Lectures on probability and statistics

I. Nagy (2020)

Contents

1 Variables and data ........................................ 3
   1.1 Data file ............................................. 3

2 Probability and random variable ......................... 5
   2.1 Probability ........................................... 5
   2.2 Conditional probability .................... 7
   2.3 Random variable ................................... 9
   2.4 Random vector .................................... 10

3 Description of random variable and vector ............... 10
   3.1 Distribution function of random variable .......... 10
   3.2 Distribution function of random vector .......... 11
   3.3 Probability and density functions ................. 11
   3.4 Probability and density functions of random vector 16
   3.5 Moments ........................................... 20
   3.6 Computation ....................................... 21

4 Important distributions .................................. 23
   4.1 Discrete random variable .................. 23
   4.2 Continuous random variable ............ 24
   4.3 Sample distributions ....................... 25
5 Regression analysis
   5.1 Linear regression .............................................. 26
   5.2 Nonlinear regression .......................................... 28

6 Data sample and point estimates
   6.1 Data sample ...................................................... 29
   6.2 Point estimation ................................................ 31

7 Statistical inference
   7.1 Confidence intervals .......................................... 32
   7.2 Test of hypotheses .............................................. 33

8 Tests with one sample ............................................. 34

9 Tests with two samples ............................................ 35

10 Tests with more samples ......................................... 36

11 Tests of independence ........................................... 37

12 Validation in regression analysis .............................. 37

13 Revision ............................................................ 39
1 Variables and data

1.1 Data file

**Example**

We measure intensity of traffic flow at a specified point of roadway. The measurements are repeated each 10 seconds so the result of measuring is a sequence of real values. The entries of the sequence are denoted by \( x_t, t = 1, 2, \cdots, N \), where integers \( t \) represent discrete time of measurements (it denotes a period in which the variable has been measured).

**Data file** \( D \) is a file of measured values of a variable

\[
D = \{ x_t \}_{t=1}^N = \{ x_1, x_2, \cdots, x_N \}
\]

where \( N \) is a number of measurements

**Examples**

The mostly used variables in transportation are “intensity of traffic flow”, “density of traffic flow” (or “occupancy”), “speed of cars” or “speed of traffic flow”, “lengths of queues in crossroads arms”, “type of traffic accident”, “number of cars taking part in an accident” etc.

From the previous example we can recognize two different types of data. The first five variables have entries as real values - they are called continuous variables, the values of the rest of them have entries as integers - they are called discrete ones.

Data can be stored and used basically in two forms

- as plain data \( x_i \),
- values \( X_i \) and frequencies \( n_i \).

**Example**

Plain data: \( x = \{4, 2, 3, 2, 2, 3, 2, 4, 3, 3\} \)

\[
\begin{array}{c|ccc}
X_i & 2 & 3 & 4 \\
n_i & 4 & 5 & 2 \\
\end{array}
\]

can be saved as \( \frac{X_i}{n_i} \) values \( X \) and frequencies \( n \). This latter form of data can be used not only for storing but also directly for computing - see later the characteristics.

Sometimes, instead of data we use their ranks

\[
\begin{array}{cccc}
\text{data } x_i & \text{ordered data} & \text{ranks } r_i \\
5,2,8,3,6 & 2,3,5,6,8 & 3,1,5,2,4 \\
\end{array}
\]

because 5 has the order 3 in the ordered data file, etc.
1.1.1 Characteristics of data file

- **average**

\[ \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i = \frac{1}{N} \sum_{X} n_i X_i = \sum_{X} X_i f_i \]

where \( f_i = \frac{n_i}{N} \) are relative frequencies

**Example:** For \( x = [1, 2, 1, 2, 2, 1, 1] \) the average is the sum of entries divided by their number; it is \( \frac{11}{8} \). There are values 1 which repeats 5 times and 2 with repetition 3. Relative frequencies are \( \frac{5}{8} \) for 1 and \( \frac{3}{8} \) for 2. Thus the average is \( 1 \times \frac{5}{8} + 2 \times \frac{3}{8} = \frac{11}{2} \) which is the same result.

- **variance, standard deviation**

\[ s^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 = \sum_{X} (X_i - \bar{x})^2 f_i \]

**Example:** For \( x = [2, 3, 1, 4] \) the average is 2.5 and the variance

\[ \left[ (2 - 2.5)^2 + (3 - 2.5)^2 + (1 - 2.5)^2 + (4 - 2.5)^2 \right] / 4 = 1.25 \]

- **quantile, critical value** \( \zeta_\alpha, z_\alpha \)

It is a border separating \( \alpha \cdot 100\% \) of the smallest values (quantile) or greatest values (critical values) of a data file.

