vdots$
A_r of them are of the "rth color"

The total number of balls in the box is $A = A_1 + \ldots + A_r$. If we draw a ball from the box, then - obviously -

$$\begin{array}{rcl} p_1 & = & \text{probability of drawing a ball of the "1st color"} = \frac{A_1}{A} \\ & \vdots & \\ p_r & = & \text{probability of drawing a ball of the "}r\text{th color"} = \frac{A_r}{A} \end{array}$$

Now let us make a given number of draws from the box *with replacement*.

Definition of the coordinates of the random variable X:

$$\begin{array}{rcl} X_1 & = & \text{the number of times we draw balls of the "1st color"} \\ & \vdots & \\ X_r & = & \text{the number of times we draw balls of the "}r\text{th color"} \end{array}$$

Now X is the r-dimensional random variable defined by these coordinates:

$$X_r = (X_1, \ldots, X_r)$$

Parameters:

$$\begin{aligned} n &= \text{the number of times we draw balls from the box} \\ A_1 &= \text{the number of times we draw balls of the "1st color"} \\ &\vdots \\ A_r &= \text{the number of times we draw balls of the "}r\text{th color"} \end{aligned}$$

2. Phenomenon: Let un consider a total system of events. The number of events in the total system is denoted by r.

   Definition of the coordinates of the random variable X:

$$\begin{aligned} X_1 &= \text{the number of times the 1st event occurs} \\ &\vdots \\ X_r &= \text{the number of times the } r\text{th event occurs} \end{aligned}$$

   Now X is the r-dimensional random variable defined by these coordinates:

$$X_r = (X_1; \dots, X_r)$$

   Parameters:

$$\begin{aligned} n &= \text{number of events in the total system} \\ p_1 &= \text{probability of the 1st event} \\ &\vdots \\ p_r &= \text{probability of the } r\text{th event"} \end{aligned}$$

Other name for the polynomial distribution is: **multinomial distribution**.

Weight function (probability function):

$$p(x_1, \dots, x_r) = \frac{n!}{x_1! \dots x_r!} \; p_1^{x_1} \dots p_r^{x_r}$$

$$\text{if } x_1, \dots, x_r \text{ are integers, } x_1 \geq 0, \dots, x_r \geq 0 \text{ , } x_1 + \dots + x_r = n$$

Using Excel. In Excel, the function `MULTINOMIAL` (in Hungarian: `MULTINOMIAL`, too) may be used for this distribution, since

$$\text{MULTINOMIAL}(x_1, \dots, x_r) = \frac{(x_1 + \dots + x_r)!}{x_1! \dots x_r!}$$

Thus, the mathematical formula

$$\frac{n!}{x_1! \dots x_r!} \; p_1^{x_1} \dots p_r^{x_r}$$

for the polynomial distribution can be composed in Excel like this:

$$\text{MULTINOMIAL}(x_1,\ldots,x_r)*\text{POWER}(p_1;x_1)*\ldots*\text{POWER}(p_r;x_r)$$

In Hungarian:

$$\text{MULTINOMIAL}(x_1,\ldots,x_r)*\text{HATVÁNY}(p_1;x_1)*\ldots*\text{HATVÁNY}(p_r;x_r)$$

Section 24
Generating a random variable with a given discrete distribution

It is an important and useful fact that a random variable with a given discrete distribution easily can be generated by a calculator or a computer. The following is a method for this.

Assume that a discrete distribution is given:

x_1	x_2	x_3	x_4	...
p_1	p_2	p_3	p_4	...

We may calculate the so called **accumulated probabilities**:

$$P_0 = 0$$

$$P_1 = P_0 + p_1 = p_1$$
$$P_2 = P_1 + p_2 = p_1 + p_2$$
$$P_3 = P_2 + p_3 = p_1 + p_2 + p_3$$
$$P_4 = P_3 + p_4 = p_1 + p_2 + p_3 + p_4$$
$$\dots$$

These probabilities clearly constitute an increasing sequence between 0 and 1. So the following definition of the random variable X based on a random number RND is correct, and obviously guarantees that the distribution of the random variable X is the given discrete distribution:

$$X = \begin{cases} x_1 & \text{if} \quad P_0 < \text{RND} < P_1 \\ x_2 & \text{if} \quad P_1 < \text{RND} < P_2 \\ x_3 & \text{if} \quad P_2 < \text{RND} < P_3 \\ x_4 & \text{if} \quad P_3 < \text{RND} < P_4 \\ \dots & \end{cases}$$

The next file uses this method:

Demonstration file: Generating a random variable with a discrete distribution
130-00-00

Section 25
Mode of a distribution

The most probable value (values) of a discrete distribution is (are) called the **mode (modes)** of the distribution. The following method is applicable in calculating the mode of a distribution in many cases.

Method to determine the mode. Assume that the possible values of a distribution constitute an interval of integer numbers. For an integer x, let the probability of x be denoted by $p(x)$. The mode is that value of x for which $p(x)$ is maximal. Finding the maximum by the method, known from calculus, of taking the derivative, equate it to 0, and then solving the arising equation is not applicable, since the function $p(x)$ is defined only for integer values of x, and thus, differentiation is meaningless for $p(x)$. However, let us compare two adjacent function values, for example $p(x-1)$ and $p(x)$ to see which of them is larger than the other:

$$p(x-1) < p(x) \quad \text{or} \quad p(x-1) = p(x) \quad \text{or} \quad p(x-1) > p(x)$$

In many cases, after simplifications, it turns out that there exists a real number c so that the above inequalities are equivalent to the inequalities

$$x < c \quad \text{or} \quad x = c \quad \text{or} \quad x > c$$

In other words:

$$\begin{array}{llll} \text{when} & x < c, & \text{then} & p(x-1) < p(x), \\ \text{when} & x = c, & \text{then} & p(x-1) = p(x), \\ \text{when} & x > c, & \text{then} & p(x-1) > p(x) \end{array}$$

This means that $p(x)$ is increasing on the left side of $\lfloor c \rfloor$, and $p(x)$ is decreasing on the right side of $\lfloor c \rfloor$, guaranteeing that the maximum occurs at $\lfloor c \rfloor$. When c itself is an integer, then there are two values where the maximum occurs: both $c-1$ and c. (Notation: $\lfloor c \rfloor$ means the integer part of c, that is, the greatest integer below c.)

Here we list the modes - without proofs -of the most important distributions. The proofs are excellent exercises for the reader.

1. **Uniform distribution on {A, A+1, . . . , B-1, B}**

   Since all the possible values have the same probability, all are modes.

2. **Hyper-geometrical distribution with parameters** A, B, n

   If $(n+1)\,\frac{A+1}{A+B+2}$ is an integer, then there are two modes:

   $$(n+1)\,\frac{A+1}{A+B+2}$$

   and

   $$(n+1)\,\frac{A+1}{A+B+2}-1$$

   If $(n+1)\,\frac{A+1}{A+B+2}$ is not an integer, then there is only one mode:

   $$\left\lfloor (n+1)\,\frac{A+1}{A+B+2} \right\rfloor$$

3. **Indicator distribution with parameter** p

   If $p < \frac{1}{2}$, then the mode is 0,

   if $p > \frac{1}{2}$, then the mode is 1,

   if $p = \frac{1}{2}$, then both values are the modes.

4. **Binomial distribution with parameters** n **and** p

   If $(n+1)\,p$ is an integer, then there are two modes:

   $$(n+1)\,p$$

   and

   $$(n+1)\,p-1$$

   If $(n+1)\,p$ is not an integer, then there is only one mode:

   $$\lfloor (n+1)\,p \rfloor$$

5. **Poisson-distribution with parameter** λ

   If λ is an integer, then there are two modes:

   $$\lambda$$

   and

   $$\lambda - 1$$

   If λ is not an integer, then there is only one mode:

   $$\lfloor \lambda \rfloor$$

6. **Geometrical distribution (optimistic) with parameter** p

   The mode is 1.

7. **Geometrical distribution (pessimistic) with parameter** p

   The mode is 0.

8. **Negative binomial distribution (optimistic) with parameters** r **and** p

   If $\frac{r-1}{p}$ is an integer, then there are two modes:

   $$\frac{r-1}{p}+1$$

   and

   $$\frac{r-1}{p}$$

   If $\frac{r-1}{p}$ is not an integer, then there is only one mode:

   $$\left\lfloor \frac{r-1}{p} \right\rfloor + 1$$

9. **Negative binomial distribution (pessimistic) with parameters** r **and** p

   If $\frac{r-1}{p}$ is an integer, then there are two modes:

   $$\frac{r-1}{p} - r + 1$$

   and

   $$\frac{r-1}{p} - r$$

   If $\frac{r-1}{p}$ is not an integer, then there is only one mode:

   $$\left\lfloor \frac{r-1}{p} \right\rfloor - r + 1$$

When a discrete distribution is given by a table in Excel, its mode can be easily identified. This is shown is the next file.

Demonstration file: Calculating - with Excel - the mode of a discrete distribution 140-01-00

The next file shows the modes of some important distributions:

Demonstration file: Modes of binomial, Poisson and negative binomial distributions 130-50-00

Section 26
Expected value of discrete distributions

Formal definition of the expected value. Imagine a random variable X which has a finite or infinite number of possible values:

$$x_1, x_2, x_3, \dots$$

The probabilities of the possible values are denoted by

$$p_1, p_2, p_3, \dots$$

The possible values and their probabilities together constitute the distribution of the random variable. We may multiply each possible value by its probability, and we get the products:

$$x_1 p_1, \quad x_2 p_2, \quad x_3 p_3, \quad \dots$$

Summarizing these products we get the series

$$x_1 p_1 + x_2 p_2 + x_3 p_3 + \dots$$

If this series is absolutely convergent, that is

$$|x_1 p_1| + |x_2 p_2| + |x_3 p_3| + \dots < \infty$$

then the value of the series

$$x_1 p_1 + x_2 p_2 + x_3 p_3 + \dots$$

is a well defined finite number. As you learned in calculus this means that rearrangements of the terms of the series do not change the value of the series. Clearly, if all the possible values are greater than or equal to 0, then absolute convergence means simple convergence. If there are only a finite number of possible values, then absolute convergence is fulfilled. In case of absolute convergence, the value of the series

$$x_1 p_1 + x_2 p_2 + x_3 p_3 + \dots$$

is called the **expected value** of the distribution or the expected value of the random variable X, and we say that the expected value **exists and it is finite**. The expected value is denoted by the letter μ or by the symbol $\mathbf{E}(X)$:

$$\mathbf{E}(X) = \mu = x_1 p_1 + x_2 p_2 + x_3 p_3 + \dots$$

Sigma-notation of a summation. The sum defining the expected value can be written like this:

$$\mathbf{E}(X) = \sum_i x_i\, p_i$$

or

$$\mathbf{E}(X) = \sum_x x\, p(x)$$

The summation, obviously, takes place for all possible values x of X.

Using Excel. In Excel, the function `SUMPRODUCT` (in Hungarian: `SZORZATÖSSZEG`) can be used to calculate the expected value of X: if the x values constitute $\texttt{array}_1$ (a row or a column) and the $p(x)$ values constitute $\texttt{array}_2$ (another row or column), then

$$\texttt{SUMPRODUCT(array}_1\texttt{;array}_2\texttt{)}$$

is the sum of the products $xp(x)$, which is the expected value of X:

$$\mathbf{E}(X) = \sum_x x\, p(x) = \texttt{SUMPRODUCT(array}_1\texttt{;array}_2\texttt{)}$$

The following file shows how the expected value of a discrete distribution can be calculated if the distribution is given by a table in Excel.

Demonstration file: Calculating the expected value of a discrete distribution
150-01-00

Mechanical meaning of the expected value: center of mass. If a point-mass distribution is considered on the real line, then - as it is known from mechanics - the center of mass is at the point:

$$\frac{x_1p_1 + x_2p_2 + x_3p_3 + \ldots}{p_1 + p_2 + p_3 + \ldots}$$

If the total mass is equal to one - and this is the case when we have a probability distribution - , then

$$p_1 + p_2 + p_3 + \ldots = 1$$

and we get that the center of mass is at the point

$$\frac{x_1p_1 + x_2p_2 + x_3p_3 + \ldots}{1} = x_1p_1 + x_2p_2 + x_3p_3 + \ldots$$

which gives that the mechanical meaning of the expected value is the center of mass.

Law of large numbers. Now we shall derive the probabilistic meaning of the expected value. For this purpose imagine that we make N experiments for X. Let the experimental results be denoted by $X_1, X_2, \ldots, X_N$. The average of the experimental results is

$$\frac{X_1 + X_2 + \ldots + X_N}{N}$$

We shall show that if the expected value exist, and it is finite, then for large N, the average of the experimental results stabilizes around the expected value:

$$\frac{X_1+X_2+\ldots+X_N}{N} \approx x_1p_1+x_2p_2+x_3p_3+\ldots = \mu = \mathbf{E}(X)$$

This fact, namely, that for a large number of experiments, the average of the experimental results approximates the expected value, is called the **law of large numbers** for the averages. In order to see that the law of large numbers holds, let the frequencies of the possible values be denoted by

$$N_1, N_2, N_3, \ldots$$

Remember that

N_1 shows how many times x_1 occurs among $X_1, X_2, \ldots, X_N$,
N_2 shows how many times x_2 occurs among $X_1, X_2, \ldots, X_N$,
N_3 shows how many times x_3 occurs among $X_1, X_2, \ldots, X_N$,
and so on.

The relative frequencies are the proportions:

$$\frac{N_1}{N}, \frac{N_2}{N}, \frac{N_3}{N}, \ldots$$

If N is large, then the relative frequencies stabilize around the probabilities:

$$\frac{N_1}{N} \approx p_1, \quad \frac{N_2}{N} \approx p_2, \quad \frac{N_2}{N} \approx p_2, \quad \ldots$$

Obviously, the sum of all experimental results can be calculated so that x_1 is multiplied by N_1, x_2 is multiplied by N_2, x_3 is multiplied by N_3, and so on, and then these products are added:

$$X_1+X_2+\ldots+X_N = x_1N_1+x_2N_2+x_3N_3+\ldots$$

This is why

$$\frac{X_1+X_2+\ldots+X_N}{N} = \frac{x_1N_1+x_2N_2+x_3N_3+\ldots}{N} = x_1\frac{N_1}{N}+x_2\frac{N_2}{N}+x_3\frac{N_3}{N}+\ldots$$

Since the relative frequencies on the right side of this equality, for large N, stabilize around the probabilities, we get that

$$\frac{X_1+X_2+\ldots+X_N}{N} \approx x_1p_1+x_2p_2+x_3p_3+\ldots = \mu = \mathbf{E}(X)$$

Remark. Sometimes it is advantageous to write the sum in the definition of the expected value like is:

$$\mathbf{E}(X) = \mu = \sum_x x\, p(x)$$

where the summation takes place for all possible values x.

The following files show how the averages of the experimental results stabilize around the expected value.

Demonstration file: Tossing a die - Average stabilizing around the expected value
150-02-10

Demonstration file: Tossing 5 false coins - Number of heads - Average stabilizing around the expected value
150-02-20

Demonstration file: Tossing a die, law of large numbers
150-02-90

Demonstration file: Tossing a die, squared, law of large numbers
150-03-00

Demonstration file: Binomial random variable, law of large numbers
150-04-00

Demonstration file: Binomial random variable squared, law of large numbers
150-05-00

The expected value may not exist! If the series

$$x_1p_1 + x_2p_2 + x_3p_3 + \dots$$

is not absolutely convergent, that is

$$|x_1p_1| + |x_2p_2| + |x_3p_3| + \dots = \infty$$

then one of the following 3 cases holds:

1. Either
$$x_1p_1 + x_2p_2 + x_3p_3 + \dots = \infty$$
2. or
$$x_1p_1 + x_2p_2 + x_3p_3 + \dots = -\infty$$
3. or the value of the series
$$x_1p_1 + x_2p_2 + x_3p_3 + \dots$$
is not well defined, because different rearrangements of the series may yield different values for the sum.

It can be shown that, in the first case, as N increases

$$\frac{X_1 + X_2 + \dots + X_N}{N}$$

will become larger and larger, and it approaches ∞. This is why we may say that the expected exists, and its value is ∞. In the second case, as N increases,

$$\frac{X_1 + X_2 + \ldots + X_N}{N}$$

will become smaller and smaller, and it approaches $-\infty$. This is why we may say that the expected exists, and its value is $-\infty$. In the third case, as N increases,

$$\frac{X_1 + X_2 + \ldots + X_N}{N}$$

does not approach to any finite or infinite value. In this case we say that the expected value does not exist.

In the following example, we give an example when the expected value is infinity, and thus, the sequence of averages goes to infinity.

Example 1. ("What I pay doubles" - sequence of averages goes to infinity) We toss a coin until the first head (first "success"), and count how many tails ("failures") we get before the first head. If this number is T , then the amount of money I pay is $X = 2^T$ forints. The amount of money is as much as if "we doubled the amount for each failure". We study the sequence of the money I pay. In the following file, we shall see that the sequence of the averages goes to infinity. Because of the large size of the file, opening or downloading may take longer.

Demonstration file: "What I pay doubles" - averages go to infinity
150-06-00

In the following example, we give an example when the expected value does not exists, and thus, the sequence of averages does not converge.

Example 2. ("What we pay to each other doubles" - averages do not stabilize) We toss a coin. If it is a head, then I pay a certain amount of money to my opponent. If it is a tail, then my opponent pays the certain amount of money to me. The amount of money is generated by tossing a coin until the first head (first "success"), and counting how many tails ("failures") we get before the first head. If this number is T , then the amount of money is $X = 2^T$ forints. The amount of money is as much as if "we doubled the amount for each failure". We study the sequence of the money I get, which is positive if my opponent pays to me, and it is negative if I pay to my opponent. In the following file, we shall see that the sequence of the averages does not converge. Because of the large size of the file, opening or downloading may take longer.

Demonstration file: "What we pay to each other doubles" - averages do not stabilize
150-06-10

Section 27
Expected values of the most important discrete distributions

Here we give a list of the formulas of the expected values of the most important discrete distributions. The proofs are given after the list. They are based mainly on algebraic identities and calculus rules. Some proofs would be easy exercises for the reader, others are trickier and more difficult.

1. **Uniform distribution on {A, A+1, . . . , B-1, B}**

$$\mathbf{E}(X) = \frac{A+B}{2}$$

2. **Hyper-geometrical distribution with parameters** A, B, n

$$\mathbf{E}(X) = n\,\frac{A}{A+B}$$

3. **Indicator distribution with parameter** p

$$\mathbf{E}(X) = p$$

4. **Binomial distribution with parameters** n **and** p

$$\mathbf{E}(X) = np$$

5. **Geometrical distribution (optimistic) with parameter** p

$$\mathbf{E}(X) = \frac{1}{p}$$

6. **Geometrical distribution (pessimistic) with parameter** p

$$\mathbf{E}(X) = \frac{1}{p} - 1$$

7. **Negative binomial distribution (optimistic) with parameters r and p**

$$\mathbf{E}(X) = \frac{r}{p}$$

8. **E**(X) = **Negative binomial distribution (pessimistic) with parameters r and p**

$$\mathbf{E}(X) = \frac{r}{p} - r$$

9. **Poisson-distribution with parameter λ**

$$\mathbf{E}(X) = \lambda$$

Proofs.

1. **Uniform distribution on {A, A+1, . . . , B-1, B}**

$$\mathbf{E}(X) = \sum_x x\, p(x) = \sum_{x=A}^{B} x\ \frac{1}{B-A+1} =$$

$$\frac{1}{B-A+1}\ \sum_{x=A}^{B} x = \frac{1}{B-A+1}\ (B-A+1)\,\frac{A+B}{2} = \frac{A+B}{2}$$

Since the distribution is symmetrical about $\frac{a+b}{2}$, it is natural that the expected value is $\frac{a+b}{2}$.

2. **Hyper-geometrical distribution with parameters A, B, n**

$$\mathbf{E}(X) = \sum_x x\, p(x) =$$

$$\sum_{x=\max(0,n-B)}^{\min(n,A)} x\,\frac{\binom{A}{x}\binom{B}{n-x}}{\binom{A+B}{n}} =$$

$$\sum_{x=\max(1,n-B)}^{\min(n,A)} x\,\frac{\binom{A}{x}\binom{B}{n-x}}{\binom{A+B}{n}} =$$

$$\sum_{x=\max(1,n-B)}^{\min(n,A)} \frac{x\,\binom{A}{x}\binom{B}{n-x}}{\binom{A+B}{n}} =$$

$$\sum_{x=\max(1,n-B)}^{\min(n,A)} \frac{A\binom{A-1}{x-1}\binom{B}{n-x}}{\frac{A+B}{n}\binom{A-1+B}{n-1}} =$$

$$n\frac{A}{A+B}\sum_{x=\max(1,n-B)}^{\min(n,A)} \frac{\binom{A-1}{x-1}\binom{B}{n-x}}{\binom{A-1+B}{n-1}} =$$

$$n\frac{A}{A+B}\sum_{x=\max(0,n-1-B)}^{\min(n-1,A-1)} \frac{\binom{A-1}{y}\binom{B}{n-1-y}}{\binom{A-1+B}{n-1}} = n\frac{A}{A+B}$$

We replaced $x-1$ by y, that is why we wrote $1+y$ instead of x, and in the last step, we used that

$$\sum_{x=\max(0,n-1-B)}^{\min(n-1,A-1)} \frac{\binom{A-1}{y}\binom{B}{n-1-y}}{\binom{A-1+B}{n-1}} = 1$$

which follows from the fact that

$$\frac{\binom{A-1}{y}\binom{B}{n-1-y}}{\binom{A-1+B}{n-1}}$$

$$(\max(0,n-1-B) \le x \le \min(n-1,A-1))$$

is the weight function of the hyper-geometrical distribution with parameters $A-1, B, n-1$.

3. **Indicator distribution with parameter** p

$$\mathbf{E}(X) = \sum_x x\,p(x) = 0(1-p)+1p = p$$

4. **Binomial distribution with parameters** n **and** p

$$\mathbf{E}(X) = \sum_x x\,p(x) =$$

$$\sum_{x=0}^{n} x\binom{n}{x}p^x(1-p)^{n-x} =$$

$$\sum_{x=1}^{n} x\binom{n}{x}p^x(1-p)^{n-x} =$$

$$\sum_{x=1}^{n} n \binom{n-1}{x-1} p\, p^{x-1} (1-p)^{n-x} =$$

$$np \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1}(1-p)^{n-x} =$$

$$np \sum_{y=0}^{n-1} \binom{n-1}{y} p^{y}(1-p)^{n-1-y} = np$$

We replaced $x-1$ by y, that is why we wrote $1+y$ instead of x, and in the last step, we used that

$$\sum_{y=0}^{n-1} \binom{n-1}{y} p^{y}(1-p)^{n-1-y} = 1$$

which follows from the fact that

$$\binom{n-1}{y} p^{y}(1-p)^{n-1-y} \quad \text{if} y = 0,1,2,\ldots,n$$

is the weight function of the binomial distribution with parameters $n-1$ and p.

5. **Geometrical distribution (optimistic) with parameter p.** We give two proofs. The first proof uses the techniques of summarizing geometrical series:

$$\mathbf{E}(X) = \sum_{x} x\, p(x) =$$

$$
\begin{array}{lllllllll}
p & + & 2pq & + & 3pq^2 & + & 4pq^3 & + & \ldots & = \\
\\
p & + & pq & + & pq^2 & + & pq^3 & + & \ldots & \\
 & + & pq & + & pq^2 & + & pq^3 & + & \ldots & \\
 & & & + & pq^2 & + & pq^3 & + & \ldots & \\
 & & & & & + & pq^3 & + & \ldots & \\
 & & & & & & & \ddots & & \\
 & & & & & & & & & = \\
\frac{p}{1-q} & & & & & & & & & \\
 & + & \frac{pq}{1-q} & & & & & & & \\
 & & & + & \frac{pq^2}{1-q} & & & & & \\
 & & & & & + & \frac{pq^3}{1-q} & & & \\
 & & & & & & & \ddots & & \\
 & & & & & & & & & = \\
1 & & & & & & & & & \\
 & + & q & & & & & & & \\
 & & & + & q^2 & & & & & \\
 & & & & & + & q^3 & & & \\
 & & & & & & & \ddots & & \\
 & & & & & & & & & =
\end{array}
$$

$$\frac{1}{1-q} = \frac{1}{p}$$

The second proof uses the techniques of power series. Using the notation $q = 1 - p$, we get that

$$\mathbf{E}(X) = \sum_x x\, p(x) =$$

$$1\, p + 2\, p(1-p) + 3\, p(1-p)^2 + 4\, p(1-p)^3 + \ldots =$$

$$1\, p + 2\, pq + 3\, pq^2 + 4\, pq^3 + \ldots =$$

$$p\, \left(1 + 2q + 3q^2 + 4q^3 + \ldots\right) = p\, \frac{1}{(1-q)^2} = p\, \frac{1}{p^2} = \frac{1}{p}$$

We used the identity

$$1 + 2q + 3q^2 + 4q^3 + \ldots = \frac{1}{(1-q)^2}$$

which is proved by first considering the given infinite series as the derivative of a geometrical series, then taking the closed form of the geometrical series, and then differentiating the closed form:

$$1 + 2q + 3q^2 + 4q^3 + \ldots =$$

$$\frac{d}{dq}\left(1+q+q^2+q^3+q^4\ldots\right)=$$

$$\frac{d}{dq}\left(\frac{1}{1-q}\right)=\frac{d}{dq}\left((1-q)^{-1}\right)=(1-q)^{-2}=\frac{1}{(1-q)^2}$$

6. **Geometrical distribution (pessimistic) with parameter p.** Since the pessimistic geometrical distribution can be derived from the optimistic by a shift of 1 unit to the left, the expected value of the pessimistic geometrical distribution is equal to the expected value of the optimistic geometrical distribution minus 1.

$$\mathbf{E}(X)=\frac{1}{p}-1$$

7. **Negative binomial distribution (optimistic) with parameters r and p**

$$\mathbf{E}(X)=\sum_x x\,p(x)=$$

$$\sum_{x=r}^{\infty} x\binom{x-1}{r-1}p^r(1-p)^{x-r}=$$

$$\sum_{x=r}^{\infty} r\binom{x}{r}\frac{p^{r+1}}{p}(1-p)^{x-r}=$$

$$\frac{r}{p}\sum_{x=r}^{\infty}\binom{x}{r}p^{r+1}(1-p)^{x-r}=$$

$$\frac{r}{p}\sum_{x=r+1}^{\infty}\binom{y-1}{r}p^{r+1}(1-p)^{y-1-r}=$$

$$\frac{r}{p}$$

$$\sum_{x=r+1}^{\infty}\binom{y-1}{r}p^{1+r}(1-p)^{y-1-r}=1$$

which follows from the fact that

$$\binom{y-1}{r}p^{1+r}(1-p)^{y-1-r}\quad\textbf{if } y=r+1,r+2,r+3,\ldots$$

is the weight function of the (optimistic) negative binomial distribution with parameters $r+1$ and p.

8. **Negative binomial distribution (pessimistic) with parameters r and p.** Since the pessimistic negative binomial distribution can be derived from the optimistic by a shift of r units to the left, the expected value of the pessimistic negative binomial distribution is equal to the expected value of the optimistic negative binomial distribution minus r.

$$\mathbf{E}(X)=\frac{r}{p}-r$$

9. **Poisson-distribution with parameter λ**

$$\mathbf{E}(X) = \sum_x x\, p(x) =$$

$$\sum_{x=0}^{\infty} x\ \frac{\lambda^x}{x!} e^{-\lambda} =$$

$$\sum_{x=1}^{\infty} x\ \frac{\lambda^x}{x!} e^{-\lambda} =$$

$$\sum_{x=1}^{\infty} \lambda\ \frac{\lambda^{x-1}}{(x-1)!} e^{-\lambda} =$$

$$\lambda\ \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} e^{-\lambda} =$$

$$\lambda\ \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} e^{-\lambda} = \lambda$$

We replaced $x-1$ by y, that is why we wrote $1+y$ instead of x, and in the last step, we used that

$$\sum_{y=0}^{\infty} \frac{\lambda^y}{y!} e^{-\lambda} = 1$$

which follows from the fact that

$$\frac{\lambda^y}{y!} e^{-\lambda} \quad (y = 0, 1, 2, \ldots)$$

is the weight function of the Poisson distribution with parameter λ.

Example 1. (Shooting stars) If you watch the sky form a peak of "Kékes-tetõ" (highest peak in Hungary) around midnight in August, you will see shooting-stars. Assume that the amount of time between two shooting-stars, which is a random quantity, is 10 minutes on the average. If we watch the sky for 15 minutes, then how much is the probability that we see exactly 2 shooting-stars?

Remark. It is not difficult to figure out wrong arguments to get the wrong answers: 0.5 or 0.75. The reader will enjoy to find out these wrong arguments.

Solution. The number of shooting-stars visible during 15 minutes is a random variable. The following facts are obvious:

1. the number of meteors in the space is vary large, and
2. for each meteor the probability that it causes a shooting-star during our 10 minute is small, and
3. the meteors cause a shooting-stars independently of each other,

These facts guarantee that the number of shooting-stars visible during 15 minutes follows a Poisson distribution. The parameter λ of this distribution is equal to the expected value of the number of shooting-stars visible during 15 minutes. Since the amount of time between two shooting-stars is 10 minutes in the average, the average number of shooting-stars visible during 15 minutes is 1.5. Thus, $\lambda = 1.5$. This is why the answer to the question is

$$\mathbf{P}(\text{We see exactly 2 shooting-stars}) = \frac{\lambda^2}{2!}e^{-1.5} =$$

$$= \texttt{POISSON(2; 1,5; FALSE)} \approx 0,251$$

The following file gives not only the expected value but also the standard deviation of geometrical, binomial and Poisson-distributions. The notion of the standard deviation will be introduced later when we will work with the continuous distributions, as well.

Demonstration file: Discrete distributions, expected value, standard deviation
150-07-00

Section 28
Expected value of a function of a discrete random variable

When a random variable X is considered, and $y = t(x)$ is a function, then $Y = t(X)$ clearly defines another random variable. The random variable Y is called the **function of the random variable** X. If we make N experiments for the random variable X, and we substitute the experimental results $X_1, X_2, \ldots, X_N$ into the function $y = t(x)$, we get the values $t(X_1), t(X_2), \ldots, t(X_N)$. Their average is

$$\frac{t(X_1) + t(X_2) + \ldots + t(X_N)}{N}$$

It can be shown that - under some conditions - if N is large, then this average also stabilizes around a non-random value. We show this fact. Obviously,

$$t(X_1) + t(X_2) + \ldots + t(X_N) = t(x_1)N_1 + t(x_2)N_2 + t(x_3)N_3 + \ldots$$

This is why

$$\frac{t(X_1) + t(X_2) + \ldots + t(X_N)}{N} = \frac{t(x_1)N_1 + t(x_2)N_2 + t(x_3)N_3 + \ldots}{N} =$$

$$= t(x_1)\frac{N_1}{N} + t(x_2)\frac{N_2}{N} + t(x_3)\frac{N_3}{N} + \ldots$$

Since the relative frequencies in this formula, for large N, stabilize around the probabilities, we get that

$$\frac{t(X_1) + t(X_2) + \ldots + t(X_N)}{N} \approx t(x_1)p_1 + t(x_2)p_2 + t(x_3)p_3 + \ldots$$

The non-random value on the right side of this formula is the expected value of $t(X)$:

$$\mathbf{E}\,(t(X)) = t(x_1)p_1 + t(x_2)p_2 + t(x_3)p_3 + \ldots$$

Sometimes it is advantageous to write the sum in the following form:

$$\mathbf{E}\,(t(X)) = \sum_i t(x_i)\; p_i$$

or

$$\mathbf{E}\left(t(X)\right) = \sum_x t(x)\ p(x)$$

where the summation takes place for all possible values x. We emphasize again that if N is large, then the average of the values $t(X_1), t(X_2), \ldots, t(X_N)$ is close to the expected value of $t(X)$:

$$\frac{t(X_1)+t(X_2)+\ldots+t(X_N)}{N} \approx \mathbf{E}\left(t(X)\right)$$

The condition for the existence and finiteness of the expected value of $t(X)$ is that the series

$$t(x_1)\ p_1 + t(x_2)\ p_2 + t(x_3)\ p_3 + \ldots$$

is absolutely convergent, which means that

$$|t(x_1)|\ p_1 + |t(x_2)|\ p_2 + |t(x_3)|\ p_3 + \ldots < \infty$$

The following file shows how the expected value of a function of a discrete random variable can be calculated if the distribution of the random variable is given by a table in Excel.

