Fundamentals of Mathematical Statistics

Pavol ORŠANSKÝ

Contents

1 Probability theory **9**
- 1.1 Random event . . . 9
  - 1.1.1 Algebraic operations and programs with events . . . 10
- 1.2 Classical definition of probability . . . 12
- 1.3 Kolmogorov definition of probability . . . 14
- 1.4 Probability of the unification of random events . . . 15
- 1.5 Probality of the opposite event . . . 17
- 1.6 Conditional probability . . . 18
- 1.7 Intersection probability of random phenomena . . . 19
- 1.8 Full probability formula . . . 20
- 1.9 Bayesov vzorec . . . 22
- 1.10 Bernoulliho vzorec . . . 23

2 Random variable **25**
- 2.1 Discrete probability distribution . . . 25
- 2.2 Distribution function of random variable . . . 27
- 2.3 Density distribution . . . 28
  - 2.3.1 Basic features of the density distribution. . . . 29
- 2.4 Numerical characteristics of random variable . . . 31
  - 2.4.1 Mean value . . . 31
  - 2.4.2 Variance (dispersion) and standard deviation . . . 33

3 Significant continuous distribution of random variable **35**
- 3.1 Normal distribution (Laplace-Gaussovo distribution) . . . 35
- 3.2 Standard normal distribution . . . 36
  - 3.2.1 The relationship between $F(x)$ and $\Phi(x)$ . . . 38

4 Descriptive statistics **41**

5 Estimates of parameters **45**
- 5.1 Point estimate . . . 45
- 5.2 Interval estimation of parameters . . . 49
  - 5.2.1 $100 \cdot (1-\alpha)\%$ bilated confidence interval for mean value μ 49
  - 5.2.2 $100 \cdot (1-\alpha)\%$ bilateral confidence interval for dispersion σ^2 53

6 Testing statistical hypotheses **57**
6.1 Parametric testing an single file 57
6.1.1 Testing parameter μ if the set is small ($n \leq 30$) 59
6.1.2 Testing parameter μ if the set is large ($n > 30$) 60
6.1.3 Testing parameter σ 61
6.2 Comparing two files 62
6.2.1 Testing equality of means of two fundamental files, when the files are large ($n > 30$), and if σ_1, σ_2 are known . . . 62
6.2.2 Testing equality of means of two fundamental files, when the files are small ($n < 30$), and if σ_1, σ_2 are known . . . 64
6.2.3 Testing equality of means of two fundamental files, when the files are large ($n > 30$), and if σ_1, σ_2 are unknown . . 64
6.2.4 Testing equality of means of two fundamental files, when the files are small ($n < 30$), and if σ_1, σ_2 are unknown . . 65

7 Correlation Analysis **69**
7.1 Coefficient covariance 69
7.2 Correlation coefficient 70
7.3 Coefficient of determination 70

8 Paired linear regression **77**
8.1 Regression line 77
8.2 Estimation of the parameters β_0 and β_1 78

9 Attachments **81**

Preface

This text was based on the inherent requirements of my students on the subject of Statistics Faculty of Management, University of Prešov, an overview of a comprehensive publication of purely on statistics, contained in their basic course. Most existing text was either too large, and thus a deterrent to the reader, or, conversely, some publications contained only secondary curriculum. Demands on university students has recently changed considerably, therefore was also a gap in the literature, which lacks books covering a kind of intermediate stage of secondary literature and literature explicitly university type, ie. kladúce books to the reader requirements less stringent than in the past. This gap, I tried the "quick fix" patch overwrite my lectures from the course Statistics in acceptable form. I have text in addition to enriching addressed but not resolved by the examples that I drew from its own resources respectively. I took them from the book [1].[1]

This text does not replace any, in my opinion, excellent publications by other authors, of which I once again mention Chajdiaks [1]. But in some way trying to bring modern students who are trying, let's face it, as far as possible to save mathematics. It is strictly text reader for inexpensive, lightweight for deeper analysis of issues.

Finally, I would like to thank my colleague and good friend Dr., Ing. Ján RYBÁRIK, PhD., For their support and valuable advice of an experienced teacher, without whom this text would not arise. Also thank my students for comment.

Trenčín, May 30, 2009 Pavol ORŠANSKÝ

[1] No results or arguments and examples used in them are not based on facts, and therefore there is no real prefiguration them, which could be an eventual usurper can look up.

Preface to the translation

This text was created, for the purpose of teaching foreign students, as a translation of teaching material originally intended for teaching at the University of Prešov. Transfer and reducing the learning material we have created a text that covers the issue of statistics needed to manage the course of numerical and discrete mathematics.

Individual chapters present the minimum necessary for basic understanding of statistical methods. The first chapter defined the very concept of probability, we deal with here its basic operations.

In the second and third chapter describes some basic probability distribution.

In the fourth chapter, the fundamentals of descriptive statistics, the necessary basis for the data processed.

The fifth chapter deals with the estimation of parameters.

The sixth chapter is devoted to testing statistical hypotheses about the parameters set.

Seventh and eighth chapter discusses the issue of linear dependency files.

Trenčín, January 2011 Pavol ORŠANSKÝ

Chapter 1

Probability theory

Probability theory is a kind of basis for statistics. Is an integral part thereof. For a deeper understanding of statistics is necessary to have knowledge of at least this chapter. Probability theory describes random events and probability of the occurrence. Statistics specific model empirical events, and statistical methods to describe these events have a fundamental right in probability theory.

1.1 Random event

Definition 1 ***Random attempt (plot)*** *is an attempt, the outcome is not clearly defined conditions, which were carried out. We have in mind attempts (even if the term fits better story), whose outcome can not be determined in advance.*

For example: Coin toss or dice, the number of dead after a moderate earthquake in China, the possibility of TB infection using trains Slovak railways, etc. ..

Definition 2 ***Random event*** *is true that the result of an accidental experiment. In other words, the outcome of the trial.*

For example: "When throwing a coin falls symbol character", "The roll of dice you get the number five", "When a moderate earthquake in China killed 3624 people" or "While driving Slovak railways would be infected with TB every one hundred fifty-nine billionth (wonderful word) the passengers' .

Random events will be marked in capitals, eg.

event A ... "Hugo Chaves wins the presidential election in Venezuela."

In our view, are also interesting attempts, which will be studied event, the increasing number of experiments show some stability. Which means that the relative frequency of the occurrence of accidental event A has an increasing

number of trials tends to be constant. Where relative abundance is expressed as the ratio of "the occurrence of the event A" to "of all attempts.

Therefore nothing other than

$$\frac{n_A}{n} \sim konst., \text{ for } n \gg 1,$$

n_A indicates how many times there was a event A,
n indicates the number of times an attempt was made.

We call this constant probability of accidental event A. (For the example above, this probability is equal to the probability of the same event, as you say later.)

For illustration, consider the dice with six sides and a random event A that "throwing the dice when you get the number five".

If we started to throw the dice and the results we would be registered, after a sufficient number of throws we would have noticed that the number of falling numbers five and the number of roll the dice in a ratio of about 1 to 6th This result would probably not surprising, since almost everyone also expects that the probability of "throwing the dice when you get the number five" is $\frac{1}{6} = 0.166\,67 \mathrel{\hat{=}} 16.7\%$.

At this point we would like to alert readers to the increasingly paid attention to the nomenclature of the probability theory, which as the years seem, also tends to mislead, and therefore it is quite likely that the reader stops to focus on the issue, only because of uncertainties signs of inconsistency and opacity.

1.1.1 Algebraic operations and programs with events

Let A, B are random events.

I. Relation equality $A = B$

event A occurs if the event is part of B and B at the same event is part of the event A.

Eg.: event A ..." In throwing dice scored number six. "

event B ..." In throwing dice will fall even number divisible by three. "

II. Operation unification $A \cup B$

Is a random event, which occurs if and only if there is a event A event or B, ie. occurs at least one of the events A or B.

Eg.: event A ..." In throwing dice scored number six. "

event B ..." In throwing dice scored number five. "

event $A \cup B$ ..." In throwing dice scored number five or number six. "

III. Operation breakthrough $A \cap B$

Is a random event, which occurs if and only if there is a event A and B at the same event occurs, ie. both events occur simultaneously.

Eg.: event A ..." In throwing dice scored number six. "

event B ..." In throwing dice scored number five. "

event $A \cap B$ ..." In throwing dice scored number five and number six at the same time. "

IV. Opposite event $\bar{A}$

Contrary to the random event accidental event and occurs when the event occurs A.

Eg.: event A ..." In throwing dice scored number six. "

event $\bar{A}$ ..." In throwing dice scored number six, ie. scored one of the numbers one, two, three, four or five. "

V. Certain event Ω

Certain event is a random event that always occurs.

Certain event Ω ..." In throwing dice scored one of the numbers one, two, three, four, five or six. "

It could still possibly be that we fall on the edge of the cube but that it is impossible.

Probably true is that $A \cup \bar{A} = \Omega$.

VI. Impossible event $\varnothing$

Impossible event is a random event that never occurs.

Eg.: Ω event Ω ..." when throwing dice, cube fall "(we think of anything more imaginative)

Impossible event is the opposite event to the same event $\overline{\Omega} = \varnothing$.

Remark 1 *If random events $A \cap B = \varnothing$, ie. if these two events never occur simultaneously, then these events are called* ***disjoint (incompatible) random events****.*
Eg.: Current tumble numbers two and three at once throwing dice.

Remark 2 *Events can also compare, if $A \subset B$, then we call that event A event is subevent B.*

Remark 3 *Eg.: event A ... "The roll of dice you get the number six."*
event B ... "In throwing dice will fall even number."

1.2 Classical definition of probability

Probability should have similar properties as relative abundance $\frac{n_A}{n}$, since it is modeled.

For further understanding it is necessary to introduce some concepts that we try to explain most empirically.

Event that can not be further broken down into detailed call elemental event, for example. when we throw dice event of falling even number decomposed into three elementary events, namely: the event of falling number two, the event of falling numbers four and six event of falling numbers, which we can no longer spread, therefore it is elementary phenomena.

Definition 3 *The system of sets α called algebra, if:*

1. $\forall A, B \in \alpha$: $A \cup B \in \alpha$,
2. $\forall A, B \in \alpha$: $A \cap B \in \alpha$,
3. $\forall A \in \alpha : \overline{A} \in \alpha$
4. $\Omega \in \alpha$,
5. $\varnothing \in \alpha$.

Definition 4 *Real function $P(A)$ defined on the algebra α subsets of Ω will be called if the following shall apply:*

1. $\forall A \in \alpha \Rightarrow P(A) \geq 0$,
2. $\forall A, B \in \alpha, : A \cap B = 0$ *(ie. disjoint)* $\Rightarrow P(A \cup B) = P(A) + P(B)$,
3. $P(\Omega) = 1$,
4. $P(\varnothing) = 0$.

Definition 5 *Assuming that the set of elementary events Ω and is final, and while each elementary event is in addition to the same probability of getting a special case, so.* ***classical definition of probability****.*

Definition 6 *Let Ω be a finite set, and let α be a algebra subset of the set of elementary events Ω.*

Then the probability $P(A)$ of set A means the ratio

$$P(A) = \frac{|A|}{|\Omega|},$$

where the symbol $|A|$ means the number of elements set A and the symbol $|\Omega|$ means the number of elements set Ω.

Remark 4 *Symbol $|A|$ can be interpreted as a number of favorable results, ie. such results, in which event A occurs.*

Symbol $|\Omega|$ represents the number of all possible outcomes when a random experiment.

Remark 5 *Trinity (Ω, α, P) will be called a* ***probability space (classical)****.*

Example 1 *Throw dice. Calculate how likely that number will fall more than 2*

Solution: $|A| = 4$, since $A = \{3; 4; 5; 6\}$.
$|\Omega| = 6$, since $\Omega = \{1, 2, 3; 4; 5; 6\}$.

$$P(A) = \frac{|A|}{|\Omega|} = \frac{4}{6} \doteq 0.67 = 67\%.\spadesuit$$

Example 2 *Throw while playing with three dice. Vypočítajme what is likely to fall by three equal numbers?*

Solution: $|A| = 6$.
$|\Omega| = C_3'(6) = \binom{n+k-1}{k} = \binom{6+3-1}{3} = 56.$

$$P(A) = \frac{|A|}{|\Omega|} = \frac{6}{56} = \frac{3}{28} 0.107 = 10.7\%.\spadesuit$$

Example 3 *V škatuli je 20 výrobkov, z toho 7 je nepodarkov. Náhodne vyberieme 5 výrobkov. Vypočítame, aká je pravdepodobnosť, že práve 2 medzi vybranými výrobkami budú nepodarky?*

Solution: $|A| = C_2(7) \cdot C_3(13) = \binom{7}{2} \cdot \binom{13}{3} = 21 \cdot 286 = 6006.$

A - medzi vybranými budú dva nepodarky (tj. $C_2(7) = \binom{7}{2}$) a zvyšné tri budú v poriadku (tj. $C_3(13) = \binom{13}{3}$).

$$|\Omega| = C_5(20) = \binom{20}{5} = 15\,504.$$

Ω - výber 5 z 20, pričom na poradí nezáleží.

$$P(A) = \frac{|A|}{|\Omega|} = \frac{6006}{15\,504} = \frac{1001}{2584} = 0.387\,38 = 38.7\,38\%.\spadesuit$$

1.3 Kolmogorov definition of probability

Tadeáš Nikolajevič Kolmogorov[1] in the thirties of last century by definition the likelihood of the foundations of modern probability theory, which is essentially still. Likelihood considered fair feature random event, regardless of whether you are able to measure this property is expressed in numbers. As a convention to indicate the likelihood of accidental event as $P(A)$. Calculation of probability is based on three basic axioms, therefore, this definition is sometimes called the *axiomatic definition of probability.*

The basic idea is that the set of elementary events Ω is *infinite*, and also that individual *elementary events don't have equal probability.* Consequently, it is necessary to take account of the limit of an infinite sequence of random events. Given these facts to extend the definition of algebra, so-called **σ-algebru**.

Definition 7 *Let Ω be any non-empty set and let α is a nonempty subset of the set system Ω.*

*Than system α we call **σ-algebra** if:*

1. $\forall A \in \alpha \Rightarrow \overline{A} \in \alpha$.
2. $\forall A_i \in \alpha$, *where* $i = 1, 2, \ldots \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \alpha$.

Remark 6 *Random event we understand each set α from σ-algeba.*

Definition 8 ***(Kolmogorovova definition of probability - I.)***

Let Ω be an nonemtz set and α let be σ-algebra of random elements (subset of set Ω) defind on set Ω.

Then the probability $P(A)$ event $A \in \alpha$ is a real function defined on α, that for all disjoint events (ie. $\forall A_i, Aj \in \alpha$; where $i, j = 1, 2, 3, \ldots$are $A_i \cap Aj = \varnothing$, for $i \neq j$) satisfies the following:

1. $P(\Omega) = 1$.
2. $P(A) \geq 0 \qquad \forall A_j \in \alpha,\ j = 1, 2, \ldots,$
3. $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P\left(A_i\right).$

Or otherwise.

[1] **Tadeáš Nikolajevič Kolmogorov** (*25. 4., 1903 – †20. 10. 1987) Soviet mathematician, founder of modern probability theory and complexity theory algorithms. He also worked in the fields of topology, logic, Fourier series, turbulence and classical mechanics.

Definition 9 ***(Kolmogorovova definition of probability - II.)***
Set of all random events (results) of trials $A_1, A_2, \ldots, A_m$ together with the impossible event $\varnothing$ denote Ω.

Probability $P(A)$ random event we call the real function defined on Ω and satisfying to axioms of probability:

1. $P(A) \geq 0 \qquad \forall A_j \in \Omega,\ j = 1, 2, \ldots, m,$
2. $P(\Omega) = 1;$
3. $P(A_1 \cup A_2 \cup \ldots \cup A_k, \ldots) = P(A_1) + P(A_2) + \ldots + P(A_k), \ldots$ *for any (final or infinite) sequence of bilaterally disjoint random events $A_1, A_2, \ldots, A_k, \ldots$.*

Based on this axiom is clear that the likelihood of additional features:

Remark 7 *Probability can take values from zero to one inclusive, ie.* $0 \leq P(A) \leq 1$.

Remark 8 *Impossible event $\varnothing$ has zero probability, ie.* $P(\varnothing) = 0$.

Remark 9 *Probability of the opposite event is equal to add to one of the original event, ie.*$P\left(\bar{A}\right) = 1 - P(A)$.

Remark 10 *If the event A is a part of the event B (ie. A is subevent of event B, $A \subset B$),than the probability of event A is less equal at most to probability of event B, ie.* $P(A) \leq P(B)$.

Remark 11 *If the event A is a part of the event B (ie. A is subevent of event B, $A \subset B$),than the probability of difference events $B - A$ is equal to probability of diference of probality of this events* $P(B - A) = P(B) - P(A)$.

1.4 Probability of the unification of random events

Probability for the unification of two random events applies

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

if we consider n random events, then the formula for the probability of them together should form

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{1 \leq i_1 < i_2 \leq n} P(A_{i_1} \cap A_{i_2}) + \sum_{1 \leq i_1 < i_2 < i_3 \leq n} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) + \ldots$$
$$\ldots + (-1)^{n+1} \cdot P(A_1 \cap A_2 \cap \ldots \cap A_n).$$

Remark 12 *Previous relationship is at first glance seem incomprehensible, but for specific values of n (most used values are 2, 3 or 4) becomes considerably more comprehensible*

i) $P(A\cup B\cup C)=P(A)+P(B)+P(C)-P(A\cap B)-P(A\cap C)-P(B\cap C)+$
$+P(A\cap B\cap C)$,

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

Consequence 1 *Consider the ending of the relationship for the unification of disjoint events, ie. if $A_i\cap A_j=\varnothing$ for $i\neq j$, than*

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

Example 4 *The consignment of 20 Chinese dairy products, of which 4 contain melanin. Randomly select seven Chinese yogurt. Let´s calculate probability that the selected fermented products is at least 5 without melanin content!*

Solution: $A:\ldots$"Among the selected products is at least 5 without melanin content."