**Example:** For dataset \( x = [5, 2, 4, 8, 2, 4, 1, 3, 6, 5] \) find quantile and critical value with \( \alpha = 0.1 \) (i.e. 10%).

First order the values of \( x \)

\[ \text{ord} (x) = [1, 2, 2, 3, 4, 4, 5, 5, 6, 8] . \]

As the number of data is 10 and we want to separate 10\% of the smallest, the border lies between 1 and 2. The border is the average, i.e. it is \( \xi_{0.1} = 1.5 \). For 10\% of the greatest the border lies between 6 and 8 so it is \( z_{0.1} = 7 \).

- **median** \( x_{0.5} \) is the \( \xi_{0.5} \), i.e. 50% quantile.

**Example:** For the above dataset, the median is \( x_{0.5} = \frac{4+4}{2} = 4 \).

- **mode** \( \hat{x} \) is the value of dataset which has the higher frequency of repetition.

**Example:** The above dataset has three modes: 2, 4 and 5.
1.1.2 Graphs

- **time graph**: plots values of $x$ in a discrete time of measurements: 1,2,3,…
- **scatter graph** (*xy*-graph): plots values of $y$ against values of $x$ (used mainly in regression)
- **bar graph**: the values of $x$ in time are plotted as columns.
- **histogram**: is similar to the bar graph but it does not plot values of $x$ it plots frequencies of individual values $X_i$.

2 Probability and random variable

2.1 Probability

Probability is introduced through several basic notions.

**Random experiment** - is a trial with correctly defined set of possible results, that occur accidentally (not a chaos - there are some rules how frequently individual results occur)

**Example**: Tossing a dice with the result 1,2,3,4,5 or 6.

**Result** - defined outcome of the experiment.

**Example**: The number that fell during the dice roll.

**Event** - set of results.

**Example**: E.g. the set {2, 4, 6} represents the event “the even number will fall”.

**Probability** - function that assigns to each event a real number. The following axioms must be fulfilled:

- the number must be nonnegative (probability cannot be negative),
- the maximum number is one (if something is sure, it as probability 1 - we say it is 100 percent)
- the function is additive: I.e. for arbitrary two events $E_1$ and $E_2$ whose product is empty set $E_1 \cap E_2 = \emptyset$, it holds $P(E_1 \cup E_2) = P(E_1) + P(E_2)$, i.e. the probability of their union is equal to the sum of their individual probabilities.

**Remark**

Probability can be viewed as an area of a set. For two disjoint sets it holds that the area of their union is equal to the sum of their areas. For two sets with nonempty product this is not true!!

**Remark**: We say that an event “has occurred” if the result we obtain is an element of this event (as a set of results). An event odd number occurred if the result was say 3 (or 1 or 5).
2.1.1 Definitions of probability

Up to now, we have only delimited the notion of probability. We have not discussed its value. This is defined through the following two definitions - classical and statistical.

**Classical definition** of probability is given by the following formula

\[ P = \frac{m}{n} \]

where \( m \) is a number of ways how to obtain a positive result and \( n \) is a number of all ways how to obtain any result.

**Example:** With the dice and the event \( E \) - “odd number” the positive results are \{1, 3, 5\} and all possible results are \{1, 2, 3, 4, 5, 6\}. So \( m = 3 \), \( n = 6 \) and the probability \( P = \frac{3}{6} = 0.5 \).

**Remark**

Notice, that the classical definition concerns possibilities, not experiments and their results.

**Statistical definition** of probability is given by a similar formula

\[ P = \frac{M}{N} \]

but here \( M \) is a number of experiments with positive result and \( N \) is a number of all experiments performed.

**Example:** For determining the probability of the result “odd number” we perform \( N = 100 \) experiments and \( M = 53 \) out of them was positive (we obtained either 1 or 3 or 5). The statistical probability is \( P = \frac{53}{100} = 0.53 \).

**Remark**

Here, instead of analyzing possibilities we just perform experiments.

Comments to the definitions

- The classical definition gives a fixed value of probability while that according to the statistical one will vary for each serial of \( N \) experiments.

- Evidently, for \( N \to \infty \) the value of probability according to the statistical definition will converge to that of the classical one.

- It is easy to use the statistical definition - we perform experiments and count those with positive result. On the other hand, the classical definition needs a throughout analysis of the experiment performed. It is possible only for the simplest experiments like throwing a dice. In practice, the experiments are complex and the “true” classical probability is estimated using the statistical one. This is the subject of the whole inference statistic.
Consequence

It holds: If some event has probability \( p \), then we can expect that in \( N \) experiments it occurs \( pN \)-times.

**Example**

Consider an example of throwing two dices. What is the probability that the sum on both dices will be greater than ten.