Demonstration file: Calculating - with Excel - the expected value of a function for a discrete distribution
160-01-00

The following files show how the averages of experimental results of a function of a random variable approximate the corresponding theoretical expected value.

Demonstration file: Tossing a die, moments, law of large numbers
160-02-00

Demonstration file: Binomial random variable, moments, law of large numbers
160-03-00

The expected value may not exist! If the infinite sum

$$t(x_1)\ p_1 + t(x_2)\ p_2 + t(x_3)\ p_3 + \ldots$$

is not absolute convergent, then its value is either ∞ or $-\infty$ or its value is not well defined, because different rearrangements of the series may yield different values for the sum. In the first case, it can be proven that as N increases,

$$\frac{t(X_1)+t(X_2)+\ldots+t(X_N)}{N}$$

will become larger and larger, and it approaches ∞. This is why we may say that the expected value of $t(X)$ is ∞. In the second case, it can be proven that as N increases,

$$\frac{t(X_1)+t(X_2)+\ldots+t(X_N)}{N}$$

will become smaller and smaller, and it approaches $-\infty$. This is why we may say that the expected value of $t(X)$ is $-\infty$. In the third case, it can be proven that as N increases,

$$\frac{t(X_1)+t(X_2)+\ldots+t(X_N)}{N}$$

does not approach to any finite or infinite value. In this case we say that the expected value of $t(X)$ does not exist.

Using Excel. In Excel, the function SUMPRODUCT (in Hungarian: SZORZATÖSSZEG) can be used to calculate the expected value of $t(X)$: if the x values constitute $\texttt{array}_1$ (a row or a column) and the $p(x)$ values constitute $\texttt{array}_2$ (another row or column) and the $t(x)$ values constitute $\texttt{array}_3$ (a third row or column), then

$$\texttt{SUMPRODUCT(array}_3\texttt{;array}_2\texttt{)}$$

is the sum of the products $t(x)p(x)$, which is the expected value of $t(X)$:

$$\mathbf{E}(t(X)) = \sum_x t(x)\ p(x) = \texttt{SUMPRODUCT(array}_3\texttt{;array}_2\texttt{)}$$

Demonstration file: Calculating - with Excel - the moments of a discrete distribution
160-04-00

Section 29
Moments of a discrete random variable

The expected value of X^n is called the ***n*th moment** of X:

$$\mathbf{E}\,(X^n)) = x_1^n\, p_1 + x_2^n\, p_2 + x_3^n\, p_3 + \ldots$$

or, using the other notations:

$$\mathbf{E}\,(X^n) = \sum_x x^n\, p(x)$$

The first moment coincides with the expected value. Among all moments, the first and the second moment play the most important role. For emphases, we repeat the definition of the second moment: the expected value of X^2 is called the **second moment** of X:

$$\mathbf{E}\left(X^2\right)) = x_1^2 p_1 + x_2^2 p_2 + x_3^2 p_3 + \ldots$$

or, using the other notations:

$$\mathbf{E}\left(X^2\right) = \sum_x x^2\, p(x)$$

The following files show how the moments of a discrete distribution can be calculated if the distribution is given by a table in Excel.

Demonstration file: Calculating the second moment of a discrete distribution
170-01-00

The following files show how the averages of experimental moments of a random variable approximate the corresponding theoretical moment.

Demonstration file: Tossing a die, second moment
150-03-00

Demonstration file: Binomial random variable, moments
150-05-00

Demonstration file: Tossing a die, moments, law of large numbers
160-02-00

Demonstration file: Binomial random variable, moments, law of large numbers
160-03-00

Using Excel. In Excel, the function SUMPRODUCT (in Hungarian: SZORZATÖSSZEG) can be used to calculate a moment of X: if the x values constitute $\texttt{array}_1$ (a row or a column) and the $p(x)$ values constitute $\texttt{array}_2$ (another row or column) and the nth powers of the x values constitute $\texttt{array}_3$ (a third row or column), then

$$\texttt{SUMPRODUCT(array}_3\texttt{;array}_2\texttt{)}$$

is the sum of the products $x^n p(x)$, which is the nth moment of X:

$$\mathbf{E}(X^n) = \sum_x x^n\, p(x) = \texttt{SUMPRODUCT(array}_3\texttt{;array}_2\texttt{)}$$

Using Excel. In Excel, the second moment of X can be calculated also like this: if the x values constitute $\texttt{array}_1$ (a row or a column) and the $p(x)$ values constitute $\texttt{array}_2$ (another row or column), then

$$\texttt{SUMPRODUCT(array}_1\texttt{;array}_1\texttt{;array}_2\texttt{)}$$

is the sum of the products $x\,x\,p(x)$, which is the second moment of X:

$$\mathbf{E}(X^2) = \sum_x x^2\, p(x) = \sum_x x\,x\,p(x) = \texttt{SUMPRODUCT(array}_1\texttt{;array}_1\texttt{;array}_2\texttt{)}$$

As an example, we calculate here the second moment of the binomial distribution.

Second moment of the binomial distribution with parameters n and p

$$np + n^2p^2 - np^2$$

Proof.

$$\mathbf{E}(X^2) = \sum_x x^2\, p(x) =$$

$$\sum_{x=0}^{n} x^2 \binom{n}{x} p^x (1-p)^{n-x} =$$

$$\sum_{x=1}^{n} x^2 \binom{n}{x} p^x (1-p)^{n-x} =$$

$$\sum_{x=1}^{n} n\,x \binom{n-1}{x-1} p^x\,(1-p)^{n-x} =$$

$$np \sum_{x=1}^{n} x \binom{n-1}{x-1} p^{x-1} (1-p)^{n-x} =$$

Now we replace $x-1$ by y, that is we write $1+y$ instead of x, and we get:

$$np \sum_{y=0}^{n-1} (1+y) \binom{n-1}{y} p^y (1-p)^{n-1-y} =$$

Now the sum splits into the sum of two sums:

$$np \left[\left(\sum_{y=0}^{n-1} 1 \binom{n-1}{y} p^y (1-p)^{n-1-y} \right) + \left(\sum_{y=0}^{n-1} y \binom{n-1}{y} p^y (1-p)^{n-1-y} \right) \right] =$$

Here, inside the bracket, the first sum is the sum of the terms of the binomial distribution with parameters $n-1$ and p, so it is qual to 1. The second sum is the expected value of the binomial distribution with parameters $n-1$ and p, so it is equal to $(n-1)p$. This is why we get:

$$np \ [\, 1 \ + \ (n-1)\, p\,] = np \ + \ n^2 p^2 \ - \ np^2$$

Section 30
Projections and conditional distributions for discrete distributions

If X and Y are discrete random variables, then putting them together we get the two-dimensional random variable (X,Y). Let the weight function of the two-dimensional random variable (X,Y) be denoted by p(x,y), the weight function of the random variable X be denoted by $p_1(x)$, and the weight function of the random variable Y be denoted by $p_2(y)$.

Projection. If the weight function of the two-dimensional random variable (X,Y) is p(x,y), then the weight function $p_1(x)$ of the random variable X can be calculated by summation:

$$p_1(x) = \sum_y p(x,y)$$

Similarly, the weight function $p_2(y)$ of the random variable Y is

$$p_2(y) = \sum_x p(x,y)$$

Case of independent random variables. If X and Y are independent, and their weight functions are $p_1(x)$ and $p_2(y)$ respectively, then the multiplication rule obviously implies that the weight function $p(x,y)$ of (X,Y) is the direct product of the weight functions $p_1(x)$ and $p_2(y)$:

$$p(x,y) = p_1(x)\, p_2(y)$$

Conditional weight function. If a two-dimensional random variable (X,Y) is considered, and somehow the actual value x of X is known for us, but the value of Y is unknown, then we may need to know the conditional distribution of Y under the condition that $X = x$. The conditional weight function can be calculated by division:

$$p_{2|1}(y|x) = \frac{p(x,y)}{p_1(x)}$$

Similarly, the conditional weight function of X under the condition that $Y = y$ is

$$p_{1|2}(x|y) = \frac{p(x,y)}{p_2(y)}$$

It often happens that the weight function of (X,Y) is calculated from one of the product-rules:

$$p(x,y) = p_1(x)\; p_{2|1}(y|x)$$

$$p(x,y) = p_2(y)\; p_{1|2}(x|y)$$

Conditional probability. The conditional probability of an interval for Y, under the condition that $X = x$, can be calculated from the conditional weight by summation:

$$\mathbf{P}(\, y_1 < Y < y_2 \mid X = x \,) = \sum_{y_1}^{y_2} p_{2|1}(y|x)$$

Similarly, the conditional probability of an interval for X, under the condition that $Y = y$, can be calculated from the other conditional weight by summation:

$$\mathbf{P}(\, x_1 < X < x_2 \mid Y = y \,) = \sum_{x_1}^{x_2} p_{1|2}(x|y)$$

Conditional expected value. The conditional expected value is the expected value of the conditional distribution:

$$\mathbf{E}(Y|X = x) = \mu_{2|1}(|x) = \sum_{-\infty}^{\infty} y\; p_{2|1}(y|x)$$

$$\mathbf{E}(X|Y = y) = \mu_{1|2}(|y) = \sum_{-\infty}^{\infty} x\; p_{1|2}(x|y)$$

Files to study construction of a two-dimensional discrete distribution using conditional distributions:

Demonstration file: Construction from conditional distributions, discrete case (version A) 200-75-00

Demonstration file: Construction from conditional distributions, discrete case (version B) 200-76-00

Demonstration file: Projections and conditional distributions, discrete case (version A) 200-77-00

Demonstration file: Projections and conditional distributions, discrete case (version B) 200-78-00

Section 31
Transformation of discrete distributions

Assume that a function $y = t(x)$ is given. When the values of a random variable X are plugged into the argument of the function $y = t(x)$, then then arising $t(X)$ values define a new random variable Y: $Y = t(X)$. It may be important for us to determine the distribution of Y from the distribution of X. When we determine the distribution of Y from the distribution of X, we say that we transform the distribution of X by transformation $Y = t(X)$.

When the one-dimensional random variable X has a discrete distribution $p(x)$, then the distribution $r(z)$ of $Z = t(X)$ is calculated by summation:

$$r(z) = \sum_{x:\, t(x)=z} p(x)$$

With a two-dimensional discrete random variable (X,Y) having a discrete distribution $p(x,y)$, we can calculate the distribution $r(z)$ of $Z = t(X,Y)$, basically, the same way:

$$r(z) = \sum_{(x,y):\, t(x,y)=z} p(x,y)$$

Part - III.

Continous distributions in one-dimension

Section 32
Continuous random variables

In this chapter, we start studying random variables which, contrary to discrete random variables, may take any numbers from an interval of the real line. Some examples:

1. $X =$ the amount of time someone has to wait for the bus when he or she goes to work. The waiting time may be any positive number not exceeding a certain upper limit.

2. $X =$ the weight of a new-born baby. The weight may be any number between certain lower and upper limits.

3. $X =$ the height of a randomly chosen Hungarian man. The height may be any number between certain lower and upper limits.

Obviously, when we observe such a random variable, that is, we check the time or measure the weight or the height,
- we get the time rounded, for example, to minutes, possibly with some decimals,
- we get the weight rounded, for example, to kilograms or grams, or to pounds, possibly with some decimals,
- we get the height rounded, for example, to centimeters, or to inches, possibly with some decimals, as well.

Using more decimals, we get more precise results. However, what we get, are always rational numbers. We emphasize that, in spite of the fact that the measurement results we get are rational numbers, the waiting time, the weight, the height themselves may be not only rational numbers but any real numbers between certain lower and upper limits.

Each specific possible value has zero probability. An important property of such random variables is, that for any fixed real number x, the probability that the random variable is equal to that specific number x is equal to 0:

$$\mathbf{P}(X = x) = 0$$

A random variable X is called to be **continuous** if for any fixed x value, the probability that the random X value is equal to the given x value is equal to 0.

Section 33
Distribution function

The notion of the (cumulative) distribution function (often abbreviated in text-books as **c.d.f.**) plays an important role both in theory and in practice. The **(cumulative) distribution function** $F(x)$ of a random variable X is defined by

$$F(x) = \mathbf{P}(X \leq x)$$

In some text-books, the definition of distribution function may be different

$$F(x) = \mathbf{P}(X < x)$$

However, for a continuous distribution this is not a real difference, because for a continuous distribution:

$$\mathbf{P}(X \leq x) = \mathbf{P}(X < x)$$

Clearly, for any real number x, the event $X > x$ is the complement of the event $X \leq x$, so

$$\mathbf{P}(X > x) = 1 - \mathbf{P}(X \leq x) = 1 - F(x)$$

For any real numbers $a < b$, the event $a < X \leq b$ is the difference of the events $X \leq b$ and $X \leq a$, so

$$\mathbf{P}(a < X \leq b) = \mathbf{P}(X \leq b) - \mathbf{P}(X \leq a) = F(b) - F(a)$$

For a continuous random variable

$$\mathbf{P}(X < x) = \mathbf{P}(X \leq x) = F(x)$$

and

$$\mathbf{P}(X \geq x) = \mathbf{P}(X > x) = 1 - F(x)$$

and

$$\mathbf{P}(a < X < b) = F(b) - F(a)$$
$$\mathbf{P}(a \leq X < b) = F(b) - F(a)$$
$$\mathbf{P}(a < X \leq b) = F(b) - F(a)$$
$$\mathbf{P}(a \leq X \leq b) = F(b) - F(a)$$

Characteristic properties of a distribution function:

1. Any distribution function $F(x)$ is an increasing function (not necessarily strictly increasing), that is, if $x_1 < x_2$, then $F(x_1) \leq F(x_2)$.
2. $F(x)$ has a limit equal to 0 at $-\infty$: $\lim_{x\to-\infty} F(x) = 0$.
3. $F(x)$ has a limit equal to 1 at $(+\infty)$: $\lim_{x\to\infty} F(x) = 1$.
4. The distribution function of a continuous random variable is continuous.

These four properties are characteristics for distribution functions, because, on one side, they are true for the distribution function of any continuous random variable, and on the other side, if a function $F(x)$ is given which has these four properties, then it is possible to define a random variable X so that its distribution function is the given function $F(x)$.

Sometimes it is advantageous to use the so called **tail function** of a distribution:

$$T(x) = \mathbf{P}(X > x)$$

which is the "complement" of the distribution function:

$$T(x) = 1 - F(x)$$

or equivalently:

$$F(x) = 1 - T(x)$$

The tail function is obviously a decreasing function: if $x_1 < x_2$, then $T(x_1) \geq T(x_2)$.

Section 34
*** Empirical distribution function

It is an important fact that the distribution function of a random variable X can be approximated from experimental results as described here. Imagine that we make N experiments for X, and we get the experimental results $X_1, X_2, \ldots, X_N$. Using these experimental results, let us consider the horizontal line segments (closed from the left and open from the right) defined by the point-pairs:

$$(-\infty;0),\ \ (X_1;0)$$
$$(X_1;\frac{1}{N}),\ \ (X_2;\frac{1}{N})$$
$$(X_2;\frac{2}{N}),\ \ (X_3;\frac{2}{N})$$
$$(X_3;\frac{3}{N}),\ \ (X_4;\frac{3}{N})$$
$$(X_4;\frac{4}{N}),\ \ (X_5;\frac{4}{N})$$
$$\vdots$$
$$(X_{N-2};\frac{N-2}{N}),\ \ (X_{N-1};\frac{N-2}{N})$$
$$(X_{N-1};\frac{N-1}{N}),\ \ (X_N;\frac{N-1}{N})$$
$$(X_N;1),\ \ (\infty;1)$$

These line segments constitute the graph of a function called **empirical distribution function** fitted to the experimental results $X_1, X_2, \ldots, X_N$.

For technical purposes, it is more convenient to draw not only the horizontal but the vertical line segments, as well, which will yield a broken line connecting the points

$$(-\infty;0),\ \ (X_1;0),\ \ (X_1;\frac{1}{N}),\ \ (X_2;\frac{1}{N}),\ \ (X_2;\frac{2}{N}),\ \ (X_3;\frac{2}{N}),\ \ \ldots$$

$$\ldots,\ \ (X_{N-1};\frac{N-1}{N}),\ \ (X_N;\frac{N-1}{N}),\ \ (X_N;1),\ \ (\infty;1)$$

It is convenient to think of this broken line as a representation of the graph of the **empirical distribution function** fitted to the experimental results $X_1, X_2, \ldots, X_N$.

The following file shows how the empirical distribution function approximates the theoretical distribution function. Making experiments again and again, that is, pressing the F9 key you can see how the empirical distribution function oscillates around the theoretical distribution.

Demonstration file: Empirical distribution function
200-01-00

Imagine that we make now more and more experiments for X, and we get the experimental results $X_1, X_2, \ldots$. Using these experimental results, we may construct the empirical distribution function or the broken line representing the graph of the empirical distribution function fitted to the experimental results $X_1, X_2, \ldots, X_N$ for all N, and we get a sequence of functions or graphs. The so called **basic theorem of mathematical statistics** states that as N goes to infinity, the sequence of functions or graphs approaches uniformly to the (graph of the) theoretical distribution function of the random variable X. Obviously, the question of the precision of the approximation opens several questions, which we do not discuss here.

Section 35
Density function

The **(probability) density function** (often abbreviated in text-books as **p.d.f.**) is the function $f(x)$ which has the property that for any interval $[a,b]$

$$\mathbf{P}(a < X < b) = \int_a^b f(x)\,dx$$

If $a = x$ and $b = x + \Delta x$, then

$$\mathbf{P}(x < X < x + \Delta x) = \int_x^{x+\Delta x} f(x)\,dx$$

For small Δx, the integral can be approximated by $f(x)\Delta x$, and we get:

$$\mathbf{P}(x < X < x + \Delta x) \approx f(x)\Delta x$$

that is

$$f(x) \approx \frac{\mathbf{P}(x < X < x + \Delta x)}{\Delta x}$$

We emphasize that the value $f(x)$ of the density function does not represent any probability value. If x is a fixed value, then $f(x)$ may be interpreted as a constant of approximate proportionality: for small Δx, the interval $[x, x + \Delta x]$ has a probability approximately equal to $f(x)\Delta x$:

$$\mathbf{P}(x < X < x + \Delta x) \approx f(x)\Delta x$$

The following files show that scattering a point-cloud vertically under the graph of the density function yields a uniformly distributed point-cloud.

Demonstration file: Density, constant on intervals - points scattered vertically / 1

Demonstration file: Density, constant on intervals - points scattered vertically / 2
200-03-00

Mechanical meaning of the density. While learning probability theory, it is useful to know about the mechanical meaning of a density function. For many mechanical problems, if

the density function of a mass-distribution is given, then we are able to imagine the mass-distribution. We study one-dimensional distributions in this chapter, but since we live in a 3-dimensional world, the mechanical meaning of the density function will be first introduced in the 3-dimensional space. The reader must have learned the following arguments in mechanics.

First imagine that mass is distributed in the 3-dimensional space. If we take a point x and a small set (for example, sphere) A around x, then we may compare the amount of mass located in A to the volume of A:

$$\frac{\text{amount of mass located in } A}{\text{volume of } A}$$

This ratio is the average mass-density inside A. Now, if A is getting smaller and smaller, then the average density inside A will approach a number, which is called the mass-density at x. If $f(x,y,z)$ is the density function of mass-distribution in the space, and A is a region in the space, then the total amount of mass located in the region A is equal to the integral

$$\iiint_A f(x,y,z)\,dxdydy$$

Now imagine that mass is distributed on a surface, or specifically on a plane. If we take a point on the surface and a small set (for example, circle) A around the point on the surface, then we may compare the amount of mass located in A to the surface-area of A:

$$\frac{\text{amount of mass located in } A}{\text{surface-area of } A}$$

This ratio is the average density inside A. Now, if A is getting smaller and smaller, then the average density inside A will approach a number, which is called the mass-density at the point on the surface, or specifically on the plane. If $f(x,y)$ is the density function of mass-distribution on the plane, and A is a region in the plane, then the total amount of mass located in the region A is equal to the integral

$$\iint_A f(x,y)\,dxdy$$

Finally, imagine that mass is distributed on a curve, or specifically on a straight line. If we take a point x on the curve and a small set (for example, interval) A around x on the curve, then we may compare the amount of mass located in A to the length of A:

$$\frac{\text{amount of mass located in } A}{\text{length of } A}$$

This ratio is the average density inside A. Now, if A is getting smaller and smaller, then the average density inside A will approach a number, which is called the mass-density at x on the curve, or specifically on the straight line. If $f(x)$ is the density function of mass-distribution on the real line, and $[a,b]$ is an interval, then the total amount of mass located in the interval $[a,b]$ is equal to the integral

$$\int_a^b f(x)\,dx$$

It is clear that a probability density function corresponds to a mass distribution when the total amount of mass is equal to 1.

Characteristic properties of a probability density function:

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

These two properties are characteristic for density functions, because, on one side, they are true for all density functions of any continuous random variables, and on the other side, if a function $f(x)$ is given which has these two properties, then it is possible to define a random variable X so that its density function is the given function $f(x)$.

Relations between the distribution function and the density function. The relations between the distribution function and the density function can be given by integration and differentiation. Integrating the density function from $-\infty$ to x, we get the distribution function at x:

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

However, when $f(x) = 0$ outside an interval $[A,B]$, then to get $F(x)$ for any x between A and B, instead of integrating from $-\infty$ to x, we may integrate from A:

$$F(x) = \int_{A}^{x} f(x)\,dx \quad \text{if} \quad A < x < B$$

and

$$F(x) = \begin{cases} 0 & \text{if} \quad x < A \\ 1 & \text{if} \quad B < x \end{cases}$$

Differentiating the distribution function, we get the density function:

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

Uniform distribution under the graph of the density function. If X is a random variable, and $f(x)$ is its density function, then we may plug the random X value into the density function. We get the random value $f(X)$. It is clear that the random point $(X, f(X))$ will always be on the graph of the density function. Now let us take a random number RND, independent of X, and uniformly distributed between 0 and 1, and let us consider the random point $(X, f(X)\text{RND})$. It is easy to see that the random point $(X, f(X)\text{RND})$ is uniformly distributed in the region under the graph of the density function.

The following files shows the most important distributions: their distribution function and density functions are graphed.

Demonstration file: Continuous distributions
200-57-50

Section 36
*** Histogram

It is an important fact that the density function of a random variable X over an interval $(A;B)$ can be approximated from experimental results as described here. First of all let us divide the interval $(A;B)$ into n small intervals by the points

$$x_0 = A,\ \ x_1,\ \ x_2,\ \ \ldots x_{N-1},\ \ x_N = B$$

Let us imagine that we make N experiments for X. We get the experimental results $X_1, X_2, \ldots, X_N$. Using these experimental results, let us us calculate how many of the experimental results fall into each of the small intervals, and let us consider the relative frequency of each small interval. Then let us draw a rectangle above each small interval so that the area of the small interval is equal to the relative frequency of that small interval, which means that the height of the small interval is equal to the relative frequency of that small interval divided by its length:

$$\text{height} = \frac{\text{relative frequency}}{\text{length}}$$

The upper horizontal sides of the rectangles constitute the graph of a function, which we call the **histogram** constructed from the experimental results to the given small intervals. The histogram approximates the graph of the density function, so the name **empirical density function** is also justified. Obviously, the question of the precision of the approximation opens several questions, which we do not discuss here. One thing, however, must be mentioned: in order to get an acceptable approximation, the number of the experiments must be much larger than the number of the small intervals. In order to get an acceptable figure, the number of small intervals may be around 10, and the number of experiments may be around 1000.

The following files show how a histogram is constructed from experimental results.

Demonstration file: Histogram
200-04-00

Demonstration file: Histogram, standard normal
200-05-00

Section 37
Uniform distributions

1. Special case: Uniform distribution on $(0;1)$

Applications: 1. When something happens during a time interval of unit length so that it may happen in any small part of the interval with a probability equal to the length of that small part, then the time-instant when it occurs follows uniform distribution on $(0;1)$.

2. Random numbers generated by calculators or computers follow uniform distribution on $(0;1)$.

Density function:

$$f(x) = 1 \quad \text{if} \quad 0 < x < 1$$

Distribution function:

$$F(x) = x \quad \text{if} \quad 0 < x < 1$$

The following file shows uniform distributions on $(0;1)$.

Demonstration file: Uniform distribution, random variable and point-cloud on [0,1] 200-07-00

2. General case: Uniform distribution on $(A;B)$

Applications: 1. When something happens during the time interval (A,B) so that it may happen in any small part of the interval with a probability proportional to the length of that small part, then the time-instant when it occurs follows uniform distribution on (A,B).

2. Take a circle with radius 1, and choose a direction, that is, choose a radius of the circle. Then choose a point on the circumference of the circle at random so that none of the parts of the circle have any preference compared to other parts. Then the angle, measured in radians, determined by the fixed radius and the radius drawn to the chosen point has a uniform distribution on the interval $(0;\pi)$. If, instead of radians, the angle is measured in degrees, we get a random variable uniformly distributed between 0 and 360.

3. Random numbers generated by calculators or computers follow uniform distribution on $(0;1)$. If we multiply them by $B-A$, then we get random numbers which follow uniform distribution on $(0;B-A)$. If we add now A to them, then we get random numbers which follow uniform distribution on $(A;B)$.

Density function:

$$f(x)=\frac{1}{B-A} \quad \text{if} \quad A<x<B$$

Distribution function:

$$F(x)=\frac{x-A}{B-A} \quad \text{if} \quad A<x<B$$

Parameters: $-\infty<A<B<\infty$

The following files show uniform distributions.

Demonstration file: Uniformly distributed point on the perimeter of a circle
200-06-00

Demonstration file: Uniform distribution, random variable and point-cloud on [A,B]
200-08-00

Section 38
Distributions of some functions of random numbers

In this chapter, a few examples are given to show how random variables can be constructed by transforming random numbers generated by a calculator or a computer. Since most students have a calculator or a computer, they themselves can generate such random variables, and study them and their distributions. In the next chapter, we shall see that not only the random variables listed in this chapter but all random variables can be simulated by some transformation of a random number generated by a calculator or a computer.

1. **Distribution of RND^2 (random number squared)**

   Density function:
   $$f(x) = \frac{1}{2\sqrt{x}} \quad \text{if} \quad 0 < x < 1$$
   Distribution function:
   $$F(x) = \sqrt{x} \quad \text{if} \quad 0 < x < 1$$

   Files to study RND^2 (random number squared).

   Demonstration file: Point-cloud for RND^2
   200-47-00

   Demonstration file: Density of RND^2
   200-48-00

   Proof. The possible values of RND^2 constitute the interval $(0,1)$. Thus, in the following calculation, we assume that x satisfies $0 < x < 1$.

   (a) Distribution function:
   $$F(x) = \mathbf{P}(X \leq x) = \mathbf{P}\left(\text{RND}^2 \leq x\right) =$$
   $$\mathbf{P}\left(\text{RND} \leq \sqrt{x)}\right) = \sqrt{x} \qquad textif 0 < x < 1$$

(b) Density function:

$$f(x) = F'(x) = \left(\sqrt{x}\right)' = \left(x^{\frac{1}{2}}\right)' =$$

$$\frac{1}{2}\left(x^{-\frac{1}{2}}\right) = \frac{1}{2\sqrt{x}} \qquad \text{if } 0 < x < 1$$

2. **Distribution of $\sqrt{\text{RND}}$ (square root of a random number)**

Density function:

$$f(x) = 2x \quad \text{if} \quad 0 < x < 1$$

Distribution function:

$$F(x) = x^2 \quad \text{if} \quad 0 < x < 1$$

Files to study point-clouds for $\sqrt{\text{RND}}$ (square root of a random number).

Demonstration file: Point-cloud for $\sqrt{RND}$
200-50-00

Demonstration file: Density of $\sqrt{RND}$
200-50-10

Proof. The possible values of $\sqrt{\text{RND}}$ constitute the interval $(0,1)$. Thus, in the following calculation, we assume that x satisfies $0 < x < 1$.

(a) Distribution function:

$$F(x) = \mathbf{P}(X \le x) = \mathbf{P}\left(\sqrt{\text{RND}} \le x\right) =$$

$$\mathbf{P}\left(\text{RND} \le x^2\right) = x^2 \qquad \text{if } 0 < x < 1$$

(b) Density function:

$$f(x) = F'(x) = \left(x^2\right)' = 2x \qquad \text{if } 0 < x < 1$$

3. **Distribution of RND^c $(c > 0)$**

Density function:

$$f(x) = \frac{1}{c}x^{\frac{1}{c}-1} \quad \text{if} \quad 0 < x < 1$$

Distribution function:

$$F(x) = x^{\frac{1}{c}} \quad \text{if} \quad 0 < x < 1$$

File to study point-clouds for RND^c $(c>0)$ (positive power of a random number).

Demonstration file: Point-cloud for RND^c $(c>0)$
200-51-00

Proof. The proof is left for the reader as an exercise.

4. **Distribution of $a\,\mathrm{RND}^c$ $(a>0, c>0)$**

Density function:

$$f(x)=\frac{1}{c}\left(\frac{x}{a}\right)^{\frac{1}{c}-1}\frac{1}{a} \quad \text{if} \quad 0<x<a$$

Distribution function:

$$F(x)=\left(\frac{x}{a}\right)^{\frac{1}{c}} \quad \text{if} \quad 0<x<a$$

File to study point-clouds for $a\mathrm{RND}^c$ $(a>0, c>0)$ (a positive constant times a positive power of a random number).

Demonstration file: Point-cloud for a RND^c $(a>0, c>0)$
200-51-50

Proof. The proof is left for the reader as an exercise.

5. **Distribution of $1/\mathrm{RND}$**

Density function:

$$f(x)=\frac{1}{x^2} \quad \text{if} \quad 1<x<\infty$$

Distribution function:

$$F(x)=1-\frac{1}{x} \quad \text{if} \quad 1<x<\infty$$

File to study point-clouds for $1/\mathrm{RND}$ (reciprocal of a random number).

Demonstration file: Point-cloud for $1/\mathrm{RND}$
200-52-00

Proof. The possible values of $1/\mathrm{RND}$ constitute the interval $(1,\infty)$. Thus, in the following calculation, we assume that x satisfies $0<x<\infty$.

(a) Distribution function:

$$F(x) = \mathbf{P}(X \le x) = \mathbf{P}\left(\frac{1}{\mathrm{RND}} \le x\right) = \mathbf{P}\left(\frac{1}{x} \le \mathrm{RND}\right) =$$

$$\mathbf{P}\left(\mathrm{RND} \ge \frac{1}{x}\right) = 1 - \frac{1}{x} \qquad (x > 1)$$

(b) Density function:

$$f(x) = F'(x) = \left(1 - \frac{1}{x}\right)' = \left(1 - x^{-1}\right)' = x^{-2} = \frac{1}{x^2} \qquad (x > 1)$$

6. **Distribution of $\mathbf{RND}^c$** $(c < 0)$

Density function:

$$f(x) = -\frac{1}{c} x^{\frac{1}{c}-1} \quad \text{if} \quad 1 < x < \infty$$

Distribution function:

$$F(x) = 1 - x^{\frac{1}{c}} \quad \text{if} \quad 1 < x < \infty$$

File to study point-clouds for RND^c $(c < 0)$ (negative power of a random number).

Demonstration file: Point-cloud for RND^c $(c < 0)$
200-53-00

Proof. The proof is left for the reader as an exercise.

7. **Distribution of $a\,\mathbf{RND}^c$** $(a > 0, c < 0)$

Density function:

$$f(x) = -\frac{1}{c}\left(\frac{x}{a}\right)^{\frac{1}{c}-1}\frac{1}{a} \quad \text{if} \quad a < x < \infty$$

Distribution function:

$$F(x) = 1 - \left(\frac{x}{a}\right)^{\frac{1}{c}} \quad \text{if} \quad a < x < \infty$$

Proof. The proof is left for the reader as an exercise.

File to study point-clouds for $a\mathrm{RND}^c$ $(a > 0, c < 0)$ (a positive constant times a negative power of a random number).

Demonstration file: Point-cloud for $aRND^c$ $(a > 0, c < 0)$
200-54-00

8. **Distribution of the product $\mathrm{RND}_1\mathrm{RND}_2$**

Density function:

$$f(x) = -\ln x \quad \text{if} \quad 0 < x < 1$$

Distribution function:

$$F(x) = x - x\ln x \quad \text{if} \quad 0 < x < 1$$

File to study point-clouds for $\mathrm{RND}_1\mathrm{RND}_2$, the product of two random numbers.

Demonstration file: Point-cloud for RND_1RND_2
200-55-00

Proof. The possible values of $\mathrm{RND}_1\mathrm{RND}_2$ constitute the interval $(0,1)$. Thus, in the following calculation, we assume that x satisfies $0 < x < 1$.