$\left.\begin{array}{l}A_1:\ldots\text{"Among the chosen is just 5 without melanin content."}\\ A_2:\ldots\text{"Among the chosen is just 6 without melanin content.."}\\ A_3:\ldots\text{"Among the chosen is just 7 without melanin content."}\end{array}\right\}$disjoint

$$A=A_1\cup A_2\cup A_3,$$

$$P(A)=P(A_1\cup A_2\cup A_3)=P(A_1)+P(A_2)+P(A_3)=\ldots$$

$$P\left(A_1\right)=\frac{|A_1|}{|\Omega|}=\frac{\binom{16}{5}\cdot\binom{4}{2}}{\binom{20}{7}}=\frac{546}{1615}=0.338\,08;$$

$$P\left(A_2\right)=\frac{|A_2|}{|\Omega|}=\frac{\binom{16}{6}\cdot\binom{4}{1}}{\binom{20}{7}}=\frac{2002}{4845}=0.413\,21;$$

$$P\left(A_3\right)=\frac{|A_3|}{|\Omega|}=\frac{\binom{16}{7}\cdot\binom{4}{0}}{\binom{20}{7}}=\frac{143}{969}=0.147\,57;$$

$$\ldots=\frac{546}{1615}+\frac{2002}{4845}+\frac{143}{969}=\frac{871}{969}=0.898\,86\doteq 90\%\spadesuit$$

Example 5 *The disused hockey stadium metropolis East could be deployed for up to three television cameras, in the case of a live television broadcast of a hockey game. These detect low independently. For the first (central) camera is likely that they will shoot at any given time 60% for the second and third (covering a third or home. Guests) is the same probability equal to 80%. Let´s calculate probability that they will shoot at any given moment what is happening on the ice surface at least one of the cameras!*

Solution: $A : \ldots$"Will shoot at least one of the cameras."
$A_1 : \ldots$ "It will shoot the first camera."
$A_2 : \ldots$ "Will shoot second camera."
$A_3 : \ldots$ "Will capture third camera."
} nie sú disjunktné!!!

$$A = A_1 \cup A_2 \cup A_3,$$

$$P(A) = P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - \ldots$$

$$\ldots - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3) = \ldots$$

$P(A_1) = 0.6, \quad P(A_2) = 0.8, \quad P(A_3), = 0.8,$
$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2) = 0.6 \cdot 0.8 = 0.48,$
$P(A_1 \cap A_3) = P(A_1) \cdot P(A_3) = 0.6 \cdot 0.8 = 0.48,$
$P(A_2 \cap A_3) = P(A_2) \cdot P(A_3) = 0.8 \cdot 0.8 = 0.64,$
$P(A_1 \cap A_2 \cap A_3) = 0.6 \cdot 0.8 \cdot 0.8 = 0.384.$
} are independent!!![2]

$$\ldots = 0.6 + 0.8 + 0.8 - 0.48 - 0.48 - 0.64 + 0.384 = 0.984 \ = 98.4\%\spadesuit$$

1.5 Probality of the opposite event

Theorem 1 *Probality of the opposite event $\overline{A}$ to event A satisfies the equation*

$$P(\overline{A}) = 1 - P(A).$$

Dôkaz: $\Omega = A \cup \overline{A}$,
$P(\Omega) = P\left(A \cup \overline{A}\right)$,
$1 = P(A) + P\left(\overline{A}\right)$,
$P\left(\overline{A}\right) = 1 - P(A)$.■

Example 6 *Set of 32 cards we pulled 4 cards.*

Let´s calculate what is the probability that the card withdrawal will be at least one ace (ie 1 to 4)!

[2] That are independent events we guarantee that the intersection probability calculation we use the above relationship. For a more detailed explanation see. chapter on the probability of intersection of random variables, respectively. independence of random events.

Solution: A : ..."Between card withdrawal will be at least one ace."
$\overline{A}$: ..."Between card withdrawal will not be one ace."

$$P\left(\overline{A}\right) = \frac{\left|\overline{A}\right|}{|\Omega|} = \frac{\binom{28}{4}}{\binom{32}{4}} = 0.569\,38,$$

$$P(A) = 1 - P(\overline{A}) = 1 - 0.569\,38 = 0.430\,62 \triangleq 43\%.\spadesuit$$

Example 7 *Throw while playing with two dice.*
Calculate the probability that the total scored 12! [97.2%]

Example 8 *In the fate there are 2 white, 3 black and 5 blue chips. Randomly select 3 tokens.*
Calculate the probability that among them are at least 2 chips of the same color! [75%]

1.6 Conditional probability

If you are not on the occurrence of the event and placed no conditions, the probability $P(A)$ of A event called the **unconditional probability**.

Often, however, is contingent upon the occurrence of the event of the occurrence of another event, ie. A event can occur only if an event B, the probability is $P(B) > 0$. In this case we are talking about conditional probability. The concept of conditional probability $P\left(A \mid B\right)$ two events defines the following relationship

$$P\left(A \mid B\right) = \frac{P(A \cap B)}{P\left(B\right)}.$$

Example 9 *The telephone exchange is among 110 cable 67 red and those involved is 45. Randomly select the red wire. Calculate the probability that this will be involved?*

Solution: A : ..."Cable is plugged."
B : ..."Cable is red."

$$P\left(A \mid B\right) = \frac{P(A \cap B)}{P\left(B\right)} = \frac{\frac{45}{110}}{\frac{67}{110}} = \frac{45}{67} = 0.671\,64 = 67.164\%.\spadesuit$$

Example 10 *Throw while playing with two dice.*

Example 11 *Calculate the probability that the sum will fall more than 5, if at first glance the figure fell 3!* [16.7%]

1.7 Intersection probability of random phenomena

For intersection probability of random phenomena $A_1, A_2, \ldots, A_n$ applies

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1) \cdot P\left(A_2 \mid A_1\right) \cdot P\left(A_3 \mid A_1 \cap A_2\right) \cdot \ldots \cdot P\left(A_3 \mid \bigcap_{i=1}^{n-1} A_i\right).$$

In the special case of the intersection of two events A and B, ie. case if these event occur simultaneously true that the probability of intersection is equal probability of one of the events times the conditional probability of the second event

$$P(A \cap B) = P\left(A\right) \cdot P\left(B \mid A\right) = P\left(B\right) \cdot P\left(A \mid B\right).$$

Similarly we could define the probability of three events

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P\left(A_2 \mid A_1\right) \cdot P\left(A_3 \mid A_1 \cap A_2\right).$$

Event A and B will be called **independent events** if the occurrence of events and event depend on the occurrence of the event of B and B while the incidence of the event does not depend on the occurrence of event A.

Due to the previous it is clear that for independent events A and B holds[3]

$$\begin{aligned} P\left(A \mid B\right) &= P\left(A\right), \\ P\left(B \mid A\right) &= P\left(B\right). \end{aligned}$$

For intersection of independent events will probably be the following relationship

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1) \cdot P\left(A_2\right) \cdot P\left(A_3\right) \cdot \ldots \cdot P\left(A_n\right).$$

Example 12 *The fates are 3 white, 5 red and 7 blue balls. Randomly select four balls in a row.*

What is the probability that 1 ball will be white, 2 red, 3 red and 4 blue?

Solution: A : ..."1. white, 2 red, 3 red, 4 blue".
A_1 : ..."1. white",
A_2 : ..."2. red",
A_3 : ..."3. red",
A_4 : ..."4. blue".

$$A = A_1 \cap A_2 \cap A_3 \cap A_4$$

$$P(A_1 \cap A_2 \cap A_3 \cap A_4) = P(A_1) \cdot P\left(A_2 \mid A_1\right) \cdot P\left(A_3 \mid A_1 \cap A_2\right) \cdot P\left(A_4 \mid A_1 \cap A_2 \cap A_3\right) =$$

[3] Once again, however, stressed that this applies only if they are independent event.

$$P(A_1) = \frac{|A|}{|\Omega|} = \frac{\binom{3}{1}}{\binom{15}{1}} = \frac{3}{15} = \frac{1}{5},$$
$$P\left(A_2 \mid A_1\right) = \frac{5}{14},$$
$$P\left(A_3 \mid A_1 \cap A_2\right) = \frac{4}{13},$$
$$P\left(A_4 \mid A_1 \cap A_2 \cap A_3\right) = \frac{7}{12},$$
$$\ldots = \frac{1}{5} \cdot \frac{5}{14} \cdot \frac{4}{13} \cdot \frac{7}{12} = \frac{1}{78}.\spadesuit$$

Example 13 *Labourer serves three devices, which operate independently. The probability that the machine does not need within one hour of intervention is equal to the first machinery 90%, the 2nd machine and 80% in the third 85% of the machine.*

What is the probability that one hour will require the intervention of workers or one machine?

Solution: A : ..."For one hour will require the intervention of workers or one machine"

A_1 : ..."For one hour will not need workers hit first machine",

A_2 : ..."For one hour will not need workers hit second machine",

A_3 : ..."For one hour will not need workers hit third machine".

$$A = A_1 \cap A_2 \cap A_3,$$

$$P\left(A\right) = P(A_1) \cdot P\left(A_2\right) \cdot P\left(A_3\right) = 0.9 \cdot 0.8 \cdot 0.85 \doteq 0,61 = 61\%.\spadesuit$$

Example 14 *We have three shipments of a product. The first shipment contains 8 good and 3 defective products. The second consignment contains 10 good and 5 bad product. Third consignment contains 7 good and 4 defective products. For each shipment we choose (independently), one product. Calculate the probability that all three selected products are safe!* [30.8%]

Example 15 *V osudí sú 3 modré, 4 červené a 5 bielych guličiek. Náhodne vyberieme 5 guličiek po sebe (vyberáme po sebe a nevkladáme naspäť, a teda každý výber je závislý na predchádzajúcom). Vypočítajte pravdepodobnosť, že 1. guličkа bude biela, 2. biela, 3. červená, 4. červená a 5. modrá!* [0.758%]

1.8 Full probability formula

First we define the notion of a complete system of events.

Definition 10 *The complete system of events is called system of events $H_1, H_2, \ldots, H_n$, which are by two disjoint, ie. $H_i \cap H_j = \varnothing$ for $i \neq j$, while their union is the certain events, ie. $H_1 \cup H_2 \cup \ldots \cup H_n = \Omega$.*

Consequence 2 *While events $H_1, H_2, \ldots, H_n$ form a complete system of events, then the probality of any event and can be determined using what is known as full probability formula*

$$P(A) = \sum_{i=1}^{n} P(H_i) \cdot P(A \mid H_i) \qquad \textbf{\textit{(Full probability formula)}}$$

or else

$$P(A) = P(H_1) \cdot P(A \mid H_1) + P(H_2) \cdot P(A \mid H_2) + \ldots + P(H_n) \cdot P(A \mid H_n).$$

The formula is the culmination of a clear definition full probability, where any event which is fully distributed in complete system of events to disjoint parts, which, since the system is complete, unification gives us the whole event A.

Remark 13 *Full probability formula is true if we replace n by infinite, ie. $n \to \infty$.*

Remark 14 *Events $H_1, H_2, \ldots, H_n$ commonly we called a hypothesis and apply for them*

$$P(H_1) + P(H_2) + \ldots + P(H_n) = 1.$$

Example 16 *Come into the store three production companies of the same type products represented 2:3:4. Probability of a flawless product is first holding 82% of 2 holding 93% and 90% of 3.podniku.*

What is the probability that a flawless product we buy?

Solution: A : ..."We buy a perfect product."
H_i : ..."The product is the i-th company."

$$\begin{aligned} P(A) &= P(H_1) \cdot P(A \mid H_1) + P(H_2) \cdot P(A \mid H_2) + P(H_3) \cdot P(A \mid H_3) = \\ &= \frac{2}{9} \cdot 0.82 + \frac{3}{9} \cdot 0.93 + \frac{4}{9} \cdot 0.9 \doteq 0.892 \ = 89.2\%\spadesuit \end{aligned}$$

Example 17 *Two consignments containing products. The first shipment contains 18 good and 3 bad. The second shipment includes 9 good and 1 bad. The second consignment choose a product and give it to first. Then randomly select one product from the first shipment.*

Let´s calculate probability that this is good!

Solution: A : ..."The product is good."
H_1 : ..."The product is the second shipment is defective."
H_2 : ..."The product is the second shipment is good."

$$P(A) = P(H_1) \cdot P(A \mid H_1) + P(H_2) \cdot P(A \mid H_2) = \frac{1}{10} \cdot \frac{18}{22} + \frac{9}{10} \cdot \frac{19}{22} = 85.9\%\spadesuit$$

Example 18 *Until fate with three balls to put a white ball. Assume that any number of white balls in fate is likely. Then drawn at random from pulled one ball.*

How likely is it white?

Solution: A : ..."The selected ball is white."
H_1 : ..."In the original fates are all three white balls."
H_2 : ..."In the original fates are two white balls."
H_3 : ..."In the initial fate is a white ball."
H_4 : ..."In the initial fate is no white ball. "

$$\begin{aligned} P(A) &= P(H_1)\cdot P(A \mid H_1) + P(H_2)\cdot P(A \mid H_2) + \\ &\quad P(H_3)\cdot P(A \mid H_3) + P(H_4)\cdot P(A \mid H_4) \\ &= \frac{1}{4}\cdot 1 + \frac{1}{4}\cdot\frac{3}{4} + \frac{1}{4}\cdot\frac{2}{4} + \frac{1}{4}\cdot\frac{1}{4} = \frac{5}{8} = 0.625 = 62.5\%\spadesuit \end{aligned}$$

Example 19 *Three plants produced bulbs.*

The first one covers 60% of all consumption region, the second covers 25% and third 15%.

Of every 100 first-tales factory default is 92, of 2 race 78 and of 3 their factory is only 63 standard.

Calculate what is the probability that the bulb was purchased in this region is the standard! [84.1%]

Example 20 *Two consignments containing products. The first shipment contains 18 good and 3 bad. The second shipment includes 9 good and 1 bad. The second consignment choose 2 products and let him into the first. Then randomly select one product from the first shipment.*

Calculate probability that this is good! [86.1%]

1.9 Bayesov vzorec

Let events $H_1, H_2, \ldots, H_n$ form a complete system of events. If a result of random trial is an event A, then the conditional probability of event H_j with respect to event A we calculat by using **Bayes formula**

$$P(H_j \mid A) = \frac{P(H_j)\cdot P(A \mid H_j)}{P(A)} = \frac{P(H_j)\cdot P(A \mid H_j)}{\sum_{i=1}^{n} P(H_i)\cdot P(A \mid H_i)}.$$

Remark 15 *A is known event that occurs when by a random experiment. Bayes formula refers to the probability of hypotheses, provided that the event occurs A.*

Remark 16 *Bayes formula is true also if we replace n infinite, ie.*

Example 21 *An insurance company classifies drivers into three categories, namely A_1 - good, A_2 - moderately good and A_3 - bad. In the experience of knowing that the group A_1 is about 20% of the insured, group A_2 30% and is the most numerous group A_3 with with 50% representation. Likelihood that the driver of the group A_1 will have an accident is 0.01, the driver of the group A_2 this option 0:03, but the driver the last category A_3 it is 0.1.*

The insurance company shall indemnify Mr. Z. Mr. Z then has an accident, what is the probability that Mr. Z is from category A_3?

Solution: A : . . ."Has an accident."
A_1:. . . "Belongs to A_1."$\longleftrightarrow P(A_1) = 0.2$
A_2:. . . "Belongs to A_2."$\longleftrightarrow P(A_2) = 0.3$
A_3:. . . "Belongs to A_3."$\longleftrightarrow P(A_3) = 0.5$

$A_3 \mid A$: . . ."Patrí do A_3 a má nehodu. "

$$P(A_3 \mid A) = \frac{P(A_3) \cdot P(A \mid A_3)}{P(A)} = \frac{0.5 \cdot 0.1}{0.2 \cdot 0.01 + 0.3 \cdot 0.03 + 0.5 \cdot 0.1} = 0.819\,67 = 82\%$$

Example 22 *The plant produces some components of which two quality control inspectors. The probability that the component will be checked first sampler is 0.7 to 0.3 second. The probability that the first component sampler deems acceptable is 0.92, the second is the probability of 0.98. When an exit clearance, it was found that the component has been satisfactory.*

Calculate the probability that it controlled the first controller! [68.6%]

Example 23 *Before the disease breaks out of its existence can be determined by biological tests, the outcome is not entirely clear. When the sick person is the likelihood of a positive test result 0.999, contrary to a healthy person, a positive test appears with a 0.01. Test condition therefore not be detected. We assume that the disease affects approximately 10% of the population.*

XY person has tested positive.

Calculate the probability that the disease has actually Z! [91.7%, *ie. hope is* 8.3%]

1.10 Bernoulliho vzorec

Theorem 2 *Let the probability of realization of the phenomenon and in each experiment is the same and equal to p.*

Then the probability k-multiple occurrence of event A in n independent experiments is given by Bernoulli[4] *formula*

$$P_n(k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k},$$

[4] **Daniel Bernoulli** (* 8. 2. 1700 – † 17. 3. 1782) was a Swiss mathematician, physicist and medic. Significant contribution in the theory of differential calculus, like mathematical, statistical and probabilistic number theory and mechanics.

or, if we put $q = 1 - p$ (ie. probability of the opposite event), then

$$P_n(k) = \binom{n}{k} \cdot p^k \cdot q^{n-k}.$$

Example 24 *Let´s Calculate the probability that the 5-fold attempt to get the number just six 2-times?*

Solution: $p = \frac{1}{6}; \qquad k = 2; \qquad n = 5.$

$$\begin{aligned} P_2(5) &= \binom{5}{2} \cdot \left(\frac{1}{6}\right)^2 \cdot \left(1 - \frac{1}{6}\right)^{5-2} = \binom{5}{2} \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{5}{6}\right)^3 = \\ &= \binom{5}{2} \cdot \frac{1}{36} \cdot \frac{125}{216} = \frac{625}{3888} \doteq 0.160 = 16\%.\spadesuit \end{aligned}$$

Example 25 *The company produces 70% of production I. class.*

Determine the probability that in a series of 100 products in number of products I-class quality is between 60 to 62.