We would like to use classical definition of probability, however, the results (sums on both dices) are not equally probable. E.g. the result 2 can be got only by 1 and 1. The result 2 can occur in two ways 1 and 2 or 2 and 1, etc. What is unique is the couple of numbers at first and second dice. The situation can be depicted as follows (vertical axis represents first dice and horizontal second - inside are sums)

\[
\begin{array}{cccccc}
1 & 2 & 3 & 4 & 5 & 6 \\
1 & 2 & 3 & 4 & 6 & 6 & 7 \\
2 & 3 & 4 & 6 & 6 & 7 & 8 \\
3 & 4 & 6 & 6 & 7 & 8 & 9 \\
4 & 5 & 6 & 7 & 8 & 9 & 10 \\
5 & 6 & 7 & 8 & 9 & 10 & 11 \\
6 & 7 & 8 & 9 & 10 & 11 & 12 \\
\end{array}
\]

Here we can see that the total number of different results is \( 6 \times 6 = 36 \) and the number of positive ones is 3. So the probability will be

\[
P(x > 10) = \frac{3}{36} = \frac{1}{12}.
\]

**2.2 Conditional probability**

The definition of conditional probability is

\[
P(E_1|E_2) = \frac{P(E_1, E_2)}{P(E_2)}
\]

where \( P(E_1, E_2) \) is the probability of intersection \( E_1 \cap E_2 \).

From this definition we can get so called **chain rule**

\[
P(E_1, E_2) = P(E_1|E_2) P(E_2)
\]

**Example:** For the experiment of throwing a dice we take \( E_1 = \{2, 4, 6\} \) \( \cdots \) “even number” and \( E_2 = \{1, 2, 3\} \) \( \cdots \) “less than 4”. Then \( E_1 \cap E_2 = \{2, 4, 6\} \cap \{1, 2, 3\} = \{2\} \). \( P(\{2\}) = \frac{1}{6}; P(E_2) = P(\{1, 2, 3\}) = \frac{1}{2}. P(E_1|E_2) = \frac{1}{6}/\frac{1}{2} = \frac{1}{3}. \)
Remark
The result can be logically verified. The condition is \{1, 2, 3\}, i.e. nothing else could appear. From it, only the result 2 meets the event \(E_1\). So, one positive and three possible results gives probability \(\frac{1}{3}\).

Example
We have 3 white and 5 blue beans. We draw two beans without returning the first drawn. What is the probability that the second drawn bean will be blue?

The probability of the second draw cannot be computed directly because we do not know what is the color of the first draw. We must proceed as follows. We denote \(b_1\) is the first draw, \(b_2\) is the second one. The colors will be \(w\) - white and \(b\) - blue. What we want to know is

\[ P(b_2 = b) = P(b_2 = b, b_1 = w) + P(b_2 = b, b_1 = b) = (*) \]

i.e. the second draw is blue after either first draw was white or first draw was blue

\[(*) = P(b_2 = b|b_1 = w) P(b_1 = w) + P(b_2 = b|b_1 = b) P(b_1 = b) \]

where we used the definition of the conditional probability (we used chain rule).

These probabilities can be directly evaluated

\[ P(b_2 = b|b_1 = w) = \frac{5}{7} \]
\[ P(b_1 = w) = \frac{3}{8} \]
\[ P(b_2 = b|b_1 = b) = \frac{4}{7} \]
\[ P(b_1 = b) = \frac{5}{8} \]

where in numerators are numbers of positive present beans and in numerators are total numbers of present beans. If some bean has been drawn in the first draw it is missing in the second draw.

The final result is

\[ P(b_2 = b) = \frac{5}{7} \cdot \frac{3}{8} + \frac{4}{7} \cdot \frac{5}{8} = 0.625 \]

The computations can be demonstrated in the following graph
In the graph, circles represent the current state of beans and the numbers at the arrows are probabilities. The leftmost circle is the beginning the second column is the first draw and the third one is the second draw. The result is sum of probabilities where the second draw is blue.

2.3 Random variable

We have spoken about experiment and its results. Further on we will need to use operation line averaging. If the results are non numerical, e.g. the colors at the signal light, we are in a trouble - what is average color? That is why we introduce the notion of random variable. It is equivalent to the random experiment with its results always numerical. If they are numbers it is OK. If they are not numerical, we simply assign them numbers.

**Example:** The signs on signal lights (red, yellow and green) can be denoted e.g. by numbers 1, 2 and 3.

The following definition is sufficient for us:

**Random variable** is a variable whose values occur randomly - similarly as the results of random experiment. We can imagine that it is a standard variable whose values are affected by some noise.