(a) Distribution function: Since the random point $(\mathrm{RND}_1, \mathrm{RND}_2)$ follows uniform distribution on the unit square, the probability of an event related to $(\mathrm{RND}_1, \mathrm{RND}_2)$ can be calculated as the area of the set corresponding to the event divided by the area of the unit square. However, the area of the unit square is 1, so the probability of an event related to $(\mathrm{RND}_1, \mathrm{RND}_2)$ is equal to the area of the set corresponding to the event.
The event $\{\mathrm{RND}_1\mathrm{RND}_2 \le x\}$ is the union of two exclusive events:

$$\{\mathrm{RND}_1\mathrm{RND}_2 \le x\}\} =$$

$$\{\mathrm{RND}_1 \le x\} \cup \{x < \mathrm{RND}_1 \ \text{ and } \ \mathrm{RND}_1\mathrm{RND}_2 \le x\}$$

Thus, we get that

$$F(x) = \mathbf{P}(X \le x) = \mathbf{P}(\mathrm{RND}_1\mathrm{RND}_2 \le x) =$$

$$\mathbf{P}(\mathrm{RND}_1 \le x) + \mathbf{P}(x < \mathrm{RND}_1 \ \text{ and } \ \mathrm{RND}_1\mathrm{RND}_2 \le x) =$$

$$\mathbf{P}(\mathrm{RND}_1 \le x) + \mathbf{P}(x \le \mathrm{RND}_1 \ \text{ and } \ \mathrm{RND}_2 \le x/\mathrm{RND}_1) =$$

The first term is equal to x. The second term is equal to

$$\text{area of } \{(u,v) : \ x \le u \le 1 \ \text{ and } \ 0 \le v \le x/u\} =$$

$$\int_{u=x}^{u=1} \left(\int_{v=0}^{v=x/u} 1 \, dv \right) du = \int_{u=x}^{u=1} x/u \ du = -x\ln x$$

The two terms together yield the given formula for $F(x)$.

(b) Density function:

$$f(x) = F'(x) = (x - x\ln x)' = 1 - \ln x - x\,\frac{1}{x} = -\ln x \qquad (0 < x < 1)$$

9. **Distribution of the ratio $\mathrm{RND}_2/\mathrm{RND}_1$**

Density function:

$$f(x) = \begin{cases} \frac{1}{2} & \text{if } 0 < x < 1 \\ \\ \frac{1}{2x^2} & \text{if } x > 1 \end{cases}$$

Distribution function:

$$F(x) = \begin{cases} \frac{x}{2} & \text{if } 0 < x < 1 \\ \\ 1 - \frac{1}{2x} & \text{if } x > 1 \end{cases}$$

Files to study point-clouds for $\mathrm{RND}_2/\mathrm{RND}_1$, the ratio of two random numbers.

Demonstration file: Point-cloud for RND_2/RND_1
200-56-00

Proof. The possible values of $\mathrm{RND}_2/\mathrm{RND}_1$ constitute the interval $(0,\infty)$. Thus, in the following calculation, we assume that x satisfies $0 < x < \infty$.

(a) Distribution function: The same way as in the previous problem the probability of an event related to $(\mathrm{RND}_1, \mathrm{RND}_2)$ is equal to the area of the set corresponding to the event.

$$F(x) = \mathbf{P}(X \le x) = \mathbf{P}(\mathrm{RND}_2/\mathrm{RND}_1 \le x) =$$

$$\mathbf{P}(\mathrm{RND}_2 \le x\,\mathrm{RND}_1)$$

If $x \le 1$, then this probability is equal to

$$\text{area of } \{(u,v):\ 0 \le u \le 1 \text{ and } 0 \le v \le xu\} =$$

This set is a right triangle with vertices $(0,0)$, $(1,0)$, $(1,x)$, so its area is equal to $x/2$. This why $F(x) = x/2$.
If $x > 1$, then it is advantageous to take the complement, and we get that the above probability is equal to

$$1 - \mathbf{P}(\mathrm{RND}_2 > x\,\mathrm{RND}_1) =$$

$$1 - \mathbf{P}(\mathrm{RND}_2/x > \mathrm{RND}_1) =$$

$$1 - \mathbf{P}(\mathrm{RND}_1 < \mathrm{RND}_2/x) =$$

$$1 - \text{area of } \{(u,v):\ 0 \le v \le 1 \text{ and } 0 \le u \le v/x\}$$

The set whose area appears here is a right triangle with vertices $(0,0)$, $(0,1)$, $(1/x,1)$, so its area is equal to $1/(2x)$. This why $F(x) = 1/(2x)$.

(b) Density function:

$$f(x) = F'(x) = \begin{cases} (\frac{x}{2})' = \frac{1}{2} & \text{if } \ 0 < x < 1 \\ (1 - \frac{1}{2x})' = \frac{1}{2x^2} & \text{if } \ x > 1 \end{cases}$$

Section 39
*** Arc-sine distribution

Applications: 1. Choose a point at random on the circumference of a circle with radius 1 according to uniform distribution, and project the point perpendicularly onto a fixed diameter. The point we get on the diameter has the arc-sine distribution.

2. Consider one of the Galilean moons of Jupiter (Io, Europa, Ganymede, Callisto), for example, Callisto. If you look at it with a telescope, then you will see the projection of its circular movement. So if your observation is performed at a random time instant, then the point you see will follow the arc-sine distribution.

Density function:

$$f(x) = \frac{1}{\pi}\frac{1}{\sqrt{1-x^2}} \quad \text{if} \quad -1 < x < 1$$

Distribution function:

$$F(x) = \frac{1}{2} + \frac{1}{\pi}\sin^{-1}(x) \quad \text{if} \quad -1 < x < 1$$

The proofs are left for the reader as an exercise.

The following file shows the arc-sine distribution.

Demonstration file: Arc-sine random variable as projection from a circle
200-09-00

Demonstration file: Arc-sine distribution, random variable and point-cloud
200-10-00

Section 40
*** Cauchy distribution

Applications: 1. Choose a point at random on the circumference of a circle with radius 1 according to uniform distribution, and project the point, from the center of the circle, onto a fixed tangent line. The point we get on the tangent line has the Cauchy distribution.

2. Imagine that a source of light is rotating in front of a long wall so that the trace of the light can be considered as a point on the wall. Observing this point at a random time-instant, the position of the point follows Cauchy distribution.

Density function:

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2} \quad \text{if} \quad -\infty < x < \infty$$

Distribution function:

$$F(x) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(x) \quad \text{if} \quad -\infty < x < \infty$$

The proofs are left for the reader as an exercise.

The following files show the Cauchy distribution.

Demonstration file: Cauchy random variable as projection from a circle
200-11-00

Demonstration file: Cauchy distribution, random variable and point-cloud
200-12-00

Section 41
*** Beta distributions

1. Special case: Beta distribution on the interval $[0;1]$

Applications: 1. If n people arrive between noon and 1pm independently of each other according to uniform distribution, and we are interested in the time instant when the kth person arrives, then this arrival time follows the beta distribution related to size n and rank k.

2. If we generate n independent random numbers by a computer, and we consider the kth smallest of the generated n values, then this random variable follows the beta distribution related to size n and rank k.

Density function:

$$f(x) = \frac{n!}{(k-1)!(n-k)!}\, x^{k-1}(1-x)^{n-k} \quad \text{if} \quad 0<x<1$$

Distribution function:

$$F(x) = \sum_{i=k}^{n} \binom{n}{i} x^i (1-x)^{n-i} \quad \text{if} \quad 0<x<1$$

or, equivalently,

$$F(x) = 1 - \sum_{i=0}^{k-1} \binom{n}{i} x^i (1-x)^{n-i} \quad \text{if} \quad 0<x<1$$

Parameters: k and n are positive integers so that $k \le n$. n can be called as the size, k as the rank of the distribution.

Remark. In order to remember the exponents in the formula of the density function, we mention that $k-1$ is the number of random numbers before the kth smallest, and $n-k$ is the number of random numbers after the kth smallest.

Proof of the formula of the density function. Let us generate n uniformly distributed independent random points between 0 and 1. Let X be the kth smallest among them. We calculate here the density function of the random variable X. Let $0 < x < 1$, and let $\Delta x = [x_1, x_2]$ be a small interval around x. By the meaning of the density function:

$$f(x) \approx \frac{\mathbf{P}(X \in \Delta x)}{x_2 - x_1}$$

The event $X \in \Delta x$, which stands in the numerator, means that the kth smallest point is in $[x_1, x_2)$, which means that

there is	at least one	point X	in	$[x_1, x_2)$,	and
there are	$k-1$	points	in	$[0, X)$,	and
there are	$n-k$	points	in	$[X, 1]$.	

This, with a very good approximation, means that

there are	$k-1$	points	in	$[0, x_1)$,	and
there is	1	point	in	$[x_1, x_2)$,	and
there are	$n-k$	points	in	$[x_2, 1]$.	

Using the formula of the poly-hyper-geometrical distribution, we get that the probability of the event $X \in \Delta x$ is approximately equal to

$$\frac{n!}{(k-1)!\ 1!\ (n-k)!}\ x_1^{k-1}(x_2 - x_1)^1(1-x_2)^{n-k}$$

Since $1! = 1$, we may omit some unnecessary factors and exponents, and the formula simplifies to

$$\frac{n!}{(k-1)!\ (n-k)!}\ x^{k-1}\ (x_2 - x_1)\ (1-x_2)^{n-k}$$

Dividing by $(x_2 - x_1)$, we get that the density function, for $0 < x < 1$, is

$$f(x) = \frac{n!}{(k-1)!\ (n-k)!}\ x^{k-1}\ (1-x)^{n-k} \quad \text{if } 0 < x < 1$$

Proof of the formulas of the distribution function. The proof of the first formula is based on the fact that the kth point is on the left side of x if and only if there are k or $k+1$ or $\dots$ n points on the left side of x. So we use the binomial distribution with parameters n and $p = x$, and summarize its kth, $(k+1)$th $\dots$ nth terms. The second formula follows from the complementary rule of probability.

Relations between the density and the distribution functions. Since the density function is equal to the derivative of the distribution function, and the distribution function is equal to the integral of the density function, we get the equalities:

$$\frac{d}{dx}\left(\sum_{i=k}^{n}\binom{n}{i}x^i(1-x)^{n-i}\right) = \frac{n!}{(k-1)!(n-k)!}\,x^{k-1}(1-x)^{n-k}$$

and

$$\int_0^x \frac{n!}{(k-1)!(n-k)!}\, x^{k-1}(1-x)^{n-k}\, dx = \sum_{i=k}^{n} \binom{n}{i} x^i (1-x)^{n-i}$$

The first equality can be derived by simple differentiation and then simplification of the terms which cancel out each other. The second can be derived by integration by parts.

Using Excel. In Excel, the function `BETADIST` (in Hungarian: `BÉTA.ELOSZLÁS`) is associated to the distribution function of the beta distribution:

$$F(x) = \texttt{BETADIST(x;k;n-k+1;0;1)}$$

We may omit the parameters 0 and 1, and write simply

$$F(x) = \texttt{BETADIST(x;k;n-k+1)}$$

There is no special Excel function for the density function, so if you need the density function, then - studying the mathematical formula of the density function - you yourself may construct it using the Excel functions `COMBIN` and `POWER` (in Hungarian: `KOMBINÁCIÓ` and `HATVÁNY`). In Excel, the inverse of the distribution function `BETADIST(x;k;n-k+1;0;1)` is `BETAINV(x;k;n-k+1;0;1)` (in Hungarian: `INVERZ.BÉTA(x;k;n-k+1;0;1)`).

The following files show beta distributions on the interval $[0;1]$.

Demonstration file: Beta distribution, n=5, k=2
200-13-00

Demonstration file: Beta distribution
200-14-00

2. General case: Beta distribution on the interval $[A;B]$

Applications:

1. If n people arrive between the time instants A and B independently of each other according to uniform distribution, and we are interested in the time instant when the kth person arrives, then this arrival time follows the beta distribution on the interval $[A,B]$ related to size n and rank k.

2. If we have n uniformly distributed, independent random values between A and B, and we consider the kth smallest of the generated n values, then this random variable follows the beta distribution on the interval $[A,B]$ related to size n and rank k.

Density function:

$$f(x) = \frac{1}{B-A}\, \frac{n!}{(k-1)!\,(n-k)!} \left(\frac{x-A}{B-A}\right)^{k-1} \left(\frac{B-x}{B-A}\right)^{n-k}$$

$$\text{if} \quad A < x < B$$

Distribution function:

$$F(x) = \sum_{i=k}^{n} \binom{n}{i} \left(\frac{x-A}{B-A}\right)^i \left(\frac{B-x}{B-A}\right)^{n-i} \quad \text{if} \quad A < x < B$$

or, equivalently

$$F(x) = 1 - \sum_{i=0}^{k-1} \binom{n}{i} \left(\frac{x-A}{B-A}\right)^i \left(\frac{B-x}{B-A}\right)^{n-i} \quad \text{if} \quad A < x < B$$

Parameters: k and n are positive integers so that $k \leq n$. A and B are real numbers so that $A < B$. n can be called **the size**, and k **the rank of the distribution**.

The proofs of the above formulas are similar to the special case $A = 0$, $B = 1$, and are left for the reader as an exercise.

Using Excel. In Excel, the function `BETADIST` (in Hungarian: `BÉTA.ELOSZLÁS`) is associated to the beta distribution. The distribution function of the beta distribution on the interval $[A,B]$ related to size n and rank k in Excel is:

$$F(x) = \texttt{BETADIST(x;k;n-k+1;A;B)} =$$

There is no special Excel function for the density function. If you need an Excel formula for the density function, then - studying the mathematical formula of the density function - you yourself may construct it using the Excel functions `COMBIN` and `POWER` (in Hungarian: `KOMBINÁCIÓ` and `HATVÁNY`). In Excel, the inverse of the distribution function `BETADIST(x;k;n-k+1;A;B)` is `BETAINV(x;k;n-k+1;A;B)` (in Hungarian: `INVERZ.BÉTA(x;k;n-k+1;A;B)`).

The following file shows beta distributions on the interval $[A;B]$.

Demonstration file: Beta distribution on (A;B)
200-15-00

Section 42
Exponential distribution

Notion of the memoryless property: We say that the random life-time X of an object has the memoryless property if

$$\mathbf{P}(\, X > a+b \,|\, X > a) = \mathbf{P}(\, X > b \,) \quad \text{for all positive } a \text{ and } b$$

The meaning of the memoryless property in words: if the object has already lived a units of time, then its chances to live more b units of time is equal to the chance of living b units of time for a brand-new object of this type after its birth. In other words: if the object is still living, then it has the same chances for its future as a brand-new one. We may also say that the memoryless property of an object means that the past of the object does not have an effect on the future of the object.

Example and counter-example for the the memoryless property:

1. The life-time of a mirror hanging on the wall has the memoryless property.

2. The life-time of a tire on a car dos not have the memoryless property.

Applications of the exponential distribution:

1. *The life-time of objects having the memoryless property have an exponential distribution.*

2. *Under certain circumstances waiting times follow an exponential distribution. (We learn about these circumstances in Chapter 20 entitled "Poisson process") For example, the amount of time until the next serious accident in a big city where the traffic has the same intensity day and night continuously follows an exponential distribution.*

The following file interprets the memoryless property of exponential distributions.

Demonstration file: Memoryless property visualized by point-clouds
200-17-00

Density function:

$$f(x) = \lambda\, e^{-\lambda x} \quad \text{if} \quad x \geq 0$$

Distribution function:

$$F(x) = 1 - e^{-\lambda x} \quad \text{if} \quad x \geq 0$$

Parameter: $\lambda > 0$

Remark. We will see later that the reciprocal of the parameter λ shows how much the theoretical average life-time (or the theoretical average waiting time) is.

Remark. In some real-life problems the memory-less property is only approximately fulfilled. In such cases, the the application of the exponential distribution is only an approximate model for the problem.

The following file shows exponential distributions.

Demonstration file: Exponential distribution, random variable and point-cloud 200-16-00

Proof of the formula of the distribution and density functions. The memoryless property

$$\mathbf{P}(\, X > a+b \,|\, X > a) = \mathbf{P}(\, X > b\,)$$

can be written like this:

$$\frac{\mathbf{P}(\, X > a+b\,)}{\mathbf{P}(\, X > a)} = \mathbf{P}(\, X > b\,)$$

Using the tail function $T(x) = \mathbf{P}(X > x)$, we may write the equation like this:

$$\frac{T(a+b)}{T(a)} = T(b)$$

that is

$$T(a+b) = T(a)T(b)$$

It is obvious that the exponential functions $T(x) = e^{cx}$ with an arbitrary constant c satisfy this equation. On the contrary, it can be shown that if a function is monotonous and satisfies this equation, then it must be an exponential functions of the form $T(x) = e^{cx}$. Since a tail function is monotonously decreasing, the memoryless property really implies that $T(x) = e^{cx}$ with a negative constant c, so we may write $c = -\lambda$, that is, $T(x) = e^{-\lambda x}$, which means that the distribution function is $F(x) = 1 - T(x) = 1 - e^{-\lambda x}$. Differentiating the distribution function, we get the density function: $f(x) = (F(x))' = \left(1 - e^{-\lambda x}\right)' = \lambda e^{-\lambda x}$.

Using Excel. In Excel, the function EXPONDIST (in Hungarian: EXP.ELOSZLÁS) is associated to this distribution. If the last parameter is FALSE, we get the density function of the exponential distribution:

$$f(x) = \lambda\, e^{-\lambda x} = \texttt{EXPONDIST(x;n;}\lambda\texttt{;FALSE)}$$

If the last parameter is TRUE, and the third parameter is the reciprocal of λ, we get the distribution function of the exponential distribution:

$$F(x) = 1 - e^{-\lambda x} = \texttt{EXPONDIST(x;n;}\lambda\texttt{;TRUE)}$$

You may use also the function EXP (in Hungarian: KITEVÕ) like this:

$$f(x) = \lambda\, e^{-\lambda x} = \lambda\ \texttt{EXP(-}\lambda\texttt{x)}$$

The distribution function then looks like as this:

$$F(x) = 1 - e^{-\lambda x} = 1 - \texttt{EXP(-}\lambda\texttt{x)}$$

Section 43
*** Gamma distribution

Application: In a big city where the traffic has the same intensity days and nights continuously, the amount of time until the nth serious accident follows a gamma distribution.

Density function:

$$f(x) = \frac{\lambda^n x^{n-1}}{(n-1)!}\, e^{-\lambda x} \quad \text{if} \quad x \geq 0$$

Distribution function:

$$F(x) = 1 - \sum_{i=0}^{k-1} \frac{(\lambda x)^i}{i!} e^{-\lambda x} \quad \text{if} \quad x \geq 0$$

Parameters: n is a positive integer, and $\lambda > 0$

Remark. The reciprocal of the parameter λ shows how much the theoretical average waiting time for the first accident is. Thus, the reciprocal of the parameter λ multiplied by n shows how much the theoretical average waiting time for the nth accident is.

The proof of the formulas of the density and distribution functions is omitted here. The formulas can be derived after having learned about Poisson-processes in Chapter 20.

The following files show gamma distributions.

Demonstration file: Gamma, n=3 distribution, random variable and point-cloud
200-25-00

Demonstration file: Gamma distribution, random variable and point-cloud
200-26-00

Demonstration file: Exponential point-cloud
200-19-00

Demonstration file: Gamma point-cloud of order 2
200-28-00

Demonstration file: Gamma point-cloud of order 3
200-29-00

Demonstration file: Gamma point-cloud of order 4
200-30-00

Demonstration file: Gamma point-cloud of order 5
200-31-00

Demonstration file: Gamma point-cloud of order k
200-32-00

Remark. If $n = 1$, then the gamma distribution reduces to the exponential distribution.

Using Excel. In Excel, the function `GAMMADIST` (in Hungarian: `GAMMA.ELOSZLÁS`) is associated to this distribution. If the last parameter is `FALSE` and the third parameter is the reciprocal of λ (unfortunately, in the `GAMMADIST` function of Excel, the third parameter should be the reciprocal of λ, and not λ), then we get the density function of the gamma distribution with parameters n and λ:

$$f(x) = \frac{\lambda^n x^{n-1}}{(n-1)!}\, e^{-\lambda x} = \texttt{GAMMADIST(x;n;}\frac{1}{\lambda}\texttt{;FALSE)}$$

If the last parameter is `TRUE`, and the third parameter is the reciprocal of λ, then we get the distribution function of the gamma distribution with parameters λ and λ:

$$F(x) = \int_0^x \frac{\lambda^n x^{n-1}}{(n-1)!}\, e^{-\lambda x}\, dx = \texttt{GAMMADIST(x;n;}\frac{1}{\lambda}\texttt{;TRUE)}$$

Using Excel. If $n = 1$, then the Excel function GAMMADIST returns the exponential distribution. This means that, with the `FALSE` option

$$\texttt{GAMMADIST(x;1;}\frac{1}{\lambda}\texttt{;FALSE)} = \lambda\, e^{-\lambda x}$$

is the exponential density function with parameter λ, and, with the `TRUE` option

$$\texttt{GAMMADIST(x;1;}\frac{1}{\lambda}\texttt{;TRUE)} = 1 - e^{-\lambda x}$$

is the exponential distribution function with parameter λ.

Section 44
Normal distributions

Application: If a random variable can be represented as the sum of many, independent, random quantities so that each has a small standard deviation, specifically, the sum of many, independent, small random quantities, then this random variable follows a normal distribution with some parameters μ and σ. Such random variable are, for example, the amount of electricity used by the inhabitants of a town, or the total income of a shop during a day.

1. Special case: Standard normal distribution

Density function:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \; e^{-\frac{x^2}{2}} \quad \text{if } -\infty < x < \infty$$

Distribution function:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \; e^{-\frac{x^2}{2}} \; dx \quad \text{if } -\infty < x < \infty$$

The following file shows standard normal distributions.

Demonstration file: Standard normal distribution, random variable and point-cloud 200-35-05

Remark. The usage of the Greek letters φ and Φ for the density and distribution functions of the standard normal distribution is so wide-spread as, for example, sin and cos for the sine- and cosine-functions. The symmetry of the density function about the origin φ implies the equality:

$$\int_{-\infty}^{-x} \frac{1}{\sqrt{2\pi}} \; e^{-\frac{x^2}{2}} \; dx = \int_{x}^{\infty} \frac{1}{\sqrt{2\pi}} \; e^{-\frac{x^2}{2}} \; dx$$

that is

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

Since $\Phi(x)$ is a strictly increasing function of x, its inverse exists. It is denoted by $\Phi^{-1}(y)$.

Using Excel. In Excel, the function NORMDIST (in Hungarian: NORM.ELOSZL) with the special parameter values 0 and 1 corresponds to this distribution. If the last parameter is FALSE, then we get the density function of the normal distribution:

$$\texttt{NORMDIST(x;0;1;FALSE)} = \varphi(x)$$

If the last parameter is TRUE, then we get the distribution function of the normal distribution:

$$\texttt{NORMDIST(x;0;1;TRUE)} = \Phi(x)$$

Since the function $\Phi(x)$ plays a very important role, there is a special simple Excel function for $\Phi(x)$, namely, NORMSDIST (in Hungarian: STNORMELOSZL), which has only one variable and no parameters. You may remember that the letter S stands (in Hungarian: the letters ST stand) for "standard":

$$\texttt{NORMSDIST(x)} = \Phi(x)$$

In Excel, the notation for the inverse of the function Φ is NORMSINV or NORMINV (in Hungarian: INVERZ.STNORM, or INVERZ.NORM) with the special parameter values $(\mu =)\ 0$ and $(\sigma =)\ 1$

$$\texttt{NORMSINV(y)} = \texttt{NORMINV(y;0;1)} = \Phi^{-1}(y)$$

The following file shows a standard normal point-cloud.

Demonstration file: Standard normal point-cloud
200-33-00

2. General case: Normal distribution with parameters μ and σ

The following file simulates the height of men as a normally distributed random variable.

Demonstration file: Height of men
200-64-00

Density function:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\ e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \quad \text{if } -\infty < x < \infty$$

Distribution function:

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}\ dx \quad \text{if } -\infty < x < \infty$$

Parameters: μ may be any real number, and σ may be any positive number.

Remark. We will see later that the parameters μ and σ are the expected value and the standard deviation, respectively. The normal distribution with parameters $\mu = 0$ and $\sigma = 1$ is the standard normal distribution.

The following files show normal distributions.

Demonstration file: Normal distribution, random variable and point-cloud
200-35-00

Demonstration file: Density function of the normal distribution
200-36-00

The distribution function of the normal distribution with parameters μ and σ can be expressed in terms of the standard normal distribution function:

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \qquad \text{if } -\infty < x < \infty$$

Here $\frac{x-\mu}{\sigma}$ is called the z**-value** associated with x, or the **standardized value** of x. The transformation $y = \frac{x-\mu}{\sigma}$ is called **standardization**.

If a random variable X follows a normal distribution with parameters μ and σ, then the following rules, called 1**-sigma rule**, 2**-sigma rule**, 3**-sigma rule**, are true:

$$\begin{array}{rcccl} \mathbf{P}(\mu-\sigma<X<\mu+\sigma) & \approx & 0.68 & = & 68\% \\ \mathbf{P}(\mu-2\sigma<X<\mu+2\sigma) & \approx & 0.95 & = & 95\% \\ \mathbf{P}(\mu-3\sigma<X<\mu+3\sigma) & \approx & 0.997 & = & 99.7\% \end{array}$$

These rules, in words, sound like this:

1**-sigma rule**: the value of a normally distributed random variable X falls in the interval $(\mu-\sigma,\mu+\sigma)$ with a probability 0.68.

2**-sigma rule**: the value of a normally distributed random variable X falls in the interval $(\mu-2\sigma,\mu+2\sigma)$ with a probability 0.95.

3**-sigma rule**: the value of a normally distributed random variable X falls in the interval $(\mu-3\sigma,\mu+3\sigma)$ with a probability 0.95.

Using Excel. In Excel, the function `NORMDIST` (in Hungarian: `NORM.ELOSZL`) corresponds to this distribution. If the last parameter is `FALSE`, then we get the density function of the normal distribution with parameters μ and σ:

$$\texttt{NORMDIST(x;}\mu\texttt{;}\sigma\texttt{;FALSE)} = f(x)$$

If the last parameter is `TRUE`, then we get the distribution function of the normal distribution with parameters μ and σ:

$$\texttt{NORMDIST(x;}\mu\texttt{;}\sigma\texttt{;TRUE)} = F(x)$$

In Excel, the notation for the inverse $F^{-1}(y)$ of the distribution function $F(x)$ of the normal distribution with parameters μ and σ is `NORMINV` (in Hungarian: `INVERZ.NORM`):

$$F^{-1}(y) = \texttt{NORMINV(y;}\mu\texttt{;}\sigma\texttt{)}$$

Section 45
*** Distributions derived from normal

In this chapter, some distributions which are derived from normal distributions are only visualized by Excel files. The reader can find the formulas of the density functions, the formulas of the distribution functions, applications of these distributions in many of the text-books on probability and statistics.

Demonstration file: Log-normal distribution
200-38-00

Demonstration file: Chi-square distribution, n=3
200-39-00

Demonstration file: Chi-square distribution
200-40-00

Demonstration file: Chi-distribution, n=3
200-41-00

Demonstration file: Chi-distribution
200-42-00

Demonstration file: Student-distribution (T-distribution) and random variable, n=3
200-43-00

Demonstration file: Student-distribution (T-distribution) and random variable
200-44-00

Demonstration file: F-distribution, m=3, n=4
200-45-00

Demonstration file: F-distribution
200-46-00

Section 46
***Generating a random variable with a given continuous distribution

It is important for a given continuous distribution that a random variable can be generated by a calculator or a computer so that its distribution is the given continuous distribution. The following is a method to define such a random variable.

Assume that a continuous distribution function $F(x)$ is given so that the function $F(x)$ is strictly increasing on an interval (A,B), and it is 0 on the left side of A, and it is 1 on the right side of B. If either $A=-\infty$ or $B=+\infty$, then we do not have to even mention them. The restriction of $F(x)$ onto the interval (A,B) has an inverse $F^{-1}(y)$ which is defined for all $y\in[0,1]$. The way how we find a formula for $F^{-1}(y)$ is that we solve the equation

$$y=F(x)$$

for x, that is, we express x from the equation in terms of y:

$$x=F^{-1}(y)$$

We may consider the random variable X defined by

$$X=F^{-1}(\text{RND})$$

It is easy to be convinced that the distribution function of the random variable X is the given function $F(x)$.

Some examples:

1. **Uniform random variable on** $(A;B)$

   Distribution function:

   $$y=F(x)=\frac{x-A}{B-A}\quad \text{if } A<x<B$$

   Inverse of the random variable function:

   $$x=F^{-1}(y)=A+(B-A)y\quad \text{if } 0<y<1$$

Simulation:

$$X = A + (B - A)\text{RND}$$

2. **Exponential random variable with parameter λ**

   Distribution function:

   $$y = F(x) = 1 - e^{-\lambda x} \quad \text{if } x \geq 0$$

   Inverse of the distribution function:

   $$x = F^{-1}(y) = -\frac{\ln(1-y)}{\lambda} \quad \text{if } 0 < y < 1$$

   Simulations:

   $$X = -\frac{\ln(1-\text{RND})}{\lambda}$$

   Obviously, the substraction from 1 can be omitted, and

   $$X = -\frac{\ln(\text{RND})}{\lambda}$$

   is also exponentially distributed.

   For some distributions, an explicit formula for the inverse of the distribution function is not available. In such cases, the simulation may be based on other relations. We give some examples for this case:

3. **Gamma random variable with parameters n and λ**

   Simulation:

   $$X = \sum_{i=1}^{n} X_i^0 = X_1^0 + X_2^0 + \text{ł...} + X_n^0$$

   where $X_1^0, X_2^0, \text{ł...}, X_n^0$ are independent, exponentially distributed random variables with parameter λ.

4. **Standard normal random variable**

   Simulations:

   $$X = \texttt{NORMSINV(RND)}$$

   $$X = \texttt{NORMINV(RND;0;1)}$$

   Accept that the following simulations are correct, and that they are more efficient:

   $$X = \left(\sum_{i=1}^{12} \text{RND}_i\right) - 6$$

   where $\text{RND}_1, \text{RND}_2, \text{ł...}, \text{RND}_{12}$ are independent random variables, uniformly distributed between 0 and 1.

   $$X = \sqrt{(2\ln(\text{RND}_1))}\,\cos(\text{RND}_2)$$

   where $\text{RND}_1, \text{RND}_2$ are independent random variables, uniformly distributed between 0 and 1.

5. **Normal random variable with parameters μ and σ**

   Simulations:

$$X = \texttt{NORMINV}(\text{RND}; \mu; \sigma)$$

$$X = \mu + \sigma X^0$$

   where X^0 is a standard normal random variable.

Section 47
Expected value of continuous distributions

In Chapter 15 of Part II. we learned that, by performing a large number of experiments for a discrete random variable, the average of the experimental results $X_1, X_2, \ldots, X_N$ stabilizes around the expected value of X:

$$\frac{X_1 + X_2 + \ldots + X_N}{N} \approx \mathbf{E}(X)$$

The same stabilization rule is true in case of a continuous random variable. In this chapter, we define the notion of the expected value for continuous distributions, and we list the formulas of the expected values of the most important continuous distributions.