Solution:

$$P_{100}(60) + P_{100}(61) + P_{100}(62) = \binom{100}{60} \cdot 0.7^{60} \cdot 0.3^{40} +$$

$$+ \binom{100}{61} \cdot 0.7^{61} \cdot 0.3^{39} + \binom{100}{62} \cdot 0.7^{62} \cdot 0.3^{38} \doteq$$

$$\doteq 0.0085 + 0.0130 + 0.0191 = 0.040\,6 = 4.1\%.\spadesuit$$

Example 26 *Two equal opponents play chess, a draw and as a result exclude.*

Calculate whether it is more likely to win 3 games of 5 or 4 batches of 7! [31.25% > 27.344%, *1. possibility*]

Chapter 2

Random variable

Sometimes it is useful to describe the random phenomenon by using some of its numerical characteristics (eg size, weight, number of features, rates, etc..), which we call the random variable.

Definition 11 *Concept of random variable denote variable whose value is determined by the result of an accidental experiment. Repeat the experiment occurs due to changes in random processes random variable and its value can not be determined before carrying out the experiment, and therefore is a random variable specified probability distribution.*

Random variable we denote as ξ[1] and its value as small Arabic letters eg. x, y, z etc..

Remark 17 *Sometimes it is the practice of random variables labeled in capital letters, for example X, Y, Z and its values with corresponding lowercase ie. x, y, z. However, inattention are often confused with the variable value (ie. lowercase with capital letter), and therefore we will stick to the next sign ξ_1, ξ_2, ξ_3 etc..*

Remark 18 *Symbol $P(\blacksquare = x)$ (resp. $P(\blacksquare < x)$) will mean the probability of such random variable $\blacksquare$ takes the value of x (or enter into a value less than x).*

2.1 Discrete probability distribution

Definition 12 *We say that a random variable $\blacksquare$ has a discrete probability distribution, if there is a final or countable set of real numbers $\{x_1, x_2, x_3, ...\}$, such that pays*

$$\sum_{i=1}^{\infty} P\left(\xi_i = x\right) = 1.$$

[1] small Greek letter "xi"

Probability distribution random variable is given values and probabilities with which these values shall!

Example 27 *In the fate are 2 white and 3 red balls. Randomly select 3 balls. Determine the probability distribution of white balls in random sampling!*

Solution: Random variable ξ represents in this case number of white balls in random sampling, the number of white balls in the selection of three balls from the fate. In order to define the distribution, as we have pointed out, among all values of a random variable (obviously there may be three cases, and the random selection will be two, one or no white ball) we need also the probability of such an attempt, these results may occur.

probability that the random variable takes the value $x = 0$:

$$P(\xi = 0) = \frac{\binom{3}{3}}{\binom{5}{3}} = \frac{1}{10},$$

probability that the random variable takes the value $x = 1$:

$$P(\xi = 1) = \frac{\binom{2}{1} \cdot \binom{3}{2}}{\binom{5}{3}} = \frac{6}{10},$$

probability that the random variable takes the value $x = 2$:

$$P(\xi = 2) = \frac{\binom{2}{2} \cdot \binom{3}{1}}{\binom{5}{3}} = \frac{3}{10}.$$

Obtained results stored into the table:

$x:$	0	1	2
$P(\xi = x):$	$\frac{1}{10}$	$\frac{6}{10}$	$\frac{3}{10}$

. ♠

In the previous example, we tried as much detail to explain what constitutes the concept of probability distribution of random variable and how it can present.

Notice also the fact that the sum of probabilities over all values of the random variable is equal to the number 1, ie. is equal to the probability of a full system of events.

$$\sum_{i=1}^{3} P(\xi_i = x) = P(\xi = 0) + P(\xi = 1) + P(\xi = 3) = \frac{1}{10} + \frac{6}{10} + \frac{3}{10} = 1.$$

2.2 Distribution function of random variable

Definition 13 *Let the random variable* ■ *be defined on probability space* (Ω, α, P).
Then the real function:

$$F(x) = P(\xi < x), \qquad pre\ x \in (-\infty; \infty),$$

is called the distribution function of random variable.

Remark 19 *Distribution function of random variable is a function*

i) *non-decreasing,*

ii) *left continuous,*

iii) *satisfies:* $\lim_{x \to -\infty} F(x) = 0$ *and* $\lim_{x \to \infty} F(x) = 1.$

Example 28 *Throw a coin three times.*
For the random variable ■*, which represents the number of coins fit the character side up, and we determine the probability distribution then the distribution function.*

Solution: Probability distribution can be determined similarly as in the previous example using Bernoulli's formula. Results are entered in the table below which are referred to the calculations.

$x:$	0	1	2	3
$P(\xi = x):$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

$$P(\xi = 0) = P_3(0) = \binom{3}{0} \cdot p^0 \cdot (1-p)^3 = \dots$$

probability of falling character is probably $p = 0.5 = \frac{1}{2}$

$$\dots = \binom{3}{0} \cdot \left(\frac{1}{2}\right)^0 \cdot \left(\frac{1}{2}\right)^3 = \frac{1}{8},$$

$$\begin{aligned} P(\xi = 1) &= P_3(1) = \binom{3}{1} \cdot \left(\frac{1}{2}\right)^1 \cdot \left(\frac{1}{2}\right)^2 = \frac{3}{8}, \\ P(\xi = 2) &= P_3(2) = \binom{3}{2} \cdot \left(\frac{1}{2}\right)^2 \cdot \left(\frac{1}{2}\right)^1 = \frac{3}{8}, \\ P(\xi = 3) &= P_3(3) = \binom{3}{3} \cdot \left(\frac{1}{2}\right)^3 \cdot \left(\frac{1}{2}\right)^0 = \frac{1}{8}. \end{aligned}$$

Distribution function can be determined at intervals defined by the values of random variable

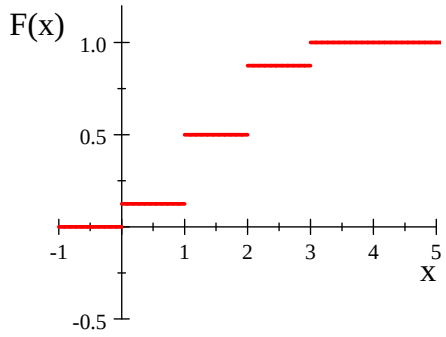
$x < 0:$ $\quad F(x) = P(\xi < 0) = 0$ (impossible event),

$$
\begin{aligned}
0 \le x < 1: \quad & F(x) = P(\xi < 1) = P(\xi = 0) = \frac{1}{8}, \\
1 \le x < 2: \quad & F(x) = P(\xi < 2) = P[(\xi = 0) \cup (\xi = 1)] = \\
& = P(\xi = 0) + P(\xi = 1) = \frac{1}{8} + \frac{3}{8} = \frac{4}{8}, \\
2 \le x < 3: \quad & F(x) = P(\xi < 3) = P[(\xi = 0) \cup (\xi = 1) \cup (\xi = 2)] = \\
& = P(\xi = 0) + P(\xi = 1) + P(\xi = 2) = \\
& = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{7}{8}, \\
3 \le x: \quad & F(x) = P[(\xi = 0) \cup (\xi = 1) \cup (\xi = 2) \cup (\xi = 3)] = \\
& = P(\xi = 0) + P(\xi = 1) + P(\xi = 2) + P(\xi = 3) = \\
& = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = 1.
\end{aligned}
$$

Distribution function can thus be written as

$$
F(x) = \begin{cases} 0 & \text{pre } x < 0 \\ \frac{1}{8} & \text{pre } 0 \le x < 1 \\ \frac{4}{8} & \text{pre } 1 \le x < 2 \\ \frac{7}{8} & \text{pre } 2 \le x < 3 \\ 1 & \text{pre } 3 \le x \end{cases} \quad \spadesuit
$$

Distribution function can be graphically illustrated as follows



graph of distribution function

2.3 Density distribution

As for the discrete distribution, we define the distribution function, in the case of **continuous distribution** is equivalent to the concept of so-called **density distribution**.

Definition 14 *Says that a random variable* ■ *has a continuous distribution if there is a nonnegative function* $f(x)$ *for which an*

$$
F(x) = P(\xi < x) = \int_{-\infty}^{x} f(t)dt,
$$

where $F(x)$ is the distribution function pertaining random variable ■.

Function $f(x)$ is called the probability distribution density of random variable ■.

Remark 20 *Previous relation clearly determines the density distribution. If that is the distribution density function $f(x)$ continuous everywhere (ie the interval $(-\infty, \infty)$), we can also define the relationship*

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

Remark 21 *If the density distribution f (x) everywhere continuous function, is the distribution function F (x) everywhere continuous function.*

2.3.1 Basic features of the density distribution.

$$F(b) - F(a) = \int_a^b f(x)dx,$$

from where directly derives important relationships for calculating the prbability of a random variable which has continuous distribution

$$P(a < \xi < b) = \int_a^b f(x)dx = F(b) - F(a),$$

$$P(a \leq \xi < b) = \int_a^b f(x)dx = F(b) - F(a),$$

$$P(a < \xi \leq b) = \int_a^b f(x)dx = F(b) - F(a),$$

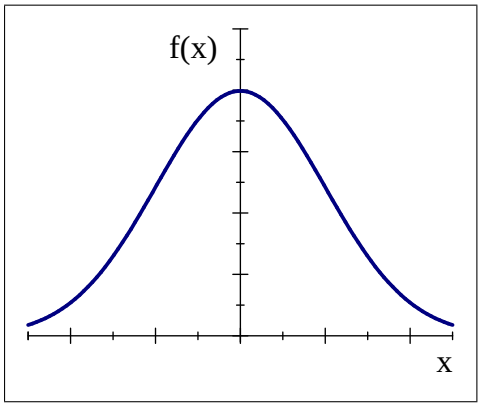
$$P(a \leq \xi \leq b) = \int_a^b f(x)dx = F(b) - F(a).$$

In the previous formulas we have seen that the calculation of definite integral to our border points do not matter and therefore it does not matter whether the relations we use a sharp or blurred inequality.

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

This relationship is sort of the equivalent relation $\sum_{i=1}^{\infty} P(\xi_i = x) = 1$ for discrete random variable distribution.

Graphically it means, that the area between the x-axis and the density function is equal to 1.



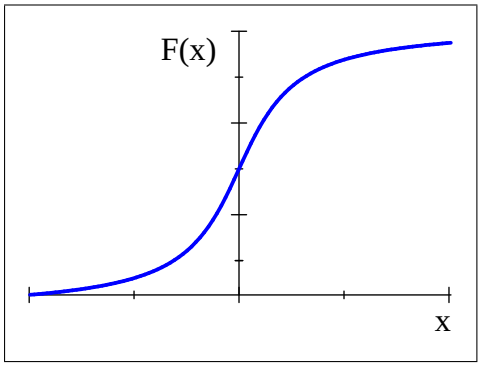
Density distribution

When calculating the probability that a random variable will be included in such interval $\langle -1, 2\rangle$, we calculated the definite integral

$$P(-1 \leq \xi \leq 2) = \int_{-1}^{2} f(x)dx,$$

which result would be the difference values of the distribution function, ie.

$$P(-1 \leq \xi \leq 2) = \int_{-1}^{2} f(x)dx = F\,(2) - F(-1).$$



Distribution function

Example 29 *Determine the constant c so that the function $f(x)$ would be the density distribution of random variable* ■*, and consequently calculate the probality, that the random variable* ■ *will take values from interval $\langle -1; 1\rangle$, ie. calculate $P(-1 \leq \xi \leq 1)$.*

$$f(x) = \begin{cases} cx^2 \cdot e^{-x^3} & x \geq 0 \\ 0 & x < 0 \end{cases}.$$

Solution: Must satisfy $\int_{-\infty}^{\infty} f(x)dx = 1$.

$$\int_0^{\infty} f(x)dx = \int_0^{\infty} cx^2 \cdot e^{-x^3}dx = c \cdot \int_0^{\infty} x^2 \cdot e^{-x^3}dx = \left/ \begin{array}{c} t = x^3 \\ dt = 3x^2dx \\ 0 \to 0 \\ \infty \to \infty \end{array} \right/ =$$

$$= \frac{c}{3}\int_0^{\infty} e^{-t}dt = \frac{c}{3}\left[-e^{-t}\right]_0^{\infty} = \frac{c}{3}\left[0 + e^0\right]_0^{\infty} = \frac{c}{3},$$

and therefore

$$\frac{c}{3} = 1.$$

Thus, where true

$$c = 3.$$

Density distribution will thus in shape

$$f(x) = \begin{cases} 3x^2 \cdot e^{-x^3}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Probability $P(-1 \leq \xi \leq 1)$ we calculate as follows

$$P(-1 \leq \xi \leq 1) = \int_{-1}^{1} f(x)dx = 0 + \int_0^{1} f(x)dx =$$

$$= \int_0^{1} 3x^2 \cdot e^{-x^3}dx = \left[-e^{-1} + e^0\right] = -\frac{1}{e} + 1 \doteq 0.632 = 63.2\%.\spadesuit$$

2.4 Numerical characteristics of random variable

Distribution function, probability distribution and density distribution accurately describe the likelihood of conduct random variable. In practice, we often used certain numeric characteristics to describe a random variable.

2.4.1 Mean value

Mean value of random variable is a kind of "location characteristic". We can seen at it as a "center" of the probability distribution, which is centered around all the values of random variable ■. Označujeme ju $E(\xi)$ alebo μ.

Mean value is for:

1. discrete random variable ■ and its probability function p_i defined as

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

continuous random variable ■ and its probability density $f(x)$ defined as

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

Properties of mean value

i) $E(c) = c$, where c is constant,

ii) $E(a \cdot \xi + b) = a \cdot E(\xi) + b$, where a, b are constants,

iii) $E(\xi_1 + \xi_2 + \ldots + \xi_n) = E(\xi_1) + E(\xi_2) + \ldots + E(\xi_n)$,

iv) $E(\xi_1 \cdot \xi_2) = E(\xi_1) \cdot E(\xi_2)$, if the random variable ξ_1, ξ_2 independent.

Example 30 *Random variable has a distribution of the table*

ξ :	-1	0	1	2	3
p_i :	0.1	0.2	0.2	0.4	0.1

.

Let´s calculate mean value $E(\xi)$!

Solution:

$$E(\xi) = -1 \cdot 0.1 + 0 \cdot 0.2 + 1 \cdot 0.2 + 2 \cdot 0.4 + 3 \cdot 0.1 = 1.2.\spadesuit$$

Example 31 *Density distribution of random variable* ■ *is*

$$f(x) = \begin{cases} \frac{1}{2} \cdot \sin x & 0 < x < \pi \\ 0 & else \end{cases}.$$

Let´s calculate mean value $E(\xi)$!

Solution:

$$E(\xi) = \int_{-\infty}^{\infty} x \cdot f(x)dx = \frac{1}{2} \cdot \int_{0}^{\pi} x \cdot \sin x dx = \ldots = -\frac{\pi}{2}.\spadesuit$$

2.4.2 Variance (dispersion) and standard deviation

Variability reflects the variability of distribution of random variable values, their fluctuations around some constant. We will be particularly interested in how these values scatter around the mean value. The most common expression of this variability is the **variance (dispersion)**, which we signify $D(\xi)$ or σ^2.

Generally, we can the dispersion express with relation

$$D(\xi) = E\left(\left[\xi - E\left(\xi\right)\right]^2\right),$$

and using the properties of the mean change in shape

$$D(\xi) = E\left(\xi^2\right) - \left[E\left(\xi\right)\right]^2.$$

Dispersion for:

1. discrete random variable ■ is defined as

$$D(\xi) = \sum_{i=1}^{n} \left[x_i - E\left(\xi\right)\right]^2 \cdot p_i,$$

2. continuous random variable ■ and its probability density $f(x)$ defined as

$$D(\xi) = \int_{-\infty}^{\infty} \left[x - E\left(\xi\right)\right]^2 f(x)dx.$$

Dispersion characteristics

i) $D(c) = 0$, where c is constant,

ii) $D(a \cdot \xi + b) = a^2 \cdot E(\xi)$, where a, b are constants,

iii) Let the random variable $\xi_1, \xi_2, \ldots, \xi_n$ be independent, then satisfy

$$D(\xi_1 + \xi_2 + \ldots + \xi_n) = D(\xi_1) + D(\xi_2) + \ldots + D(\xi_n),$$
$$D(\xi_1 - \xi_2 - \ldots - \xi_n) = D(\xi_1) - D(\xi_2) - \ldots - D(\xi_n).$$

Another feature of variability is the standard deviation ■, which is defined by dispersion as

$$\sigma = \sqrt{D(\xi)} \qquad \text{or} \qquad \sigma^2 = D(\xi).$$

Example 32 *Random variable has a distribution of the table*

ξ :	-1	0	1	2	3
p_i :	0.1	0.2	0.2	0.4	0.1

.

Let´s calculate dispersion $D(\xi)$ and standard deviation σ!