Generally, we have two types of random variables:

**Discrete** random variable - it has a finite number of values (mostly integers).

**Example:** Flipping a coin, throwing a dice, drawing colored balls from an urn etc. are discrete random variables.

**Continuous** random variable - which has uncountable many possible real values (intervals).

**Example:** Speed of a passing car, waiting time for bus etc. They are examples of continuous random variables.
2.4 Random vector

is a vector of random variables

\[ X = [X_1, X_2, \ldots, X_n] \]

Remark

The contribution of introducing random vector is twofold:

1. The random variables involved in the vector are treated together - e.g. they are measured in one object under investigation (traffic intensities in four arms of a crossroads).

2. New characteristics can be introduced - association between variables. We can investigate if the variables are mutually influenced. The tool for evaluation this dependency is covariance (will be introduced later).

3 Description of random variable and vector

3.1 Distribution function of random variable

A general description of random variable (both discrete and continuous) is distribution function, defined through probability as follows

\[ F_X (x) = P(X \leq x) \]

where \( X \) is a random variable, \( x \) is a realization (number).

Remark

The description of random variable cannot concern its values which are generated with some degree of randomness. It specifies only probability that the value occurs in a given interval. Here, the intervals are \(( -\infty, x)\).

The following two pictures show examples of distribution functions for both the cases.
3.2 Distribution function of random vector

For two random variables $X, Y$ we define joint distribution function by the formula

$$F_{X,Y}(x,y) = P \left( X \leq x \quad \text{and} \quad Y \leq y \right).$$

The probability evaluates all values of random variable $X$ that are less or equal to the number $x$ and all values of $Y$ that are $Y \leq y$. The area of all such points $[x, y]$ is depicted in the picture.

3.3 Probability and density functions

Distribution function provides thorough description of random variable and moreover it is common to both types of random variable. However, for further work another description of random variable, based on the distribution function, is more convenient. It is probability function and density function. These functions must be defined for discrete and continuous case, separately.

3.3.1 Probability function

For discrete random variable we can introduce the description directly as a discrete function with the function values given by probabilities of the values of random variable $X = x$.

**Probability function** is defined by the following formula
\( f_X(x) = P(X = x), \ x \in X \)

Here \( X \) is random variable and \( x \) is a number (value of \( X \)).

Each probability function must have nonnegative values and their sum must be equal to one.

**Example:** For the random variable defined by the experiment throwing a dice, the probability function is defined at \( x \in \{1, 2, 3, 4, 5, 6\} \) and its values are all the same and equal to \( \frac{1}{6} \) (the probability of each side of the dice).

General form of probability function in graph is in the following picture

![Probability Function Graph](image)

The definition of probability function can be given by formula - e.g

\[ f(x) = p^x (1-p)^{1-x}, \ \text{for} \ x = 0, 1 \]

where \( p \in (0, 1) \) is a probability\(^1\).

The most frequently used form of definition of probability function is through a table. For the previous case the tale will be

<table>
<thead>
<tr>
<th>( x )</th>
<th>( f(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1 - ( p )</td>
</tr>
<tr>
<td>1</td>
<td>( p )</td>
</tr>
</tbody>
</table>

### 3.3.2 Characteristics of discrete random variable

In difference to the probability function that gives full stochastic description of a random variable, the characteristics give only partial but very simple information. They speak either about a level or the variability of the values of the random variable.

- expectation
  \[ E[X] = \sum_X x_i f(x_i) \]

\(^1\)It is so called Bernoulli distribution.
• variance, standard deviation

\[ D[X] = \sum_X (x_i - E[X])^2 f(x_i) \]

• quantile, critical value

\[ \sum_{x_i \leq \zeta_\alpha} f(x_i) = \alpha, \quad \sum_{x_i \geq \zeta_\alpha} f(x_i) = \alpha \]

• mode, median: \( \arg \max \) from \( f(x) \); quantile or critical value for \( \alpha = 0.5 \)

**Example**

Random variable \( X \) is defined through the following table

<table>
<thead>
<tr>
<th>( x )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f(x) )</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Compute its expectation \( E[X] \), variance \( D[X] \) and standard deviation.

**Expectation**

\[ E[X] = 1 \cdot 0.2 + 2 \cdot 0.1 + 3 \cdot 0.1 + 4 \cdot 0.3 + 5 \cdot 0.2 + 6 \cdot 0.1 = 3.5 \]

**Variance**

\[ D[X] = (1 - 3.5)^2 \cdot 0.2 + (2 - 3.5)^2 \cdot 0.1 + (3 - 3.5)^2 \cdot 0.1 +
\ + (4 - 3.5)^2 \cdot 0.3 + (5 - 3.5)^2 \cdot 0.2 + (6 - 3.5)^2 \cdot 0.1 = 2.65 \]

**Standard deviation**

\[ \sqrt{D[X]} = \sqrt{2.65} = 1.628 \]

### 3.3.3 Density function

For continuous random variable it holds, that its each single value has zero probability - its total number of values is \( \infty \), then according to the classical definition of probability \( \frac{1}{\infty} = 0 \). That is why we cannot follow the definition of probability function in the discrete case and must define the density separately, as follows.