The definition of the **expected value** for continuous distributions is:

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

Remark. We shall give here some motivation for the declared formula of the expected value. For this purpose, let us take a continuous random variable X, and let $X_1, X_2, \ldots, X_N$ be experimental results for X. We will show that the average of the experimental results is close to the above integral:

$$\frac{X_1 + X_2 + \ldots + X_N}{N} \approx \int_{-\infty}^{\infty} x f(x)\, dx$$

In order to show this, we choose the fixed points $\ldots, y_i, y_{i+1}, \ldots$ on the real line so that all the differences $\Delta y_i = y_{i+1} - y_i$ are small. Then we introduce a discrete random variable Y, so that the value of Y is derived from the value of X by rounding down to the closest y_i value which is on the left side of X, that is,

$$Y = y_i \quad \text{if and only if} \quad y_i \leq X < y_{i+1}$$

Applying the rounding operation to each experimental result, we get the values $Y_1, Y_2, \ldots, Y_N$. Since all the differences $\Delta y_i = y_{i+1} - y_i$ are small, we have that

$$\frac{X_1 + X_2 + \ldots + X_N}{N} \approx \frac{Y_1 + Y_2 + \ldots + Y_N}{N}$$

Obviously, Y is a discrete random variable with the possible values $\ldots, y_i, \ldots$, so that the probability of y_i is

$$p_i = \int_{y_i}^{y_{i+1}} f(x)\,dx \approx f(y_i)\Delta y_i$$

and thus, the expected value of Y is

$$\sum_i y_i\, p_i = \sum_i y_i \int_{yi}^{yi+1} f(x)\,dx \approx \sum_i y_i\, f(y_i)\Delta y_i \approx \int_{-\infty}^{\infty} x f(x)\,dx$$

We know that the average of the experimental results of a discrete random variable is close to the expected value, so

$$\frac{Y_1 + Y_2 + \ldots + Y_N}{N} \approx \sum_i y_i\, p_i$$

From all these approximations we get that

$$\frac{X_1 + X_2 + \ldots + X_N}{N} \approx \int_{-\infty}^{\infty} x f(x)\,dx$$

Remark. (It may happen that the expected value does not exist!) If the integral

$$\int_{\infty}^{\infty} x f(x)\,dx$$

is not absolutely convergent, that is

$$\int_{\infty}^{\infty} |x| f(x)\,dx = \infty$$

then one of the following 3 cases holds:

1. Either
$$\int_0^{\infty} x f(x)\,dx = \infty \text{ and } \int_{-\infty}^{0} x f(x)\,dx > -\infty$$
2. or
$$\int_0^{\infty} x f(x)\,dx < \infty \text{ and } \int_{-\infty}^{0} x f(x)\,dx = -\infty$$
3. or
$$\int_0^{\infty} x f(x)\,dx = \infty \text{ and } \int_{-\infty}^{0} x f(x)\,dx = -\infty$$

It can be shown that, in the first case, as N increases

$$\frac{X_1 + X_2 + \ldots + X_N}{N}$$

will become larger and larger, and it approaches ∞. This is why we may say that the expected exists, and its value is ∞. In the second case, as N increases,

$$\frac{X_1 + X_2 + \ldots + X_N}{N}$$

will become smaller and smaller, and it approaches $-\infty$. This is why we may say that the expected exists, and its value is $-\infty$. In the third case, as N increases,

$$\frac{X_1 + X_2 + \ldots + X_N}{N}$$

does not approach to any finite or infinite value. In this case we say that the expected value does not exist.

Here we give a list of the formulas of the expected values of the most important continuous distributions. The proofs are given after the list.

1. **Uniform distribution on an interval** $(A;B)$

$$\mathbf{E}(X) = \frac{A+B}{2}$$

2. **Arc-sine distribution**

$$\mathbf{E}(X) = 0$$

3. **Cauchy distribution**

   The expected value does not exist.

4. **Beta distribution related to size** n **and** k

$$\mathbf{E}(X) = \frac{k}{n+1}$$

5. **Exponential distribution with parameter** λ

$$\mathbf{E}(X) = \frac{1}{\lambda}$$

6. **Gamma distribution of order** n **with parameter** λ

$$\mathbf{E}(X) = \frac{n}{\lambda}$$

7. **Normal distribution with parameters** μ **and** σ

$$\mathbf{E}(X) = \mu$$

Proofs.

1. **Uniform distribution on an interval** $(A;B)$**.** Since the distribution is concentrated on a finite interval, the expected value exists. Since the density function is symmetrical about $\frac{A+B}{2}$, the expected value is

$$\mathbf{E}(X) = \frac{A+B}{2}$$

We may get this result by calculation, too:

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} xf(x)\,dx = \int_A^B x\,\frac{1}{B-A}\,dx =$$

$$\left[\frac{x^2}{2}\right]_A^B \frac{1}{B-A}\,dx = \left[\frac{B^2-A^2}{2}\right]\frac{1}{B-A}\,dx = \frac{A+B}{2}$$

2. **Arc-sine distribution.** Since the distribution is concentrated on the interval $(-1,1)$, the expected value exists. Since the density function is symmetrical about 0, the expected value is

$$\mathbf{E}(X) = 0$$

3. **Cauchy distribution.** Since the density function is symmetrical about 0, the 0 is a candidate for being the expected value. However, since

$$\int_0^{\infty} xf(x)\,dx = \int_0^{\infty} x\,\frac{1}{\pi}\frac{1}{1+x^2}\,dx = \left[\frac{1}{2\pi}\ln(1+x^2)\right]_0^{\infty} = \infty$$

and

$$\int_{-\infty}^{0} xf(x)\,dx = -\infty$$

the expected value does not exist.

4. **Beta distribution related to size** n **and** k

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

$$\int_0^1 x\;\frac{n!}{(k-1)!(n-k)!}\;x^{k-1}(1-x)^{n-k}\,dx =$$

$$\frac{k}{n+1}\;\int_0^1 \frac{(n+1)!}{k!(n-k)!}\;x^k(1-x)^{n-k}\,dx = \frac{k}{n+1}$$

In the last step, we used the fact that

$$\int_0^1 \frac{(n+1)!}{k!(n-k)!}\;x^k(1-x)^{n-k}\,dx = 1$$

which follows from the fact that

$$\frac{(n+1)!}{k!(n-k)!}\ x^k(1-x)^{n-k}$$

that is

$$\frac{(n+1)!}{k!(n-k)!}\ x^k(1-x)^{(n+1)-(k+1)}$$

is a the density function of the beta distribution related to size $n+1$ and $k+1$.

5. **Exponential distribution with parameter λ.** Using integration by parts with $u=x$, $v'=\lambda\ e^{-\lambda x}$, $u'=1$, $v=-e^{-\lambda x}$, we get that

$$\mathbf{E}(X)=\int_{-\infty}^{\infty} xf(x)\,dx=\int_0^{\infty} x\,\lambda\ e^{-\lambda x}\,dx=$$

$$\left[x\ \left(-e^{-\lambda x}\right)\right]_0^{\infty}-\int_0^{\infty}1\ \ \left(-e^{-\lambda x}\right)\,dx=$$

$$0+\int_0^{\infty}e^{-\lambda x}\,dx=\int_0^{\infty}e^{-\lambda x}\,dx=\left[\frac{e^{-\lambda x}}{-\lambda}\right]_0^{\infty}=\frac{1}{\lambda}$$

6. **Gamma distribution of order n with parameter λ**

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

$$\int_0^{\infty} x\,\frac{x^{n-1}\lambda^n}{(n-1)!}\,e^{-\lambda x}\,dx=$$

$$\frac{n}{\lambda}\ \int_0^{\infty}\frac{x^n\ \lambda^{n+1}}{n!}\,e^{-\lambda x}\,dx=\frac{n}{\lambda}$$

In the last step, we used the fact that

$$\int_0^{\infty}\frac{x^n\ \lambda^{n+1}}{n!}\,e^{-\lambda x}\,dx=1$$

This follows from the fact that

$$\frac{x^n\ \lambda^{n+1}}{n!}\,e^{-\lambda x}$$

is a density function of the gamma distribution of order $n+1$ with parameter λ.

7. **Normal distribution with parameters μ and σ.** Since the improper integrals

$$\int_0^{\infty} xf(x)\,dx=\int_0^{\infty} x\ \ \frac{1}{\sqrt{2\pi}}\ \ e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}\,dx$$

$$\int_{-\infty}^{0} xf(x)\,dx=\int_{-\infty}^{0} x\ \ \frac{1}{\sqrt{2\pi}}\ \ e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}\,dx$$

are obviously convergent, the expected value exists. Since the density function is symmetrical about μ, the expected value is

$$\mathbf{E}(X)=\mu$$

File to study the expected value of several continuous distributions.

Demonstration file: Continuous distributions, expected value, standard deviation
200-57-60

Minimal property of the expected value. If X is a continuous random variable with the density function $f(x)$, and c is a constant, then distance between X and c is $|X-c|$, the distance squared is $(X-c)^2$, the expected value of the squared distance is

$$\mathbf{E}\left((X-c)^2\right) = \int_{-\infty}^{\infty} (x-c)^2 f(x)\,dx$$

This integral is minimal if c is the expected value of X.

Proof. The value of the integral depends on c, so the integral defines a function:

$$h(c) = \int_{-\infty}^{\infty} (x-c)^2 f(x)\,dx$$

Expanding the square, we get:

$$h(c) = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \int_{-\infty}^{\infty} 2xc\, f(x)\,dx + \int_{-\infty}^{\infty} c^2 f(x)\,dx =$$

$$\int_{-\infty}^{\infty} x^2 f(x)\,dx - 2c \int_{-\infty}^{\infty} x f(x)\,dx + c^2 \int_{-\infty}^{\infty} 1 f(x)\,dx =$$

$$\int_{-\infty}^{\infty} x^2 f(x)\,dx - 2c\,\mathbf{E}(X) + c^2$$

Since the integral in the last line does not depend on c, differentiating with respect to c, we get that

$$h'(c) = -2\,\mathbf{E}(X) + 2c$$

Equating the derivative to 0, we get that the minimum occurs at $c = \mathbf{E}(X)$.

Section 48
Expected value of a function of a continuous random variable

We leared in Chapter 17 of Part II that if we make N experiments for a discrete random variable X, and we substitute the experimental results $X_1, X_2, \ldots, X_N$ into the function $y = t(x)$, and we consider the values $t(X_1), t(X_2), \ldots, t(X_N)$, then their average is is close to the expected value of $t(X)$:

$$\frac{t(X_1) + t(X_2) + \ldots + t(X_N)}{N} \approx \mathbf{E}\,(t(X))$$

The same stabilization rule is true in the case of a continuous random variable. Let X be a continuous random variable, and $t(x)$ a continuous function. The expected value of the random variable $t(X)$ is calculated by the integral:

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

Motivation of the declared formula. We give here some motivation of the declared formula of the expected value of $t(X)$. For this purpose, let us take a continuous random variable X, and a continuous function $t(x)$, and and let $X_1, X_2, \ldots, X_N$ be the experimental results for X. We will show that the average of the function values of the experimental results is close to the above integral:

$$\frac{t(X_1) + X_2 + \ldots + t(X_N)}{N} \approx \int_{-\infty}^{\infty} t(x) f(x)\, dx$$

In order to show this, we choose the fixed points $\ldots, y_i, y_{i+1}, \ldots$ on the real line so that all the differences $\Delta y_i = y_{i+1} - y_i$ are small. Then we introduce a discrete random variable, so that the value of Y is derived from the value of X by rounding down to the closest y_i value which is on the left side of X, that is,

$$Y = y_i \quad \text{if and only if} \quad y_i \leq X < y_{i+1}$$

Applying the rounding operation to each experimental result, we get the values

$$Y_1, Y_2, \ldots, Y_N$$

Since all the differences $\Delta y_i = y_{i+1} - y_i$ are small, and the function $t(x)$ is continuous, we have that

$$\frac{t(X_1)+t(X_2)+\ldots+t(X_N)}{N} \approx \frac{t(Y_1)+t(Y_2)+\ldots+t(Y_N)}{N}$$

Obviously, Y is a discrete random variable with the possible values $\ldots, y_i \ldots$, so that the probability of y_i is

$$p_i = \int_{y_i}^{y_{i+1}} f(x)\,dx \approx f(y_i)\Delta y_i$$

and thus, the expected value of $t(Y)$ is

$$\sum_i t(y_i)\,p_i = \sum_i t(y_i) \int_{yi}^{yi+1} f(x)\,dx \approx \sum_i t(y_i)\,f(y_i)\Delta y_i \approx \int_{-\infty}^{\infty} t(x)f(x)\,dx$$

We know that the average of the function values of the experimental results of a discrete random variable is close to its expected value, so

$$\frac{t(Y_1)+t(Y_2)+\ldots+t(Y_N)}{N} \approx \sum_i t(y_i)\,p_i$$

From all these approximations we get that

$$\frac{t(X_1)+t(X_2)+\ldots+t(X_N)}{N} \approx \int_{-\infty}^{\infty} t(x)f(x)\,dx$$

The expected value of X^n is called the n**th moment** of X:

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

specifically, the **second moment** of X is:

$$\mathbf{E}\left(X^2\right) = \int_{-\infty}^{\infty} x^2 f(x)\,dx$$

The expected value of $(X-c)^n$ is called the n**th moment** about a the point c:

$$\mathbf{E}\left((X-c)^n\right) = \int_{-\infty}^{\infty} (x-c)^n f(x)\,dx$$

specifically, the **second moment** about a point c is:

$$\mathbf{E}\left((X-c)^2\right) = \int_{-\infty}^{\infty} (x-c)^2 f(x)\,dx$$

Second moment of some continuous distributions:

1. **Uniform distribution on an interval** $(A;B)$

$$\mathbf{E}\left(X^2\right) = \frac{A^2+AB+B^2}{3}$$

2. **Exponential distribution**

$$\mathbf{E}\left(X^2\right) = \frac{2}{\lambda^2}$$

Proofs.

1.

$$\mathbf{E}(X^2) = \int_{-\infty}^{\infty} x^2 \, f(x) \, dx = \int_A^B x^2 \, \frac{1}{B-A} \, dx =$$

$$\frac{1}{B-A} \left[\frac{x^3}{3}\right]_A^B = \frac{1}{B-A} \, \frac{B^3-A^3}{3} = \frac{A^2+AB+B^2}{3}$$

2. Using integration by parts with $u = x^2$, $v' = \lambda \, e^{-\lambda x}$, $(u^2)' = 2u$, $v = -e^{-\lambda x}$, we get that

$$\mathbf{E}(X^2) = \int_{-\infty}^{\infty} x^2 \, f(x) \, dx = \int_0^{\infty} x^2 \;\; \lambda \, e^{-\lambda x} \, dx =$$

$$\left[x^2 \, \left(-e^{-\lambda x}\right)\right]_0^{\infty} - \int_0^{\infty} 2x \;\; \left(-e^{-\lambda x}\right) \, dx =$$

$$0 + 2 \int_0^{\infty} x \, e^{-\lambda x} \, dx =$$

$$2 \, \frac{1}{\lambda} \int_0^{\infty} x \;\; \lambda \, e^{-\lambda x} \, dx =$$

Here we recognize that the integral in the last line is the expected value of the λ-parametric exponential distribution, which is equal to $\frac{1}{\lambda}$, so we get

$$2 \, \frac{1}{\lambda} \, \frac{1}{\lambda} = \frac{2}{\lambda^2}$$

as it was stated.

File to study the expected value of several functions of RND.

Demonstration file: E (t(RND)), expected value of functions of a random number
200-58-00

Section 49
***Median

In this chapter, we learn about the notion of the median, which is a kind of a "center" of a data-set or of a distribution. In the next chapter, we will learn the notion of the expected value also for continuous random variables and distributions, which is a kind of "center", too, and then we will be able to compare them.

If a data-set consists of n numbers, then we may find
the smallest of these numbers, let us denote it by z_1^*,
the second smallest, let us denote it by z_2^*,
the third smallest, let us denote it by z_3^*,
and so on,
the kth smallest, let us denote it by z_k^*,
and so on,
the nth smallest, which is actually the largest, let us denote it by z_n^*.

Using Excel. In Excel, for a data-set, the function `SMALL` (in Hungarian: `KICSI`) can be used to find the kth smallest element in an `array`:

$$z_k^* = \texttt{SMALL(array;k)}$$

Now we may arrange the numbers $z_1, z_2, \ldots, z_n$ in the increasing order: $z_1^*, z_2^*, \ldots, z_n^*$. If the number n is odd, then there is a well defined center element in the list $z_1^*, z_2^*, \ldots, z_n^*$. This center element is called the **median of the data-set**. If n is even, then there are two center elements. In this case, the average of these two center elements is the **median of the data-set**.

Using Excel. In Excel, for a data-set, the function `MEDIAN` (in Hungarian: `MEDIÁN`) is used to calculate the median of a data-set:

`MEDIAN(array)`

The **median** of a continuous random variable or distribution is the value c for which it is true that both the probability of being less than c and the probability of being greater than c is equal to $\frac{1}{2}$:

$$P((-\infty, c)) = \frac{1}{2}$$

$$P((c,\infty)) = \frac{1}{2}$$

The median is the solution to the equation

$$F(x) = \frac{1}{2}$$

For a continuous distribution, this equation has a solution. If the inverse of $F(x)$ exists, and it is denoted by $F^{-1}(y)$, then

$$c = F^{-1}\left(\frac{1}{2}\right)$$

Using the density function, the median can be characterized obviously by the property

$$\int_{-\infty}^{c} f(x)\,dx = \frac{1}{2}$$

or, equivalently,

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

The notion of the median can be defined for discrete distributions, too, but the definition is a little bit more complicated. The **median** of a discrete random variable or distribution is the value c for which it is true that both the probability of being less than c at least $\frac{1}{2}$ and the probability of being greater than c at least $\frac{1}{2}$:

$$P((-\infty,c)) \geq \frac{1}{2}$$

$$P((-\infty,c)) \geq \frac{1}{2}$$

In a long sequence of experiments, the median of the experimental results for a random variable stabilizes around the median of the distribution of the random variable: if $X_1, X_2, \ldots, X_N$ are experimental results for a random variable X, and N is large, then the median of the dataset $X_1, X_2, \ldots, X_N$, the so called experimental median is close to the median of the distribution of the random variable.

Here is a file to study the notion of the median.

Demonstration file: Median of the exponential distribution
200-57-00

Minimal property of the median. If X is continuous random variable with the density function $f(x)$, and c is a constant, then the expected value of the distance between X and c is

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

This integral is minimal if c is the median.

Proof. Let us denote the value of the integral, which depends on c, by $h(c)$

$$h(c) = \int_{-\infty}^{\infty} |x-c|\, f(x)\, dx =$$

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

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

$$\int_{-\infty}^{c} x\, f(x)\, dx - c\int_{-\infty}^{c} 1\, f(x)\, dx + c\int_{c}^{\infty} 1\, f(x)\, dx - \int_{c}^{\infty} x\, f(x)\, dx$$

Let us take the derivative of each term with respect to c:

$$\left(\int_{-\infty}^{c} x\, f(x)\, dx\right)' = c\, f(c)$$

$$\left(-c\int_{-\infty}^{c} f(x)\, dx\right)' = -1\int_{-\infty}^{c} 1\, f(x)\, dx - c\, f(c) = -F(x) - c\, f(c)$$

$$\left(c\int_{c}^{\infty} 1\, f(x)\, dx\right)' = 1\int_{c}^{\infty} 1\, f(x)\, dx - c\, f(c) = 1 - F(x) - c\, f(c)$$

$$\left(-\int_{c}^{\infty} x\, f(x)\, dx\right)' = c\, f(c)$$

Now adding the 6 terms on the right sides, the terms $c\, f(c)$ cancel each other, and what we get is

$$h'(c) = 1 - 2F(c)$$

Since the equation

$$1 - 2F(c) = 0$$

is equivalent to the equation

$$F(c) = 1/2$$

and the solution to this equation is the median, we get that

$$h'(c) = 1 - 2F(c) = 0 \quad \text{if } c = \text{median}$$

$$h'(c) = 1 - 2F(c) < 0 \quad \text{if } c < \text{median}$$
$$h'(c) = 1 - 2F(c) > 0 \quad \text{if } c > \text{median}$$

which means that the minimum of $h(c)$ occurs if $c = \text{median}$.

Section 50
Standard deviation, etc.

Both the median and the average define a kind of "center" for a data-set, or for a distribution. It is important to have some characteristics to measure the deviation from the center for a data-set and for a distribution. In this chapter, we shall learn such characteristics.

If $z_1, z_2, \ldots, z_N$ is a data-set, consisting of numbers, then their average is a well-known characteristic of the data-set, which will be denoted by $\overline{z}_N$ or, for simplicity, by $\overline{z}$:

$$\overline{z}_N = \overline{z} = \frac{z_1 + z_2 + \ldots + z_N}{N}$$

The average shows where the center of the data-set is.

It is important for us to know how far the data are from the average. This is why we consider the distance (that is, the absolute value of the difference) between the data elements and the average:

$$|z_1 - \overline{z}| \ , \ \ |z_2 - \overline{z}| \ , \ \ \ldots \ \ , \ |z_N - \overline{z}|$$

The average of these distances is a characteristic of how far the data elements, in the average, are from their average:

$$\frac{|z_1 - \overline{z}| + |z_2 - \overline{z}| + \ldots + |z_N - \overline{z}|}{N}$$

This quantity is called the **average distance from the average**.

Using Excel. In Excel, for a data-set, the function `AVEDEV` (in Hungarian: `ÁTL.ELTÉRÉS`) calculates the average distance from the average.

If, before taking the average, instead of the absolute value, we take the square of each difference, we get another characteristic of the data-set, the average squared distance from the average, which is called the **variance** of the data-set:

$$\frac{(z_1 - \overline{z})^2 + (z_2 - \overline{z})^2 + \ldots + (z_N - \overline{z})^2}{N}$$

The square root of the variance is called the **standard deviation** of the data-set:

$$\sqrt{\frac{(z_1 - \overline{z})^2 + (z_2 - \overline{z})^2 + \ldots + (z_N - \overline{z})^2}{N}}$$

Using a calculator. Most calculators have a key to determine not only the average of a data-set, but the average distance from the average, the variance and the standard deviation, as well.

Using Excel. In Excel, for a data-set, the function `VARP` (in Hungarian: `VARP`, too) calculates the variance, and the function `STDEVP` (in Hungarian: `SZÓRÁSP`) calculates the standard deviation.

Here are some files to study the notion of the standard deviation for a data-set and for a random variable.

Demonstration file: Standard deviation for a data-set
200-59-00

Demonstration file: Standard deviation for a random variable
200-60-00

Sample variance and sample standard deviation in Excel. In Excel, the functions `VAR` (in Hungarian: `VAR`, too) and `STDEV` (in Hungarian: `SZÓRÁS`) calculate the so called **sample variance** and **sample standard deviation**. The sample variance and sample standard deviation are defined almost the same way as the variance and standard deviation, but the denominator is $N-1$ instead of N:

$$\frac{(z_1-\overline{z})^2+(z_2-\overline{z})^2+\ldots+(z_N-\overline{z})^2}{N-1}$$

$$\sqrt{\frac{(z_1-\overline{z})^2+(z_2-\overline{z})^2+\ldots+(z_N-\overline{z})^2}{N-1}}$$

The advantage of taking $N-1$ instead of N becomes clear in statistics. We will not use the functions `VAR` and `STDEV`.

Recall that if we make a large number of experiments for a random variable X, then the average of the experimental results, in most cases, stabilizes around a non-random value, the expected value of the random variable, which we denote by μ:

$$\overline{X}=\frac{X_1+X_2++X_N}{N}\approx$$

$$\mu=\begin{cases}\sum\limits_x xp(x) & \text{in the discrete case}\\ \int_{-\infty}^{\infty} x\,p(x) & \text{in the continuous case}\end{cases}$$

The average distance from the average of the experimental results, long sequence of experiments, also stabilizes around a non-random value, which we call the **average distance from the average** of the random variable or of the distribution, which we denote by d:

$$\frac{|X_1-\overline{X}|+|X_2-\overline{X}|+\ldots+|X_N-\overline{X}|}{N}\approx$$

$$\frac{|X_1-\mu|+|X_2-\mu|+\ldots+|X_N-\mu|}{N} \approx$$

$$d=\begin{cases} \sum\limits_x |x-\mu|\ p(x) & \text{in the discrete case} \\ \int_{-\infty}^{\infty} |x-\mu|\ f(x)\,dx & \text{in the continuous case} \end{cases}$$

Here is a file to study the notion of the average deviation from the average of a discrete distribution.

Demonstration file: Calculating the average deviation from the average of a discrete distribution
170-02-00

The variance of the experimental results, in a long sequence of experiments, also stabilizes around a non-random value, which we call the **variance** of the random variable or of the distribution, which we denote by σ^2:

$$\frac{(X_1-\overline{X})^2+(X_2-\overline{X})^2+\ldots+(X_N-\overline{X})^2}{N} \approx$$

$$\frac{(X_1-\mu)^2+(X_2-\mu)^2+\ldots+(X_N-\mu)^2}{N} \approx$$

$$\mathbf{VAR}(X)=\sigma^2=\begin{cases} \sum\limits_x (x-\mu)^2\ p(x) & \text{in the discrete case} \\ \int_{-\infty}^{\infty} (x-\mu)^2\ f(x)\,dx & \text{in the continuous case} \end{cases}$$

The standard deviation of the experimental results, which is the square root of the variance, in a long sequence of experiments, obviously stabilizes around the square root of σ^2, that is, σ, which we call the **standard deviation** of the random variable or of the distribution, which we denote by σ:

$$\sqrt{\frac{(X_1-\overline{X})^2+(X_2-\overline{X})^2+\ldots+(X_N-\overline{X})^2}{N}} \approx$$

$$\sqrt{\frac{(X_1-\mu)^2+(X_2-\mu)^2+\ldots+(X_N-\mu)^2}{N}} \approx$$

$$\mathbf{SD}(X)=\sigma=\begin{cases} \sqrt{\sum\limits_x (x-\mu)^2\ p(x)} & \text{in the discrete case} \\ \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2\ f(x)\,dx} & \text{in the continuous case} \end{cases}$$

These non-random values are characteristics of the random variable and of the distribution. Among these three characteristics the variance and the standard deviation play a much more important theoretical and practical role than the average distance from the average.

Mechanical meaning of the variance. The mechanical meaning of the variance is the inertia about the center, because it is calculated by the same formula as the variance:

$$\begin{cases} \sum\limits_x (x-\mu)^2\ p(x) & \text{in the discrete case} \\ \int_{-\infty}^{\infty} (x-\mu)^2\ f(x)\,dx & \text{in the continuous case} \end{cases}$$

Remark. It may seem a little bit strange that the notion of the variance and the standard deviation play a more important role than the notion of the average distance from the average. The reason is that the variance and the standard deviation satisfy a rule which is very important both for the theory and the practice. Namely, it is true that the variance of the sum of independent random variables equals the sum of the variances of the random variables, or equivalently the standard deviation of the sum of independent random variables equals to the sum of the squares of the standard deviations of the random variables. Such a general rule does not hold for the average distance from the average.

The variance is very often calculated on the basis of the following relation.

The variance equals the second moment minus the expected value squared:

$$\sum_x (x-\mu)^2\ p(x) = \sum_x x^2\ p(x) - \mu^2$$

in the discrete case, and

$$\int_{-\infty}^{\infty} (x-\mu)^2\ f(x)\ dx = \int_{-\infty}^{\infty} x^2\ f(x)\ dx - \mu^2$$

in the continuous case.

The proof of these relations is quite simple. In the continuous case:

$$\int_{-\infty}^{\infty} (x-\mu)^2\ f(x)\ dx =$$

$$\int_{-\infty}^{\infty} \left(x^2 - 2x\mu + \mu^2\right)\ f(x)\ dx =$$

$$\int_{-\infty}^{\infty} x^2\ f(x)\ dx - \int_{-\infty}^{\infty} 2x\mu\ f(x)\ dx + \int_{-\infty}^{\infty} \mu^2\ f(x)\ dx =$$

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

$$\int_{-\infty}^{\infty} x^2\ f(x)\ dx - 2\,\mu\ \mu + \mu^2\ 1 =$$

$$\int_{-\infty}^{\infty} x^2\ f(x)\ dx - \mu^2$$

In the discrete case, the integral is replaced by summation, $f(x)$ is replaced by $p(x)$.

Using Excel. In Excel, the variance of a discrete distribution given by numerical values can be calculated like this: if the distribution is arranged in a table-form so that the x values constitute

$\texttt{array}_1$ (a row or a column) and the associated $p(x)$ values constitute $\texttt{array}_2$ (another row or column) then we may calculate the expected value μ by the

$$\mu = \texttt{SUMPRODUCT(array}_1\texttt{;array}_2\texttt{)}$$

command, and then we may calculate $(x-\mu)^2$ for each x, and arrange these squared distances into $\texttt{array}_3$. Then the variance is

$$\sigma^2 = \texttt{SUMPRODUCT(array}_3\texttt{;array}_2\texttt{)}$$

and the standard deviation is

$$\sigma = \texttt{SQRT(SUMPRODUCT(array}_3\texttt{;array}_2\texttt{))}$$

Here are some files which show how the standard deviation, the variance, and the average deviation from the average of a discrete distribution can be calculated in Excel.

Demonstration file: Calculating - with Excel - the variance and the standard deviation of a discrete distribution
200-63-00

Demonstration file: Calculating - with Excel - the average deviation from the average of a discrete distribution
200-62-00

Steiner's equality. The second moment of a distribution about a point c is equal to the variance plus the difference between the expected value and c squared:

$$\sum_x (x-c)^2\ p(x) = \sum_x (x-\mu)^2\ p(x) + (\mu-c)^2 = \sigma^2 + (\mu-c)^2$$

in the discrete case, and

$$\int_{-\infty}^{\infty} (x-c)^2\ f(x)\ dx = \int_{-\infty}^{\infty} (x-\mu)^2\ f(x)\ dx + (\mu-c)^2 = \sigma^2 + (\mu-c)^2$$

in the continuous case.

Steiner's inequality. The second moment of a distribution about any point c is greater than the variance, and equality holds only if $c = \mu$. In other words, the second moment of a distribution about any point c is minimal, if $c = \mu$, and the minimal value is σ^2:

$$\sum_x (x-c)^2\ p(x) \geq \sigma^2$$

in the discrete case, and

$$\int_{-\infty}^{\infty} (x-c)^2\ f(x)\ dx \geq \sigma^2$$

in the continuous case. Equality holds if and only if $c = \mu$.

The proof of Steiner's equality, for the continuous case:

$$\int_{-\infty}^{\infty}(x-c)^2\,f(x)\,dx=$$

$$\int_{-\infty}^{\infty}\big((x-\mu)+(\mu-c)\big)^2\,f(x)\,dx=$$

$$\int_{-\infty}^{\infty}\big((x-\mu)^2+2(x-\mu)(\mu-c)+(\mu-c)^2\big)\,f(x)\,dx=$$

$$\int_{-\infty}^{\infty}(x-\mu)^2\,f(x)\,dx+\int_{-\infty}^{\infty}2(x-\mu)(\mu-c)f(x)\,dx+\int_{-\infty}^{\infty}(\mu-c)^2f(x)\,dx=$$

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

$$\sigma^2+2(\mu-c)\,0+(\mu-c)^2\,1=$$

$$\sigma^2+0+(\mu-c)^2=$$

$$\sigma^2+(\mu-c)^2$$

In the above argument, we used the fact that

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

$$\int_{-\infty}^{\infty}xf(x)\,dx-\int_{-\infty}^{\infty}\mu f(x)\,dx=$$

$$\int_{-\infty}^{\infty}xf(x)\,dx-\mu\int_{-\infty}^{\infty}f(x)\,dx=$$

$$\mu-\mu\,1=0$$

In the discrete case, the integral is replaced by summation, $f(x)$ is replaced by $p(x)$. The Steiner's inequality is an obvious consequence of the Steiner's equality.

Steiner's equality in mechanics. Steiner's equality in mechanics is well-known: the inertia about a point c is equal to the inertia about the center of mass plus inertia about the point c as if the total amount of mass were in the center of mass.

Steiner's inequality in mechanics. Steiner's inequality in mechanics is well-known: the inertia about a point c which is different from the center of mass is greater than the inertia about the center of mass.

Variance and standard deviation of some distributions:

1. **Binomial distribution.**

   The second moment of the binomial distribution is

   $$\mathbf{E}\left(X^2\right) = n^2p^2 - np^2 + np$$

   The expected value of the binomial distribution is

   $$\mathbf{E}\left(X\right) = np$$

   So, the variance is

   $$\mathbf{VAR}(X) = \left(n^2p^2 - np^2 + np\right) - (np)^2 = np - np^2 = np(1-p)$$

   Thus, the standard deviation of the binomial distribution is

   $$\mathbf{SD} = \sqrt{np(1-p)}$$

2. **Uniform distribution**

   The second moment of the uniform distribution on an interval $(A;B)$ is

   $$\frac{A^2 + AB + B^2}{3}$$

   The expected value of the uniform distribution is

   $$\frac{A+B}{2}$$

   So, the variance is

   $$\mathbf{VAR}(X) = \left(\frac{A^2 + AB + B^2}{3}\right) - \left(\frac{A+B}{2}\right)^2 = \frac{(B-A)^2}{12}$$

   Thus, the standard deviation of the uniform distribution is

   $$\mathbf{SD} = \frac{(B-A)}{\sqrt{12}}$$

   Files to study the expected value, the standard deviation and the second moment of uniform distributions:

   Demonstration file: RND, expected value, standard deviation, second moment
   200-58-40

   Demonstration file: Uniform distribution on $(A;B)$, expected value, standard deviation, second moment
   200-58-60

3. **Exponential distribution**

   The second moment of the exponential distribution is

   $$\mathbf{E}\left(X^2\right) = \frac{2}{\lambda^2}$$

   The expected value of the exponential distribution is

   $$\mathbf{E}\left(X\right) = \frac{1}{\lambda}$$

   So, the variance is

   $$\mathbf{VAR}(X) = \left(\frac{2}{\lambda^2}\right) - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

   Thus, the standard deviation of the exponential distribution is

   $$\mathbf{SD} = \frac{1}{\lambda}$$

4. **Normal distributions**

   We know that the expected value of the normal distribution with parameters μ and σ is μ. The variance of the normal distribution with parameters μ and σ is

   $$\mathbf{VAR}(X) = \sigma^2$$

   Thus, the standard deviation of the normal distribution is

   $$\mathbf{SD} = \sigma$$

   This is why the normal distribution with parameters μ and σ is also called the normal distribution with expected value μ and standard deviation σ.