Solution:

$$D(\xi) = (-1 - 1.2)^2 \cdot 0.1 + (0 - 1.2)^2 \cdot 0.2 + (1 - 1.2)^2 \cdot 1.2 +$$

$$+ (2 - 1.2)^2 \cdot 0.4 + (3 - 1.2)^2 \cdot 0.1 = 1.4,$$

$$\sigma = \sqrt{D(\xi)} = \sqrt{1.4} = 1.1832.\spadesuit$$

Example 33 *Density distribution of random variable* ■ *is*

$$f(x) = \begin{cases} \sin x & 0 < x < \dfrac{\pi}{2} \\ 0 & else \end{cases} .$$

Let´s calculate dispersion $D(\xi)$ *and standard deviation* σ*, if the mean value is* $E(\xi) = a!$

Solution:

$$D(\xi) = \int_{-\infty}^{\infty} (x - a)^2 \cdot f(x)dx = \int_{0}^{\frac{\pi}{2}} (x - a)^2 \cdot \sin x dx = \ldots = a^2 - 2a + \pi - 2.\spadesuit$$

Chapter 3

Significant continuous distribution of random variable

3.1 Normal distribution (Laplace-Gaussovo distribution)

Normal distribution (Laplace-Gauss distribution)[1][2] we use wherever fluctuations caused by the random variable is the sum of many small and independent of each other impacts, eg. on the production dimensions of a product affects fluctuating raw material quality, uniformity machine processing, attention to different worker, etc. ..

parameters:	μ, σ^2
density distribution:	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad \text{for } x \in (-\infty, \infty)$
distribution function:	$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt, \quad \text{for } t \in (-\infty, \infty)$
mean value	$E(\xi) = \mu$
dispersion:	$D(\xi) = \sigma^2$

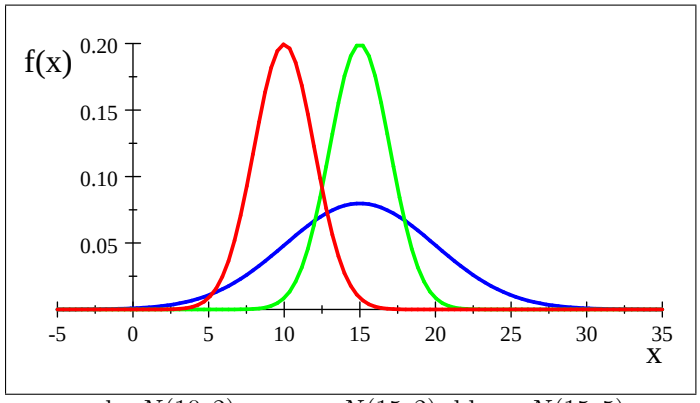
The fact that the random variable ■ has a normal probability distribution with mean μ and standard deviation $\sigma = \sqrt{D(\xi)} = \sqrt{\sigma^2}$ we denote as $\xi \sim$

[1] **Pierre Simon de Laplace** (* 23. 3. 1749 – † 5. 3. 1827) was a French mathematician, physicist, astronomer and politician. He concentrated on mathematical analysis, probability theory and celestial mechanics.

[2] **(Johann) Carl Friedrich Gauß** (Latin form of name Carolus Fridericus Gauss) (* 30. 4. 1777 – † 23. 2. 1855) was one of the greatest mathematicians and physicists of all time. He dealt with number theory, mathematical analysis, geometry, surveying, magnetism, astronomy and optics.

$N(\mu;\sigma^2)$. (Note the fact that is often neglected in solving tasks, and that the registration $N(\mu,\sigma^2)$ appears squared standard deviation σ^2, respectively dispersion$D(\xi)$!!!)

Graph density normal distribution is *Gaussian curve* (attentive reader will understand that this is the "tilde", not dissimilar to bell)



red - $N(10;2)$; green - $N(15;2)$; blue - $N(15;5)$

Mean represents the location of the peak Gaussian curve and the dispersion indicates the degree of variance (indicates the degree of latitude of the curve)

Remark 22 *For normal distribution can be apply the* ***three sigma rule****, which says, that in interval* $\langle \mu - 3\cdot\sigma; \mu + 3\cdot\sigma \rangle$ *are 99.7% all values, ie.*

$$P\left(\mu - 3\cdot\sigma \leq \xi \leq \mu + 3\cdot\sigma\right) = 0.997 = 99.7\%$$

3.2 Standard normal distribution

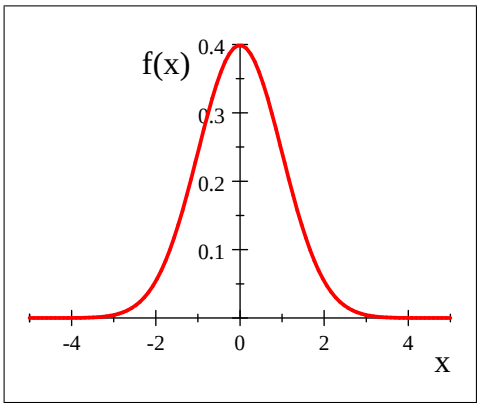
If the parameters μ and ■ the normal distribution are equal $\mu = 0$ and $\sigma = 1$, we say that the random variable ■ has a normal probability distribution in a standardized form, ie. has **standard normal distribution** of probality (in some literature referred to as the Cauchy distribution).

The fact that the random variable has a standardized normal distribution specified as follows $\xi \sim N(0;1)$.

For the standardized normal distribution appears to apply:

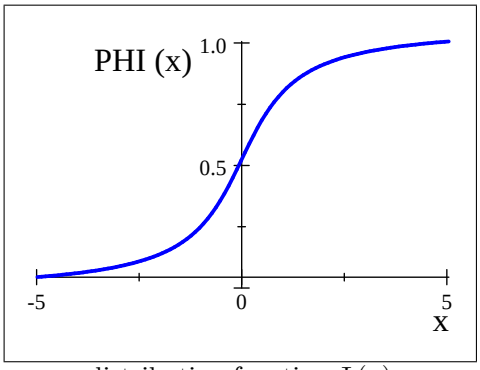
parameters:	$\mu = 0, \sigma^2 = 1$
density distribution:	$f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{\frac{-x^2}{2}}, \quad \text{for } x \in (-\infty, \infty)$
distribution function:	$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}} dt, \quad \text{for } t \in (-\infty, \infty)$
mean value:	$E(\xi) = \mu = 0$
dispersion:	$D(\xi) = \sigma^2 = 1$

Graph density standard normal distribution is Gaussian curve centered around the number zero



Graph density standard normal distribution

Graph of the distribution function $\Phi(x)$ looks like this



distribution function $\Phi(x)$

Integral values $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-x^2}{2}}\, dt$ can not be determined (the integral is unsolvable because they can not be expressed using elementary functions), and therefore are used in practice approximate value calculated numerical calculations. These values are listed in the table annexed to the document, where the probability for each value of the random variable $\xi \sim N(0;1)$. This valoues are also known as quantiles.

Remark 23 *Function* $\Phi(x)$ *satisfy*

$$\boxed{\Phi(x) = 1 - \Phi(-x)} \qquad \textit{pre } x \in (-\infty;\infty)\,.$$

Example 34 *Determine we the value* $\Phi(-1.18)$.

Solution: Table quantiles of the distribution function $\Phi(x)$ aren´t the quantiles for the negative random variable, therefore we will use previous relationship

$$\Phi(-1.18) = 1 - \Phi(1.18) = 1 - 0.881 = 0.119.\spadesuit$$

3.2.1 The relationship between $F(x)$ and $\Phi(x)$

Since the values are tabulated only for a standard normal distribution with normal distribution we need to obtain the quantiles used a conversion, which we now derive

$$\begin{aligned} F(x) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt = \left/ \begin{array}{c} u = \dfrac{t-\mu}{\sigma} \\ du = \dfrac{1}{\sigma} dt \\ -\infty \to -\infty \\ x \to \dfrac{x-\mu}{\sigma} \end{array} \right/ = \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\boxed{\frac{x-\mu}{\sigma}}} e^{\frac{-u^2}{2}} du = \Phi\left(\frac{x-\mu}{\sigma}\right). \end{aligned}$$

Therefore

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

Example 35 *Random variable ξ has distribution $N(0,1)$. Calculate:*

a) $P(2 \le \xi \le 10)$,

b) $P(\xi \ge 0)$.

Solution:

a) $P(2 \le \xi \le 10) = \Phi(10) - \Phi(2) = 1 - 0.9772 = 0.0228 = 2.28\%$,

b) $P(\xi \ge 0) = P(0 \le \xi < \infty) = \Phi(\infty) - \Phi(0) = 1 - 0.5 = 0.5 = 50\%.\spadesuit$

Example 36 *Random variable ξ has distribution $N(0.8;4)$. Calculate :*

a) $P(\xi \ge 1)$,

b) $P(\xi \le -1.16)$.

Solution:

a) $P(\xi \ge 1) = P(1 \le \xi < \infty) = F(\infty) - F(1) = 1 - \Phi\left(\frac{1-0.8}{2}\right) =$

$= 1 - \Phi(0.1) = 1 - 0.5398 \doteq 54\%$,

b) $P(\xi \leq -1.16) = P(-\infty < \xi \leq -1.16) = F(-1.16) - F(-\infty) =$

$= \Phi\left(\frac{-1.16-0.8}{2}\right) - 0 = \Phi(-0.98) = 1 - \Phi(-0.98) =$

$= 1 - 0.8365 = 0.1635$♠

Example 37 *Produced thick plates can be seen as a random variable* ξ*. mean value* $E(\xi) = 10\,\mathrm{mm}$ *and standard deviation* $\sigma = 0.02\,\mathrm{mm}$.

Determine what percentage can be expected if we assume that they are wrong if

a) *plates are thinner as* $9.97\,\mathrm{mm}$,

b) *plates are thicker as* $10.024\,\mathrm{mm}$.

Solution:

a) $P(\xi \leq 9.97) = P(-\infty < \xi \leq -1.16) = F(9.97) - F(-\infty) =$

$= \Phi\left(\frac{9.97-10}{0.02}\right) - 0 = \Phi(-1.5) = 1 - \Phi(1.5) =$

$= 1 - 0.332 = 0.668 = 66.8\%$,

b) $P(\xi \geq 10.024) = P(10.024 \leq \xi < \infty) = F(\infty) - F(10.024) = 1 - \Phi\left(\frac{10.024-10}{0.02}\right) =$

$= 1 - \Phi(1.2) = 1 - 0.864 = 0.136 = 16.6\%$.♠

Example 38 *The IQ level of intelligence we know that the normal distribution with mean* $IQ = 90$ *and a standard deviation of* ± 15 *points.*

Calculate the probability that

a) *andom passerby has IQ lower or equal to 60 points,*

b) *random student has enough IQ to be able to handle this course, ie. IQ is a measure of his more than 80 points.*

Solution:

a) $P(\xi \leq 60) = P(-\infty < \xi \leq 60) = F(60) - F(-\infty) =$

$= \Phi\left(\frac{60-90}{15}\right) - 0 = \Phi(-2) = 1 - 0.977 = 0.023 = 2.3\%$,

b) $P(\xi \geq 80) = P(80 \leq \xi < \infty) = F(\infty) - F(80) = 1 - \Phi\left(\frac{80-90}{15}\right) =$

$= 1 - \Phi\left(-0.6\overline{6}\right) = 1 - \left(1 - \Phi\left(0.6\overline{6}\right)\right) = 0.716 = 71.6\%$.♠

Chapter 4

Descriptive statistics

In this chapter we consider a set of statistical N scale units (random tests), for which we have found the value of the investigated random variable ξ for each statistical unit (random experiment), which means that we know the values x_i (for $i = 1, 2, \ldots, N$) examined variable ξ.

x_i represents a particular value of random variable ξ by i-th try.

Probability distribution of values of random variable ξ we get by so called **sorting**, ie. creating sort of classes similar statistical units (trials).

Of course that is most similar to the units with the same type of random variable being examined, but not always be classified as evaluated in this way, whether because of the diversity of file (eg, all values are different each other and therefore the number of classes would be equal to the scope file) or the nominal characteristics (subjective assessment, for example. the word). In such cases, the class represents a class representative (one specific value of the variable). In the case of continuous distribution of the class represented by the interval values (or the center of this interval).

When sorting is necessary to comply with the *principle of completeness* (ie, every element must be included in any class) and the *principle of clarity* (ie, every element must be included in one class).

Now would be followed by a detailed description of the compilation of statistical tables for the proliferation of file extent n. This part we omit and we will try to illustrate what most telling in the following example.

Example 39 *Investigated the population of thirty apartments. Values are:*
$3; 2; 4; 5; 2; 2; 3; 2; 4; 5; 1; 3; 4; 4; 5; 4; 1; 3; 6; 2; 3; 4; 6; 2; 3; 4; 1; 3; 5; 4.$
Put together a table of the probability distribution.

Solution: First highlights the basic concepts and sign:
measured values x_i, for $i = 1, 2, \ldots, N$,
absolute abundance n_i (how many times a character in the selection is),

relative abundance $p_i = \dfrac{n_i}{N}$,

cumulative abundance $N_i = \sum_{i=1}^{N} n_i$,

kumulatívna relatívna početnosť $M_i = \dfrac{N_i}{N}$.

x_i :	n_i :	p_i :	N_i :	M_i :
1	3	$\frac{3}{30}$	3	$\frac{3}{30}$
2	6	$\frac{6}{30}$	8	$\frac{8}{30}$
3	7	$\frac{7}{30}$	16	$\frac{16}{30}$
4	8	$\frac{8}{30}$	24	$\frac{24}{30}$
5	4	$\frac{4}{30}$	28	$\frac{28}{30}$
6	2	$\frac{2}{30}$	30	$\frac{30}{30}$
	$\sum = 30$	$\sum = 1$		

We hope that the illustration was sufficient.♠

Example 40 *For the same flats were determined and their size:*
82.6; 57.3; 70.4; 65; 48.4; 103.8; 73.6; 43.5; 66.1; 93; 52.6; 70; 84.2; 55; 81.3; 61.5; 75.1; 34.8; 62.4; 116; 70.1; 63.6; 93; 59.2; 65.9; 77.2; 52.8; 68.7; 79.2; 87.4.
Find a group frequency distribution for the number of classes $k = 9$.

Solution: Number of classes is obviously smaller than the number of character values, we therefore have to create the interval probability distribution.

First, we determine on the basis of *variation margin*

$$R = x_{\max} - x_{\min} = 116 - 34.8 = 81.2$$

length of the interval

$$h = \frac{R}{k} = \frac{\text{variačného rozpätia}}{\text{počet tried}} = \frac{81.2}{9} = 9.022\,2 \doteq 10.$$

With regard to the principle of completeness always rounded up the length of the interval![1]

Representative of the class is now the middle of k-th interval x_j

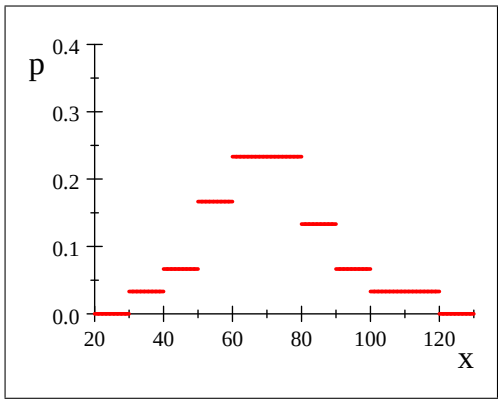
	x_j :	n_j :	p_j :	N_j :	M_j :
$\langle 30; 40)$	35	1	$\frac{1}{30}$	1	$\frac{1}{30}$
$\langle 40; 50)$	45	2	$\frac{2}{30}$	3	$\frac{3}{30}$
$\langle 50; 60)$	55	5	$\frac{5}{30}$	8	$\frac{8}{30}$
$\langle 60; 70)$	65	7	$\frac{7}{30}$	15	$\frac{15}{30}$
$\langle 70; 80)$	75	7	$\frac{7}{30}$	22	$\frac{22}{30}$
$\langle 80; 90)$	85	4	$\frac{4}{30}$	26	$\frac{26}{30}$
$\langle 90; 100)$	95	2	$\frac{2}{30}$	28	$\frac{28}{30}$
$\langle 100; 110)$	105	1	$\frac{1}{30}$	29	$\frac{29}{30}$
$\langle 110; 120)$	115	1	$\frac{1}{30}$	30	$\frac{30}{30} = 1$
		$\sum = 30$	$\sum = 1$		

[1] But really and truly always up.

Basically it was a similar procedure as in the previous example, but please note the index with respect to j, where $j = 1, 2, \ldots, k$, where k represents number of classes.♠

For visual understanding is the best graphically represent results. For these purposes, is commonly used **histogram**.

Histogram is essentially columnal diagram where the x-axis value is applied random variable representing the class and the y-axis is applied corresponding to an absolute (or relative frequencies).



Histogram

Chapter 5

Estimates of parameters

Estimate we mean a statistical method by which the approximately determined (estimated) unknown parameters of statistical files.

Let's random selection $\xi_1, \xi_2, \ldots, \xi_n$ of a distribution, which depends on unknown parameters Θ[1], then Θ parameter can take only certain values of the areaΩ. Through the estimation theory we are trying to create statistics $T(\xi_1, \xi_2, \ldots, \xi_n)$, which distribution comes closer to that parameter $\Theta \subset \Omega$.

Odhady, pri ktorých hľadáme určitý parameter, nazývame *parametrické odhady. Neparametrickými odhadmi* nazývame odhady, pri ktorých nie je požadovaná parametrická špecifikácia typu pravdepodobnostného rozdelenia.

5.1 Point estimate

Point estimate lies in replacing the unknown parameter values of the population, or its functions, the value of the selection characteristics.

At some point estimate, we pay claims as to its consistency and unbias.

Consistent (undisputed) point estimators Θ call such a set of basic statistics $T_n = T(\xi_1, \xi_2, \ldots, \xi_n)$, that for sufficiently large values of index n satisfies the condition

$$P(|T_n - \Theta| \leq \varepsilon) < 1 - \eta,$$

for any $\varepsilon > 0$ and $\eta > 0$, ie.require that the parameter belongs to the interval, whose radius is less than the arbitrarily small but positive ε with probability $1 - \eta$, where η is any positive number, which usually is chosen close to zero as possible. In other words, the point estimate consistent if it lie in the smallest possible interval with most probability as can.