**Density function** is a real function defined as a derivative of the distribution function.
\[ f(x) = \frac{dF(x)}{dx} \rightarrow F(x) = \int_{-\infty}^{x} f(t) dt. \]

The second (right) definition is implicit. It follows from the first (left) form and has integral form. Using this integral form we can easily derive formula for probability of an interval \((a, b)\)

\[ P(X \in (a, b)) = \int_{a}^{b} f(x) dx = F(b) - F(a) \]

because it holds \( F(x) = P(X \leq x) \) and \( \int_{-\infty}^{b} - \int_{-\infty}^{a} = \int_{a}^{b} \).

**Remark**

The difference of density function from the probability function is that in density function we speak not about probabilities of points but about probabilities of intervals.

An example of density function is in the following picture\(^2\).

3.3.4 Characteristics of continuous random variable

The same characteristics as for the discrete random variable are defined but instead of sum there is integral.

Expectation

\[ E[X] = \int_{-\infty}^{\infty} x f(x) dx \]

Variance

\[ D[X] = \int_{-\infty}^{\infty} (x - E[X])^2 f(x) dx \]

\(^2\)The distribution used in the picture as normal or Gaussian one.
Quantile $\zeta_\alpha$ and critical value $z_\alpha$

$$
\int_{-\infty}^{\zeta_\alpha} f(x) \, dx = \alpha, \quad \int_{z_\alpha}^{\infty} f(x) \, dx = \alpha
$$

Mode $\hat{x}$ and median $x_{0.5}$

$$
\hat{x} = \arg \max (f(x)), \quad x_{0.5} = \zeta_{0.5}
$$

**Example 1**

Determine $E[X]$ and $D[X]$ if the density function is

$$
f(x) = \frac{3}{4} \left( 1 - (x - 1)^2 \right) \text{ for } x \in (0, 2)
$$

The density function is in the following picture

![density function](image.png)

**Expectation**

$$
E[X] = \int_0^2 x \frac{3}{4} \left( 2x - x^2 \right) \, dx = \frac{3}{4} \int_0^2 \left( 2x^2 - x^3 \right) \, dx = \frac{3}{4} \left[ \frac{2}{3}x^3 - \frac{1}{4}x^4 \right]_0 = \frac{3}{4} \left( \frac{2}{3} - \frac{1}{4} \right) = 4 - 3 = 1
$$

which also follows directly from the graph.

**Variance**

$$
D[X] = \int_0^2 (x - 1)^2 \frac{3}{4} \left( 2x - x^2 \right) \, dx = \frac{3}{4} \int_0^2 \left( -x^4 + 4x^3 - 5x^2 + 2x \right) \, dx = \frac{3}{4} \left[ -\frac{1}{5}x^5 + x^4 - \frac{5}{3}x^3 + x^2 \right]_0 = \frac{1}{5}
$$

**Remark**
Computation of 0.05-quantil:

\[
\int_{0}^{\zeta_{0.05}} \frac{3}{4} (2x - x^2) \, dx = 0.05
\]

\[
\frac{3}{4} \left[ x^2 - \frac{1}{3} x^3 \right]_{0}^{\zeta_{0.05}} = 0.05 \quad \Rightarrow \quad \zeta_{0.05}^2 - \frac{1}{3} \zeta_{0.05}^3 = \frac{4}{3} 
\]

\[
\zeta_{0.05}^3 - 3\zeta_{0.05}^2 + 0.2 = 0
\]

It cannot be solved analytically.

**Example 2**

*Random variable* \(X\) *is defined by the formula*

\[
f(x) = p(1-p)^{x-1}, \quad x = 1, 2, \cdots
\]

*Find* \(\alpha\)-quantile \(\zeta_\alpha\) *and* \(\alpha\)-critical value \(z_\alpha\) *of this distribution with* \(\alpha = 0.05\).*

*The probability function has a form of geometrical sequence with the first term* \(a_1 = p\) *and the quotient* \(q = 1 - p\). *The sum of its first* \(n\) *terms is*

\[
s_n = a_1 \frac{1 - q^n}{1 - q} = p \frac{1 - (1-p)^n}{1 - (1-p)} = 1 - (1-p)^n
\]

*Now, quantile is equal to* \(n\) *for which* \(s_n = \alpha\). *Taking logarithm of the previous equation we get*

\[
\ln(1 - \alpha) = n \ln(1 - p)
\]

\[
\frac{\ln(1 - \alpha)}{\ln(1 - p)}, \text{ for } \alpha = 0.05
\]

*The result is not integer. Its round up is an approximation of the quantile.*

*The critical value can be obtained by setting* \(\alpha = 0.95\).