Section 51
*** Poisson-processes

In this chapter we study special processes. The main goal is to discover that, under certain circumstances, waiting times in such processes follow exponential and gamma distributions.

In a city, unfortunately accidents occur during each day. It may happen that no accident occurs during a day, but on other days, their number may be large. The number of accidents during a day is a random variable following Poisson distribution. We may be interested not only in the total number of accidents during a day, but we may look at the process in a more detailed form: we may be interested in each of the time instants when an accident occurs. These time instants constitute a set of random points on the real line, so we refer to the process as a **(random) point process**. We emphasize that in a point process not only the number of the points but their position on the real line is random.

If we choose a time-interval $(a;b)$, then the number of accidents, that is, the number of random points in that interval is a random variable following a Poisson distribution. The parameter, that is, the expected value of the number of points in that interval obviously depends on a and b. If the interval $(a;b)$ is longer, then the expected value of the number of points is larger. If the expected value of the number of points on any interval is proportional to the length of the interval, then the process is called a **homogeneous Poisson process**, otherwise it is called non-homogeneous.

For any Poisson process, the expected value of the number of points in the interval $(0;x)$ defines the **expected value function:**

$$\Lambda(x) = \text{expected value of the number of points in } (0;x)$$

The derivative of $\Lambda(x)$, denoted by $\lambda(x)$, for an arbitrary interval $(a;b)$ obviously satisfies that

$$\int_a^b \lambda(x)\,dx = \Lambda(b) - \Lambda(a) = \text{expected value of the number of points in } (a;b)$$

$\lambda(x)$ is called the **intensity function**. The expected value function of a homogeneous Poisson process is a linear function:

$$\Lambda(x) = \lambda x \quad \text{if} \quad x \geq 0 \quad \text{(linear function)}$$

the intensity function of a homogeneous Poisson process is a constant:

$$\lambda(x) = \lambda \quad \text{(constant)}$$

In a point process, we may study the occurrence of the first point after a given point. The same way, we may study the occurrence of the second point, or the occurrence of the third point after a given point, and so on. If the given point is 0, then the distribution function of the first occurrence is

$$F(x) = \mathbf{P}(\text{ first occurrence } < x) = \mathbf{P}(\text{ at least one point in } (0,x))$$

$$= 1 - \mathbf{P}(\text{ no points in } (0,x)) = 1 - e^{-\Lambda(x)} \quad \text{if} \quad x \geq 0$$

Taking the derivative, we get the density function:

$$f(x) = \lambda(x)\, e^{-\Lambda(x)} \quad \text{if} \quad x \geq 0$$

When the process is a homogeneous Poisson process, then these formulas simplify to

$$F(x) = 1 - e^{-\lambda x} \quad \text{if} \quad x \geq 0$$

and the density function is:

$$f(x) = \lambda\, e^{-\lambda x} \quad \text{if} \quad x \geq 0$$

showing that the first occurrence after 0, that is, the waiting time for the first accident follows an exponential distribution with parameter λ. It can be shown that, in a homogeneous Poisson process, the second accident after 0, that is, the waiting time for the second accident follows a second order gamma distribution with parameter λ, and the third accident after 0, that is, the waiting time for the third accident follows a third order gamma distribution with parameter λ, and so on.

The following files simulate Poisson processes with different intensity functions. First, second and third occurrences are observed in them.

Demonstration file: First (second and third) occurrence, homogeneous Poisson-process
200-21-00

Demonstration file: First (second and third) occurrence, "trapezoid shaped" intensity function
200-22-00

Demonstration file: First (second and third) occurrence, linearly increasing intensity function
200-23-00

Demonstration file: First (second and third) occurrence, decreasing intensity function
200-24-00

Section 52
***Transformation from line to line

Assume that the distribution function of a continuous random variable X is $F(x)$, and $y = t(x)$ is a function having a continuously differentiable inverse $x = t^{-1}(y)$. If we plug the random X value into the function $t(x)$, we get a new random variable: $Y = t(X)$. We may want to know both the distribution function $G(y)$ and the density function $g(y)$ of this new random variable.

If $y = t(x)$ is an increasing function, then

$$G(y) = F\left(t^{-1}(y)\right)$$

$$g(y) = f\left(t^{-1}(y)\right)\ \left(t^{-1}(y)\right)'$$

Sketch of the first proof.

$$G(y) = \mathbf{P}(Y < y) = \mathbf{P}\left(t(X) < y\right) = \mathbf{P}\left(X < t^{-1}(y)\right) = F\left(t^{-1}(y)\right)$$

Taking the derivative with respect to y on both sides of the equation $G(y) = F\left(t^{-1}(y)\right)$, we get $g(y)$ on the left side, and using the chain-rule , we get $f\left(t^{-1}(y)\right)\ \left(t^{-1}(y)\right)'$ on the right side.

Sketch of the second proof. The event $y < Y < y + \Delta y$ is equivalent to the event $x < X < x + \Delta x$. Thus, $\mathbf{P}(y < Y < y + \Delta y) = \mathbf{P}(x < X < x + \Delta x$. Using this fact, we get

$$g(y) \approx \frac{\mathbf{P}(y < Y < y + \Delta y)}{\Delta y} = \frac{\mathbf{P}(x < X < x + \Delta x)}{\Delta y} =$$

$$\frac{\mathbf{P}(x < X < x + \Delta x)}{\Delta x}\ \frac{\Delta x}{\Delta y} \approx f(x)\ \left(t^{-1}(y)\right)' = f\left(t^{-1}(y)\right)\ \left(t^{-1}(y)\right)'$$

If $y = t(x)$ is a decreasing function, then

$$G(y) = 1 - F\left(t^{-1}(y)\right)$$

$$g(y) = -f\left(t^{-1}(y)\right)\ \left(t^{-1}(y)\right)'$$

or

$$g(y) = f\left(t^{-1}(y)\right) \left|\left(t^{-1}(y)\right)'\right|$$

Sketch of proof.

$$G(y) = \mathbf{P}(Y < y) = \mathbf{P}\left(t(X) < y\right) = \mathbf{P}\left(X > t^{-1}(y)\right) =$$

$$1 - \mathbf{P}\left(X < t^{-1}(y)\right) = 1 - F\left(t^{-1}(y)\right)$$

Taking the derivative with respect to y on both sides of $G(y) = 1 - F\left(t^{-1}(y)\right)$, we get $g(y) = -f\left(t^{-1}(y)\right)\ \left(t^{-1}(y)\right)'$ on the right side. Since $x = t^{-1}(y)$ is a decreasing function, $\left(t^{-1}(y)\right)' \leq 0$, and $-\left(t^{-1}(y)\right)' = \left|\left(t^{-1}(y)\right)'\right|$.

Obviously, the formula

$$g(y) = f\left(t^{-1}(y)\right) \left|\left(t^{-1}(y)\right)'\right|$$

is applicable both for the increasing and decreasing case, as well.

Linear transformations. If a continuous distribution is transformed by an increasing linear transformation

$$y = ax + b \quad (a > 0)$$

then

$$G(y) = F\left(\frac{y}{a}\right)$$

$$g(y) = f\left(\frac{y}{a}\right)\ \frac{1}{a}$$

If a continuous distribution is transformed by a decreasing linear transformation

$$y = ax + b \quad (a < 0)$$

then

$$G(y) = 1 - F\left(\frac{y}{a}\right)$$

$$g(y) = -f\left(\frac{y}{a}\right)\ \frac{1}{a}$$

The formula

$$g(y) = f\left(\frac{y}{a}\right)\ \frac{1}{|a|}$$

is applicable in both cases.

Linear transformation of normal distributions. If a normal distribution is transformed by a linear transformation

$$y = ax + b$$

then the new distribution is a normal distribution, too. If the expected value of the old distribution is μ_{old}, then the expected value of the new distribution is

$$\mu_{\text{new}} = a\,\mu_{\text{old}} + b$$

If the variance of the old distribution is σ^2_{old}, then the variance of the new distribution is

$$\sigma^2_{\text{new}} = \sigma^2_{\text{old}}\,a^2$$

If the standard deviation of the old distribution is σ_{old}, then the standard deviation of the new distribution is

$$\sigma_{\text{new}} = \sigma_{\text{old}}\,|a|$$

Here are some files to study transformations from line to line.

Demonstration file: Uniform distribution transformed by a power function with a positive exponent
200-95-00

Demonstration file: Uniform distribution transformed by a power function with a negative exponent
200-96-00

Demonstration file: Uniform distribution on (A,B) transformed by a power function
300-01-00

Demonstration file: Exponential distribution transformed by a power function with a positive exponent
200-97-00

Demonstration file: Exponential distribution transformed by a power function with a negative exponent
200-98-00

Demonstration file: Exponential distribution transformed by a power function
200-99-00

Part - IV.

Two-dimensional continous distributions

Section 53
Two-dimensional random variables and distributions

In this chapter, we start to work with two-dimensional continuous random variables and distributions. When two random variables, say X and Y are considered, then we may put them together to get a pair of random numbers, that is, a random point (X,Y) in the two-dimensional space. Some examples:

1. Let us choose a Hungarian man, and let

$$X = \text{ his height}$$

$$Y = \text{ his weight}$$

   Then $(X,Y) =$ (height,weight) is a two-dimensional random variable.

2. Let us generate three independent random numbers, and let

$$X = \text{ the smallest of them}$$

$$Y = \text{ the largest of them}$$

   Then $(X,Y) =$ (smallest,biggest) is a two-dimensional random variable.

Density function. Two-dimensional continuous random variables are described mainly by their density function $f(x,y)$, which integrated on a set A gives the probability of the event that the value of (X,Y) is in the set A:

$$\mathbf{P}(A) = \mathbf{P}\left((X,Y) \in A\right) = \iint\limits_A f(x,y)\,dx\,dy$$

The characteristic properties of two-dimensional density functions are:

$$f(x,y) \geq 0$$

$$\iint\limits_{R^2} f(x,y)\,dx\,dy = 1$$

These two properties are characteristic for two-dimensional density functions, because, on one side, they are true for two-dimensional density functions of any continuous random variables, and on the other side, if a function $f(x,y)$ is given which has these two properties, then it is possible to define a two-dimensional random variable (X,Y) so that its density function is the given function $f(x,y)$.

The density function as a constant of approximate proportionality. If A is a small set around a point (x,y), then

$$\mathbf{P}\left((X,Y)\in A\right) = \iint\limits_{A} f(x,y)\,dx\,dy \approx f(x,y)\ \times\ \text{area of } A$$

and

$$f(x,y) \approx \frac{\mathbf{P}\left((X,Y)\in A\right)}{\text{area of } A}$$

We emphasize that the value $f(x,y)$ of the density function does not represent any probability value. If (x,y) is a fixed point, then $f(x,y)$ may be interpreted as a constant of approximate proportionality: if A is a small set around a point (x,y), then the probability that the point (X,Y) is in A is approximately equal to $f(x,y)\times$ area of A:

$$\mathbf{P}\left((X,Y)\in A\right) \approx f(x,y)\ \times\ \text{area of } A$$

Approximating the density function. If A is a small rectangle with sides of lengths Δx and Δy, then we get that

$$f(x,y) \approx \frac{\mathbf{P}\left(x < X < x+\Delta x \text{ and } y < Y < y+\Delta y\right)}{\Delta x\,\Delta y}$$

This formula is useful to determine the density function in some problems.

Conditional probability. If A and B are subsets of the plane, then both $(X,Y)\in A$ and $(X,Y)\in B$ defines an event. The conditional probability of the event $(X,Y)\in B$ on condition that the event $(X,Y)\in A$ occurs is denoted by $\mathbf{P}\left((X,Y)\in B|(X,Y)\in A\right)$ or $\mathbf{P}(B|A)$ for short. This conditional probability can be be calculated obviously as tho ratio of two integrals:

$$\mathbf{P}(B|A) = \frac{\iint\limits_{A\cap B} f(x,y)\,dx\,dy}{\iint\limits_{A} f(x,y)\,dx\,dy}$$

If $B\subseteq A$, then $A\cap B = B$, and we get that

$$\mathbf{P}(B|A) = \frac{\iint\limits_{B} f(x,y)\,dx\,dy}{\iint\limits_{A} f(x,y)\,dx\,dy}$$

Conditional density function. If A is a subset of the plane, which has a positive probability, and we know that the condition $(X,Y) \in A$ is fulfilled, then the density function of (X,Y) under this condition is, obviously

$$f(x,y|A) = \frac{f(x,y)}{\mathbf{P}(A)} = \frac{f(x,y)}{\iint\limits_A f(x,y)\,dx\,dy} \quad \text{if } (x,y) \in A$$

Multiplication rule for independent random variables. If X and Y are independent, and their density functions are $f_1(x)$ and $f_2(y)$ respectively, then the density function $f(x,y)$ of (X,Y) is the direct product of the density functions $f_1(x)$ and $f_2(y)$:

$$f(x,y) = f_1(x)\, f_2(y)$$

Proof.

$$f(x,y) \approx \frac{\mathbf{P}(x < X < x+\Delta x \text{ and } y < Y < y+\Delta y)}{\Delta x\,\Delta y} \approx$$

$$\frac{\mathbf{P}(x < X < x+\Delta x)}{\Delta x}\ \frac{\mathbf{P}(y < Y < y+\Delta y)}{\Delta y} \approx$$

$$f_1(x)\, f_2(y)$$

General multiplication rule. If the density function of X is $f_1(x)$ and the density function of Y under the condition that $X = x$ is $f_{2|1}(y|x)$, then the density function of (X,Y) is

$$f(x,y) = f_1(x)\, f_{2|1}(y|x)$$

Similarly:

$$f(x,y) = f_2(y)\, f_{1|2}(x|y)$$

Proof. We give the proof of the first formula:

$$f(x,y) \approx \frac{\mathbf{P}(x < X < x+\Delta x \text{ and } y < Y < y+\Delta y)}{\Delta x\,\Delta y} =$$

$$\frac{\mathbf{P}(x < X < x+\Delta x)}{\Delta x}\ \frac{\mathbf{P}(y < Y < y+\Delta y \mid x < X < x+\Delta x)}{\Delta y} \approx$$

$$\frac{\mathbf{P}(x < X < x+\Delta x)}{\Delta x}\ \frac{\mathbf{P}(y < Y < y+\Delta y \mid X \approx x)}{\Delta y} \approx$$

$$f_1(x)\, f_{2|1}(y|x)$$

Distribution function. The distribution function of a two-dimensional random variable is defined by

$$F(x,y) = \mathbf{P}(X < x, Y < y)$$

The distribution function can be calculated from the density function by integration:

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(x,y)\,dx\,dy$$

The density function can be calculated from the distribution function by differentiation:

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x\,\partial y}$$

The probability of a rectangle can be calculated from the distribution function like this:

$$P(x_1 < X < x_2 \quad \text{and} \quad y_1 < Y < y_2) =$$
$$= P(X < x_2 \quad \text{and} \quad Y < y_2) - P(X < x_1 \quad \text{and} \quad Y < y_2) -$$
$$-P(X < x_2 \quad \text{and} \quad Y < y_1) + P(X < x_1 \text{ and } Y < y_1) =$$
$$= F(x_2,y_2) - F(x_1,y_2) - F(x_2,y_1) + F(x_1,y_1)$$

Expected value of a function of (X,Y). If we make N experiments for a two-dimensional random variable (X,Y), and we substitute the experimental results

$$(X_1,Y_1),(X_2,Y_2),\ldots,(X_N,Y_N)$$

into the function $y = t(x,y)$, and we consider the values

$$t(X_1,Y_1),t(X_2,Y_2),\ldots,t(X_N,Y_N)$$

then their average is close to the expected value of $t(X,Y)$:

$$\frac{t(X_1,Y_1)+t(X_2,Y_2)+\ldots+t(X_N,Y_N)}{N} \approx \mathbf{E}\,(t(X,Y))$$

where $\mathbf{E}\,(t(X,Y))$ is the expected value of $t(X,Y)$, which is calculated by a double integral:

$$\mathbf{E}\,(t(X,Y)) = \iint\limits_{R^2} t(x,y)f(x,y)\,dx\,dy$$

Expected value of the product. As an example, and because of its importance, we mention here that the expected value of the product XY of the random variables X and Y is calculated by a double integral:

$$\mathbf{E}\,(XY) = \iint\limits_{R^2} x\,y\;\; f(x,y)\,dx\,dy$$

which means that if N is large, then

$$\frac{X_1Y_1+X_2Y_2+\ldots+X_NY_N}{N} \approx \mathbf{E}\,(XY) = \iint\limits_{R^2} x\,y\;\; f(x,y)\,dx\,dy$$

Covariance and covariance matrix. The notion of the covariance is an auxiliary notion in two-dimensions. For a two-dimensional data-set, it is

$$\frac{(X_1-\overline{X})(Y_2-\overline{Y})+(X_2-\overline{X})(Y_2-\overline{Y})+\ldots+(X_N-\overline{X})(Y_N-\overline{Y})}{N}=$$

$$\frac{X_1Y_2+X_2Y_2+\ldots+X_NY_N}{N}-\overline{X}\,\overline{Y}$$

For a two-dimensional random variable and distribution the covariance $\mathbf{COV}(X,Y)$ is defined by

$$\mathbf{COV}(X,Y)=\mathbf{E}\left((X-\mathbf{E}(X))(Y-\mathbf{E}(Y))\right)=\mathbf{E}(XY)-\mathbf{E}(X)\mathbf{E}(Y)=$$

$$\iint\limits_{R^2}(x-\mu_1)\,(y-\mu_2)\,f(x,y)\,dx\,dy=\iint\limits_{R^2}x\,y\,f(x,y)\,dx\,dy-\mu_1\mu_2$$

The covariance and the variances of the coordinates can be arranged into a matrix. This matrix is called the **covariance matrix**:

$$C=\begin{pmatrix}\sigma_1^2 & c\\ c & \sigma_2^2\end{pmatrix}$$

This matrix can be considered as a two-dimensional generalization or the notion of the variance.

Correlation coefficient. The notion of the correlation coefficient plays an important role in describing the relation between the coordinates of a two-dimensional data-set of random variable. Its definition is:

$$\mathbf{CORR}(X,Y)=\frac{\mathbf{COV}(X,Y)}{\sigma_1\sigma_2}$$

Its value is always between -1 and 1: $-1\le r\le 1$. If $r>0$, then larger x-coordinates mostly imply larger y-coordinates, if $r<0$, then larger x-coordinates mostly imply smaller y-coordinates. In the first case, we say that the coordinates have a **positive correlation**, in the second case, we say that the coordinates have a **negative correlation**. If $|r|$ is close to 1, then the coordinates are in a **strong correlation**, if $|r|$ is close to 0, then the coordinates are in a **loose correlation**.

Using a calculator. More sophisticated calculators have a key to determine the covariance and the correlation coefficient of a two-dimensional data-set, as well.

Section 54
Uniform distribution on a two-dimensional set

If S is a set in the two-dimensional plane, and S has a finite area, then we may consider the density function equal to the reciprocal of the area of S inside S, and equal to 0 otherwise:

$$f(x,y) = \frac{1}{\text{area of } S} \quad \text{if } (x,y) \in S$$

The distribution associated to this density function is called **uniform distribution on the set** S. Since the integral of a constant on a set A is equal to the area of A multiplied by that constant, we get that

$$\mathbf{P}(A) = \iint_A f(x,y)\,dx\,dy = \iint_A \frac{1}{\text{area of } S}\,dx\,dy = \frac{\text{area of } A}{\text{area of } S}$$

for any subset A of S. Thus, uniform distribution on S means that, for any subset A of S, the probability of A is proportional to the area of A.

The reader probably remembers that in Chapter 6 of Part I, under the title "Geometrical problems, uniform distributions", we worked with uniform distributions. Now it should become clear that the uniform distribution on the set S is a special continuous distribution whose density function is equal to a constant on the set S.

Section 55
*** Beta distributions in two-dimensions

Assume that n people arrive between noon and 1pm independently of each other according to uniform distribution, and let X be the ith, and let Y be the jth arrival time. This real-life problem can be simulated like this: we generate n uniformly distributed independent random points between 0 and 1, and X = ith smallest, and Y = jth smallest among them. We calculate here the density function of the two-dimensional random variable (X,Y). Let $0 < x < y < 1$, let $[x_1,x_2]$ be a small interval around x, and let $[y_1,y_2]$ be a small interval around y. We assume that $x_2 < y_1$. By the meaning of the density function:

$$f(x,y) \approx \frac{\mathbf{P}(X \in \Delta x, Y \in \Delta y)}{(x_2 - x_1)\,(y_2 - y_1)}$$

The event $X \in \Delta x$, $Y \in \Delta y$, which stands in the numerator, means that the ith smallest point is in $[x_1,x_2)$, and the jth smallest point is in $[y_1,y_2)$, which means that

there is	at least one	point X	in	$[x_1,x_2)$,	and
there is	at least one	point Y	in	$[y_1,y_2)$,	and
there are	$i-1$	points	in	$[0,X)$,	and
there are	$j-i-1$	points	in	$[X,Y)$,	and
there are	$n-j$	points	in	$[Y,1]$.	

This, with a very good approximation, means that

there are	$i-1$	points	in	$[0,x_1)$,	and
there is	1	point	in	$[x_1,x_2)$,	and
there are	$j-i-1$	points	in	$[x_2,y_1)$,	and
there is	1	point	in	$[y_1,y_2)$,	and
there are	$n-j$	points	in	$[y_2,1]$.	

Using the formula of the poly-hyper-geometrical distribution, we get that the probability of the event $X \in \Delta x, Y \in \Delta y$ is approximately equal to

$$\frac{n!}{(i-1)!\;1!\;(j-i-1)!\;1!\;(n-j)!}\;x_1^{i-1}(x_2-x_1)^1(y_1-x_2)^{j-i-1}(y_2-y_1)^1(1-y_2)^{n-j}$$

Since $1! = 1$, we may omit some unnecessary factors and exponents, and the formula simplifies to

$$\frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!}\; x^{i-1}\,(x_2-x_1)\,(y-x)^{j-i-1}\,(y_2-y_1)\,(1-y)^{n-j}$$

Dividing by $(x_2-x_1)\,(y_2-y_1)$, we get that the density function, for $0<x<y<1$, is

$$f(x,y) = \frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!}\; x^{i-1}\,(y-x)^{j-i-1}\,(1-y)^{n-j}$$

Special cases:

1. **Three independent random numbers** (uniformly distributed between 0 and 1) are generated.

   (a) $X =$ **the smallest,** $Y =$ **the biggest** of them.
   Replacing $n=3, i=1, j=3$, we get:

   $$f(x,y) = 6(y-x) \quad \text{if } 0<x<y<1$$

   (b) $X =$ **the smallest,** $Y =$ **the second smallest** of them.
   Replacing $n=3, i=1, j=2$, we get:

   $$f(x,y) = 6(1-y) \quad \text{if } 0<x<y<1$$

   (c) $X =$ **the second smallest,** $Y =$ **the biggest** of them.
   Replacing $n=3, i=2, j=3$, we get:

   $$f(x,y) = 6x \quad \text{if } 0<x<y<1$$

2. **Four independent random numbers** (uniformly distributed between 0 and 1) are generated.

   (a) $X =$ **the smallest,** $Y =$ **the biggest** of them.
   Replacing $n=4, i=1, j=4$, we get:

   $$f(x,y) = 12\;(y-x)^2 \quad \text{if } 0<x<y<1$$

   (b) $X =$ **the smallest,** $Y =$ **the second smallest** of them.
   Replacing $n=4, i=1, j=2$, we get:

   $$f(x,y) = 12(1-y)^2 \quad \text{if } 0<x<y<1$$

   (c) $X =$ **the second smallest,** $Y =$ **the third smallest** of them.

   $$f(x,y) = 24x(1-y)x \quad \text{if } 0<x<y<1$$

3. **Ten independent random numbers** (uniformly distributed between 0 and 1) are generated.

   $X =$ **the 3rd smallest,** $Y =$ **the 7th smallest** of them.

   Replacing $n = 10, i = 3, j = 7$, we get:

$$f(x,y) = \frac{10!}{2!\,3!\,3!}\; x^2\,(y-x)^3\,(1-y)^3 \quad \text{if } 0 < x < y < 1$$

More general two-dimensional beta distributions. If the people arrive between A and B instead of 0 and 1, that is, the n independent, uniformly distributed random numbers are generated between A and B, and

$$X = \text{the } i\text{th smallest of them}$$

$$Y = \text{the } j\text{th smallest of them}$$

of them, then for $A < x < B$, the density function of (X,Y) is

$$f(x,y) =$$

$$\frac{1}{(B-A)^2}\,\frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!}\left(\frac{x-A}{B-A}\right)^{i-1}\left(\frac{y-x}{B-A}\right)^{j-i-1}\left(\frac{B-y}{B-A}\right)^{n-j}$$

Files to study two-dimensional beta point-clouds:

Demonstration file: Two-dimensional beta point-cloud related to size 2 *and ranks* 1 *and* 2
200-69-00

Demonstration file: Two-dimensional beta point-cloud related to size 3 *and ranks* 1 *and* 2
200-70-00

Demonstration file: Two-dimensional beta point-cloud related to size 3 *and ranks* 1 *and* 3
200-71-00

Demonstration file: Two-dimensional beta point-cloud related to size 3 *and ranks* 2 *and* 3
200-72-00

Demonstration file: Two-dimensional beta point-cloud related to size 5 *and ranks* k_1 *and* k_2
200-73-00

Demonstration file: Two-dimensional beta point-cloud related to size 10 *and ranks* k_1 *and* k_2
200-74-00

File to study two-dimensional point-clouds for arrival times:

Demonstration file: Two-dimensional gamma distribution
200-68-00

Section 56
Projections and conditional distributions

Projections. If the density function of the two-dimensional random variable (X,Y) is f(x,y), then the density function $f_1(x)$ of the random variable X can be calculated by integration:

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$

Sketch of proof. The interval $[x,x+\Delta x]$ on the horizontal line defines a vertical strip in the plane:

$$S_{[x,x+\Delta x]} = \{(x,y) : x \in [x,x+\Delta x]\}$$

so that the event $X \in [x,x+\Delta x]$ is equivalent to $(X,Y) \in S_{[x,x+\Delta x]}$. Using this fact we get that

$$f_1(x) \approx \frac{\mathbf{P}\left(X \in [x,x+\Delta x]\right)}{\Delta x} = \frac{\mathbf{P}\left((X,Y) \in S_{[x,x+\Delta x]}\right)}{\Delta x} =$$

$$\frac{\iint\limits_{S_{[x,x+\Delta x]}} f(x,y)\,dx\,dy}{\Delta x} = \frac{\int\limits_{x}^{x+\Delta x}\left(\int\limits_{-\infty}^{\infty} f(x,y)\,dy\right)dx}{\Delta x} \approx \int_{-\infty}^{\infty} f(x,y)\,dy$$

Similarly, the density function $f_2(x)$ of the random variable Y is

$$f_2(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

Conditional distributions. If a two-dimensional random variable (X,Y) is considered, and somehow the actual value x of X is known, but the value of Y is unknown, then we may need to know the conditional distribution of Y under the condition that $X = x$. The conditional density function can be calculated by division:

$$f_{2|1}(y|x) = \frac{f(x,y)}{f_1(x)}$$

Similarly, the conditional density function of X under the condition that $Y = y$ is

$$f_{1|2}(x|y) = \frac{f(x,y)}{f_2(y)}$$

Sketch of proof. We give the proof of the first formula.

$$f_{2|1}(y|x) \approx \frac{1}{\Delta y}\ \mathbf{P}\left(Y \in [y, y+\Delta y] \ \mid\ X = x\right) \approx$$

$$\frac{1}{\Delta y}\ \mathbf{P}\left(Y \in [y, y+\Delta y] \ \mid\ X \in [x, x+\Delta x]\right) =$$

$$\frac{1}{\Delta y}\ \frac{\mathbf{P}\left(X \in [x, x+\Delta x] \text{ and } Y \in [y, y+\Delta y]\right)}{\mathbf{P}\left(X \in [x, x+\Delta x]\right)} =$$

$$\frac{1}{\Delta y}\ \frac{\frac{\mathbf{P}(X \in [x, x+\Delta x] \text{ and } Y \in [y, y+\Delta y])}{\Delta x}}{\frac{\mathbf{P}(X \in [x, x+\Delta x])}{\Delta x}} =$$

$$\frac{\frac{\mathbf{P}(X \in [x, x+\Delta x] \text{ and } Y \in [y, y+\Delta y])}{\Delta x\ \Delta y}}{\frac{\mathbf{P}(X \in [x, x+\Delta x])}{\Delta x}} \approx \frac{f(x,y)}{f_1(x)}$$

Product rules. It often happens that the density function of (X,Y) is calculated from one of the product rules:

$$f(x,y) = f_1(x)\ f_{2|1}(y|x)$$
$$f(x,y) = f_2(y)\ f_{1|2}(x|y)$$

Conditional distribution function. The distribution function of the conditional distribution is calculated from the conditional density function by integration:

$$F_{2|1}(y|x) = \int_{\infty}^{y} f_{2|1}(y|x)\ dy = P\left(Y < y \mid X = x\right)$$

Similarly,

$$F_{1|2}(x|y) = \int_{\infty}^{x} f_{1|2}(x|y)\ dx = P\left(X < x \mid Y = y\right)$$

On the contrary, the conditional density function is a partial derivative of the conditional distribution function:

$$f_{2|1}(y|x) = \frac{\partial F_{2|1}(y|x)}{\partial y}$$

Similarly,

$$f_{1|2}(x|y) = \frac{\partial F_{1|2}(x|y)}{\partial x}$$

Conditional probability. The conditional probability of an interval for Y, under the condition that $X = x$, can be calculated from the conditional density by integration:

$$\mathbf{P}(\ y_1 < Y < y_2 \mid X = x\ ) = \int_{y_1}^{y_2} f_{2|1}(y|x)\ dy$$

Similarly, the conditional probability of an interval for X, under the condition that $Y = y$, can be calculated from the other conditional density by integration:

$$\mathbf{P}(\, x_1 < X < x_2 \mid Y = y \,) = \int_{x_1}^{x_2} f_{1|2}(x|y)\, dx$$

The conditional probability of an interval for Y, under the condition that $X = x$, can be calculated from the conditional distribution function as a difference:

$$\mathbf{P}(\, y_1 < Y < y_2 \mid X = x \,) = F_{2|1}(y_2|x) - F_{2|1}(y_1|x)$$

Similarly, the conditional probability of an interval for X, under the condition that $Y = y$, can be calculated from the other conditional distribution function as a difference:

$$\mathbf{P}(\, x_1 < X < x_2 \mid Y = y \,) = F_{1|2}(x_1|y) - F_{1|2}(x_1|y)$$

Remark. Notice that in the conditional probability

$$\mathbf{P}(\, y_1 < Y < y_2 \mid X = x \,)$$

the probability of the condition is zero:

$$\mathbf{P}(X = x\,) = 0$$

Thus, the definition

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

would not be applicable to define $\mathbf{P}(\, y_1 < Y < y_2 \mid X = x \,)$.

Conditional median. Solving the equation

$$F_{2|1}(y|x) = \frac{1}{2}$$

for y, that is, expressing y in terms of x, we get the conditional median of Y on condition that $X = x$.

Similarly, solving the equation

$$F_{1|2}(x|y) = \frac{1}{2}$$

for x, that is, expressing x in terms of y, we get the conditional median of X on condition that $Y = y$.