Unbiased point estimate of the parameter Θ is called the basic set of statistics $T_n = T(\xi_1, \xi_2, \ldots, \xi_n)$, which mean value holds $E(T_n) = \Theta$. Otherwise, we talk about estimating the distortion (bias). Difference $b(\Theta) = E(T_n) - \Theta$ we call the bias parameter estimation Θ. As with a growing range of n random distortion

[1] Θ is a large Greek letter "theta"

is reduced, then the statistics is asymptotically unbiased estimate of parameter Θ.

The best undistorted point estimate of the average of the basic set μ is **sample mean**

$$\boxed{\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i}$$

The best undistorted point estimate of the dispersion of the basic set $D(\xi) = \sigma^2$ is **sample covariance**

$$\boxed{S_x^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

The best undistorted point estimate of the standard deviation of the basic set $\sigma = \sqrt{D(\xi)}$ je **sample standard deviation**

$$\boxed{S_x = \sqrt{S_x^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}}.$$

In the case of group probability distribution, where the measured values represented by the centers of intervals and their abundance, ie. if we do not have values directly but only the group distribution table, use the following relations

for **sample mean**

$$\boxed{\overline{x} = \frac{1}{N} \cdot \sum_{j=1}^{k} n_j \cdot x_j},$$

for **sample covariance**

$$\boxed{S_x^2 = \frac{1}{N-1} \cdot \sum_{j=1}^{k} n_{j\cdot} (x_j - \overline{x})^2}$$

an for **sample standard deviation**

$$\boxed{S_x = \sqrt{\frac{1}{N-1} \cdot \sum_{j=1}^{k} n_{j\cdot} (x_j - \overline{x})^2}}$$

where x_j is value of the representative of j-th class, n_j is his abundance and k represents the number of classes of statistics.

Remark 24 *Since this is the parameter estimates and not precise characteristics of a random variable, we can write*

$$\begin{aligned} \mu &\approx \overline{x}, \\ D(\xi) &\approx S_x^2, \\ \sigma &\approx S_x, \end{aligned}$$

but not equality.

Example 41 *From Example 40 let us take only the middles of intervals (as representaive) and estimate sample mean and sample covariance.*

Solution: Values

$x_j:$	35	45	55	65	75	85	95	105	115
$n_j:$	1	2	5	7	7	4	2	1	1

substitute into the formulas for sample mean

$$\begin{aligned}\overline{x} &= \frac{1}{N}\cdot\sum_{j=1}^{k} n_j\cdot x_j = \frac{1}{30}\cdot\sum_{j=1}^{9} n_j\cdot x_j = \\ &= \frac{1}{30}\cdot(1\cdot 35+2\cdot 45+5\cdot 55+7\cdot 65+7\cdot 75+4\cdot 85+2\cdot 95+1\cdot 105+1\cdot 115) = \\ &= 71.\end{aligned}$$

and sample covariance

$$S_x^2 = \frac{1}{N-1}\cdot\sum_{j=1}^{k} n_{j\cdot}\,(x_j-\overline{x})^2 = \frac{1}{30-1}\cdot\sum_{j=1}^{9} n_{j\cdot}\,(x_j-71)^2 =$$

$$= \frac{1}{30-1}\cdot\Big[1\cdot(35-71)^2+2\cdot(45-71)^2+5\cdot(55-71)^2+7\cdot(65-71)^2+7\cdot(75-71)^2+$$

$$+4\cdot(85-71)^2+2\cdot(95-71)^2+1\cdot(105-71)^2+1\cdot(115-71)^2\Big] =$$

$$= \frac{1}{30-1}\cdot(1296+1352+1280+252+112+784+1152+1156+1936) =$$

$$= 321.38$$

Sample standard deviation is so equal

$$S_x = \sqrt{S_x^2} = \sqrt{321.38.} = 17.927$$

For comparison, giving the values of the sample mean and sample covariance calculated from original values

$$\begin{aligned}\overline{x} &= \frac{1}{n}\cdot\sum_{i=1}^{n} x_i = 70.457,\\ S_x^2 &= \frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-\overline{x})^2 = 310.89,\\ S_x &= \sqrt{S_x^2} = \sqrt{310.89} = 17.632.\end{aligned}$$

We see that the values are almost identical.♠

Other estimates of the mean value and *median* are *modus*.

Definition 15 ***Modus*** *is the most frequently occurring value of the character is "most likely" value of the random variable for the character (random experiment)*

Definition 16 ***Median*** *is represented by a value that is particularly "mean" (in this case the quotes are indeed legitimate, and really ask the reader to the concept did not confuse with sample mean) variable in some experiment. The median divides the range of values for two of the almost equally likely. For strongly asymmetric distribution reflects the distribution middle median better than mean.*

Example 42 *For power lines requires a high tensile strength cables.*
Values were measured for two types of cables:
I. kind
$302, 310, 312, 310, 313, 318, 305, 309, 301, 309, 310, 307, 313, 229,$
$315, 312, 310, 308, 314, 333, 305, 310, 309, 314$
II. kind
$300, 310, 320, 309, 312, 311, 31, 317, 309, 313, 315, 314, 307, 322,$
$313, 313, 311, 316, 315, 314, 308, 319, 313, 312$
Estimate the mean value strength of both types of wire through

a) *sample mean,*

b) *modus,*

c) *median.*

Solution: We calculate estimates for the first kind of cable, the second type we leave to the reader with capabilities to repeat the same process with other values.

a) I. druh $\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = 307.42$

b) I. druh modus as most occurring character is the value 310, that occurred 5-th times,

c) I. druh median is also the value 310, because the arrangement of characters in ascending order value is in the middle of such an arrangement.

a) II. druh $[\overline{x} = 301.]$,

b) II. druh $[\text{mod} = 313;\ \text{4-th times}]$,

c) II. druh $[\text{median} = 313]$.

5.2 Interval estimation of parameters

In the previous section, we estimated the unknown parameter points, ie. unknown parameter was "replacing" the particular values that best estimated. It is understandable that using a larger sample we get more accurate results than a smaller sample, but the point estimate method disregard this fact.

Another possibility is to estimate the parameter interval estimation. Unknown parameter is estimated interval, meaning that lies between two values.

The center of interval is a kind estimation of parameter of mean value and width of the interval represents the degree of dispersion of the values.

This interval is called the **confidence interval** (T_a, T_b) and unknown parameter Θ this interval will contain the probability $(1-\alpha)\%$, where number α called the *significance level*. This level is chosen in advance, and expresses the required "degree of accuracy" with which we are looking for the interval in which the parameter is located.

If the selection is repeated many times (given the trend of stability of stochastic processes), then the unknown parameter Θ "falls" into the confidence interval (T_a, T_b) in about the $100 \cdot (1-\alpha)\%$ cases (ie. the probability that the parameter is within the interval (T_a, T_b) is equal to the number $1-\alpha$).

We are talking about so called $100 \cdot (1-\alpha)\%$ confidence interval and write

$$P(T_a \leq \Theta \leq T_b) = 1 - \alpha$$

In the event that the boundaries T_a, T_b are finite, we say about *bilateral confidence interval*.

If $T_a = -\infty$, ie. $(-\infty, T_b)$...*right-sided confidence interval*.

If $T_b = \infty$, ie. (T_a, ∞)...*left-sided confidence interval*.

Boundaries depend on the estimated parameter, the random selection and distribution. However, we further restrict ourselves to a normal distribution $N\left(\mu, \sigma^2\right)$, for which we estimate the parameters μ and σ^2.

5.2.1 $100 \cdot (1-\alpha)\%$ bilated confidence interval for mean value μ

Sample mean shoud have a normal distribution with mean $E(\xi) = \mu$ and dispersion $D(\overline{x}) = \dfrac{\sigma^2}{n}$, and so $\overline{x} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$, which values we find by substitution using the distribution function $\Phi\left(\dfrac{\overline{x}-\mu}{\sigma} \cdot \sqrt{n}\right)$ the standardized normal distribution $N(0,1)$.

For chosen α we can find values $u_{\left(1-\frac{\alpha}{2}\right)}$ and $-u_{\left(1-\frac{\alpha}{2}\right)}$, which correspond to the extreme values of the Gauss curve, respectively. probability that the mean value will be outside the chosen accuracy.

With probality $1-\alpha$ would be the mean value in interval $\left(-u_{1-\frac{\alpha}{2}}; u_{1-\frac{\alpha}{2}}\right)$:

$$P\left(-u_{1-\frac{\alpha}{2}} \leq \frac{\overline{x}-\mu}{\sigma} \cdot \sqrt{n} \leq u_{1-\frac{\alpha}{2}}\right) = 1-\alpha,$$

$$\begin{aligned} -\frac{\sigma}{\sqrt{n}} \cdot u_{\left(1-\frac{\alpha}{2}\right)} &\leq \overline{x} - \mu \leq \frac{\sigma}{\sqrt{n}} \cdot u_{\left(1-\frac{\alpha}{2}\right)} \\ -\overline{x} - \frac{\sigma}{\sqrt{n}} \cdot u_{1-\frac{\alpha}{2}} &\leq -\mu \leq -\overline{x} + \frac{\sigma}{\sqrt{n}} \cdot u_{1-\frac{\alpha}{2}} \end{aligned}$$

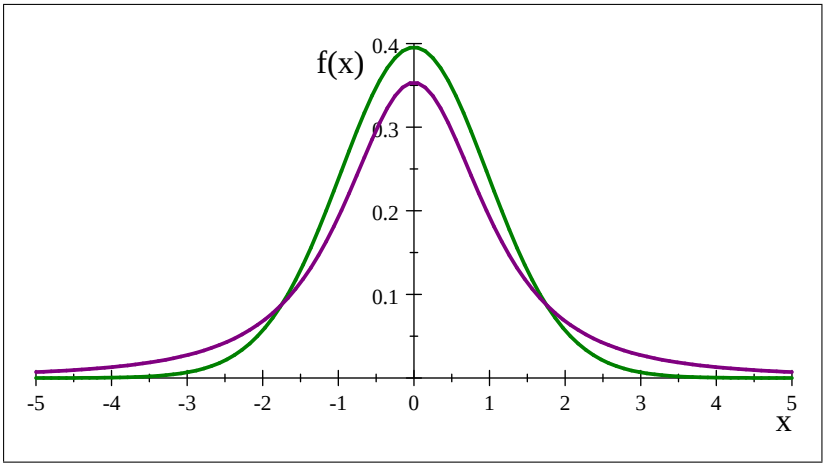
$$\boxed{\overline{x} - \frac{\sigma}{\sqrt{n}} \cdot u_{1-\frac{\alpha}{2}} \leq \mu \leq \overline{x} + \frac{\sigma}{\sqrt{n}} \cdot u_{1-\frac{\alpha}{2}}}.$$

So far we have assumed that we know the value of σ. If we did not know her well, we can for a sufficiently large statistical set ($n \geq 30$) the value of standard deviation σ point estimate with sample standard deviation S_x.

In practice, however, often occur files, whether for economic or purely practical reasons, not too large (eg when determining the quality of the product is destroyed or if it is monitoring some phenomenon of time or otherwise economically too demanding). And if the file is small, we would be using a point estimate committed significant errors.

In 1908 the English chemist Arthur Guinness Brewery & Son Brewery, W. S. Gosset[2] ppublished under the pseudonym Student work, which dealt with just small sample. Derived sample statistics for the distribution of small files, ie. $n < 30$. This division is called Student's t-distribution.

Parameter of this distribution is the number of degrees of freedom. If we have a random file with a range of n elements, then the Student t-distribution $(n-1)$ degrees of freedom. Values of Student's t-distribution $t_{1-\frac{\alpha}{2}}^{(n-1)}$ are different degrees of freedom and significance level are tabulated and listed in the Appendix.



Graf hustoty pravdepodobnosti Studentovho rozdelenia pre 30 st. vonosti (zelen) a $N(0,1)$ rozdelenia (fialov)

[2] **William Sealy Gosset** (*13. 6. 1876 – †16. 9., 1937) English mathematician and chemist. The contract with the Guinness brewery, for reasons of industrial secrecy, did not allow employees to publish any work, because almost all their publications published under his pseudonym "Student".

Graf rozdelenia pravdepodobnosti je veľmi podobný grafu normovaného normálneho rozdelenia, avšak krivka Studentovho rozdelenia je viac "zaoblená" okolo strednej hodnoty. Nie je ťažké porovnaním hustôt pravdepodobností týchto dvoch rozdelení dokázať, že platí:

$$\lim_{n\to\infty} t^{(n)} = N(0;1).$$

So if we don´t known the value of standard deviation σ we use the sample standard deviation S_x. S_x.

For large sets ($n > 30$) we use for confidence interval for mean value μ on significance level α, values standardized normal distribution $u_{(1-\frac{\alpha}{2})}$, than

$$\boxed{\overline{x} - \frac{S_x}{\sqrt{n}} \cdot u_{1-\frac{\alpha}{2}} \leq \mu \leq \overline{x} + \frac{S_x}{\sqrt{n}} \cdot u_{1-\frac{\alpha}{2}}}.$$

For small sets ($n < 30$) we use for confidence interval for mean value μ on significance level α, values Student t-distribution $t^{(n-1)}_{1-\frac{\alpha}{2}}$ with $(n-1)$ degrees of freedom, than

$$\boxed{\overline{x} - \frac{S_x}{\sqrt{n}} \cdot t^{(n-1)}_{1-\frac{\alpha}{2}} \leq \mu \leq \overline{x} + \frac{S_x}{\sqrt{n}} \cdot t^{(n-1)}_{1-\frac{\alpha}{2}}}.$$

Example 43 *The airline estimates the average number of passengers. Within 20 days, the average number of passengers 112 with sample variance 25.*

Find a 95% bilateral confidence interval for the average number of passengers μ.

Solution: $\overline{x} = 112$; $S_x^2 = 25$ (tj.: $S_x = \sqrt{25} = 5$); $n = 20$; $\alpha = 0.05$.

Since the file is small ($n = 20 < 30$), we use Student t-distribution, where $t^{(20-1)}_{1-\frac{5\%}{2}} = t^{(20-1)}_{1-\frac{5\%}{2}} = t^{(19)}_{1-0.025} = t^{(19)}_{0.975} = 2.1$,

and the interval estimate will apply:

$$\begin{aligned}
\overline{x} - \frac{S_x}{\sqrt{n}} \cdot t^{(n-1)}_{1-\frac{\alpha}{2}} &\leq \mu \leq \overline{x} + \frac{S_x}{\sqrt{n}} \cdot t^{(n-1)}_{1-\frac{\alpha}{2}}, \\
112 - \frac{5}{\sqrt{20}} \cdot 2.1 &\leq \mu \leq 112 + \frac{5}{\sqrt{20}} \cdot 2.1, \\
112 - 2.35 &\leq \mu \leq 112 + 2.35, \\
109.65 &\leq \mu \leq 114.35.\spadesuit
\end{aligned}$$

Example 44 *The automatic production lines producing rings of ball bearings sample was taken and found 50 pieces radius rings. From the measured values we obtain the average execution $\overline{x} = 70.012\,\text{mm}$. Find a 99-percent confidence interval for the mean radius of the ring produced when the measured value is equal to the execution statistics $S_x^2 = 0,00723$.*

Solution: $\overline{x} = 70.012$; $S_x^2 = 0.00723$ (tj.: $S_x = \sqrt{0.00723} = 0.08503$); $n = 50$; $\alpha = 0.01$.

Since the file is large ($n = 50 > 30$) we use standardized normal distribution, where $u_{0.99} = 2.33$, and the interval estimate will apply:

$$\begin{aligned} \overline{x} - \frac{S_x}{\sqrt{n}} \cdot u_{0.995} &\leq \mu \leq \overline{x} + \frac{S_x}{\sqrt{n}} \cdot u_{0.995}, \\ 70.012 - \frac{0.08503}{\sqrt{50}} \cdot 2.57 &\leq \mu \leq 70.012 + \frac{0.08503}{\sqrt{50}} \cdot 2.57, \\ 70.012 - 0.0309 &\leq \mu \leq 70.012 + 0.0309, \\ 69.981 &\leq \mu \leq 70.0429.\spadesuit \end{aligned}$$

Example 45 *Measuring the resistance cable from eight randomly selected samples obtained the following values:*

0.139; 0.144; 0.139; 0.140; 0.136; 0.143; 0.141; 0.136.

Assume that the measured values can be considered a random realization from a normal distribution with unknown mean and unknown variance.

Find a 95% confidence interval for the mean value.

Solution: Since we do not know the mean value $E(X)$ or standard deviation σ, we use instead samprle mean $\overline{x}$ and sample standard deviatinon S_x.

$$\begin{aligned} \overline{x} &= \frac{1}{8} \cdot (0.139 + 0.144 + 0.139 + 0.140 + 0.136 + 0.143 + 0.141 + 0.136) = \\ &= 0.13975 \end{aligned}$$

$$\begin{aligned} S_x^2 = \frac{1}{7} \cdot &(0.139 - 0.13975)^2 + (0.144 - 0.13975)^2 + (0.139 - 0.13975)^2 + \\ &+ (0.140 - 0.13975)^2 + (0.136 - 0.13975)^2 + (0.143 - 0.13975)^2 + \\ &+ (0.141 - 0.13975)^2 + (0.136 - 0.13975)^2 = 8,5 \times 10^{-6} = 0.0000085, \end{aligned}$$

and so

$$S_x = \sqrt{S_x^2} = \sqrt{8,5 \times 10^{-6}} = 2.915\,5 \times 10^{-3} = 0.0029155$$

The file is small, therefore use the t-distribution quantiles, where $t_{0.975}^{(7)} = 2.365$.