*For* \(p = 0.6\) *we get* \(\zeta_{0.05} = \text{round}(0.05) = 0\) *and* \(z_{0.05} = \text{round}(3.269) = 3\).*

### 3.4 Probability and density functions of random vector

Similarly as for a single random variable we need first to introduce the notion of distribution function. Then, separately for discrete and continuous cases the notions of probability function and density function can be introduced.
3.4.1 Probability function of random vector

For two random variables $X, Y$ we define joint probability function as

$$f_{X,Y} (x,y) = P \left( X = x, \quad Y = y \right)$$

where inside the probability the “,” (comma) means logical “and”.

For two random variables we are in a plane $x, y$ and given $x$ and given $y$ is a point $[x, y]$ in this plane.

3.4.2 Density function of random vector

For two random variables $X, Y$ we define joint density function as a derivative of the distribution function as follows

$$f_{X,Y} (x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y} (x,y)$$

or in an integral form

$$F_{X,Y} (x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y} (u,v) dudv$$

**Remark**

As $F (x,y) = P ([x,y] \in (-\infty,x), (-\infty,y))$ again it holds that the probability of random vector belonging to some area is equal to the integral of density function over this area.

3.4.3 Factorization of random vector

The description $f(x,y)$ is called joint distribution. It can be factorized according to the chain rule in the following way

$$f(x,y) = f(x|y) f(y) \quad \text{or} \quad f(y|x) f(x)$$

where $f(y)$ or $f(x)$ are marginal distributions. They can be computed:

for discrete random variable (e.g. for $X$)

$$f(x) = \sum_y f(x,y)$$
for continuous random variable

\[ f(x) = \int_{-\infty}^{\infty} f(x,y) \, dy \]

\( f(x_1|x_2) \) or \( f(x_2|x_1) \) are conditional pf. (meaning)

The second distribution in the factorization is conditional one

\[ f(y|x) = \frac{f(x,y)}{f(x)}. \]

### 3.4.4 Covariance

Covariance is a characteristics of two random variables \( X \) and \( Y \). It is defined for:

Discrete random variable

\[ C[X,Y] = \sum_x \sum_y (x - E[X])(y - E[Y]) \, f(x,y) \]

Continuous random variable

\[ C[X,Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - E[X])(y - E[Y]) \, f(x,y) \, dx \, dy \]

For two correlated variables it holds that if \( x \) grows \( y \) mostly also grows or if \( x \) falls \( y \) also falls. In this case the covariance is positive and the variables are called positively correlated. If \( x \) grows and \( y \) falls or \( x \) falls \( y \) grows than their covariance is negative and they are called negatively correlated.

### 3.4.5 Uncorrelated random variables

Random variables \( X \) and \( Y \) are called uncorrelated if their covariance is equal to zero

\[ C[X,Y] = 0. \]

It means no correlation exists between them.
3.4.6 Independent random variables

Random variables $X$ and $Y$ are called independent if it holds

$$f(x, y) = f(x) f(y)$$

Independence means that the random variables do not influence one another. From the known behavior of $X$ you can learn nothing about $Y$ and vice versa.

**Remark**

The two lately introduced notions of independent and uncorrelated variables are similar. However, independence is stronger. It holds that independent variables are always uncorrelated. The reverse is not true. Uncorrelated variables can be dependent.

**Example**

We have two discrete random variables $X$ and $Y$ with the joint probability function

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.4</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>

Determine covariance $C[X, Y]$.

First we must to compute marginal probability functions

$$f(x) = \sum_y f(x, y) = [0.5, 0.5]$$

$$f(y) = \sum_x f(x, y) = [0.6, 0.4]$$

Then expectations can be determined

$$E[X] = \sum_x x f(x) = 1 \cdot 0.5 + 2 \cdot 0.5 = 1.5$$

$$E[Y] = \sum_y y f(y) = 1 \cdot 0.6 + 2 \cdot 0.4 = 1.4$$

Now, according to the definition formula for covariance it is

$$C[X, Y] = \sum_x \sum_y (x - E[X]) (y - E[Y]) f(x, y)$$

$$= (1 - 1.5) (1 - 1.4) \cdot 0.4 + (1 - 1.5) (2 - 1.4) \cdot 0.1 + (1 - 1.5) (2 - 1.4) \cdot 0.3 = 0.1$$
3.5 Moments

Here we introduce notions of general and central moments. Also, you can notice here, how moments (and at the same time other characteristics) are computed for: “data”, “discrete random variable” and “continuous random variable”. For data you use only summation (no probability or density functions exist). For random variables the probability or density functions must be used. In discrete case, summation is used (over all different values of the variable), in continuous case you use integration.