Conditional expected value. The conditional expected value is the expected value of the conditional distribution:

$$\mathbf{E}(Y|X = x) = \mu_{2|1}(|x) = \int_{-\infty}^{\infty} y\, f_{2|1}(y|x)\, dy$$

$$\mathbf{E}(X|Y=y)=\mu_{1|2}(|y)=\int_{-\infty}^{\infty} x\, f_{1|2}(x|y)\, dx$$

Conditional variance. The variance of the conditional distribution is the conditional variance:

$$\mathbf{VAR}(Y|X=x)=\sigma^2_{2|1}(|x)=$$

$$\int_{-\infty}^{\infty}\big(y-\mu_{2|1}(|x)\big)^2\, f_{2|1}(y|x)\, dy=\int_{-\infty}^{\infty} y^2\, f_{2|1}(y|x)\, dy\ -\big(\,\mu_{2|1}(|x)\,\big)^2$$

$$\mathbf{VAR}(X|Y=y)=\sigma^2_{1|2}(|y)=$$

$$\int_{-\infty}^{\infty}\big(x-\mu_{1|2}(|y)\big)^2\, f_{1|2}(x|y)\, dx=\int_{-\infty}^{\infty} x^2\, f_{1|2}(x|y)\, dx\ -\big(\,\mu_{1|2}(|y)\,\big)^2$$

Conditional standard deviation. The standard deviation of the conditional distribution is the conditional standard deviation:

$$\mathbf{SD}(Y|X=x)=\sigma_{2|1}(|x)=$$

$$\sqrt{\int_{-\infty}^{\infty}\big(y-\mu_{2|1}(|x)\big)^2\, f_{2|1}(y|x)\, dy}=\sqrt{\int_{-\infty}^{\infty} y^2\, f_{2|1}(y|x)\, dy-\big(\mu_{2|1}(|x)\big)^2}$$

$$\mathbf{SD}(X|Y=y)=\sigma_{1|2}(|y)=$$

$$\sqrt{\int_{-\infty}^{\infty}\big(x-\mu_{1|2}(|y)\big)^2\, f_{1|2}(x|y)\, dx}=\sqrt{\int_{-\infty}^{\infty} x^2\, f_{1|2}(x|y)\, dx-\big(\mu_{1|2}(|y)\big)^2}$$

Remark. The notion of the conditional variance and the conditional standard deviation can obviously be introduced and calculated for discrete distributions as well: in the above formulas, instead of integration summation is taken.

Example 1. We choose a random number between 0 and 1 according to uniform distribution, let it be X. If $X=x$, then we choose another random number between 0 and x according to uniform distribution, let it be Y. We shall calculate now all the density functions. By the definition of X and Y, we may write:

$$f(x)=1\quad\text{if}\quad 0<x<1$$

$$f_{2|1}(y|x)=\frac{1}{x}\quad\text{if}\quad 0<y<x<1$$

By the product rule:

$$f(x,y) = f_1(x)\, f_{2|1}(y|x) = 1\,\frac{1}{x} = \frac{1}{x} \quad \text{if} \quad 0 < y < x < 1$$

By the integration rule:

$$f_2(y) = \int_{-\infty}^{\infty} f(x,y)\, dx = \int_y^y \frac{1}{x}\, dx = -\ln(y) \quad \text{if} \quad 0 < y < 1$$

By the division rule:

$$f_{1|2}(x|y) = \frac{f(x,y)}{f_2(y)} = \frac{\frac{1}{x}}{-\ln(y)} = -\frac{1}{x\ln(y)} \quad \text{if} \quad 0 < y < x < 1$$

The conditional distribution function is

$$F_{1|2}(x|y) = \int_{\infty}^{x} f_{1|2}(x|y)\, dx =$$

$$\int_y^x \left(-\frac{1}{x\ln(y)}\right) dx = \frac{1}{-\ln(y)} \int_y^x \frac{1}{x}\, dx =$$

$$\frac{1}{-\ln(y)}\,(\ln(x) - \ln(y)) = 1 - \frac{\ln(x)}{\ln(y)} \quad \text{if} \quad 0 < y < 1$$

The conditional median is the solution for x of the equation

$$F_{1|2}(x|y) = 1 - \frac{\ln(x)}{\ln(y)} = \frac{1}{2}$$

that is

$$\frac{\ln(x)}{\ln(y)} = \frac{1}{2}$$

$$\ln(x) = \frac{1}{2}\ln(y)$$

$$\ln(x) = \ln(\sqrt{y})$$

$$x = \sqrt{y}$$

The conditional expected value is

$$\mathbf{E}(X|Y=y) = \mu_{1|2}(|y) = \int_{-\infty}^{\infty} x\, f_{1|2}(x|y)\, dx =$$

$$\int_y^1 x \left(-\frac{1}{x\ln(y)}\right) dx = \int_y^1 \left(-\frac{1}{\ln(y)}\right) dx =$$

$$\left(-\frac{1}{\ln(y)}\right)(1-y) = \frac{y-1}{\ln(y)} \quad \text{if} \quad 0 < y < 1$$

Files to visualize projections and conditional distributions:

Demonstration file: $X = RND_1$, $Y = X\ RND_2$, projections and conditional distributions
200-79-00

Demonstration file: Two-dim beta distributions, $n = 10$, projections and conditional distributions
200-80-00

Demonstration file: Two-dim beta distributions, $n \leq 10$, projections and conditional distributions
200-81-00

Files to study construction of a two-dimensional continuous distribution using conditional distributions:

Demonstration file: Conditional distributions, uniform on parallelogram
200-84-00

Demonstration file: Conditional distributions, $(RND_1, RND_1 RND_2)$
200-85-00

Demonstration file: Conditional distributions, uniform on triangle
200-86-00

Demonstration file: Conditional distributions, Bergengoc bulbs
200-87-00

Demonstration file: Conditional distributions, standard normal
200-88-00

Demonstration file: Conditional distributions, normal
200-89-00

Section 57
Normal distributions in two-dimensions

Standard normal distribution. The two-dimensional standard normal distribution is defined on the whole plane by its density function:

$$f(x,y) = \frac{1}{2\pi}\, e^{-\frac{1}{2}(x^2+y^2)}$$

Since the value of the density function depends on x and y only through x^2+y^2, the distribution is circular symmetrical. Actually, the surface defined by the density function resembles a hat or a bell.

It is easy to check that the projections of the standard normal distribution onto both axes are one-dimensional standard normal distributions.

The conditional distributions both on the vertical and the horizontal lines are one-dimensional standard normal distributions, as well.

General two-dimensional normal distributions. General two-dimensional normal distributions are defined by their density function:

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}}\, e^{-\frac{1}{2}\frac{\left(\frac{x-\mu_1}{\sigma_1}\right)^2+\left(\frac{y-\mu_2}{\sigma_2}\right)^2-2r\frac{x-\mu_1}{\sigma_1}\frac{y-\mu_2}{\sigma_2}}{\sqrt{1-r^2}}}$$

where the parameters μ_1, μ_2 are real numbers, σ_1, σ_2 are positive numbers, and r is a number between -1 and 1, equality permitted. The surface defined by the density function resembles a hat which is compressed in one direction so that looking at it from above it has an elliptical shape.

It can be shown that the level curves of the density function are ellipses, whose center is at the point (μ_1,μ_2), the directions of the axes are determined by the directions of the eigenvectors of the covariance matrix, and the sizes of the axes are proportional to the square roots of the eigen-values.

Projections. It is easy to check that the projections of a two-dimensional normal distribution onto both axes are normal distributions. The projection onto the horizontal axis is a normal

distribution with parameters μ_1 and σ_1. The projection onto the vertical axis is a normal distribution with parameters μ_2 and σ_2. The density function of the projection onto the horizontal axis is:

$$f_1(x) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}$$

The density function of the projection onto the vertical axis is:

$$f_2(y) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}$$

Correlation coefficient . The value of the correlation coefficient can be calculated according to its definition. It turns out that its value is equal to the value of the parameter r.

Conditional distributions. The conditional distributions are normal distributions. The conditional density function along the vertical line, when $X = x$, is:

$$f_{2|1}(y|x)\frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-r^2}} e^{-\frac{1}{2}\left(\frac{y-\left(\mu_2+r\frac{\sigma_2}{\sigma_1}(x-\mu_1)\right)}{\sigma_2\sqrt{1-r^2}}\right)^2}$$

The conditional density function along the horizontal line, when $Y = y$, is:

$$f_{1|2}(x|y)\frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-r^2}} e^{-\frac{1}{2}\left(\frac{x-\left(\mu_1+r\frac{\sigma_1}{\sigma_2}(y-\mu_2)\right)}{\sigma_1\sqrt{1-r^2}}\right)^2}$$

Conditional median and expected value. Both conditional medians and expected values depend on the condition linearly. Namely, the conditional median and expected value of Y, when $X = x$, is:

$$\mu_2 + r\frac{\sigma_2}{\sigma_1}(x-\mu_1)$$

The straight line defined by this formula is the so called **first regression line**.

Similarly, the conditional median and expected value of X, when $Y = y$, is:

$$\mu_1 + r\frac{\sigma_1}{\sigma_2}(y-\mu_2)$$

The straight line defined by this formula is the so called **second regression line**. The regression lines mark the place of the conditional medians and expected values.

Conditional standard deviation. Both conditional standard deviations do not depend on the condition, they are a constants. Namely, the conditional standard deviation of Y, when $X = x$, is:

$$\sigma_2\sqrt{1-r^2}$$

If we move the first regression line up and down by an amount equal to the value of the conditional standard deviation we get the so called conditional **standard deviation lines** associated to the first regression line.

Similarly, the conditional standard deviation of X, when $Y = y$, is:

$$\sigma_1\sqrt{1-r^2}$$

If we move the second regression line to the right and to the left by an amount equal to the value of the conditional standard deviation we get the so called conditional **standard deviation lines** associated to the second regression line.

Standard deviation rectangle. In order to get a better view of a normal distribution, let us consider the rectangle defined by the direct product of the intervals

$$(\mu_1-\sigma_1;\mu_1+\sigma_1) \text{ and } (\mu_2-\sigma_2;\mu_2+\sigma_2)$$

We may take also the rectangles defined by the direct product of the intervals

$$(\mu_1-s\sigma_1;\mu_1+s\sigma_1) \text{ and } (\mu_2-s\sigma_2;\mu_2+s\sigma_2)$$

where s is a positive number. This rectangle may be called the **standard deviation rectangle** of size s. Let us put a scale from -1 to 1 on each of the sides of the standard deviation rectangle so that the -1 is at the left end, and the is 1 at the right end on the horizontal sides, and the -1 is at the lower end, and the is 1 at the upper end on the vertical sides. The points on the sides which correspond to r, the value of the correlation coefficient, play an interesting role, since the following facts are true:

1. The sides of the standard deviation rectangles are *tangents to the ellipses* which arise as level curves, and the common points of the ellipses and the standard deviation rectangles correspond to r, the value of the *correlation coefficient*, on the scales on the sides of the standard deviation rectangles.

2. The *regression lines* intersect the standard deviation rectangles at points which correspond to r on the scales on the sides of the standard deviation rectangles.

3. If we draw the standard deviation rectangle of size 1, and consider the ellipse which arises as a level curve, touching the standard deviation rectangle at the point having a position r on the on the scales on the sides of the standard deviation rectangle, then we may draw the tangent lines to the ellipse, parallel to the regression lines. These lines that we get are the *standard deviation lines* associated to the regression lines.

Files to study the height and weight of men as a two-dimensional normal random variable:

Demonstration file: Height and weight
200-65-00

Demonstration file: Height and weight, ellipse, eigen-vectors (projections and conditional distributions are also studied)
200-82-00

Demonstration file: Two-dim normal distributions normal distributions, projections and conditional distributions
200-83-00

File to study voltages as a two-dimensional normal random variable:

Demonstration file: Measuring voltage
200-66-00

Section 58
Independence of random variables

The discrete random variables X and Y are **independent**, if any (and then all) of the following relations hold:

$$p_{2|1}(y|x) = p_2(y) \quad \text{for all } x \text{ and } y$$
$$p_{1|2}(x|y) = p_1(x) \quad \text{for all } x \text{ and } y$$
$$p(x,y) = p_1(x)p_2(y) \quad \text{for all } x \text{ and } y$$

The continuous random variables X and Y are **independent**, if any (and then all) of the following relations hold:

$$f_{2|1}(y|x) = f_2(y) \quad \text{for all } x \text{ and } y$$
$$f_{1|2}(x|y) = f_1(x) \quad \text{for all } x \text{ and } y$$
$$f(x,y) = f_1(x)f_2(y) \quad \text{for all } x \text{ and } y$$

If some random variables are not independent, then we call them **dependent**.

File to study the notion of dependence and independence:

Demonstration file: Lengths dependent or independent
200-91-00

Section 59
Generating a two-dimensional random variable

It is important that for a given two-dimensional distribution that a two-dimensional random variable can be generated by a calculator or a computer so that its distribution is the given two-dimensional continuous distribution. If the distribution is continuous, then the method described below defines such a two-dimensional random variable. (The discrete case is left for the reader as an exercise.)

In order to find the desired distribution in the continuous case, let the distribution function of its projection onto the horizontal axis be denoted by $F_1(x)$. Let $F^{-1}(u)$ be its inverse. (If $F_1(x)$ not strictly increasing on the the whole real-line, but only on an interval (A,B), then $F^{-1}(u)$ should be the inverse of restriction of $F(x)$ onto that interval.) The way how we technically find a formula for $F_1^{-1}(u)$ is that we solve the equation

$$u = F_1(x)$$

for x, that is, we express x from the equation in terms of u:

$$x = F_1^{-1}(u)$$

In a similar way, let the distribution function of the conditional distributions on the vertical axes be denoted by $F_{2|1}(y|x)$, and their inverse be denoted by $F_{2|1}^{-1}(v|x)$. The way how we find technically a formula for $F_{2|1}^{-1}(v|x)$ is that we solve the equation

$$v = F_{2|1}(y|x)$$

for y, that is, we express y from the equation in terms of v:

$$y = F_{2|1}^{-1}(v|x)$$

In this calculation x plays the role of a parameter.

Now we define the random value X by

$$X = F_1^{-1}(\mathrm{RND}_1)$$

It is easy to be convinced that the distribution function of the random variable X is the function $F_1(x)$. Then we define the random variable Y by

$$Y = F_{2|1}^{-1}(\mathrm{RND}_2|X)$$

It is easy to be convinced that the conditional distribution function of the random variable Y on condition that $X = x$ is the function $F_{2|1}(y|x)$. The two facts that

1. the distribution function of the random variable X is the function $F_1(x)$
2. the conditional distribution function of the random variable Y on condition that $X = x$ is the function $F_{2|1}(y|x)$

mean that the random variable (X,Y) has the given two-dimensional continuous distribution.

Section 60
Properties of the expected value, variance and standard deviation

1. **Expected value of a sum:**

$$\mathbf{E}(X+Y) = \mathbf{E}(X) + \mathbf{E}(Y)$$

$$\mathbf{E}(X_1 + X_2 + \ldots + X_n) = \mathbf{E}(X_1) + \mathbf{E}(X_2) + \ldots + \mathbf{E}(X_n)$$

2. **Expected value of constant times a random variable:**

$$\mathbf{E}(cX) = c\mathbf{E}(X)$$

3. **Linearity of the expected value:**

$$\mathbf{E}(a\,X + b\,Y) = a\,\mathbf{E}(X) + b\,\mathbf{E}(Y)$$

$$\mathbf{E}(a_1X_1 + a_2X_2 + \ldots + a_n + X_n) =$$
$$a_1\mathbf{E}(X_1) + a_2\mathbf{E}(X_2) + \ldots + a_n\mathbf{E}(X_n)$$

4. **Expected value of the sum:**
   If $X_1, X_2, \ldots, X_n$ have an identical expected value μ, then:

$$\mathbf{E}(X_1 + X_2 + \ldots + X_n) = n\mu$$

5. **Expected value of the average:**

$$\mathbf{E}(\frac{X_1 + X_2 + \ldots + X_n}{n}) = \mu$$

6. **Expected value of the product of independent random variables:**
   If X and Y are independent, then:

$$\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y)$$

7. **Variance of a sum:**

$$\mathbf{VAR}(X+Y) = \mathbf{VAR}(X) + \mathbf{VAR}(Y) + 2\mathbf{COV}(X,Y)$$

where $\mathbf{COV}(X,Y)$ is the covariance between X and Y (see the definition of the covariance later).

8. **Variance of the sum of independent random variables:**

If X and Y are independent, then

$$\mathbf{VAR}(X+Y) = \mathbf{VAR}(X) + \mathbf{VAR}(Y)$$

9. **Variance of constant times a random variable:**

$$\mathbf{VAR}(cX) = c^2\,\mathbf{VAR}(X)$$

10. **Variance of the sum and average of several independent terms:**

If $X_1, X_2, \ldots, X_n$ are independent and have a common variance σ^2, then:

$$\mathbf{VAR}(X_1 + X_2 + \ldots + X_n) = n\,\sigma^2$$

$$\mathbf{VAR}\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \frac{\sigma^2}{n}$$

11. **Standard deviation of constant times a random variable:**

$$\mathbf{SD}(cX) = |c|\,\mathbf{SD}(X)$$

12. **Square root law of the standard deviation for the sum:**

If $X_1, X_2, \ldots, X_n$ are independent and have a common standard deviation σ, then

$$\mathbf{SD}(X_1 + X_2 + \ldots + X_n) = \sqrt{n}\,\sigma$$

13. **Square root law of the standard deviation for the average**:

If $X_1, X_2, \ldots, X_n$ are independent and have a common standard deviation σ, then

$$\mathbf{SD}\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \frac{\sigma}{\sqrt{n}}$$

File to study the "average"-property of the standard deviation:

Demonstration file: Standard deviation of the average
200-61-00

Section 61
Transformation from plane to line

When the two-dimensional random continuous variable (X,Y) has a density function $f(x,y)$, and $z = t(x,y)$ is a given function, then the distribution function $R(z)$ of the random variable $Z = t(X,Y)$ is

$$R(z) = \iint\limits_{A_z} f(x,y)\,dx\,dy$$

where the set A_z is the inverse image of the interval $(-\infty, z)$ at the transformation $z = t(x,y)$:

$$A_z = \{(x,y) : t(x,y) < z)\}$$

Sketch of proof. The event $Z < z$ is equivalent to the event $(X,Y) \in A_z$, so

$$R(z) = \mathbf{P}(Z < z) = \mathbf{P}(\ (X,Y) \in A_z) = \iint\limits_{A_z} f(x,y)\,dx\,dy$$

Taking the derivative with respect to z on both sides, we get the density function:

$$r(z) = R'(z)$$

Files to study transformations from plane to line:

Demonstration file: Transformation from square to line by product
300-02-00

Demonstration file: Transformation from square to line by ratio
300-03-00

Demonstration file: Transformation from plane into chi distribution
300-04-00

Demonstration file: Transformation from plane into chi-square distribution
300-05-00

Projections from plane onto the axes. If $t(x,y) = x$, then the transformation means the projection onto the x-axis. Recall that the density function $f_1(x)$ can be calculated from f(x,y) by integration with respect to y:

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$

Similarly, if $t(x,y) = y$, then the transformation means the projection onto the y-axis. The density function $f_2(x)$ can be calculated from f(x,y) by integration with respect to x:

$$f_2(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

Files to study projections from plane to axes:

Demonstration file: Projection from triangle onto axes: $(\max(RND_1, RND_2); \min(RND_1, RND_2))$
300-06-00

Demonstration file: Projection from triangle onto axes: $(RND_1; RND_1 RND_2)$
300-07-00

Demonstration file: Projection from sail onto axes: $(RND_1 RND_2; RND_1/RND_2)$
300-08-00

Demonstration file: Projection from sale onto axes: $(\sqrt{RND_1 RND_2}; \sqrt{RND_1/RND_2})$
300-09-00

Section 62
*** Transformation from plane to plane

General case. Assume that a the density function of a distribution on the plane is $f(x,y)$. Consider a one-to-one smooth transformation t from the (x,y)-plane onto the (u,v)-plane given by a pair of functions:

$$u = u(x,y)$$
$$v = v(x,y)$$

Let the inverse of the transformation be given by the pair of functions

$$x = x(u,v)$$
$$y = y(u,v)$$

The Jacobian matrix of the inverse transformation plays an important role in the formula we will state. This is why we remind the reader that the Jacobian matrix of the inverse transformation is a two by two matrix consisting of partial derivatives:

$$\frac{\partial(x,y)}{\partial(u,v)} = \begin{pmatrix} \frac{\partial x(u,v)}{\partial u} & \frac{\partial x(u,v)}{\partial v} \\ \\ \frac{\partial y(u,v)}{\partial u} & \frac{\partial y(u,v)}{\partial v} \end{pmatrix}$$

As the (x,y)-plane is transformed into the (u,v)-plane, the distribution on the (x,y)-plane is also transformed into a distribution on the (u,v)-plane. Let the density function of the arising distribution on the (u,v)-plane denoted by $s(u,v)$. Then the value of the new density function is equal to the value of the old density function multiplied by the absolute value of the determinant of the Jacobian matrix of the inverse transformation:

$$s(u,v) = f\left(x(u,v), y(u,v)\right) \left| \det\left(\frac{\partial(x,y)}{\partial(u,v)}\right)\right|$$

Sketch of proof. Let us consider a small rectangle B at the point (u,v). Its inverse image on the (x,y) plane is approximately a small parallelogram-like set A at the point (x,y) so that the ratio of their areas is approximately equal to the absolute value of the Jacobian matrix of the inverse transformation:

$$\frac{\text{area of } A}{\text{area of } B} \approx \left| \det\left(\frac{\partial(x,y)}{\partial(u,v)}\right)\right|$$

Using this fact we get that

$$s(u,v) \approx \frac{\mathbf{P}((U,V)\in B)}{\text{area of } B} = \frac{\mathbf{P}((X,Y)\in A)}{\text{area of } A}\ \frac{\text{area of } A}{\text{area of } B} \approx f(x,y)\ \left|\det\left(\frac{\partial(x,y)}{\partial(u,v)}\right)\right|$$

Files to study a transformation from plane to plane:

Demonstration file: Projection from square onto a "sail"
300-09-50

Special case: Multiplying by a matrix. Let us consider the special case, when the transformation is a linear transformation

$$u = a_{11}x + a_{12}y$$
$$v = a_{21}x + a_{22}y$$

that is

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$$

Introducing the matrix notations

$$\mathbf{u} = \begin{pmatrix} u \\ v \end{pmatrix}$$

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

$$\mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix}$$

the linear transformation can be written briefly as

$$\mathbf{u} = \mathbf{A}\mathbf{x}$$

Then the inverse transformation is

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{u}$$

where $\mathbf{A}^{-1}$ is the inverse of $\mathbf{A}$ and the Jacobian matrix of the inverse transformation is the inverse matrix $\mathbf{A}^{-1}$ itself.

So the new density function $s(\mathbf{u})$ expressed in terms of the old density $f(\mathbf{x})$ looks like this:

$$s(\mathbf{u}) = f(\mathbf{A}^{-1}\mathbf{u})\ \left|\det\left(\mathbf{A}^{-1}\right)\right|$$

Special case: Multiplying by a matrix and adding a vector. Let us consider the special case, when the transformation is a linear transformation

$$u = a_{11}x + a_{12}y + b_1$$
$$v = a_{21}x + a_{22}y + b_2$$

Introducing the notation

$$\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

the linear transformation can be written briefly as

$$\mathbf{u} = \mathbf{A}\mathbf{x} + \mathbf{b}$$

Then the inverse transformation is

$$\mathbf{x} = \mathbf{A}^{-1}(\mathbf{u} - \mathbf{b})$$

So the new density function $s(\mathbf{u})$ expressed in terms of the old density $f(\mathbf{x})$ looks like as this:

$$s(\mathbf{u}) = f(\ \mathbf{A}^{-1}(\mathbf{u} - \mathbf{b})\ )\ \left|\det\left(\mathbf{A}^{-1}\right)\right|$$

Linear transformation of two-dimensional normal distributions. If a two-dimensional normal distribution is transformed by a linear transformation

$$\mathbf{u} = \mathbf{A}\,\mathbf{x} + \mathbf{b}$$

then the new distribution is a normal distribution, too. If the expected value of the old distribution is $\mathbf{m}_{\text{old}}$, then the expected value of the new distribution is

$$\mathbf{m}_{\text{new}} = \mathbf{A}\,\mathbf{m}_{\text{old}} + \mathbf{b}$$

If the covariance matrix of the old distribution is $\mathbf{C}_{\text{old}}$, then the covariance matrix of the new distribution is

$$\mathbf{C}_{\text{new}} = \mathbf{A}\,\mathbf{C}_{\text{old}}\,\mathbf{A}^T$$

Demonstration file: Linear transformation of the standard normal point-cloud
300-10-00

Demonstration file: Linear transformation of normal distributions
300-11-00

Section 63
*** Sums of random variables. Convolution

Discrete random variables, general case. Assume that the two-dimensional random variable (X,Y) has a distribution $p(x,y)$. Let Z denote the sum of X and Y, that is, $Z = X+Y$. Then the distribution $r(z)$ of the sum is:

$$r(z) = \sum_{(x,y):\, x+y=z} p(x,y) = \sum_x p(x,z-x) = \sum_y p(z-y,y)$$

Notice that

- in the first summation, for a given value of z, the summation takes place for all possible values of (x,y) for which $x+y=z$,

- in the second summation, for a given value of z, the summation takes place for all possible values of x.

- in the third summation, for a given value of z, the summation takes place for all possible values of y.

Remark. If $p(x,y)$ is zero outside a region S, then, in the first summation, for a given value of z, the summation can be restricted to the set $\{(x,y) : (x,y) \in S\}$:

$$r(z) = \sum_{(x,y)\in S:\, x+y=z} p(x,y)$$

In the second summation it can be restricted to the set $A_z = \{x : (x,z-x) \in S\}$:

$$r(z) = \sum_{x\in A_z} p(x,z-x)$$

In the third summation it can be restricted to the set $B_z = \{y : (z-y,y) \in S\}$:

$$r(z) = \sum_{y\in B_z} p(z-y,y)$$

Discrete, independent random variables. Assume now that the discrete random variables X and Y are independent. Since $p(x,y) = p_1(x)p_2(y)$, the above formulas reduce to:

$$r(z) = \sum_{(x,y):\, x+y=z} p_1(x)\, p_2(y) = \sum_x p_1(x)\, p_2(z-x) = \sum_y p_1(z-y)\, p_2(y)$$

We recognize that $r(z)$ is the **convolution** of the distributions $p_1(x)$ and $p_2(y)$.

Example 1. (Convolving binomial distributions) If we convolve a binomial distribution with parameters n_1 and p with a binomial distribution with parameters n_2 and p (the parameter p is the same for both distributions), then we get a binomial distribution with parameters n_1+n_2 and p.

Example 2. (Convolving Poisson distributions) If we convolve a Poisson distribution with parameter λ_1 with a Poisson distribution with parameter λ_2, then we get a Poisson distribution with parameters $\lambda_1+\lambda_2$.

Example 3. (Convolving geometrical distributions) If we convolve a geometrical distribution with parameter p with a geometrical distribution with parameter p (the parameter p is the same for both distributions), then we get a second order negative binomial distribution with parameter p.

Example 4. (Convolving negative binomial distributions) If we convolve a negative binomial distribution with parameters r_1 and p with a negative binomial distribution with parameters r_2 and p (the parameter p is the same for both distributions), then we get a negative binomial distribution with parameters r_1+r_2 and p.

Files to study how the distribution of the sum can be calculated:

Demonstration file: Summation of independent random variables, fair dice
300-12-00

Demonstration file: Summation of independent random variables, unfair dice
300-13-00

Continuous random variables, general case. Assume that the two-dimensional random variable (X,Y) has a density function $f(x,y)$. Let us consider the sum of X and Y: $Z=X+Y$. Then the density function $r(z)$ of the sum is:

$$r(z)=\int_{-\infty}^{\infty} f(x,z-x)\,dx=\int_{-\infty}^{\infty} f(z-y,y)\,dy$$

Sketch of proof. We shall perform the transformation in two steps. First we transform the distribution onto the (u,v)-plane by the linear transformation given by the equations

$$\begin{array}{ccccc} u & = & x & + & y \\ v & = & & & y \end{array}$$

and then we project onto the horizontal axis. Since the first coordinate of the above transformation is $u=x+y$, after the projection, we get what we need. The inverse transformation is

$$\begin{array}{ccccc} x & = & u & - & v \\ y & = & & & v \end{array}$$

The Jacobian matrix is

$$\begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ \\ & 1 \end{pmatrix}$$

so the Jacobian determinant is equal to 1. Thus

$$s(u,v)=f(\,u-v\,,v\,)\quad 1=f(\,u-v\,,v\,)$$

Now projecting onto the horizontal axis, the value of the density function at u turns out to be

$$\int_{-\infty}^{\infty} s(u,v)\,dv=\int_{-\infty}^{\infty} f(\,u-v\,,v\,)\,dv$$

Replacing formally the letter u by the letter z, and the integration variable v by y, we get that the value of the density function at z is

$$r(z)=\int_{-\infty}^{\infty} f(\,z-y\,,y\,)\,dy$$

Remark. If $f(x,y)$ is zero outside a region S, then, for a given z value, the interval $(-\infty,\infty)$ in the first integral can be replaced by the set $A_z=\{x:(x,z-x)\in S\}$:

$$r(z)=\int_{A_z} f(x,z-x)\,dx$$

Similarly, the interval $(-\infty,\infty)$ in the second integral can be replaced by the set $B_z=\{y:(z-y,y)\in S\}$:

$$r(z)=\int_{B_z} f(z-y,y)\,dy$$

Continuous, independent random variables. Assume now that the continuous random variables X and Y are independent. Since $f(x,y) = f_1(x)f_2(y)$, the above formulas reduce to:

$$r(z) = \int_{-\infty}^{\infty} f_1(x)\, f_2(z-x)\, dx = \int_{-\infty}^{\infty} f_1(z-y)\, f_2(y)\, dy$$

We recognize that $r(z)$ is the **convolution** of the density functions $f_1(x)$ and $f_2(y)$.

Example 5. (Convolving exponential distributions) If we convolve an exponential distribution with parameter λ with an exponential distribution with parameter λ (the parameter λ is the same for both distributions), then we get a second order gamma distribution with parameter λ.

Example 6. (Convolving gamma distributions) If we convolve a gamma distribution with parameters n_1 and λ with a gamma distribution with parameters n_2 and λ (the parameter λ is the same for both distributions), then we get a gamma distribution with parameters n_1+n_2 and λ.

Example 7. (Convolving normal distributions) If we convolve a normal distribution with parameters μ_1 and σ_1 with a normal distribution with parameters μ_2 and σ_2 (the parameter p is the same for both distributions), then we get a normal distribution with parameters $\mu_1+\mu_2$ and $\sqrt{\sigma_1^2+\sigma_2^2}$.

Section 64
Limit theorems to normal distributions

Moivre-Laplace theorem. If, for a fixed p value, we consider a binomial distribution with parameters n and p so that n is large, then the binomial distribution can be well approximated by a normal distribution. The parameters, that is, the expected value and the standard deviation of the normal distribution should be taken to be equal to the expected value and the standard deviation of the binomial distribution, that is, np and $\sqrt{np(1-p)}$.

If we standardize the binomial distribution, then the arising standardized binomial distribution will be close to the standard normal distribution.

Here is a file to study binomial approximation of normal distribution:

Demonstration file: Binomial approximation of normal distribution
300-14-00

Central limit theorem. If we add many independent random variables, then the distribution of the sum can be calculated from the distributions of the random variables by convolution. It can be shown that, under very general conditions (which we do not give here), the distribution of the sum will approximate a normal distribution. The parameters, that is, the expected value and the standard deviation of the normal distribution should be taken to be equal to the expected value and the standard deviation of the sum.

If we standardize the sum, the arising standardized value will be distributed approximately according to the standard normal distribution.

Central limit theorem in two-dimensions. If we add many independent two-dimensional random variables (random vectors), then the distribution of the sum, under very general conditions (which we do not give here), the distribution of the sum will approximate a two-dimensional normal distribution.

Files to study how convolutions approximate normal distributions:

Demonstration file: Convolution with uniform distribution
300-15-00

Demonstration file: Convolution with asymmetrical distribution
300-16-00

Demonstration file: Convolution with U-shaped distribution
300-17-00

Demonstration file: Convolution with randomly chosen distribution
300-18-00

File to study how gamma distributions approximate normal distributions:

Demonstration file: Gamma distribution approximates normal distribution
300-19-00

Files to study the two-dimensional central limit theorem:

Demonstration file: Two-dimensional central-limit theorem, rectangle
300-20-00

Demonstration file: Two-dimensional central-limit theorem, parallelogram
300-21-00

Demonstration file: Two-dimensional central-limit theorem, curve
300-22-00

Part - V.

Statistics

Section 65
Regression in one-dimension

Imagine that we work with a random variable X. Let us make N experiments, and let the observed values be $X_1, X_2, \ldots, X_N$. If we replace each observed value by a constant c, then we make an error at each replacement.