Thus for the confidence interval is valid:

$$\begin{aligned} \overline{x} - \frac{S_x}{\sqrt{n}} \cdot t_{0,975}^{(7)} &\leq \mu \leq \overline{x} + \frac{S_x}{\sqrt{n}} \cdot t_{0,975}^{(7)}, \\ 0.13975 - \frac{0.0029155}{\sqrt{8}} \cdot 2.365 &\leq \mu \leq 0.13975 + \frac{0.0029155}{\sqrt{8}} \cdot 2.365, \\ 0.13975 - 2.437\,8 \times 10^{-3} &\leq \mu \leq 0.13975 + 2.437\,8 \times 10^{-3}, \\ 0.137\,31 &\leq \mu \leq 0.142\,19.\spadesuit \end{aligned}$$

5.2.2 $100 \cdot (1-\alpha)\%$ bilateral confidence interval for dispersion σ^2

In many cases it is important to monitor not only the reliability of the average value of μ set, but the degree of variability file. TThis we have in mind particularly the efforts to reduce deviations from the average. It is clear that the manufacturer of screws with the standard 5 cm, would hardly succeed with half of production and the other 5.5 cm 4.5 cm.

To calculate the standard deviation of the confidence interval we use more selection division called chi-square distribution (χ^2).

χ^2-distribution[3]

Let the random variable have n random variables $x_1, x_2, \ldots, x_n$ and each has a normal distribution with mean μ and standard deviation σ. This random variables corresponding standardized random variables:

$$Z_1 = \frac{x_1 - \mu}{\sigma}, \ldots, Z_n = \frac{x_n - \mu}{\sigma}.$$

Statistics

$$W = Z_1^2 + Z_2^2 + Z_3^2 + \ldots + Z_n^2 = \frac{\sum\limits_{i=1}^{n} (x_1 - \mu)^2}{\sigma^2}$$

has χ^2-distribution with number of degrees of freedom n.

χ^2-distribution can also apply for distribution sampling variance S_x^2. If a random selection from a fundamental set with normal distribution, create the samples size n, then the random variable

$$\chi^2 = \frac{(n-1) \cdot S_x^2}{\sigma^2}$$

has χ^2-distribution with $(n-1)$ degrees of freedom.

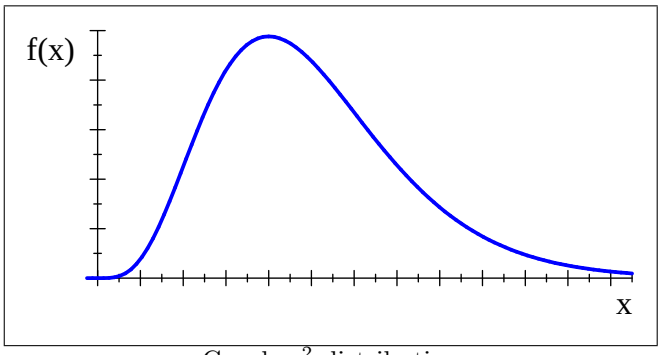
For random variabl $\chi^2 = \frac{(n-1) \cdot S_x^2}{\sigma^2}$ we can with confidence $(1-\alpha)$ determine the confidence interval

$$\chi^2_{\frac{\alpha}{2}} \leq \chi^2 \leq \chi^2_{1-\frac{\alpha}{2}},$$

where $\chi^2_{\frac{\alpha}{2}}$ and $\chi^2_{1-\frac{\alpha}{2}}$, are $\frac{\alpha}{2}$ and $\left(1-\frac{\alpha}{2}\right)$ *quantiles* χ^2-distribution.

It is important to note that χ^2-distribution, compared to the normal distribution is asymmetric distribution (see picture), and therefore $\chi^2_{\frac{\alpha}{2}} \neq \chi^2_{1-\frac{\alpha}{2}}$.

[3] χ the small Greek letter "chi"



Graph χ^2-distribution

Quantiles χ^2-distribution are tabulated and can be found in the annexes to this document.

The value of random variable χ^2-therefore we replace:

$$\chi^2_{\frac{\alpha}{2}} \leq \frac{(n-1)\,S_x^2}{\sigma^2} \leq \chi^2_{1-\frac{\alpha}{2}},$$

adjustments will be the confidence interval for dispersion σ^2:

$$\boxed{\frac{(n-1)\,S_x^2}{\chi^2_{1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)\,S_x^2}{\chi^2_{\frac{\alpha}{2}}}},$$

after extracting get the confidence interval for standard deviation σ

$$\boxed{\sqrt{\frac{(n-1)\,S_x^2}{\chi^2_{1-\frac{\alpha}{2}}}} \leq \sigma \leq \sqrt{\frac{(n-1)\,S_x^2}{\chi^2_{\frac{\alpha}{2}}}}}.$$

Example 46 *Let the systematic error measuring device is zero. Under the same conditions was carried out ten independent measurements of one and the same values μ, where $\mu = 1000m$.Specific details are given below:*

i :	1	2	3	4	5	6	7	8	9	10
$x_i\,[m]$:	92	1010	1005	994	998	1000	1002	999	1000	997

Find a 90% confidence interval for standard deviation σ.

Solution: We needs to calculate the value of the sampling variance S_x^2

$$S_x^2 = \frac{1}{9} \cdot ((992-1000)^2 + (1010-1000)^2 + (1005-1000)^2 +$$
$$+ (998-1000)^2 + (1000-1000)^2 + (1002-1000)^2 +$$
$$+ (999-1000)^2 + (1006-1000)^2 + (997-1000)^2) = 27.$$

From the tables quantiles χ^2-distribution we get for $\alpha = 0,1$ and $n = 10$, $\chi^2_{1-\frac{\alpha}{2}}(9) = \chi^2_{0,95}(9) = 16.919$ a $\chi^2_{\frac{\alpha}{2}}(9) = \chi^2_{0,05}(9) = 3.325$.

90% confidence interval for standard deviation σ is

$$\begin{aligned}
\sqrt{\frac{(n-1)S_x^2}{\chi^2_{1-\frac{\alpha}{2}}}} &\leq \sigma \leq \sqrt{\frac{(n-1)S_x^2}{\chi^2_{\frac{\alpha}{2}}}},\\
\sqrt{\frac{(10-1)S_x^2}{\chi^2_{0,95}}} &\leq \sigma \leq \sqrt{\frac{(10-1)S_x^2}{\chi^2_{0,05}}},\\
\sqrt{\frac{9\cdot 27}{16.919}} &\leq \sigma \leq \sqrt{\frac{9\cdot 27}{3.325}},\\
\sqrt{14.363} &\leq \sigma \leq \sqrt{73.08},\\
3.789\,9 &\leq \sigma \leq 8.549.\spadesuit
\end{aligned}$$

Example 47 *In laboratory experiments were needed, to maintain the standard temperature in the laboratory* 26.5 °C.

In one working week was measured 46 measurements, from which the sample average value $\overline{x} = 26.33\,^\circ\mathrm{C}$ *and sample standard deviation* $S_x = 0.748$.

Determine the 95% confidence interval for μ, σ^2 *and* σ.

$[(26.11; 26.55); (0.3; 0.62); (0.59; 0.79)]$

Chapter 6

Testing statistical hypotheses

The notion of statistical hypothesis understand some claim on the distribution of basic statistical file, respectively. its parameters (for parametric tests). Verification of the veracity of such claims on the merits of random sampling is called hypothesis testing.

In its later being limited to parametric testing. We tested the parameters mean μ and variance σ^2 (respectively standard deviation σ). Generally the parameter we denote Θ.

6.1 Parametric testing an single file

Challenged the two disjoint hypotheses:

null hypothesis

$H_0 : \Theta = \Theta_0$

alternative hypothesis (the opposite claim to the null hypothesis)

$H_1 : \Theta \neq \Theta_0$ (bilateral test).

If we oppose unilateral test the null hypothesis is $H_0 : \Theta \neq \Theta_0$ and alternative hypothesis:

for right-side test $H_1 : \Theta < \Theta_0$,

for left-side test $H_1 : \Theta > \Theta_0$.

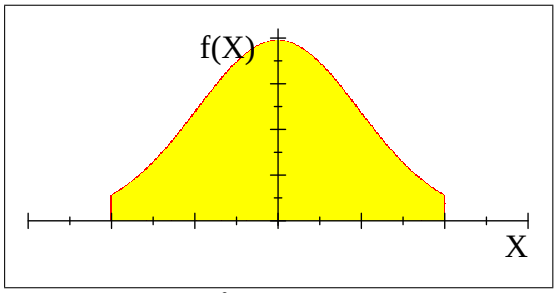
Remark 25 *For one-sided hypothesis would be more correct to formulate the null hypothesis as $H_0 : \Theta \geq \Theta_0$ against the alternative hypothesis $H_1 : \Theta < \Theta_0$ (respectively $H_0 : \Theta \leq \Theta_0$ against the alternative hypothesis $H_1 : \Theta > \Theta_0$). Below, however, dropped from the sign and also for one-sided tests the null hypothesis we formulate $H_0 : \Theta = \Theta_0$.*

To use appropriate decision-making function of a random variable, which we call the test statistic (criterion). Field values of the test statistics divided into two disjoint (one precludes the other) parts:

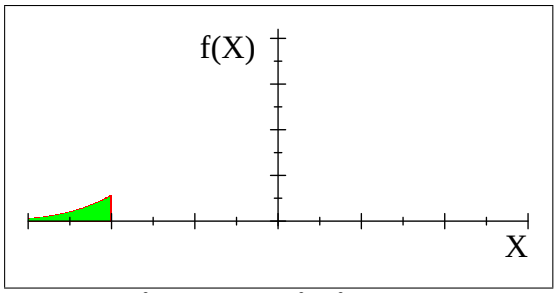
area of acceptance hypothesis H_0

area of refusal (nonacceptance) hypothesis H_0

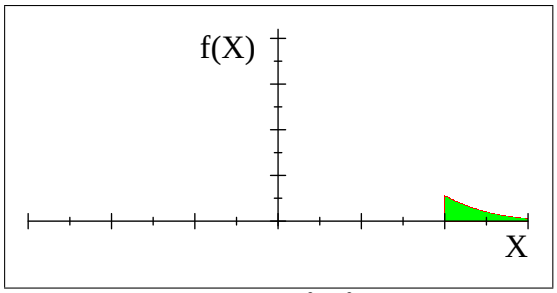
Separating points area of acceptance and area of refusal are quantiles, which are identical with the quantile at a distribution function with given level of significance.(see Figures)



area of acceptance H_0.



left part area of refusal H_0.



right part area of refusal H_0.

When testing statistical hypotheses, we proceed as follows:

1. we determine the null hypothesis H_0 and alternative hypothesis H_0,
2. choose the test statistic,
3. we determine the level of significance α if it the corresponding area of refusal,

4. calculate the test statistic,
5. decision.

Here are some of parametric tests for parameters μ and σ.

6.1.1 Testing parameter μ if the set is small ($n \leq 30$)

Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic:

$$\boxed{T = \frac{\overline{x} - \mu_0}{\sigma} \cdot \sqrt{n}, \qquad \text{respectively } T = \frac{\overline{x} - \mu_0}{S_x} \cdot \sqrt{n}}$$

a) alternative hypothesis: $H_1 : \mu \neq \mu_0$

area of refusal H_0: $|T| \geq t_{1-\frac{\alpha}{2}}^{(n-1)}$

b) alternative hypothesis: $H_1 : \mu < \mu_0$

area of refusal H_0: $T \leq -t_{1-\alpha}^{(n-1)}$ (left-side test)

c) alternative hypothesis: $H_1 : \mu > \mu_0$

area of refusal H_0: $T \geq t_{1-\alpha}^{(n-1)}$ (right-side test)

Example 48 *According to Japanese national agency, rises the average price of land central part of Tokyo for the first six months in 1986 by 49%. Suppose that the international real estate company wants to determine whether the agency-claim is true or not. The company randomly selected 18 owners in downtown Tokyo, where land prices have been known to start in mid 1986. Based on the data that was available found that the average increase in land prices in the selected 18 owners represented in the first half of the reference year 38% with a sample standard deviation 14.*

Significance level $\alpha = 0.01$ verify accuracy of that agency.

Solution: Null hypothesis and alternative hypothesis can be written as:
$H_0 : \mu = 49$
$H_1 : \mu \neq 49$

Sample size is small ($n = 18$) and since standart deviation σ we don´t known we use sample standart deviation S_x. Test statistic is

$$T = \frac{\overline{x} - \mu_0}{S_x} \cdot \sqrt{n} = \frac{38 - 49}{14} \cdot \sqrt{18} = -3.33.$$

To determine the area of refusal we use the values of quantiles of t-distribution, which has $(n-1) = 17$ degrees of freedom with significance level $\alpha = 0.01$,

$$t_{1-\frac{\alpha}{2}}^{(n-1)} = t_{1-\frac{0.01}{2}}^{(17)} = t_{0.995}^{(17)} = 2.898232.$$

Area of refusal H_0 is $|T| \geq t_{1-\frac{\alpha}{2}}^{(n-1)}$,

$$|-3.33| \geq 2.898232.$$

Therefore reject the null hypothesis H_0, since the value -3.33 lies in the area of refusal of H_0.

Agency's argument that the average price of land central part of Tokyo for the first six months of 1986 increased by 49%, we reject.♠

Example 49 *The manufacturer states that the average lifetime it produced reflectors is 70 hours. Competitive firm believes that it is in fact lower, so decided to prove that the manufacturer's claim is not correct. Randomly selected 20 reflectors and found that their average life was 67 hours and the standard deviation was 5 hours. Significance level* $\alpha = 0.05$ *verify if manufacturer's claim is actually incorrect.*

Solution: We determine the null hypothesis $H_0 : \mu = 70$ and alternative hypothesis $H_1 : \mu < 70$. Now we calculate the test statistic

$$T = \frac{\overline{x} - \mu_0}{S_x} = \frac{67 - 70}{5} \cdot \sqrt{20} = -2.683\,3.$$

Test statistic we compare with critical value

$$-t_{1-\alpha}^{(n-1)} = -t_{0,95}^{(19)} = -1.729131.$$

On this basis, we reject the hypothesis H_0, since $T < -t_{0,95}^{(19)}$, and so we accept H_1. Thus the presumption of competitive firms is confirmed.♠

6.1.2 Testing parameter μ if the set is large $(n > 30)$

Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic:

$$\boxed{T = \frac{\overline{x} - \mu_0}{\sigma} \cdot \sqrt{n}, \qquad \text{respectively } T = \frac{\overline{x} - \mu_0}{S_x} \cdot \sqrt{n}}$$

a) alternative hypothesis: $H_1 : \mu \neq \mu_0$

area of refusal H_0: $|T| \geq u_{1-\frac{\alpha}{2}}$

b) alternative hypothesis: $H_1 : \mu < \mu_0$

area of refusal H_0: $T \leq -u_{1-\alpha}$ (left-side test)

c) alternative hypothesis: $H_1 : \mu > \mu_0$

area of refusal H_0: $T \geq u_{1-\alpha}$ (right-side test)

Example 50 *The average time of making certain computational tasks on the computer is 4.56 seconds. Group of researchers tested several new algorithms for which assume that might increase the computing speed of the task. The algorithm, developed, could not say whether the average speed of implementation of tasks increased, decreased or remained unchanged. It therefore decided to test the hypotheses that the average execution time calculation remained unchanged compared with the scenario that the average time of calculation has changed. Conducted 200 random experiments of calculating the various tasks, which found an average time of various tasks, which found the average time of 4.23 seconds for their implementation ($\overline{x} = 4.23$)and sample standard deviation* 2.5 *seconds ($S_x = 2.5$).*

Can we with significance level $\alpha = 0.1$ to conclude that the average time of the new algorithm has been reduced?

Solution: Letś write zero and the alternative hypothesis
$H_0 : \mu = 4.56$
$H_1 : \mu < 4.56$
We use statistics

$$T = \frac{\overline{x} - \mu_0}{S_x} \cdot \sqrt{n} = \frac{4.23 - 4.56}{2.5} \cdot \sqrt{200} = -1.866\,8.$$

The range is greater than 30, so we determine whether it falls within the field acceptance, which pays for

$$T \leq -u_{1-\alpha} = -u_{0.90} = -1.28.$$

We see that outside the field acceptance of hypothesis H_0, therefore, must fall within the field of refusal hypothesis H_0, and therefore the significance level of 0.1, it can be argued that the average time of the new algorithm is accelerated.♠

6.1.3 Testing parameter σ

Null hypothesis: $H_0 : \sigma = \sigma_0$

Test statistics:

$$\boxed{\chi^2 = \frac{(n-1)\,S_x^2}{\sigma_0^2}}$$

a) alternative hypothesis: $H_1 : \sigma \neq \sigma_0$

area of refusal H_0: $\chi^2 \leq \chi^2_{\frac{\alpha}{2};(n-1)}$ alebo $\chi^2 \geq \chi^2_{1-\frac{\alpha}{2};(n-1)}$

b) alternative hypothesis: $H_1 : \sigma < \sigma_0$

area of refusal H_0: $\chi^2 \leq \chi^2_{\alpha;(n-1)}$ (left-side test)

c) alternative hypothesis: $H_1 : \sigma > \sigma_0$

area of refusal H_0: $\chi^2 \geq \chi^2_{1-\alpha;(n-1)}$ (right-side test)

Example 51 *Standard deviation of a particular substance in tablets manufactured by the pharmaceutical company, shall not exceed 0.45 milligrams. If you exceed this amount, a correction must be made in setting the production line. Inspector randomly selected 25 tablets and found that the dispersion of the contents of the substance being studied is 0.2383. What should be concluded if acknowledging the probality of error I. kind is 2.05 (significance level $\alpha = 0.05$)?*

Solution: $n = 25, \qquad \alpha = 0.05, \qquad \sigma_0 = 0.45, \qquad S_x^2 = 0.2383$
$H_0 : \sigma = 0.45$
$H_1 : \sigma > 0.45$
Letś calculate the test statistic

$$\chi^2 = \frac{(n-1)\, S_x^2}{\sigma_0^2} = \frac{(25-1)\cdot 0.2383}{(0.45)^2} = 28.243$$

Quantile is $\chi^2_{1-\alpha;(n-1)} = \chi^2_{0.95;(24)} = 36.415$.

$$28.243 < 36.415.$$

Test statistic is the acceptance area of hypothesis H_0, on the significance level of 0.05 can be argued that the variability of the substance in the tablets isn´t higher than the permissible standard values, and therefore it isn´t necessary to make correct settings of the production line.♠

6.2 Comparing two files

Often in practice we encounter a situation where we want to compare two files. This we mean to compare these two sets of parameters, ie. whether one or more or less than the second. or equal. We will probably compare the mean values of these files, ie. parameter μ_1 of the first file with parameter μ_2 of the second file. How do we best estimates of parameter μ is sample mean $\overline{x}$.