3.5.1 $k$-th general moments

- data

$$M'_k = \frac{1}{N} \sum_{i=1}^{N} x_i$$

where $x_i$ are measured data, $N$ is the number of data.

- discrete random variable

$$m' = \sum_{x} x^k f(x_i)$$

where $x_i$ are different values of rv, $X$ is the set of all different values of rv, $f(x_i)$ is the value of probability function at the point $x_i$.

- continuous random variable

$$m'_k = \int_{X} x^k f(x) \, dx$$

where $x$ are real numbers, $X$ is the support of rv and $f(x)$ is the density function.

3.5.2 $k$-th central moments

- data

$$M_k = \frac{1}{N} \sum_{i=1}^{N} \left( x_i - M'_1 \right)^k$$

where $x_i$ are measured data, $N$ is the number of data.

- discrete random variable

$$m'_k = \sum_{x} \left( x_i - m'_1 \right)^k f(x_i)$$

where $x_i$ are different values of rv, $X$ is the set of all different values of rv, $f(x_i)$ is the value of probability function at the point $x_i$. 

20
- continuous random variable

\[ m'_k = \int_X \left( x - m'_1 \right)^k f(x) \, dx \]

where \( x \) are real numbers, \( X \) is the support of rv and \( f(x) \) is the density function.

**Remark**

*It holds: expectation is the first general moment; variance is the second central moment.*

### 3.6 Computation

#### 3.6.1 Example

Compute expectation and variance of the of the random variable \( X \) with categorical distribution defined by probability function set by the table

<table>
<thead>
<tr>
<th>( x )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f(x) )</td>
<td>( p )</td>
<td>( 2p )</td>
<td>( 0.3 + p )</td>
<td>( 0.2 )</td>
<td>( p )</td>
</tr>
</tbody>
</table>

Solution

First we determine \( p \):

\[ p + 2p + (0.3 + p) + 0.2 + p = 5p + 0.5 = 1 \]

\( \rightarrow p = 0.1 \). So the table will be

<table>
<thead>
<tr>
<th>( x )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f(x) )</td>
<td>0.1</td>
<td>0.2</td>
<td>0.4</td>
<td>0.2</td>
<td>0.1</td>
</tr>
</tbody>
</table>

\[ E[X] = 1 \cdot 0.1 + 2 \cdot 0.2 + 3 \cdot 0.4 + 4 \cdot 0.2 + 5 \cdot 0.1 = 3 \]

\[ D[X] = (1 - 3)^2 \cdot 0.1 + (2 - 3)^2 \cdot 0.2 \cdots = 1.2 \]

#### 3.6.2 Example

Determine probability that the value of random variable with Poisson distribution \( f(x) = \exp\left\{-\frac{4x}{x!}\right\} \) will be greater than 3.

Solution

\[ P(X > 3) = 1 - P(X \leq 3) = \]

\[ = 1 - \exp\left\{-4\right\} \frac{4^0}{0!} - \exp\left\{-4\right\} \frac{4^1}{1!} - \exp\left\{-4\right\} \frac{4^2}{2!} - \exp\left\{-4\right\} \frac{4^3}{3!} = 0.566 \]
3.6.3 Example

Compute a) expectation, b) variance, c) distribution function, d) 0.05-quantil and e) median of exponential distribution with density function

\[ f(x) = a \exp\{-ax\}, \ x \geq 0, \ a > 0 \]

Solution

- expectation

\[ E[X] = \int_0^\infty x a \exp\{-ax\} \, dx = \frac{1}{a} \int_0^\infty y \exp\{-y\} \, dy \text{ per-partes = } \frac{1}{a} \]

where \( y = ax; \ \text{dy} = a \cdot dx. \)

- variance

\[ D[X] = \int_0^\infty (x - E[X])^2 f(x) \, dx = \int_0^\infty x^2 f(x) \, dx - (E[X])^2 \]

\[ \int_0^\infty x^2 a \exp\{-ax\} \, dx = \frac{1}{a^2} \int_0^\infty y^2 \exp\{-y\} \, dy = \text{two per-partes = } \frac{2}{a^2} \]

\[ D[X] = \frac{2}{a^2} - \frac{1}{a^2} = \frac{1}{a^2} \]

where again \( y = ax; \ \text{dy} = a \cdot dx \text{ and } x = \frac{y}{a}. \)

- distribution function

\[ F(x) = \int_0^x f(t) \, dt = \int_0^x a \exp\{-at\} \, dt = \left[-\exp\{-at\}\right]_0^x = \]

\[ = 1 - \exp\{-ax\}, \ x \geq 0 \text{ otherwise } 0. \]