Minimizing the expected value of the absolute error. The absolute values of the errors are:

$$|X_1 - c|, |X_2 - c|, \ldots, |X_N - c|$$

The average of the absolute errors is

$$\frac{|X_1 - c| + |X_2 - c| + \ldots + |X_N - c|}{N}$$

For large N, this average is approximated by the expected value of $|X - c|$:

$$\frac{|X_1 - c| + |X_2 - c| + \ldots + |X_N - c|}{N} \approx$$

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

We learned in Part III that this integral is minimal if c is the median of X.

Minimizing the expected value of the squared error. The squares of the errors are:

$$(X_1 - c)^2, (X_2 - c)^2, \ldots, (X_N - c)^2$$

The average of the squared errors is

$$\frac{(X_1 - c)^2 + (X_2 - c)^2 + \ldots + (X_N - c)^2}{N}$$

For large N, this average is approximated by the expected value of $(X - c)^2$:

$$\frac{(X_1 - c)^2 + (X_2 - c)^2 + \ldots + (X_N - c)^2}{N} \approx$$

$$\mathbf{E}\left((X - c)^2\right) = \int_{-\infty}^{\infty} (x - c)^2 \, f(x) \, dx$$

We learned in Part III that this integral is minimal if c is the expected value of X.

Section 66
Regression in two-dimensions

Imagine that we work with a two-dimensional random variable (X,Y). Let us make N experiments, and let the observed values be $(X_1,Y_1),(X_2,Y_2),\ldots,(X_N,Y_N)$. If we replace each observed Y-value by a function of the X-value, that is, Y is replaced by $k(X)$, then we make an error at each replacement.

Minimizing the expected value of the absolute error. The absolute values of the errors are:

$$|Y_1-k(X_1)|,|Y_2-k(X_2)|,\ldots,|Y_N-k(X_N)|$$

The average of the absolute errors is

$$\frac{|Y_1-k(X_1)|+|Y_2-k(X_2)|+\ldots+|Y_N-k(X_N)|}{N}\approx$$

For large N, this average is approximated by the expected value of $|Y-k(X)|$:

$$\frac{|Y_1-k(X_1)|+|Y_2-k(X_2)|+\ldots+|Y_N-k(X_N)|}{N}\approx$$

$$\mathbf{E}(|Y-k(X)|)=\iint\limits_{R^2}|y-k(x)|\,f(x,y)\,dx\,dy=$$

$$\iint\limits_{R^2}|y-k(x)|\,f_1(x)\,f_{2|1}(y|x))\,dx\,dy=$$

$$\int\limits_{-\infty}^{\infty}\left(\int\limits_{-\infty}^{\infty}|y-k(x)|\,f_{2|1}(y|x))\,dy\right)f_1(x)\,dx$$

For all x, the inner integral is minimal if $k(x)$ is the conditional median, that is, the median of the conditional distribution associated to the condition $X=x$. The conditional median can be calculated from the equation

$$F_{2|1}(y|x)=\frac{1}{2}$$

so that we express y in term of x to get the function $y=k(x)$.

Minimizing the expected value of the squared error. The squares of the errors are:

$$(Y_1 - k(X_1))^2, (Y_2 - k(X_2))^2, \ldots, (Y_N - k(X_N))^2$$

The average of the squared errors is

$$\frac{(Y_1 - k(X_1))^2 + (Y_2 - k(X_2))^2 + \ldots + (Y_N - k(X_N))^2}{N}$$

For large N, this average is approximated by the expected value of $(Y - k(X))^2$:

$$\frac{(Y_1 - k(X_1))^2 + (Y_2 - k(X_2))^2 + \ldots + (Y_N - k(X_N))^2}{N} \approx$$

$$\mathbf{E}\left((Y - k(X))^2\right) = \iint_{R^2} (y - k(x))^2 \, f(x,y)\, dx\, dy =$$

$$\iint_{R^2} (y - k(x))^2 \, f_1(x) \, f_{2|1}(y|x))\, dx\, dy =$$

$$\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} (y - k(x))^2 \, f_{2|1}(y|x))\, dy \right) f_1(x)\, dx$$

For all x, the inner integral is minimal if $k(x)$ is the conditional expected value, that is, the expected value of the conditional distribution associated to the condition $X = x$. The conditional expected value can be calculated by integration:

$$k(x) = m_1(x) = \int_{-\infty}^{\infty} y \, f_{2|1}(y|x)\, dy$$

Files to study regression problems:

Demonstration file: Regression
200-92-00

Demonstration file: Conditional distributions, -expected value, -median, $(RND_1 RND_2, RND_1)$
200-93-00

Demonstration file: Conditional distributions, -expected value, -median, $(RND_1 RND_2, RND_1/RND_2)$
200-94-00

Section 67
Linear regression

Imagine that we work with a two-dimensional random variable (X,Y). Let us make N experiments, and let the observed values be $(X_1,Y_1),(X_2,Y_2),\ldots,(X_N,Y_N)$. If we replace each observed Y-value by a linear function of the X-value, that is, Y is replaced by $aX+b$, then at each of the replacements we make an error. The squares of the errors are:

$$(Y_1-(aX_1+b))^2,\ (Y_2-(aX_2+b))^2,\ \ldots,\ (Y_N-(aX_N+b))^2$$

The average of the squared errors is

$$\frac{(Y_1-(aX_1+b))^2+(Y_2-(aX_2+b))^2+\ldots+(Y_N-(aX_N+b))^2}{N}$$

For large N, this average is approximated by the expected value of $(Y-(aX+b))^2$:

$$\frac{(Y_1-(aX_1+b))^2+(Y_2-(aX_2+b))^2+\ldots+(Y_N-(aX_N+b))^2}{N}\approx$$

$$\mathbf{E}\left((Y-(aX+b))^2\right)=\iint\limits_{R^2}(y-(ax+b))^2\ f(x,y)\,dx\,dy$$

We may be interested in finding the values of a and b so that the expected value of the squared error is minimal.

Solution. Expanding the square $(y-(ax+b))^2$ in the above integral, we get six terms, so the integral is equal to the sum of six integrals as follows:

$$\mathbf{E}\left((Y-(aX+b))^2\right)=$$

$$\iint\limits_{R^2}y^2\,f(x,y)\,dx\,dy+a^2\iint\limits_{R^2}x^2\,f(x,y)\,dx\,dy+b^2\iint\limits_{R^2}f(x,y)\,dx\,dy-$$

$$-2a\iint\limits_{R^2}xy\,f(x,y)\,dx\,dy-2b\iint\limits_{R^2}y\,f(x,y)\,dx\,dy+2ab\iint\limits_{R^2}x\,f(x,y)\,dx\,dy$$

Each of these six integrals is a constant, so the formula itself is a two-variable quadratic formula. The values of a and b for which this quadratic formula is minimal can be determined by taking the partial derivatives of this quadratic formula with respect to a, and with respect to b, and then solving the arising system of equations for a and b. We omit the details of the calculation, the reader can make it or accept that the solution is

$$a_{\text{opt}} = r\,\frac{\sigma_2}{\sigma_1}$$

$$b_{\text{opt}} = \mu_2 \,-\, a_{\text{opt}}\,\mu_1 = \mu_2 \,-\, r\frac{\sigma_2}{\sigma_1}\,\mu_1$$

Thus, the equation of the line yielding the smallest expected value for the squared error is

$$y = \mu_2 + r\,\frac{\sigma_2}{\sigma_1}\,(x-\mu_1)$$

or, equivalently

$$\frac{y-\mu_2}{\sigma_2} = r\,\frac{x-\mu_1}{\sigma_1}$$

This line is called the **regression line**.

Expected value of the squared error. When we use the regression line, the value of the average of the squared errors is

$$\frac{\big(Y_1-(a_{\text{opt}}X_1+b_{\text{opt}})\big)^2+\big(Y_2-(a_{\text{opt}}X_2+b_{\text{opt}})\big)^2+\ldots+\big(Y_N-(a_{\text{opt}}X_N+b_{\text{opt}})\big)^2}{N}$$

For large N, this average is approximated by the expected value of $\big(Y-(a_{\text{opt}}X+b_{\text{opt}})\big)^2$, which is equal to

$$\iint\limits_{R^2}\big(y-(a_{\text{opt}}x+b_{\text{opt}})\big)^2\;f(x,y)\,dx\,dy$$

It can be shown that this integral is equal to

$$\sigma_2^2\left(1-r^2\right)$$

The expression $\sigma_2^2\left(1-r^2\right)$ consist of two factors. The first factor is the variance of the random variable Y. The second factor is $\left(1-r^2\right)$. Since the expected value of the squared error cannot be negative, $\left(1-r^2\right)$ cannot be negative, so r^2 cannot be larger than 1, that is, r is between -1 and 1, equality permitted. Moreover, $\left(1-r^2\right)$ is a decreasing function of r^2, so the larger r^2 is, the smaller $\left(1-r^2\right)$ is, that is, if r^2 is close to 1, then replacing Y by $a_{\text{opt}}X+b_{\text{opt}}$ causes, in most cases, a smaller error, if r^2 is close to 0, then replacing Y by $a_{\text{opt}}X+b_{\text{opt}}$ causes, in most cases, a larger error. This is why we may consider r^2 or $|r|$ as a measure of how well Y can be approximated by a linear function of X, that is, how strong the linear relationship is between X and Y.

Using a calculator. More sophisticated calculators have a key to determine the slope and the intercept of the regression line, as well.

Using Excel. In Excel, the command `SLOPE` (in Hungarian: `MEREDEKSÉG`), and `INTERCEPT` (in Hungarian: `METSZ`) give the slope and the intercept of the regression line.

Section 68
Confidence intervals

Construction of a finite confidence interval when σ is known. Let X be a normally distributed random variable with parameters μ and σ, and let $\overline{X}_n$ be the average of n experimental results. Since $\overline{X}_n$ follows a normal distribution with parameters μ and $\sigma/\sqrt{n}$, the standardized average

$$\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

follows the standard normal distribution. So

$$\mathbf{P}\left(-x < \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < x\right) = 2\Phi(x) - 1$$

If, for a given probability value p, we choose x so that

$$2\Phi(x) - 1 = p$$

that is,

$$x = \Phi^{-1}\left(\frac{1+p}{2}\right)$$

then

$$\mathbf{P}\left(-\Phi^{-1}\left(\frac{1+p}{2}\right) < \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < \Phi^{-1}\left(\frac{1+p}{2}\right)\right) = p$$

or, equivalently,

$$\mathbf{P}\left(\mu - \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right) < \overline{X}_n < \mu + \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right)\right) = p$$

which means that, with a probability p, both of the following inequalities hold:

$$\mu - \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right) < \overline{X}_n$$

$$\overline{X}_n < \mu + \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right)$$

The first inequality is equivalent to

$$\mu < \overline{X}_n + \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right)$$

The second is equivalent to

$$\overline{X}_n - \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right) < \mu$$

This is how we get that, with a probability p, both of the following inequalities hold:

$$\overline{X}_n - \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right) < \mu < \overline{X}_n + \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right)$$

which means that, with a probability p, the random interval

$$\left(\overline{X}_n - \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right), \overline{X}_n + \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right)\right)$$

called **confidence interval**, contains the parameter μ. The center of the interval is $\overline{X}_n$, the radius (the half of the length) of the interval is $\frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right)$. Notice that the radius (the half of the length) of the interval is a constant, that is, it does not depend on randomness.

The result we have can be interpreted also like this: the random point $\overline{X}_n$, with a probability p, is an estimation for μ: if μ is not known for us, then we are able to declare a random point based on experimental results, so that, with a probability p, this random point is so close to μ that their difference is less than $\frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right)$.

Construction of an infinitely long confidence interval when σ is known. Let X be again a normally distributed random variable with parameters μ and σ. Let $\overline{X}_n$ be the average if n experimental result. Since $\overline{X}_n$ follows a normal distribution with parameters μ and $\sigma/\sqrt{n}$, the standardized average

$$\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

follows the standard normal distribution. So

$$\mathbf{P}\left(\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < x\right) = \Phi(x)$$

If, for a given probability value p, we choose x so that $\Phi(x) = p$, that is, $x = \Phi^{-1}(p)$, then

$$\mathbf{P}\left(\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < \Phi^{-1}(p)\right) = p$$

or, equivalently,

$$\mathbf{P}\left(\overline{X}_n < \mu + \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right)\right) = p$$

which means that, with a probability p, the following inequality holds:

$$\overline{X}_n < \mu + \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right)$$

This inequality is equivalent to

$$\overline{X}_n - \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right) < \mu$$

This is how we get that, with a probability p, the following inequality holds:

$$\overline{X}_n - \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right) < \mu$$

which means that, with a probability p, the infinitely long random interval with the left end point

$$\overline{X}_n - \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right)$$

called **confidence interval**, contains the parameter μ.

The result we have can be interpreted also like this: the point

$$\overline{X}_n - \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right)$$

with a probability p, is a lower bound for μ: if μ is not known for us, then we are able to declare a random point based on experimental results, so that, with a probability p, this random point is less than μ.

Construction of a finite confidence interval when σ is not known. Let X be a normally distributed random variable with parameters μ and σ, and let $\overline{X}_n$ be the average of the experimental results $X_1, X_2, \ldots, X_n$. If σ is not known for us, then we may replace it by the sample standard deviation, which is

$$s_n^* = \sqrt{\frac{(X_1-\overline{X})^2 + (X_2-\overline{X})^2 + \ldots + (X_n-\overline{X})^2}{n-1}}$$

The random variable, which is a modification of the standardized average,

$$\frac{\overline{X}_n - \mu}{\frac{s_n^*}{\sqrt{n}}}$$

follows the t-distribution with degrees of freedom $n-1$. (Accept this fact without a proof.) So, using the distribution function $F(x)$ of the t-distribution with degrees of freedom $n-1$, we get that

$$\mathbf{P}\left(-x < \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < x\right) = 2F(x) - 1$$

If, for a given probability value p, we choose x so that

$$2F(x) - 1 = p$$

that is,

$$x = F^{-1}\left(\frac{1+p}{2}\right)$$

then

$$\mathbf{P}\left(-F^{-1}\left(\frac{1+p}{2}\right) < \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < F^{-1}\left(\frac{1+p}{2}\right)\right) = p$$

or, equivalently,

$$\mathbf{P}\left(\mu - \frac{s_n^*}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right) < \overline{X}_n < \mu + \frac{s_n^*}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right)\right) = p$$

which means that, with a probability p, both of the following inequalities hold:

$$\mu - \frac{s_n^*}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right) < \overline{X}_n$$

$$\overline{X}_n < \mu + \frac{s_n^*}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right)$$

The first inequality is equivalent to

$$\mu < \overline{X}_n + \frac{s_n^*}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right)$$

The second is equivalent to

$$\overline{X}_n - \frac{s_n^*}{\sqrt{n}}\Phi^{-1}\left(\frac{1+p}{2}\right) < \mu$$

This is how we get that, with a probability p, both of the following inequalities hold:

$$\overline{X}_n - \frac{s_n^*}{\sqrt{n}}F^{-1}\left(\frac{1+p}{2}\right) < \mu < \overline{X}_n + \frac{s_n^*}{\sqrt{n}}F^{-1}\left(\frac{1+p}{2}\right)$$

which means that, with a probability p, the random interval

$$\left(\overline{X}_n - \frac{s_n^*}{\sqrt{n}}F^{-1}\left(\frac{1+p}{2}\right),\ \overline{X}_n + \frac{s_n^*}{\sqrt{n}}F^{-1}\left(\frac{1+p}{2}\right)\right)$$

called **confidence interval**, contains the parameter μ. The center of the interval is $\overline{X}_n$, the radius (the half of the length) of the interval is

$$\frac{s_n^*}{\sqrt{n}} F^{-1}\left(\frac{1+p}{2}\right)$$

The radius (the half of the length) of the interval is now not a constant, but it depends on randomness.

The result we have can be interpreted also like this: the random point $\overline{X}_n$, with a probability p, is an estimation for μ: if μ is not known for us, then we are able to declare a random point based on experimental results, so that, with a probability p, this random point is so close to μ that their difference is less than $\frac{s_n^*}{\sqrt{n}} F^{-1}\left(\frac{1+p}{2}\right)$.

Files to study construction of confidence intervals:

Demonstration file: Finite confidence interval for the expected value, using standard normal distribution
300-23-00

Demonstration file: Infinitely long confidence interval for the expected value, using standard normal distribution
300-23-50

Demonstration file: Finite confidence interval for the expected value, using t-distribution
300-24-00

Section 69
U-tests

Six tests, called U-tests (also called Z-tests) will be discussed in this chapter. The six cases are:

1. U-test 1: Case of "less than", when n is given
2. U-test 2: Case of "less than", when n is calculated
3. U-test 3: Case of "equality", when n is given
4. U-test 4: Case of "equality", when n is calculated
5. U-test 5: Case of "equality", when an interval is considered instead of the point μ_0
6. U-test 6: Case of "two populations"

In all of these tests, except the last one, which is at the end of this chapter, X is a normally distributed random variable, $\overline{X}_n$ is the average of n experimental results for X.

U-test 1: Case of "less than", when n is given. X is a normally distributed random variable with parameters μ and σ, where σ is known, but μ is not known. μ_0 is a given value. On the basis of n experimental results for X (where n is given), we want to decide whether the hypothesis $\mu \leq \mu_0$ holds or does not hold. For a given p_0 probability value, which is (a little bit) less than 1, we require that

1. if $\mu < \mu_0$, then the probability of accepting the hypothesis is greater than p_0,
2. if $\mu = \mu_0$, then the probability of accepting the hypothesis is equal to p_0,
3. if $\mu > \mu_0$, then the probability of accepting the hypothesis is less than p_0,
4. if μ is "very large", then the probability of accepting the hypothesis is "very small".

Solution. For fixed μ, σ, n and b, the probability

$$\mathbf{P}\left(\overline{X}_n < b\right) = \Phi\left(\frac{b-\mu}{\frac{\sigma}{\sqrt{n}}}\right)$$

is equal to the area on the left side of b under the graph of the density function of $\overline{X}_n$. For fixed σ, n and b, this expression is a function of μ, which is called the **power function** of the U-test. The graph of the power function is constructed in the following Excel file. Study it:

Demonstration file: Power function: graph of $P\left(\overline{X}_n < b\right)$ as a function of μ
300-25-10

Now, for fixed σ, μ_0, p_0 (≈ 1) and n, we look for b so that

$$\Phi\left(\frac{b-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = p_0$$

Here is the solution to this equation:

$$\frac{b-\mu_0}{\frac{\sigma}{\sqrt{n}}} = \Phi^{-1}(p_0)$$

$$b = \mu_0 + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(p_0)$$

Since

1. for $\mu = \mu_0$, we have $\mathbf{P}\left(\overline{X}_n < b\right) = \Phi\left(\frac{b-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = p_0$,
2. the power function is a strictly decreasing function of μ, approaching 0 at ∞,

we get that the following test works as required.

U-test 1: Case of "less than", when n is given. *We take the average $\overline{X}_n$ of the experimental results, and compare it to the critical value b, which was determined above in terms of given parameter values. If $\overline{X}_n < b$, then we accept the hypothesis , otherwise we reject the hypothesis.*

Since the average of the experimental results follows the normal distribution with parameters μ and $\sigma/\sqrt{n}$, in the following simulation file, the average of the experimental results is directly simulated. The experimental results themselves are not simulated.

Demonstration file: U-test 1: Case of "less than", when n is given
300-25-20

Demonstration file: U-test 1: Case of "less than", when n is given, version B
300-25-30

Demonstration file: U-test 1: Case of "less than", when n is given, version C
300-25-40

U-test 2: Case of "less than", when n is calculated. X is a normally distributed random variable with parameters μ and σ, where σ is known, but μ is not known. $\mu_0 < \mu_1$ are given values. On the basis of n experimental results for X, we want to decide whether the hypothesis $\mu \le \mu_0$ holds or does not hold. Contrary to U-test 1, now n is not given. In this U-test, we have to determine n so that more requirements will be satisfied. Namely, for a given p_0 probability value, which is (a little bit) less than 1, and for a given p_1 (small) probability value, we require that

1. if $\mu < \mu_0$, then the probability of accepting the hypothesis is greater than p_0,
2. if $\mu = \mu_0$, then the probability of accepting the hypothesis is equal to p_0,
3. if $\mu = \mu_1$, then the probability of accepting the hypothesis is equal to p_1,
4. if $\mu > \mu_1$, then the probability of accepting the hypothesis is less than p_1,
5. if μ is "very large", then the probability of accepting the hypothesis is "very small".

Solution. Now we look for b and n so that

$$\Phi\left(\frac{b-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = p_0$$

$$\Phi\left(\frac{b-\mu_1}{\frac{\sigma}{\sqrt{n}}}\right) = p_1$$

Here is the solution to this system of equations:

$$\frac{b-\mu_0}{\frac{\sigma}{\sqrt{n}}} = \Phi^{-1}(p_0)$$

$$\frac{b-\mu_1}{\frac{\sigma}{\sqrt{n}}} = \Phi^{-1}(p_1)$$

$$b-\mu_0 = \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(p_0)$$

$$b-\mu_1 = \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(p_1)$$

$$\mu_1-\mu_0 = \frac{\sigma}{\sqrt{n}}\left(\Phi^{-1}(p_0)-\Phi^{-1}(p_1)\right)$$

$$n = \left(\frac{\sigma}{\mu_1 - \mu_0} \left(\Phi^{-1}(p_0) - \Phi^{-1}(p_1) \right) \right)^2$$

$$b = \mu_0 + \frac{\sigma}{\sqrt{n}} \Phi^{-1}(p_0)$$

$$= \mu_0 + (\mu_1 - \mu_0) \frac{\Phi^{-1}(p_0)}{\Phi^{-1}(p_0) - \Phi^{-1}(p_1)}$$

Since

1. for $\mu = \mu_0$, we have $\mathbf{P}\left(\overline{X}_n < b\right) = p_0$,
2. for $\mu = \mu_1$, we have $\mathbf{P}\left(\overline{X}_n < b\right) = p_1$,
3. the power function is a strictly decreasing function, approaching 0 at ∞,

we get that the following test works as required.

U-test 2: Case of "less than", when n is calculated. *We calculate n and b according to the above formulas, and then we take the average $\overline{X}_n$ of the experimental results, and compare it to b. If $\overline{X}_n < b$, then we accept the hypothesis , otherwise we reject the hypothesis.*

Since the average of the experimental results follows the normal distribution with parameters μ and $\sigma/\sqrt{n}$, in the following simulation file, the average of the experimental results is directly simulated. The experimental results themselves are not simulated.

Demonstration file: U-test 2: Case of "less than", when n is calculated
300-26-00

Remark. In both of the above tests, we have to compare the average $\overline{X}_n$ of the experimental results to b, that is, we have to analyze whether the inequality

$$\overline{X}_n < \mu_0 + \frac{\sigma}{\sqrt{n}} \Phi^{-1}(p_0)$$

holds or does not hold. This inequality is equivalent to the following inequality:

$$\frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} < \Phi^{-1}(p_0)$$

The expression on the left side of this inequality is called the U-value of the tests:

$$U = \frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Using the notion of the U-value, both of the above tests may be performed so that we calculate the U-value, and compare it the so called **standardized critical value**

$$U_{\text{crit}} = \Phi^{-1}(p_0)$$

Notice that the standardized critical value U_{crit} depends only on the probability p_0.

U-test 1 and 2, cases of "less than" - standardized. *We determine U from the experimental results, and we calculate U_{crit}, and then compare them. If $U < U_{crit}$, then we accept the hypothesis, otherwise we reject the hypothesis.*

U-test 3: Case of "equality", when n is given. X is a normally distributed random variable with parameters μ and σ, where σ is known, but μ is not known. μ_0 is a given value. On the basis of n experimental results for X (where n is given), we want to decide whether the hypothesis $\mu = \mu_0$ holds or does not hold. For a given p_0 probability value, which is (a little bit) less than 1, we require that

1. if $\mu = \mu_0$, then the probability of accepting the hypothesis is equal to p_0,
2. if $\mu \neq \mu_0$, then the probability of accepting the hypothesis is less than p_0,
3. if μ is "farther and farther from μ_0", then the probability of accepting the hypothesis is "smaller and smaller",
4. if μ is "very far from μ_0", then the probability of accepting the hypothesis is "very small".

Solution. For fixed μ, σ, n, a and b, the probability

$$\mathbf{P}\left(a < \overline{X}_n < b\right) = \Phi\left(\frac{b-\mu}{\frac{\sigma}{\sqrt{n}}}\right) - \Phi\left(\frac{a-\mu}{\frac{\sigma}{\sqrt{n}}}\right)$$

is equal to the area between a and b under the graph of the density function of $\overline{X}_n$. For fixed σ, n, a, b, this expression defines a function of μ, which is called the power function of the U-test. The graph of the power function is constructed in the following Excel file. Study it:

Demonstration file: Power function: graph of $P\left(a < \overline{X}_n < b\right)$ as a function of μ
300-27-10

For a given μ_0, and Δb, the numbers $a = \mu_0 - \Delta b$ and $b = \mu_0 + \Delta b$ define a symmetrical interval around μ_0, and when $\mu = \mu_0$, then

$$\mathbf{P}\left(\mu_0 - \Delta b < \overline{X}_n < \mu_0 + \Delta b\right) = 2\Phi\left(\frac{\Delta b}{\frac{\sigma}{\sqrt{n}}}\right) - 1$$

Now we look for Δb so that

$$2\Phi\left(\frac{\Delta b}{\frac{\sigma}{\sqrt{n}}}\right) - 1 = p_0$$

Here is the solution to this equation:

$$\Delta b = \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}\left(\frac{1+p_0}{2}\right)$$

Since

1. for $\mu = \mu_0$, we have $\mathbf{P}\left(\mu_0 - \Delta b < \overline{X}_n < \mu_0 + \Delta b\right) = p_0$,
2. for$\mu \neq \mu_0$, we have $\mathbf{P}\left(\mu_0 - \Delta b < \overline{X}_n < \mu_0 + \Delta b\right) < p_0$,
3. the power function is strictly increasing on the left side of μ_0, and strictly decreasing on the right side of μ_0, and it is approaching 0 both at $-\infty$ and at ∞,

we get that the following test works as required.

U-test 3: Case of "equality", when n is given. *We take the average $\overline{X}_n$ of the experimental results, and compare it to the critical values $\mu_0 - \Delta b$, $\mu_0 + \Delta b$. If $\overline{X}_n$ is between them, then we accept the hypothesis , otherwise we reject the hypothesis.*

Since the average of the experimental results follows the normal distribution with parameters μ and $\sigma/\sqrt{n}$, in the following simulation file, the average of the experimental results is directly simulated. The experimental results themselves are not simulated.

Demonstration file: U-test 3: Case of "equality", when n is given
300-27-20

Demonstration file: U-test 3: Case of "equality", when n is given, version B
300-27-30

Demonstration file: U-test 3: Case of "equality", when n is given, version C
300-27-40

U-test 4: Case of "equality", when n is calculated. X is a normally distributed random variable with parameters μ and σ, where σ is known, but μ is not known. $\mu_0 < \mu_1$ are given values. On the basis of n experimental results for X, we want to decide whether the hypothesis $\mu = \mu_0$ holds or does not hold. Contrary to U-test 3, now n is not given, we have to determine n so that, in this U-test, more requirements will be satisfied. Namely, for a given p_0 probability value, which is (a little bit) less than 1, and for a given p_1 (small) probability value, we require that

1. if $\mu = \mu_0$, then the probability of accepting the hypothesis is equal to p_0,
2. if $\mu = \mu_1$, then the probability of accepting the hypothesis is equal to p_1,
3. if μ is "farther and farther from μ_0", then the probability of accepting the hypothesis is "smaller and smaller",
4. if μ is "very far from μ_0", then the probability of accepting the hypothesis is "very small".

Solution. The probability

$$\mathbf{P}\left(a < \overline{X}_n < b\right) = \Phi\left(\frac{b-\mu}{\frac{\sigma}{\sqrt{n}}}\right) - \Phi\left(\frac{a-\mu}{\frac{\sigma}{\sqrt{n}}}\right)$$

is equal to the area between a and b under the graph of the density function of $\overline{X}_n$. For fixed σ, a and b and n, this expression is a function of μ, the so called power function of the U-test. The power function obviously increases until the center of the interval $[a,b]$, and then decreases, and approaches 0 both at $-\infty$ and ∞. For a given μ_0, and Δb, the numbers $a = \mu_0 - \Delta b$ and $b = \mu_0 + \Delta b$ define a symmetrical interval around μ_0, and

$$\mathbf{P}\left(\mu_0 - \Delta b < \overline{X}_n < \mu_0 + \Delta b\right) = \Phi\left(\frac{(\mu_0 + \Delta b) - \mu}{\frac{\sigma}{\sqrt{n}}}\right) - \Phi\left(\frac{(\mu_0 - \Delta b) - \mu}{\frac{\sigma}{\sqrt{n}}}\right)$$

When $\mu = \mu_0$, then

$$\mathbf{P}\left(\mu_0 - \Delta b < \overline{X}_n < b = \mu_0 + \Delta b\right) = 2\Phi\left(\frac{\Delta b}{\frac{\sigma}{\sqrt{n}}}\right) - 1$$

Now we look for Δb and n so that

$$2\Phi\left(\frac{\Delta b}{\frac{\sigma}{\sqrt{n}}}\right) - 1 = p_0$$

$$\Phi\left(\frac{(\mu_0 + \Delta b) - \mu_1}{\frac{\sigma}{\sqrt{n}}}\right) - \Phi\left(\frac{(\mu_0 - \Delta b) - \mu_1}{\frac{\sigma}{\sqrt{n}}}\right) = p_1$$

We can easily handle the first equation, and we get that

$$\Delta b = \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}\left(\frac{1+p_0}{2}\right)$$

which is a simple linear relation between Δb and $\frac{1}{\sqrt{n}}$. In regards to the second equation, let us notice that $(\mu_0 - \Delta b) - \mu_1$ is a negative number far enough from 0, so

$$\Phi\left(\frac{(\mu_0 - \Delta b) - \mu_1}{\frac{\sigma}{\sqrt{n}}}\right) \approx 0$$

and, omitting this term in the second equation, we get the following approximate equation:

$$\Phi\left(\frac{(\mu_0 + \Delta b) - \mu_1}{\frac{\sigma}{\sqrt{n}}}\right) = p_1$$

From here we get:

$$(\mu_0 + \Delta b) - \mu_1 = \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(p_1)$$

which is another simple linear relation between Δb and $\frac{1}{\sqrt{n}}$. The two linear equations constitute a system of linear equations for Δb and $\frac{1}{\sqrt{n}}$, which can be solved. From the solution we get n and Δb:

$$n = \left(\frac{\sigma}{\mu_1 - \mu_0} \left(\Phi^{-1}\left(\frac{1+p_0}{2} \right) - \Phi^{-1}(p_1) \right) \right)^2$$

$$\Delta b = (\mu_1 - \mu_0) \frac{\Phi^{-1}\left(\frac{1+p_0}{2} \right)}{\Phi^{-1}\left(\frac{1+p_0}{2} \right) - \Phi^{-1}(p_1)}$$

So the following test works as required.

U-test 4: Case of "equality", when n is calculated. *We calculate the value of n from the above formula, and round up what we get. We calculate Δb, too. Then we make n experiments for X, take the average $\overline{X}_n$ of the experimental results, and compare it to the critical values $\mu_0 - \Delta b$ and $\mu_0 + \Delta b$. If $\overline{X}_n$ is between the critical values, then we accept the hypothesis , otherwise we reject the hypothesis.*

Since the average of the experimental results follows the normal distribution with parameters μ and $\sigma/\sqrt{n}$, in the following simulation file, the average of the experimental results is directly simulated. The experimental results themselves are not simulated.

Demonstration file: U-test 4: Case of "equality", when n is calculated
300-28-00

Remark. In Test 3 and Test 4, we have to compare the average $\overline{X}_n$ of the experimental results to $\mu_0 - \Delta b$ and $\mu_0 + \Delta b$, that is, we have to analyze whether the inequalities

$$\mu_0 - \frac{\sigma}{\sqrt{n}} \Phi^{-1}\left(\frac{1+p_0}{2} \right) < \overline{X}_n < \mu_0 + \frac{\sigma}{\sqrt{n}} \Phi^{-1}\left(\frac{1+p_0}{2} \right)$$

hold or do not hold. These inequalities are equivalent to the following inequalities:

$$-\Phi^{-1}\left(\frac{1+p_0}{2} \right) < \frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} < \Phi^{-1}\left(\frac{1+p_0}{2} \right)$$

The expression in the middle of these inequalities is called the U-value of the tests:

$$U = \frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

and the expression on the right side is the so called **standardized critical value**:

$$U_{\text{crit}} = \Phi^{-1}\left(\frac{1+p_0}{2} \right)$$

Notice that the standardized critical value U_{crit} depends only on the probability p_0, but U_{crit} for Test 3 and Test 4 is not the same as U_{crit} for Test 1 and Test 2. Using the notions of U and U_{crit}, we may write the above inequalities like this:

$$-U_{\text{crit}} < U < U_{\text{crit}}$$

or, equivalently,

$$|U| < U_{\text{crit}}$$

We see that Test 3 and Test 4 may be performed so that we calculate the absolute value of the U-value, and compare it the standardized critical value.