When comparing the averages of two sets we tested several hypotheses. Like we can, whether or not equal to the averages (both-sided test), or whether one or less. greater than the second (one-sided test). For the situations we can formulate the following hypotheses:
$H_0 : \mu_1 = \mu_2$
against
$H_1 : \mu_1 \neq \mu_2$, or $H_1 : \mu_1 > \mu_2$, or $H_1 : \mu_1 < \mu_2$.

6.2.1 Testing equality of means of two fundamental files, when the files are large $(n > 30)$, and if σ_1, σ_2 are known

Null hypothesis:: $H_0 : \mu_1 = \mu_2$

Test statistics:

$$\boxed{U = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}$$

a) alternative hypothesis: $H_1 : \mu_1 \neq \mu_2$

area of refusal H_0: $|U| \geq u_{1-\frac{\alpha}{2}}$

b) alternative hypothesis: $H_1 : \mu_1 < \mu_2$

area of refusal H_0: $U \leq -u_{1-\alpha}$ (left-side test)

c) alternative hypothesis: $H_1 : \mu_1 > \mu_2$

area of refusal H_0: $U \geq u_{1-\alpha}$ (right-side test)

Example 52 *Site selection for new store depends on many factors. One is the level of household income in the area around the proposed site. Suppose that a large store to decide whether to build its next store in town A or town B Although construction costs are lower in city B, the company decided to build in the city and, if there are average monthly household income higher than in city B. The survey of 100 randomly selected households each city found that their average monthly income is in A € 4380, -, in town B € 4050, -. From other sources it is known that the standard deviation of monthly household income is € 520, - at the city's inhabitants A and € 600, - for city residents B.*

Can be with the significance level of 5% say that the average monthly income of households in the city and exceed the average monthly income in the household in the city of B? Assume that income in both cities have a normal distribution.

Solution: Letś formulate hypotheses
$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 > \mu_2$
Test statistics is

$$U = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(4380 - 4050)}{\sqrt{\frac{520^2}{100} + \frac{600^2}{100}}} = 4.1563$$

area of refusal H_0 for $\alpha = 0.05$ is $U \geq u_{1-\alpha}$ (where $u_{1-\alpha} = u_{0.95} = 0.1645$) .

The test statistic is found in the refusal area, so we can with significance level $\alpha = 0.05$ refuse hypothesis H_0 accept hypothesis H_1, žthat average monthly income of households in the city and are higher than in city B.♠

6.2.2 Testing equality of means of two fundamental files, when the files are small $(n < 30)$, and if σ_1, σ_2 are known

Null hypothesis:: $H_0 : \mu_1 = \mu_2$

Test statistics:

$$\boxed{U = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}$$

a) alternative hypothesis: $H_1 : \mu_1 \neq \mu_2$

area of refusal H_0: $|U| \geq t_{1-\frac{\alpha}{2}}^{(n_1+n_2-2)}$

b) alternative hypothesis: $H_1 : \mu_1 < \mu_2$

area of refusal H_0: $U \leq -t_{1-\alpha}^{(n_1+n_2-2)}$ (left-side test)

c) alternative hypothesis: $H_1 : \mu_1 > \mu_2$

area of refusal H_0: $U \geq t_{1-\alpha}^{(n_1+n_2-2)}$ (right-side test)

6.2.3 Testing equality of means of two fundamental files, when the files are large $(n > 30)$, and if σ_1, σ_2 are unknown

In the event that we do not know standard deviations σ_1, σ_2 individual files and these are large $(n = (n_1 + n_2) > 30)$, these parameters we can replace them with estimates S_{x_1}, S_{x_2} (sample standard deviations) and the test statistic will have for the null hypothesis $H_0 : \mu_1 = \mu_2$ form

$$\boxed{U = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{S_{x_1}^2}{n_1} + \frac{S_{x_2}^2}{n_2}}}}$$

For the refusal area will apply:

a) alternative hypothesis: $H_1 : \mu_1 \neq \mu_2$

area of refusal H_0: $|U| \geq u_{1-\frac{\alpha}{2}}$

b) alternative hypothesis: $H_1 : \mu_1 < \mu_2$

area of refusal H_0: $U \leq -u_{1-\alpha}$ (left-side test)

c) alternative hypothesis: $H_1 : \mu_1 > \mu_2$

area of refusal H_0: $U \geq u_{1-\alpha}$ (right-side test)

Example 53 *Several years ago, users of credit cards accounted for some segments. Generally, people with higher incomes and spending tendency prevailed to possess an American Express card, while people with lower incomes and spending more use of VISA cards. For this reason, Visa has intensified its efforts to penetrate even more into groups with higher incomes and through ads in magazines and television are trying to create a greater impression on people. After some time, asked the consulting company, to determine whether the average monthly payments through American Express Gold Card reader about equal payments made Preffered VISA VISA cards. The company did a survey in which 1,200 randomly selected Preffered Visa card holders and found that their average monthly payments were $ 452 with the selection standard deviation of $ 212. Independently of this choice randomly selected 800 Gold Card holders of cards, whose average monthly payments amounted to $ 523 with the selection standard deviation of $ 185. Holders of both cards were excluded from the survey. Survey results confirmed the difference between the average amount of payments made cards and VISA Gold Card Prefferd, overme this hypothesis significance level 0.01.*

Solution: Let´s formulate hypothesis
$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$
Test statistics is

$$U = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{S_{x_1}^2}{n_1} + \frac{S_{x_2}^2}{n_2}}} = \frac{(452 - 523)}{\sqrt{\frac{212^2}{1200} + \frac{185^2}{800}}} = -7.926\,4$$

For $\alpha = 0.01$ is $u_{1-\frac{\alpha}{2}} = u_{0.995} = 2.58$.

Area of refusal H_0 is $|U| \geq u_{1-\frac{\alpha}{2}}$; ie. $|-7.926\,4| \geq 2.58$. We see that the test statistic falls into it. On what basis can we claim that the average payments made by the two credit cards are statistically significant differences.♠

6.2.4 Testing equality of means of two fundamental files, when the files are small $(n < 30)$, and if σ_1, σ_2 are unknown

If the standard deviation is unknown, comparing μ_1, μ_2 using independent random choices of small-scale requires in addition to the independence of choices and normality of distribution essential files and the additional condition that the *variances of both random choices are equal* $(\sigma_1 = \sigma_2)$.

Denote this common dispersion of both random choices σ^2. Its value of course we also do not know, and therefore puts it at variance with the common sample variance S_p^2 of sample variances. Estimate of the variance of the population has a $(n_1 - 1)$ degrees of freedom and variance estimation of the population

2 has a $(n_2 - 1)$ degrees of freedom. Relation for calculating the relationship has form:

$$S_p^2 = \frac{(n_1 - 1) \cdot S_{x_1}^2 + (n_2 - 1) \cdot S_{x_2}^2}{n_1 + n_2 - 2}.$$

Estimate the standard error of the difference of averages: $(\overline{x}_1 - \overline{x}_2)$ is given by formula

$$S_{(\overline{x}_1 - \overline{x}_2)} = \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

The test statistic for test of conformity averages of two basic groups assuming equal variances in their small selection of files for the null hypothesis $H_0 : \mu_1 = \mu_2$ is

$$\boxed{U = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}}$$

a) alternative hypothesis: $H_1 : \mu_1 \neq \mu_2$

area of refusal H_0: $|U| \geq t_{1-\frac{\alpha}{2}}^{(n_1+n_2-2)}$

b) alternative hypothesis: $H_1 : \mu_1 < \mu_2$

area of refusal H_0: $U \leq -t_{1-\alpha}^{(n_1+n_2-2)}$ (left-side test)

c) alternative hypothesis: $H_1 : \mu_1 > \mu_2$

area of refusal H_0: $U \geq t_{1-\alpha}^{(n_1+n_2-2)}$ (right-side test)

Example 54 *Manufacturer, wich mades compact disc players, wants to see whether he proposes a small reduction in prices of its products is sufficient to increase their sales volume. Random data by 14 weekly sales per store before lowering prices found that average weekly sales were 39,600 dollars with a standard deviation of 5060 dollars. 11 random weekly sales for its products by reducing their prices, found that average weekly sales were 41,200 dollars with a standard deviation of 4010 dollars. Show the data that a small reduction in price is sufficient to increase the sales of CD players, if we use the level of significance of 0.05?*

Solution: We find that the sales volume to decrease (file no. 1) is lower than the sales volume decrease (file no. 2). it is a left-sided test, zero and alternative hypothesis are

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 < \mu_2$

We assume that both variances are equal, the test statistic is

$$U = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{(39600 - 41200)}{\sqrt{\frac{13 \cdot 5060^2 + 10 \cdot 4010^2}{23} \left(\frac{1}{14} + \frac{1}{11} \right)}} = -0.857\,17$$

The area of refusal the null hypothesis at significance level $\alpha = 0.05$ is

$$U \leq -t_{1-\alpha}^{(n_1+n_2-2)} = -t_{0.95}^{(14+11-2)} = -t_{0.95}^{(23)} = -1.714.$$

Test statistic is in the area acceptance, so we with significance level $\alpha = 0.05$ say that the producers proposed reducing the prices of CD players has not resulted in increase in sales volume.♠

Chapter 7

Correlation Analysis

Correlation, we understand each other linear relationship (dependence) of two random variables X and Y[1]. This relationship may be *direct*, ie. with increasing values of one variable increase in the value of the second variable and vice versa, or *indirect*, ie. with increasing values of one variable decreases the value of the other and vice versa.

7.1 Coefficient covariance

On whether the two variables X and Y in mutual linear relationship (direct or indirect), we can see from the coefficient covariance of variables X and Y(sign.: $\operatorname{cov} xy$), definated as

$$\boxed{\operatorname{cov} xy = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y}) = \overline{x \cdot y} - \overline{x} \cdot \overline{y}}.$$

If are variable X and Y independent, then $\operatorname{cov} xy = 0$.

If $\operatorname{cov} xy > 0$, between X and Y exists direct linear relationship.

If $\operatorname{cov} xy < 0$, between X and Y existuje indirect linear relationship.

Remark 26 *Covariance can be defined as*

$$\operatorname{cov} xy = E(XY) - E(X) \cdot E(Y),$$

which explains the next definition.

Covariance of the same variable is defined

$$\operatorname{cov} xx = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x}) \cdot (x_i - \overline{x}) = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2 = D(x) = \sigma_x^2.$$

[1] In this section, for clarity, dispensed with random variables by Greek letters ξ and instead tags $\xi_1, \xi_2, \ldots$we will use the next indication $X, Y, \ldots$.

7.2 Correlation coefficient

The strength of linear relationship between two variables in the base set is given by the **correlation coefficient** r_{XY}, which can take only values from the interval $(-1;1)$. If the variables X and Y linearly independent, the correlation coefficient is equal, respectively. very close to zero. Values close to -1 are interpreted as indirect high linear correlation and values close to 1 are interpreted as high a direct linear relationship. Values close to ± 0.5 is interpreted as a weak linear relationship.

However, if values close to zero, we can not say that variables X and Y are independent, but only that they are linear nekorelovatel'né, what we mean for example. nonlinear dependence.

Suppose we know the n pairs of pairs of values $[x_i, y_i]$ variables X a Y obtained for a random selection $i = 1, 2, \ldots, n$ statistical units of the random choice. Then the force of mutual linear dependence of variables X and Y measured correlation coefficient file r_{XY} is defined

$$r_{XY} = \frac{\operatorname{cov} xy}{\sigma_x \cdot \sigma_y},$$

substituting we get

$$r_{XY} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\sqrt{\overline{x^2} - \overline{x}^2}\sqrt{\overline{y^2} - \overline{y}^2}},$$

after a full statement will form the relationship, which we call the Pearson correlation coefficient, by Karl Pearson[2]

$$\boxed{r_{XY} = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \cdot \sqrt{n \cdot \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}}$$

Relatively high correlation coefficient ($r \geq 0.7$) indicates that between variables X and Y is a linear high **mutual** dependence, but that does not mean that there are variables between the high causal dependency, because there may be another variable, eg. Z, from which the variable Y also linearly dependent and which will better explain the variability of the variable Y.

Depending on the degree of causal variables X and Y determine the coefficient of determination and index determination.

7.3 Coefficient of determination

Degree of causal depending variable Y on the variable X expresses the coefficient of determination, defined as the square of correlation coefficient r. In the sample

[2] **Karl Pearson** (* 27. 3. 1857 – † 27. 4. 1936) was an English mathematician and philosopher, proponent of machizmus.

denoted by r^2.

Interpretation of the coefficient of determination is based on an analysis of variance (dispersion) dependent variable Y, which would largely explain the variability of independent variable X, provided that it linearly depends on the size of values YY.

If, for example $r = 0.7$, then $r^2 = 0.49$, which means that only 49% of the variability of the variable Y is explained by a linear relationship with the variable X (regression line). Because 51% of the variability remains unexplained variable Y is a linear relationship with the variable X is clear that the model was chosen improperly (instead of linear dependence be considered non-linear dependence).

Example 55 *HR staff of a company feels that there is a relationship between the number of days absence from work and the worker's age. Randomly selected 10 employees work records and obtain information about their age in years (a random variable X in years) and the number of days, who did the work during the calendar year (random variable Y).*

The data are shown in table

x_i :	27	61	37	23	46	58	29	36	64	40
y_i :	15	6	10	18	9	7	14	11	5	8

.

Assuming that the number of days of absence and worker's age is a linear relationship, consider whether direct or indirect.

Calculate the correlation coefficient and coefficient of determination.

Solution: Intermediate data obtained from the table, which we slightly modify

n	x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
1	27	15	729	225	405
2	61	6	3721	36	366
3	37	10	1690	100	370
4	23	18	529	324	414
5	46	9	2116	81	414
6	58	7	3364	49	406
7	29	14	841	196	406
8	36	11	1296	121	396
9	64	5	4096	25	320
10	40	8	1600	64	320
$\sum$	421	103	19661	1221	3817

Intermediate data write again

$$\begin{aligned} n &= 10, \\ \sum_{i=1}^{n} x_i &= 421, \\ \sum_{i=1}^{n} y_i &= 103, \\ \sum_{i=1}^{n} x_i^2 &= 19661, \\ \sum_{i=1}^{n} y_i^2 &= 1221, \\ \sum_{i=1}^{n} x_i \cdot y_i &= 3817. \end{aligned}$$

To calculate the covariance is most appropriate to use the relationship

$$\begin{aligned} \operatorname{cov} xy &= \overline{x \cdot y} - \overline{x} \cdot \overline{y} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{n} - \frac{\sum_{i=1}^{n} x_i}{n} \cdot \frac{\sum_{i=1}^{n} y_i}{n} = \\ &= \frac{3817}{10} - \frac{421}{10} \cdot \frac{103}{10} = -\frac{5193}{100} = -51.93. \end{aligned}$$

Between the number of days off in a worker's age is an indirect linear relationship (with increasing age the number of days in the year in which the worker does not start to work without giving any reason, decline).

Substituting into the relation

$$r_{XY} = \frac{\operatorname{cov} xy}{\sigma_x \cdot \sigma_y} = \frac{-51.93}{13.917 \cdot 4} = -0.932\,85.$$

where σ_x and σ_x were calculated as

$$\begin{aligned} \sigma_x^2 &= \overline{x \cdot x} - \overline{x} \cdot \overline{x} = \overline{x^2} - \overline{x}^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n} \right)^2 = \\ &= \frac{19661}{10} - \left(\frac{421}{10} \right)^2 = 193.69, \end{aligned}$$

since $\sigma_x = \sqrt{193.69} = 13.917$,

$$\sigma_y^2 = \overline{y^2} - \overline{y}^2 = \frac{\sum_{i=1}^{n} y_i^2}{n} - \left(\frac{\sum_{i=1}^{n} y_i}{n} \right)^2 = \frac{1221}{10} - \left(\frac{103}{10} \right)^2 = 16.01,$$

asince $\sigma_y^2 = \sqrt{16.01} \doteq 4$.

Another option is installed directly into the relationship

$$\begin{aligned} r_{XY} &= \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left(n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right) \cdot \left(n \cdot \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i \right)^2 \right)}} = \\ &= \frac{10 \cdot 3817 - 421 \cdot 103}{\sqrt{(10 \cdot 19661 - 421^2) \cdot (10 \cdot 1221 - 103^2)}} = -0.932\,54. \end{aligned}$$

Correlation coefficient $r = -0.93$ interpretujeme high as an indirect linear relationship between the number of days off in a worker's age.

Coefficient of determination $r^2 = (-0.93)^2 = 0.864\,9$ means, that 86% variability in the number of days off in a year is explained by the influence of age of the worker and 14% of the variability in the number of days off in a year can be explained by other causes such as the linearity between variables X and Y.♠

Example 56 *Group of 100 randomly selected couples were classified by age of wife (X) and husband's age (Y). Characterize the degree of dependence between the ages of husband and wife age coefficient of correlation.*

$X \setminus Y$	*15-25*	*25-35*	*35-45*	*45-55*	*45-60*	*65-75*
15-25	*11*	*7*				
25-35	*1*	*17*	*8*	*1*		
35-45		*2*	*18*	*5*	*1*	
45-55			*2*	*13*	*3*	
45-60				*1*	*6*	*1*
65-75					*1*	*2*

.