U-test 3 and 4, cases of "equality" - standardized. *We determine U from the experimental results, and we calculate U_{crit}, and then we compare the absolute value of U to U_{crit}. If $|U| < U_{crit}$, then we accept the hypothesis , otherwise we reject the hypothesis.*

U-test 5: Case of "equality", when an interval is considered instead of a point. X is a normally distributed random variable with parameters μ and σ, where σ is known, but μ is not known. $\mu_0 < \mu_1 < \mu_2$ are given values. Let μ_1^T and μ_2^T mean the points on the left side of μ_0 which we get if μ_1 and μ_2 are reflected about μ_0. On the basis of n experimental results for X, we want to decide whether the hypothesis $\mu = \mu_0$ holds or does not hold. Similar to U-test 4, n is not given. We have to determine n so that, in this U-test, more requirements will be satisfied. Namely, for a given p_1 probability value, which is (a little bit) less than 1, and for a given p_2 (small) probability value, we require that

1. if $\mu_1^T < \mu < \mu_1$, then the probability of accepting the hypothesis is greater than p_1,
2. if $\mu < \mu_2^T$ or $\mu > \mu_2$, then the probability of accepting the hypothesis is smaller than p_2,
3. if μ is "farther and farther from μ_0", then the probability of accepting the hypothesis is "smaller and smaller",
4. if μ is "very far from μ_0", then the probability of accepting the hypothesis is "very small".

Remark. The first item of the above list of requirements may serve as an explanation for the name of this U-test, since the interval $[\mu_1^T, \mu_1]$ is considered instead of the point μ_0.

Solution. We will try to determine b and its reflection b^T about μ_0, and n so that

$$\Phi\left(\frac{b-\mu_1}{\frac{\sigma}{\sqrt{n}}}\right)-\Phi\left(\frac{b^T-\mu_1}{\frac{\sigma}{\sqrt{n}}}\right)=p_1$$

$$\Phi\left(\frac{b-\mu_2}{\frac{\sigma}{\sqrt{n}}}\right)-\Phi\left(\frac{b^T-\mu_2}{\frac{\sigma}{\sqrt{n}}}\right)=p_2$$

Since $b^T-\mu_1$, $b^T-\mu_2$ are negative numbers far enough from 0,

$$\Phi\left(\frac{b^T-\mu_1}{\frac{\sigma}{\sqrt{n}}}\right)\approx 0$$

$$\Phi\left(\frac{b^T-\mu_2}{\frac{\sigma}{\sqrt{n}}}\right)\approx 0$$

and we get the approximate equations:

$$\Phi\left(\frac{b-\mu_1}{\frac{\sigma}{\sqrt{n}}}\right)=p_1$$

$$\Phi\left(\frac{b-\mu_2}{\frac{\sigma}{\sqrt{n}}}\right)=p_2$$

Taking the inverse of the function Φ, we get that

$$b-\mu_1=\frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(p_1)$$

$$b-\mu_2=\frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(p_2)$$

which is system of linear equations for $\frac{1}{\sqrt{n}}$ and b. The solution for n and b is:

$$n=\left(\frac{\sigma}{\mu_2-\mu_1}\,\left(\Phi^{-1}(p_1)-\Phi^{-1}(p_2)\right)\right)^2$$

$$b=\mu_1+(\mu_2-\mu_1)\frac{\Phi^{-1}(p_1)}{\Phi^{-1}(p_1)-\Phi^{-1}(p_2)}$$

So the following test works as required.

U-test 5: Case of "equality", when an interval is considered instead of a point. *We calculate the value of n from the above formula, and round up what we get. We calculate b, too. Then we make n experiments for X, take the average $\overline{X}_n$ of the experimental results, and compare*

it to the critical values b^T and b. If $\overline{X}_n$ is between the critical values, then we accept the hypothesis , otherwise we reject the hypothesis.

Since the average of the experimental results follows the normal distribution with parameters μ and $\sigma/\sqrt{n}$, in the following simulation file, the average of the experimental results is directly simulated. The experimental results themselves are not simulated.

Demonstration file: U-test 5: Case of "equality", when an interval is considered instead of the point μ_0
300-29-00

U-test 6: Case of two populations. X_1 is a normally distributed random variable with parameters μ_1 and σ_1, X_2 is a normally distributed random variable with parameters μ_2 and σ_2. We assume that σ_1 and σ_2 are known for us, but μ_1 and μ_2 are not. On the basis of n_1 experimental results for X_1 and n_2 experimental results for X_2 (n_1 and n_2 are given), we want to decide whether the hypothesis $\mu_1 = \mu_2$ holds or does not hold. For a given p_0 probability value, which is (a little bit) less than 1, we require that

1. if $\mu_1 = \mu_2$, then the probability of accepting the hypothesis is equal to p_0,
2. if $\mu_1 \neq \mu_2$, then the probability of accepting the hypothesis is less than p_0,
3. if μ_1 is "farther and farther from μ_2", then the probability of accepting the hypothesis is "smaller and smaller",
4. if μ_1 is "very far from μ_2", then the probability of accepting the hypothesis is "very small".

Solution. The average of the n_1 experimental results for X_1 is $(\overline{X_1})_{n_1}$, and the average of the n_2 experimental results for X_2 is $(\overline{X_2})_{n_2}$. Let us consider the difference of the averages: $(\overline{X_1})_{n_1} - (\overline{X_2})_{n_2}$ The expected value of this difference is $\mu_1 - \mu_2$, its standard deviation is

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

If the random variable U is defined like this:

$$U = \frac{(\overline{X_1})_{n_1} - (\overline{X_2})_{n_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

then the expected value of U is obviously

$$\frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

and its standard deviation is 1. Thus, if $\mu_1 = \mu_2$, then U follows the standard normal distribution, which means that

$$\mathbf{P}(-x < U < x) = 2\Phi(x) - 1$$

Let us choose x so that $2\Phi(x) - 1 = p_0$, that is, $x = \Phi^{-1}\left(\frac{1+p_0}{2}\right)$. Let us take this x value as a critical value U_{crit}. Thus, if $\mu_1 = \mu_2$, then we have

$$\mathbf{P}(-U_{\text{crit}} < U < U_{\text{crit}}) = p_0$$

If $\mu_1 \neq \mu_2$, then the distribution of U differs from the standard normal distribution, and

$$\mathbf{P}(-U_{\text{crit}} < U < U_{\text{crit}}) < p_0$$

If μ_1 and μ_2 are farther and farther from each other, then the distribution of U becomes more and more different from the standard normal distribution, and

$$\mathbf{P}(-U_{\text{crit}} < U < U_{\text{crit}})$$

becomes smaller and smaller. Thus, the following test works as required.

U-test 6: Case of two populations. *We determine the value of U from the experimental results, and compare its absolute value to the critical value U_{crit}. If $|U|$ is less than U_{crit}, then we accept the hypothesis , otherwise we reject the hypothesis.*

Important remark: U-tests for NOT normally distributed random variables. At the beginning of this chapter, we assumed that X was a normally distributed random variable. However, if n, the number of experiments is large enough (say it is at least 25, or so), then $\overline{X}_n$, the average of the experimental results for X is approximately normally distributed even if X is not normally distributed. Since in the above tests we were using the normality of $\overline{X}_n$, the above tests are applicable for not normal random variables, as well, if n, the number of experiments is large enough.

Files to study the U-test for "two populations":

Demonstration file: U-test, two populations (Version B)
300-29-40

Demonstration file: U-test, two populations (Version C)
300-29-50

Section 70
*** T-tests

Tests, called T-tests will be discussed in this chapter. In all of them, X is a normally distributed random variable, $\overline{X}_n$ is the average of experimental results $X_1, X_2, \ldots, X_n$, and

$$s_n^* = \sqrt{\frac{(X_1 - \overline{X}_n)^2 + (X_2 - \overline{X}_n)^2 + \ldots + (X_n - \overline{X}_n)^2}{n-1}}$$

is the sample standard deviation.

T-test 1: Case of "equality". X is a normally distributed random variable with parameters μ and σ, where neither μ nor σ is known for us. μ_0 is a given value. On the basis of n experimental results for X (where n is given), we want to decide whether the hypothesis $\mu = \mu_0$ holds or does not hold. For a given p_0 probability value, which is (a little bit) less than 1, we require that

1. if $\mu = \mu_0$, then the probability of accepting the hypothesis is equal to p_0,
2. if $\mu \neq \mu_0$, then the probability of accepting the hypothesis is less than p_0,
3. if μ is "farther and farther from μ_0", then the probability of accepting the hypothesis is "smaller and smaller",
4. if μ is "very far from μ_0", then the probability of accepting the hypothesis is "very small".

Solution. If we knew σ, we could use a standardized U-test (see the remark after U-test 2), and we could calculate

$$U = \frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Since σ is not known for us, we have to replace σ by s_n^*, the sample standard deviation. This why we consider the random variable

$$T = \frac{\overline{X}_n - \mu_0}{\frac{s_n^*}{\sqrt{n}}}$$

Based on experimental results, the value of T can be calculated. If $\mu = \mu_0$, then the random variable T follows t-distribution with degrees of freedom $n-1$, so

$$\mathbf{P}(-b < T < b) = 2F(b) - 1$$

where $F(x)$ denotes the distribution function of the t-distribution with degrees of freedom $n-1$. We choose b so that $2F(b)-1 = p_0$, that is, $b = F^{-1}(\frac{1+p_0}{2})$. This b-value will be called the critical value, and will be denoted by T_{crit}:

$$T_{\text{crit}} = F^{-1}(\frac{1+p_0}{2})$$

Thus, if $\mu = \mu_0$, then we get:

$$\mathbf{P}(-T_{\text{crit}} < T < T_{\text{crit}}) = p_0$$

If $\mu \neq \mu_0$, then the random variable T follows a distribution different from t-distribution with degrees of freedom $n-1$, so that

$$\mathbf{P}(-T_{\text{crit}} < T < T_{\text{crit}}) < p_0$$

If μ is farther and farther from μ_0, then the distribution of the the random variable T becomes more and more different from t-distribution with degrees of freedom $n-1$, and

$$\mathbf{P}(-T_{\text{crit}} < T < T_{\text{crit}})$$

becomes smaller and smaller. We shall not go into the mathematical details of this test. However, we hope that the simulation files given bellow will help to accept that the following test works as required.

T-test 1: Case of "equality". *We calculate the value of T from the experimental results, and compare its absolute value to the critical value T_{crit}. If $|T|$ is less than T_{crit}, then we accept the hypothesis , otherwise we reject the hypothesis.*

Files to study the T-test for "equality":

Demonstration file: T-test, equality, sample average, 1 experiment (version A)
300-31-00

Demonstration file: T-test, equality, sample average, 1000 experiments (version A)
300-32-00

Demonstration file: T-test, equality, sample average, 1 experiment (version B, no figures)
300-33-00

Demonstration file: T-test, equality, sample average, 1 experiment (version B with figures)
300-34-00

T-test 2: Case of "less than". X is a normally distributed random variable with parameters μ and σ, where neither μ nor σ is known for us. μ_0 is a given value. On the basis of n experimental results for X (where n is given), we want to decide whether the hypothesis $\mu \leq \mu_0$ holds or does not hold. For a given p_0 probability value, which is (a little bit) less than 1, we require that

1. if $\mu < \mu_0$, then the probability of accepting the hypothesis is greater than p_0,
2. if $\mu = \mu_0$, then the probability of accepting the hypothesis is equal to p_0,
3. if $\mu > \mu_0$, then the probability of accepting the hypothesis is less than p_0,
4. if μ is "very large", then the probability of accepting the hypothesis is "very small".

Solution. If we knew σ, we could use a standardized U-test (see the remark after U-test 4), and we could calculate

$$U = \frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Since σ is not known for us, we have to replace σ by s_n^*, the sample standard deviation. This why we consider the random variable

$$T = \frac{\overline{X}_n - \mu_0}{\frac{s_n^*}{\sqrt{n}}}$$

Based on experimental results, the value of T can be calculated. If $\mu = \mu_0$, then the random variable T follows t-distribution with degrees of freedom $n-1$, so

$$\mathbf{P}(T < b) = F(b)$$

where $F(x)$ denotes the distribution function of the t-distribution with degrees of freedom $n-1$. We choose b so that $F(b) = p_0$, that is, $b = F^{-1}(p_0)$. This b-value will be called critical value, and will be denoted by T_{crit}:

$$T_{\text{crit}} = F^{-1}(p_0)$$

Thus, if $\mu = \mu_0$, then we get:

$$\mathbf{P}(T < T_{\text{crit}}) = p_0$$

If $\mu < \mu_0$, then the random variable T follows a distribution different from t-distribution with degrees of freedom $n-1$, so that

$$\mathbf{P}(T < T_{\text{crit}}) > p_0$$

If $\mu > \mu_0$, then the random variable T follows a distribution different from t-distribution with degrees of freedom $n-1$, so that

$$\mathbf{P}(T < T_{\text{crit}}) < p_0$$

If μ is very large, then $\mathbf{P}(T < T_{\text{crit}}) < p_0$ becomes very small. We shall not go into the mathematical details of this test. However, we hope that the simulation files given bellow will help to accept that the following test works as required.

T-test 2: Case of "less than". *We calculate the value of T from the experimental results, and compare its absolute value to the critical value T_{crit}. If T is less than T_{crit}, then we accept the hypothesis , otherwise we reject the hypothesis.*

Files to study the T-test for "less-than":

Demonstration file: T-test, less-than, sample average, 1 experiment (version A)
300-37-00

Demonstration file: T-test, less-than, sample average, 1000 experiments (version A)
300-38-00

Demonstration file: T-test, less-than, sample average, 1 experiment (version B, no figures)
300-39-00

Demonstration file: T-test, less-than, sample average, 1 experiment (version B with figures)
300-40-00

Section 71
*** Chi-square-test for fitness

Imagine that you need a fair die, and your friend offers you one. You are glad to get a die, but you want to be convinced that the die is really fair. So you toss the die several times, and you get experimental results. How can you decide based on the experimental results whether to accept the hypothesis that the die is fair or to reject it? We will describe a test to decide whether to accept the hypothesis or to reject it. Obviously, since the test is based on experimental results, the decision may be wrong: even if the die is fair, randomness may cause us to reject the hypothesis, so even if the die is fair, the probability of accepting the hypotheses must be less than 1. Let this probability be denoted by p_0. This is why, for a given p_0 probability value, which is (a little bit) less than 1, we require that

1. if the hypothesis holds, then the test, with a probability p_0, will suggest to accept the hypothesis,
2. if the hypothesis does not hold, then the test will suggest to accept the hypothesis with a probability smaller than p_0,
3. if the die differs from a fair die only a "little bit", then the test will suggest to accept the hypothesis with a probability only a "little bit" smaller than p_0,
4. if the die is "far" from a fair die, then the test will suggest to accept the hypothesis with a "small" probability.

Solution. When you toss the die, and observe the number on the top, then the possible values are 1, 2, 3, 4, 5, 6. Make n tosses. Observe the count for each possible value, that is, how many times you have got a 1, how many times you have got a 2, and so on. Then calculate the so called expected counts, too, which means that the hypothetical probability of each possible value is multiplied by n. Then, for each possible value, take the difference between the observed count and the expected count, take the square of the difference, and then divide by the expected count. For each possible value you get a number. Now add these numbers. This sum will be denoted by K^2. Since the value of K^2 is affected by the experimental results, K^2 is a random variable. Suppose that the number of tosses was large enough to guarantee that all the expected counts are greater than 10. If this condition is not fulfilled, then this test is not

applicable. If this condition is fulfilled, then the random variable K^2 approximately follows the distribution called chi-square distribution with degrees of freedom $r-1$, where r is the number of possible values for the die. For a usual die, $r=6$. If $F(x)$ denotes the distribution function of the chi-square distribution with degrees of freedom $r-1$, then $\mathbf{P}(K^2<x)=F(x)$. Let us choose x so that $F(x)=p_0$, that is, $x=F^{-1}(p_0)$. This x value will be denoted by K^2_{crit}, and will be called the critical value of the test. We do not go into the mathematical details, just state that the test given here works properly. Playing with the Excel files, given below, you may learn the algorithm of the test again, and you may have an experience that the test really works as desired.

Chi-square-test for fitness. *We calculate the value of K^2 from the experimental results, and compare it to the critical value K^2_{crit}. If K^2 is less than K^2_{crit}, then we accept the hypothesis , otherwise we reject the hypothesis.*

Files to study chi-square-test for fitness:

Demonstration file: Chi-square-test for fitness (version B, no figures)
300-50-00

Demonstration file: Chi-square-test for fitness (version C, with figures)
300-51-00

Section 72
*** Chi-test for standard deviation (Chi-square-test for variance)

X is a normally distributed random variable with parameters μ and σ, where neither μ nor σ is known for us. σ_0 is a given value. On the basis of n experimental results for X (where n is given), we want to decide whether the hypothesis $\sigma = \sigma_0$ holds or does not hold. For a given p_0 probability value, which is (a little bit) less than 1, we require that

1. if $\sigma = \sigma_0$, then the probability of accepting the hypothesis is equal to p_0,
2. if $\sigma \neq \sigma_0$, then the probability of accepting the hypothesis is less than p_0,
3. if σ is "farther and farther from σ_0", then the probability of accepting the hypothesis is "smaller and smaller",
4. if σ is "very far from σ_0", then the probability of accepting the hypothesis is "very small".

Solution. Let us take the sample standard deviation, which is

$$s_n^* = \sqrt{\frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{n-1}}$$

and let the random variable K be the sample standard deviation divided by the hypothetical standard deviation σ_0:

$$K = \frac{s_n^*}{\sigma_0}$$

If the hypothesis holds, then this random variable follows the distribution called chi-distribution with degrees of freedom $n-1$. If $F(x)$ denotes the distribution function of the chi-distribution with degrees of freedom $n-1$, then $\mathbf{P}(K < x) = F(x)$. Let us choose x so that $F(x) = p_0$, that is, $x = F^{-1}(p_0)$. This x value will be denoted by K_{crit}, and will be called the critical value of the test. We do not go into the mathematical details, just state that the

test given here works properly. Playing with the Excel files, given below, you may learn the algorithm of the test again, and you may have an experience that the test really works as desired.

Chi-test for standard deviation (Chi-square-test for variance). *We calculate the value of K from the experimental results, and compare it to the critical value K_{crit}. If K is less than K_{crit}, then we accept the hypothesis , otherwise we reject the hypothesis.*

Remark. If a random variable follows a chi-distribution with degrees of freedom d, then its square follows chi-square distribution with degrees of freedom d. This is why this test is applicable to test not only the standard deviation, but the variance. If we test the variance and use the sample variance $(s_n^*)^2$, then the critical value K^2_{crit} for $(s_n^*)^2$ should be chosen using the distribution function of the chi-square distribution.

Files to study Chi-test for standard deviation (Chi-square-test for variance):

Demonstration file: Chi-test for standard deviation (Chi-square-test for variance), 1 experiment (version A)
300-52-00

Demonstration file: Chi-test for standard deviation (Chi-square-test for variance), 1000 experiments (version A)
300-53-00

Demonstration file: Chi-test for standard deviation (Chi-square-test for variance) (version B, no figures)
300-54-00

Demonstration file: Chi-test for standard deviation (Chi-square-test for variance) (version C, with figures)
300-55-00

Section 73
*** F-test for equality of variances (of standard deviations)

X_1 is a normally distributed random variable with parameters μ_1 and σ_1, X_2 is a normally distributed random variable with parameters μ_2 and σ_2. The parameters μ_1, μ_2, μ_1, μ_2 are not known for us. On the basis of n_1 experimental results for X_1 and n_2 experimental results for X_2 (n_1 and n_2 are given), we want to decide whether the hypothesis $\sigma_1 = \sigma_2$ holds or does not hold. For a given p_0 probability value, which is (a little bit) less than 1, we require that

1. if $\sigma_1 = \sigma_2$, then the probability of accepting the hypothesis is equal to p_0,
2. if $\sigma_1 \neq \sigma_2$, then the probability of accepting the hypothesis is less than p_0,
3. if σ_1 is "farther and farther from σ_2", then the probability of accepting the hypothesis is "smaller and smaller",
4. if σ_1 is "very far from σ_2", then the probability of accepting the hypothesis is "very small".

Solution. Let the sample standard deviation for X_1 be s_1^*, and the sample standard deviation for X_2 be s_2^*. Their squares are the so called sample variances: $(s_1^*)^2$ and $(s_2^*)^2$. Let us take the quotient of the sample variances:

$$F = \frac{(s_1^*)^2}{(s_2^*)^2}$$

If the hypothesis holds, then the random variable F follows the distribution called F-distribution with degrees of freedom $n_1 - 1, n_2 - 1$. Using (the inverse of) the distribution function of the F-distribution with degrees of freedom $n_1 - 1, n_2 - 1$, we choose two critical values: $F_{\text{crit(lower)}}$ and $F_{\text{crit(upper)}}$ so that

$$\mathbf{P}(F < F_{\text{crit(lower)}}) = \frac{1 - p_0}{2}$$

$$\mathbf{P}(F < F_{\text{crit(upper)}}) = \frac{1+p_0}{2}$$

With this choice of $F_{\text{crit(lower)}}$ and $F_{\text{crit(upper)}}$, we achieve that

$$\mathbf{P}(F_{\text{crit(lower)}} < F < F_{\text{crit(upper)}}) = p_0$$

We do not go into the mathematical details, just state that the test given here works properly. Playing with the Excel files, given below, you may learn the algorithm of the test again, and you may have an experience that the test really works as desired.

F-test for equality of standard deviations (F-test for equality of variances). *We calculate the value of F from the experimental results, and compare it to the critical values $F_{crit(lower)}$ and $F_{crit(upper)}$. If F is between them, then we accept the hypothesis , otherwise we reject the hypothesis.*

F-tests for variances (standard deviation):

Demonstration file: F-test for equality of variances (of standard deviations) (version B, no figures)
300-56-00

Demonstration file: F-test for equality of variances (of standard deviations) (version B with figures)
300-57-00

Demonstration file: F-test for "less-than" of variances (of standard deviations) (version B, no figures)
300-58-00

Demonstration file: F-test for "less-than" of variances (of standard deviations) (version B with figures)
300-59-00

Section 74
*** Test with ANOVA (Analysis of variance)

We have r normally distributed random variables: $X_1, X_2, \ldots, X_r$, which have a common standard deviation σ, and possibly different expected values: $\mu_1, \mu_2, \ldots, \mu_r$. We make a certain number of experiments for each: n_1 experiments for X_1, n_2 experiments for X_2, and so on, n_r experiments for X_r. On the basis of these experimental results, we want to decide whether the hypothesis $\mu_1 = \mu_2 = \ldots = \mu_r$ holds or does not hold. For a given p_0 probability value, which is (a little bit) less than 1, we require that

1. if $\mu_1 = \mu_2 = \ldots = \mu_r$, then the probability of accepting the hypothesis is equal to p_0,
2. if $\mu_1 = \mu_2 = \ldots = \mu_r$ is not true, then the probability of accepting the hypothesis is less than p_0,
3. if $\mu_1, \mu_2, \ldots, \mu_r$ are farther and farther from each other, then the probability of accepting the hypothesis is "smaller and smaller",
4. if $\mu_1, \mu_2, \ldots, \mu_r$ are "very far from each other", then the probability of accepting the hypothesis is "very small".

Solution. We remind the reader that, for a data-set $z_1, z_2, \ldots, z_N$, the average is

$$\overline{z} = \frac{z_1 + z_2 + \ldots + z_N}{N}$$

and the variance is

$$\frac{(z_1 - \overline{z})^2 + (z_2 - \overline{z})^2 + \ldots + (z_N - \overline{z})^2}{N}$$

The n_1 experiments for X_1 constitute a data-set. Its average will be denoted by Ave_1, its variance will be denoted by Var_1. Related to the random variable X_i, we get similarly the

average Ave_i and the variance Var_i $(i = 1,2,\ldots,r)$. The number of all experiments is $n = \sum_i n_i$. The proportion of the ith data-set is $p_i = n_i/n$ $(i = 1,2,\ldots,r)$. The quantity

$$\text{AVE of Ave} = \sum_i \text{Ave}_i \; p_i$$

will be called the (weighted) average of the averages, and

$$\text{AVE of Var} = \sum_i \text{Var}_i \; p_i$$

will be called the (weighted) average of the variances, and

$$\text{VAR of Ave} = \sum_i (\text{Ave}_i - \text{AVE})^2 \; p_i$$

will be called the (weighted) variance of the averages.The random variable F is now defined by

$$F = \frac{\frac{\text{VAR of Ave}}{r-1}}{\frac{\text{AVE of Var}}{n-1}}$$

If the hypothesis holds, then the random variable F follows the distribution called F-distribution with degrees of freedom $r-1, n-1$. Using (the inverse of) the distribution function of the F-distribution with degrees of freedom $r-1, n-1$, we choose the critical value F_{crit} so that

$$\mathbf{P}(F < F_{\text{crit}}) = p_0$$

We do not go into the mathematical details, just state that the test given below works properly. Playing with the Excel file, given below, you may learn the algorithm of the test again, and you may have an experience that the test really works as desired.

Test with ANOVA (Analysis of variance). *We calculate the value of F from the experimental results, and compare F to the critical value F_{crit}. If F is less than F_{crit}, then we accept the hypothesis , otherwise we reject the hypothesis.*

File to study ANOVA (Analysis of variance):

Demonstration file: ANOVA (Analysis of variance)
300-49-00

Part - VI.

List of statistical Excel functions

The Hungarian names of the Excel functions are given on the right end of the lines.

AVEDEV(array) ÁTL.ELTÉRÉS

Average absolute deviation from the average.

AVERAGE(array) ÁTLAG

Average.

BETADIST(x;α;β;A;B) BÉTA.ELOSZLÁS

Distribution function of the beta distribution on interval (A,B).

BETAINV(y;α;β;A;B) INVERZ.BÉTA

Inverse of the distribution function of the beta distribution on interval (A,B).

BINOMDIST(k;n;p;FALSE) BINOM.ELOSZLÁS

Weight function of the binomial distribution.

BINOMDIST(k;n;p;TRUE) BINOM.ELOSZLÁS

Distribution function of the binomial distribution.

CORREL(array$_1$;array$_2$) KORREL

Correlation coefficient.

COUNT(array) DARAB

Number of cells containing numerical values.

COUNTA(array) DARAB2

Number of cells containing anything. Number of non-empty cells.

COUNTBLANK(array) DARABÜRES

Number of empty cells.

COUNTIF(array;crit) DARABTELI

Number of the cells satisfying a criterion.

COVAR(array_1;array_2) KOVAR

Covariance.

CRITBINOM(n; p; y) KRITBINOM

Generalized inverse of the distribution function of binomial distribution: greatest k value for which the sum of the terms of the binomial distribution from 0 to k is still less than or equal to y.

EXPONDIST(x;λ;FALSE) EXP.ELOSZLÁS

Density function of the exponential distribution with parameter λ.

EXPONDIST(x;λ;TRUE) EXP.ELOSZLÁS

Distribution function of the exponential distribution with parameter λ.

FREQUENCY(data-array;bins-array) GYAKORISÁG

Using this function a vertical table of frequencies can be constructed. Attention! This is a so called "matrix valued function".

GAMMADIST(x;n;β;FALSE) GAMMA.ELOSZLÁS

Density function of the gamma distribution with order n and parameter $\frac{1}{\beta}$. When $\alpha = 1$, GAMMADIST returns the exponential distribution with parameter $\lambda = \frac{1}{\beta}$.

GAMMADIST(x;n;β;TRUE) GAMMA.ELOSZLÁS

Distribution function of the gamma distribution with order n and parameter $\frac{1}{\beta}$. When $\alpha = 1$, GAMMADIST returns the exponential distribution with parameter $\lambda = \frac{1}{\beta}$.

GAMMAINV(y;n;β) INVERZ.GAMMA

Inverse of the distribution function of the gamma distribution with order n and parameter $\frac{1}{\beta}$.

HYPGEOMDIST(k;n;K;N) HIPERGEOM.ELOSZLÁS

Weight function of the hyper-geometrical distribution.

INTERCEPT(known-y;known-x) METSZ

Intercept.

LARGE(array;k) NAGY

The k-th largest element in an array.

LOGINV(y;μ;σ) INVERZ.LOG.ELOSZLÁS

Inverse of the distribution function of the log-normal distribution.

LOGNORMDIST(x;μ;σ) LOG.ELOSZLÁS

Distribution function of the log-normal distribution.

MAX(array) MAX

Maximum.

MEDIAN(array) MEDIÁN

Median.

MIN(array) MIN

Minimum.

MODE(array) MÓDUSZ

Mode.

NEGBINOMDIST(k; r; p) — NEGBINOM.ELOSZL

Weight function of the negative binomial distribution ("pessimistic").

NORMDIST(x;μ;σ;FALSE) — NORM.ELOSZL

Density function of the normal distribution.

NORMDIST(x;μ;σ;TRUE) — NORM.ELOSZL

Distribution function of the normal distribution.

NORMINV(y;μ;σ) — INVERZ.NORM

Inverse of the distribution function of the normal distribution.

NORMSDIST(z) — STNORMELOSZL

Distribution function of the standard normal distribution. This is the function usually denoted by Φ.

NORMSINV(y) — INVERZ.STNORM

Inverse of the distribution function of the standard normal distribution, usually denoted by Φ^{-1}.

PERCENTILE(array;k) — PERCENTILIS

Percentile (=Quantile).

POISSON(x;λ;FALSE) — POISSON

Weight function of the Poisson-distribution.

POISSON(x;λ;TRUE) POISSON

Density function of the Poisson-distribution.

QUARTILE(array;1) KVARTILIS

Lower quartile.

QUARTILE(array;2) KVARTILIS

Inner quartile (=median).

QUARTILE(array;3) KVARTILIS

Upper quartile.

SLOPE(known-y;known-x) MEREDEKSÉG

Slope.

SMALL(array;k) KICSI

The k-th smallest element in an array.

STANDARDIZE(x;μ;σ) NORMALIZÁLÁS

Standardization: $y = \frac{x-\mu}{\sigma}$.

STDEV(array) SZÓRÁS

Sample standard deviation.

STDEVP(array) SZÓRÁSP

Population standard deviation.

TDIST(x;d;1) T.ELOSZLÁS

Tail function (right sided distribution function) of the 1-sided t-distribution (student distribution) with parameter (degree of freedom) d. This distribution is symmetrical about the origin.

TDIST(x;d;2) T.ELOSZLÁS

Tail function (right sided distribution function) of the 2-sided t-distribution (student distribution) with parameter (degree of freedom) d. This distribution is concentrated on the right side of the origin. This distribution can be derived from the 1-sided t-distribution by the absolute value function.

TINV(y;d) INVERZ.T

Inverse of the tail function (right sided distribution function) of the 2-sided t-distribution (student distribution) with parameter (degree of freedom) d.

VAR(array) VAR

Sample variance.

VARP(array) VARP

Population variance.

Part - VII.

Acknowledgements

I would like to express my thanks to my colleagues and students who helped and encouraged me to include simulation in teaching probability and helped me to write this electronic book. Most of them are at the Budapest University of Technology and Economics:

Domokos Szasz - He drew my attention to using simulation in teaching, and later invited me to give presentations on simulation at his probability courses;

Eva Tringer ("Muszertechnika", MT Training) - She supported me in introducing simulation in my probability courses held at "Muszertechnika", MT Training;

Gabor Peceli - He accepted my first special simulation course at the Faculty of Electrical and Software Engineering of Budapest University of Technology and Economics;

Balint Toth - He supported my work at our department from all points of view: giving me encouragement, offering me courses on this topic, and supplying me with all the necessary equipments;

Karoly Simon - He was always enthusiastic about my work, and involved a part of my work into his courses ;

Laszlo Csirmaz (Central European University) - He invited me to hold courses on probability theory and statistics with simulations at the Central European University;

Peter Mora - He taught me the advantages of Latex and showed me how to apply the tricks of internet in this book;

Adam Gyenge, Ferenc Halanyi, Sandor Kolumban - They were the most interested students who, with their enthusiasm, criticism and many excellent ideas, helped me very much to figure out what to teach and how to teach.

Eric Noll - He came from the USA to our university to learn for a term. He read the text and corrected my English language mistakes.