Solution:

X \ Y	15-25	25-35	35-45	45-55	45-60	65-75	$\sum$	$n_{j,X} \cdot x$	$n_{j,F} \cdot x^2$
15-25	11	7					18	$18 \cdot 20$	$18 \cdot 20^2$
25-35	1	17	8	1			27	$27 \cdot 30$	$27 \cdot 30^2$
35-45		2	18	5	1		26	$26 \cdot 40$	$26 \cdot 40^2$
45-55			2	13	3		18	$18 \cdot 50$	$18 \cdot 50^2$
45-60				1	6	1	8	$8 \cdot 60$	$8 \cdot 60^2$
65-75					1	2	3	$3 \cdot 70$	$3 \cdot 70^2$
$\sum$	12	26	28	20	11	3	100	3800	161600
$n_{j,Y} \cdot y$	$12 \cdot 20$	$25 \cdot 30$	$28 \cdot 40$	$20 \cdot 50$	$11 \cdot 60$	$3 \cdot 75$	4010		
$n_{j,Y} \cdot y^2$	$12 \cdot 20^2$	$25 \cdot 30^2$	$28 \cdot 40^2$	$20 \cdot 50^2$	$11 \cdot 60^2$	$3 \cdot 75^2$	177300		

$$\begin{aligned}
\overline{x} &= \frac{1}{N} \cdot \sum_{j=1}^{k} (n_{j,X} \cdot x_j) = \frac{1}{100} \cdot 3800 = 38.0, \\
\overline{y} &= \frac{1}{N} \cdot \sum_{j=1}^{k} (n_{j,Y} \cdot y_j) = \frac{1}{100} \cdot 4010 = 40.1, \\
\overline{x^2} &= \frac{1}{N} \cdot \sum_{j=1}^{k} \left(n_{j,X} \cdot x_j^2\right) = \frac{1}{100} \cdot 161600 = 1616.0, \\
\overline{y^2} &= \frac{1}{N} \cdot \sum_{j=1}^{k} \left(n_{j,Y} \cdot y_j^2\right) = \frac{1}{100} \cdot 177300 = 1773.0,
\end{aligned}$$

$$\begin{aligned}
\overline{xy} = \frac{1}{N} \cdot \sum_{j=1}^{k} (n_j \cdot x_j \cdot y_j) = \frac{1}{100} \cdot (&11 \cdot 20 \cdot 20 + 7 \cdot 20 \cdot 30 + \\
&+1 \cdot 30 \cdot 20 + 17 \cdot 30 \cdot 30 + 8 \cdot 30 \cdot 40 + 1 \cdot 30 \cdot 50 + \\
&+2 \cdot 40 \cdot 30 + 18 \cdot 40 \cdot 40 + 5 \cdot 40 \cdot 50 + 1 \cdot 40 \cdot 60 + \\
&+2 \cdot 50 \cdot 40 + 13 \cdot 50 \cdot 50 + 3 \cdot 50 \cdot 60 + \\
&+1 \cdot 60 \cdot 50 + 6 \cdot 60 \cdot 60 + 1 \cdot 60 \cdot 70 + \\
&+1 \cdot 70 \cdot 60 + 2 \cdot 70 \cdot 70) = 1675.0
\end{aligned}$$

$$\begin{aligned}
\operatorname{cov} xy &= \overline{xy} - \overline{x} \cdot \overline{y} = 1675 - 38 \cdot 40.1 = 151.2, \\
\sigma_x^2 &= \overline{x \cdot x} - \overline{x} \cdot \overline{x} = \overline{x^2} - \overline{x}^2 = 1616 - 38^2 = 172.0, \\
\sigma_y^2 &= \overline{y \cdot y} - \overline{y} \cdot \overline{y} = \overline{y^2} - \overline{y}^2 = 1773 - 40.1^2 = 164.99, \\
r_{xy} &= \frac{\operatorname{cov} xy}{\sigma_x \cdot \sigma_y} = \frac{151.2}{\sqrt{172.0 \cdot 164.99}} = 0.89755.
\end{aligned}$$

Correlation coefficient indicates a strong direct linear relationship between age of wife and husband.

$$r_{xy}^2 = 0.897\,552^2 = 0.805\,60.$$

The coefficient of determination, we see that 80% of the variability is explained by the linear dependence.♠

In the previous example, the value of the otherwise than we were previously accustomed. Were included in the table, where each box representing a pair of two sets of values, devolves given abundance. In addressing the proliferation of deployment we properly used. Using a spreadsheet program Microsoft Excel, respectively. OpenOffice Calc is also very likely to speed up routine calculations.

Arranged this way data is called the **correlation table**.

Chapter 8

Paired linear regression

Ak existuje medzi premennými X a Y významná lineárna závislosť, tj. koeficient korelácie je štatisticky významný, bude nás zrejme zaujímať aj rovnica priamky, ktorá túto závislosť medzi premennými X a Y.reprezentuje. Určenie regresného modelu je dôležité pre určenie neznámej závislej premennej Y pre známu hodnotu nezávislej premennej X.

Z matematiky poznáme lineárny vzťah ako súbor dvojíc $[x, y]$, ktoré ležia na priamke $Y = aX + b$. V praktických situáciách, akými sa zaoberá štatistika, vzťah medzi premennými X a Y nie je funkčne lineárny, pretože namerané hodnoty $[x_i, y_i]$ neležia na priamke, ale majú tendenciu ju vytvárať. V takomto prípade hovoríme o štatistickej lineárnej závislosti.

8.1 Regression line

Suppose that we test the two physical variables X and Y, between which there is a linear dependence

$$Y = a + b \cdot X.$$

Parameters β_0, β_1 are unknown. Therefore we do an experiment in which the detected pairs of values $[x, y]$. Measurement of the x values being quite right, it is often possible to set x to a predetermined level, while y is measured with error. Therefore, a new statistical mode

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i, \qquad \text{for } i = 1, 2, \ldots, n,$$

where

y_i is i-th value variable Y in random choice,

β_0 – the value of Y, when variable $X = 0$, in random choice,

β_1 – regression coefficient in the base set, which indicates how many changes y_i, where x_i is changed by a unit of measurement,

x_i – i-th value variable X in random choice,

ϵ_i – random error of variable Y for i-th observation with normal distribution $N\left(0, \sigma^2\right)$.

8.2 Estimation of the parameters β_0 and β_1

Parameters are estimated by the method of least squares. Minimizes the sum of squares of deviations between measured and theoretical values of Y

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 \cdot x_i)^2 \to \min,$$

gives us system of normal equations for the regression line

$$\boxed{\begin{aligned} \beta_0 \cdot n + \beta_1 \cdot \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i, \\ \beta_0 \cdot \sum_{i=1}^{n} x_i + \beta_1 \cdot \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i \cdot y_i. \end{aligned}}$$

Algebraic variety can be achieved, the coefficients line β_0, β_1 can be calculated also by other relations, which are particularly useful if the unknown value $[x_i, y_i]$ of sample, but we know its characteristics, such as averages of variables X and Y, the standard deviation of X and Y respectively. variances, covariance of variables X and Y, the correlation coefficient. Coefficients for the regression line then the

$$\boxed{\begin{aligned} \beta_1 &= \frac{\operatorname{cov} xy}{\sigma_x} = r_{XY} \cdot \sqrt{\frac{\sigma_x}{\sigma_y}}, \\ \beta_0 &= \overline{y} - \beta_1 \overline{x}. \end{aligned}} \tag{9.1}$$

Example 57 *Based on data from Example 55 let's create a point estimate of the regression line according to the number of days of absence and age of the worker, while Let's create a point estimate of the number of days off 25-year employee.*

Solution: From Example 55 we know that

$$\begin{aligned} n &= 10, \\ \sum_{i=1}^{n} x_i &= 421, \\ \sum_{i=1}^{n} y_i &= 103, \\ \sum_{i=1}^{n} x_i^2 &= 19661, \\ \sum_{i=1}^{n} y_i^2 &= 1221, \\ \sum_{i=1}^{n} x_i \cdot y_i &= 3817. \end{aligned}$$

Po dosadení do sústavy

$$\begin{aligned} \beta_0 \cdot n + \beta_1 \cdot \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i, \\ \beta_0 \cdot \sum_{i=1}^{n} x_i + \beta_1 \cdot \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i \cdot y_i, \end{aligned}$$

we get

$$\begin{aligned} \beta_0 \cdot 10 + \beta_1 \cdot 421 &= 103, \\ \beta_0 \cdot 421 + \beta_1 \cdot 19661 &= 3817, \end{aligned}$$

whose solution we get the coefficients $\beta_0 = 21.587$ a $\beta_1 = -0.268$ and the regression equation has form

$$y = 21.587 - 0.268 \cdot x.$$

Point estimate of the number of days off 25-year employee to create a simple regression model substituting value $x_i = 25$

$$y_i = 21.587 - 0.268 \cdot 25 = 14.\,887.$$

One can expect that the average number of days off 25-year employee will be approximately 15 days per calendar year.♠

Example 58 *Estimate the parameters of the regression line of Example 56.*

Solution: Using formulas (9.1) we get

$$\begin{aligned} \beta_1 &= \frac{\operatorname{cov} xy}{\sigma_x} = \frac{151.2}{172} = 0.879\,07, \\ \beta_0 &= 40.1 - 0.879\,07 \cdot 38 = 6.\,695\,3, \end{aligned}$$

and therefore the regression line has form

$$y = 6.\,695\,3 + 0.879\,07 \cdot x.♠$$

Chapter 9

Attachments

$n:$	$t^{(n)}_{0.90}$	$t^{(n)}_{0.95}$	$t^{(n)}_{0.975}$	$t^{(n)}_{0.995}$	$n:$
1	3.078	6.314	12.706	63.657	**1**
2	1.886	2.92	4.307	9.925	**2**
3	1.638	2.353	3.182	5.841	**3**
4	1.533	2.132	2.776	4.604	**4**
5	1.476	2.015	2.571	4.032	**5**
6	1.44	1.943	2.447	3.707	**6**
7	1.415	1.895	2.365	3.499	**7**
8	1.397	1.86	2.306	3.355	**8**
9	1.383	1.833	2.262	3.250	**9**
10	1.372	1.813	2.228	3.169	**10**
11	1.363	1.8	2.201	3.106	**11**
12	1.356	1.782	2.179	3.055	**12**
13	1.35	1.771	2.160	3.012	**13**
14	1.345	1.761	2.145	2.977	**14**
15	1.34	1.753	2.131	2.947	**15**
16	1.337	1.746	2.12	2.921	**16**
17	1.333	1.74	2.11	2.898	**17**
18	1.33	1.734	2.101	2.878	**18**
19	1.33	1.73	2.1	2.861	**19**
20	1.325	1.725	2.086	2.845	**20**
21	1.323	1.721	2.08	2.831	**21**
22	1.321	1.717	2.074	2.819	**22**
23	1.319	1.714	2.069	2.807	**23**
24	1.318	1.711	2.064	2.797	**24**
25	1.316	1.708	2.06	2.787	**25**
26	1.315	1.706	2.056	2.779	**26**
27	1.314	1.703	2.052	2.771	**27**
28	1.313	1.701	2.048	2.763	**28**
29	1.311	1.699	2.045	2.756	**29**
30	1.31	1.697	2.042	2.750	**30**

Student t-distribution quantiles

	0	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**	
0	0.5	0.504	0.508	0.512	0.516	0.52	0.524	0.528	0.532	0.536	**0**
0.1	0.54	0.544	0.548	0.552	0.556	0.56	0.564	0.567	0.571	0.575	**0.1**
0.2	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.606	0.61	0.614	**0.2**
0.3	0.618	0.622	0.626	0.629	0.633	0.637	0.641	0.644	0.648	0.652	**0.3**
0.4	0.655	0.659	0.663	0.666	0.67	0.674	0.677	0.681	0.684	0.688	**0.4**
0.5	0.691	0.695	0.698	0.702	0.705	0.709	0.712	0.716	0.719	0.722	**0.5**
0.6	0.726	0.729	0.732	0.736	0.739	0.742	0.745	0.749	0.752	0.755	**0.6**
0.7	0.758	0.761	0.764	0.767	0.77	0.773	0.776	0.779	0.782	0.785	**0.7**
0.8	0.788	0.791	0.794	0.797	0.8	0.802	0.805	0.808	0.811	0.813	**0.8**
0.9	0.816	0.819	0.821	0.824	0.826	0.829	0.831	0.834	0.836	0.839	**0.9**
1	0.841	0.844	0.846	0.849	0.850	0.853	0.855	0.858	0.860	0.862	**1**
1.1	0.864	0.867	0.869	0.871	0.873	0.875	0.877	0.879	0.881	0.883	**1.1**
1.2	0.885	0.887	0.889	0.891	0.893	0.894	0.896	0.898	0.900	0.901	**1.2**
1.3	0.903	0.905	0.907	0.908	0.91	0.911	0.913	0.915	0.916	0.918	**1.3**
1.4	0.919	0.921	0.922	0.924	0.925	0.926	0.928	0.929	0.931	0.932	**1.4**
1.5	0.933	0.934	0.936	0.937	0.938	0.939	0.941	0.942	0.943	0.944	**1.5**
1.6	0.945	0.946	0.947	0.948	0.949	0.951	0.952	0.953	0.954	0.954	**1.6**
1.7	0.955	0.956	0.957	0.958	0.959	0.96	0.961	0.962	0.962	0.963	**1.7**
1.8	0.964	0.965	0.966	0.966	0.967	0.968	0.969	0.969	0.97	0.971	**1.8**
1.9	0.971	0.972	0.973	0.973	0.974	0.974	0.975	0.976	0.976	0.977	**1.9**
2	0.977	0.978	0.978	0.979	0.979	0.98	0.98	0.981	0.981	0.982	**2**
2.1	0.982	0.983	0.983	0.983	0.984	0.984	0.985	0.985	0.985	0.986	**2.1**
2.2	0.986	0.986	0.987	0.987	0.987	0.988	0.988	0.988	0.989	0.989	**2.2**
2.3	0.989	0.99	0.99	0.99	0.99	0.991	0.991	0.991	0.991	0.992	**2.3**
2.4	0.992	0.992	0.992	0.992	0.993	0.993	0.993	0.993	0.993	0.994	**2.4**
2.5	0.994	0.994	0.994	0.994	0.994	0.995	0.995	0.995	0.995	0.995	**2.5**
2.6	0.995	0.995	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996	**2.6**
2.7	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	**2.7**
2.8	0.997	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	**2.8**
2.9	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.999	0.999	0.999	**2.9**

Distribution function of a standard normal probability distribution $\xi \sim N(0;1)$

$n:$	$\chi^2_{0.99;(n)}$	$\chi^2_{0.975;(n)}$	$\chi^2_{0.95;(n)}$	$\chi^2_{0.05;(n)}$	$\chi^2_{0.025;(n)}$	$\chi^2_{0.01;(n)}$	$n:$
1	6.635	5.024	3.841	0.004	0.001	0	**1**
2	9.21	7.378	5.991	0.103	0.051	0..2	**2**
3	11.345	9.348	7.815	0.352	0.216	0.115	**3**
4	13.277	11.143	9.488	0.711	0.484	0.297	**4**
5	15.086	12.833	11.07	1.145	0.831	0.554	**5**
6	16.812	14.449	12.592	1.635	1.237	0.872	**6**
7	18.475	16.013	14.067	2.167	1.69	1.239	**7**
8	20.09	17.535	15.507	2.733	2.18	1.646	**8**
9	21.666	19.023	16.919	3.325	2.7	2.088	**9**
10	23.209	20.483	18.307	3.94	3.247	2.558	**10**
11	24.725	21.92	19.675	4.575	3.816	3.053	**11**
12	26.217	23.337	21.026	5.226	4.404	3.571	**12**
13	27.688	24.736	22.362	5.892	5.009	4.107	**13**
14	29.141	26.119	23.685	6.571	5.629	4.66	**14**
15	30.578	27.488	24.996	7.261	6.262	5.229	**15**
16	32	28.845	26.296	7.962	6.908	5.812	**16**
17	33.409	30.191	27.587	8.672	7.564	6.408	**17**
18	34.805	31.526	28.869	9.39	8.231	7.015	**18**
19	36.191	32.852	30.144	10.117	8.907	7.633	**19**
20	37.566	34.17	31.41	10.851	9.591	8.26	**20**
21	38.932	35.479	32.671	11.591	10.283	8.897	**21**
22	40.289	36.781	33.924	12.338	10.982	9.542	**22**
23	41.638	38.076	35.172	13.091	11.689	10.196	**23**
24	42.98	39.364	36.415	13.848	12.401	10.856	**24**
25	44.314	40.646	37.652	14.611	13.12	11.524	**25**
26	45.642	41.923	38.885	15.379	13.844	12.198	**26**
27	46.963	43.195	40.113	16.151	14.573	12.879	**27**
28	48.278	44.461	41.337	16.928	15.308	13.565	**28**
29	49.588	45.722	42.557	17.708	16.047	14.256	**29**
30	50.892	46.979	43.773	18.493	16.791	14.953	**30**
40	63.691	63.691	55.758	26.509	24.433	22.164	**40**
50	76.154	76.154	67.505	34.764	32.357	29.707	**50**
60	88.379	88.379	79.082	43.188	40.482	37.485	**60**
70	100.425	100.425	90.531	51.739	48.758	45.442	**70**
80	112.329	112.329	101.879	60.391	57.153	53.54	**80**
120	158.95	158.95	146.567	95.705	91.573	86.923	**120**

Quantiles χ^2-distribution

Bibliography

[1] J. Chajdiak, E. Rublíková, M. Gudába - ŠTATISTICKÉ METÓDY V PRAXI, STATIS Bratislava 1994.