- 0.05 quantil

\[ \int_0^\zeta f(x) \, dx = 0.05 \rightarrow F(\zeta) = 0.05 \]

\[ 1 - \exp\{-a\zeta\} = 0.05 \]

\[ \exp\{-a\zeta\} = 0.95 \]

\[ -a\zeta = \ln\{0.95\} \]

\[ \zeta = \frac{- \ln\{0.95\}}{a} = \frac{0.051}{a} \]

- median

\[ x_{0.5} = \zeta_{0.5} = \frac{- \ln\{0.5\}}{a} = \frac{0.693}{a} \]

22
4 Important distributions

4.1 Discrete random variable

Bernoulli distribution

A single experiment with only two possible outcomes $x = 0$ (failure) or $x = 1$ (success). The probability of $x = 1$ is constant and equal to $\pi$.

**Example:** a car turns to left or right.

Probability function

$$P(x; \pi) = \pi^x (1 - \pi)^{1-x}, \ x = 0, 1$$  \hspace{1cm} (1)

Binomial distribution

$n$ times repeated Bernoulli experiment. The result $x$ is the number of successes.

**Example:** $n$ times toss a coin. $x = 3$ means we demand so that head comes three times.

Probability function

$$P(x; n, \pi) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}, \ x = 0, 1, 2, \ldots, n$$  \hspace{1cm} (2)

Poisson distribution

It is a limit case for binomial distribution for $n \to \infty$ and $\pi \to 0$ so that $\lambda = n\pi$ is a finite number (called intensity).

Probability function

$$P(x; \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}, \ x = 0, 1, 2, \ldots$$ \hspace{1cm} (3)

General discrete (categorical) distribution

It deals with a random variable $x$ that can take on one of a finite number of different values $\{x_1, x_2, \ldots, x_n\}$, each with its own probability $p_1, p_2, \ldots, p_n$.

Probability function

$$P(x; p_1, p_2, \ldots, p_n) = p_x, \ x \in \{x_1, x_2, \ldots, x_n\}$$ \hspace{1cm} (4)
where \( p_i \geq 0 \) and \( \sum_{i=1}^{n} p_i = 1 \).

Mostly it is defined in a table

| \( x \) | 1 | 2 | \cdots | \( n \) |
|\( f(x) \) | \( p_1 \) | \( p_2 \) | \cdots | \( p_n \) |

### 4.2 Continuous random variable

#### Uniform distribution

It describes a random variable with no preferences for any values, but with fix lower and upper borders. Its discrete version can be obtained from the general discrete distribution for equal probabilities \( p_1 = p_2 = \cdots = p_n = 1/n \).

Probability density function

\[
f(x; a, b) = \begin{cases} \frac{1}{b-a}, & x \in (a, b), \\ 0, & \text{otherwise} \end{cases}
\]

#### Normal distribution

It is the most frequently used distribution describing e.g. errors in repetitive measurements. Its standard version has zero expectation \( \mu = 0 \) and the variance equal to one \( \sigma^2 = 1 \).

Probability density function

\[
f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2}, \quad x \in \mathbb{R}
\]

#### Lognormal distribution

This distribution is similar to normal one for big \( \mu \) but for small \( \mu \) it is unsymmetrical, it means, it guaranties that \( x > 0 \). It is suitable for modeling variables that are naturally non-negative.

Probability density function

\[
f(x; \mu, \sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \left( \frac{\log(x)-\mu}{\sigma} \right)^2}, \quad x > 0
\]
Exponential distribution

This distribution describes processes that monotonously evolve in time. Its application is e.g. in description of failure free state of some product.

Probability density function

\[ f(x; \delta) = \frac{1}{\delta} e^{-\frac{x}{\delta}}, \quad x > 0 \]  \hspace{1cm} (8)

4.3 Sample distributions

They arise in connection with a statistics for estimation (it will be introduced later).

\( \chi^2 \) distribution

It is used in confidence intervals and testing hypotheses about variance or other quadratic characteristics.

Generation

\[ \chi^2 (n) = \sum_{i=1}^{n} (N_i (0, 1))^2 \]  \hspace{1cm} (9)

where \( N (0, 1) \) denotes a realization of standard normal random variable.

Student distribution

It is used in confidence intervals and testing of hypotheses about expectation.

Generation

\[ St (n) = \frac{N (0, 1)}{\chi^2 (n) / n} \]  \hspace{1cm} (10)

where \( n \) is so called number of degrees of freedom.

\( F \) distribution

It is used in confidence intervals and testing of hypotheses about ratio of two variances.

Generation

\[ F (m, n) = \frac{\chi^2 (m) / m}{\chi^2 (n) / n} \]  \hspace{1cm} (11)

where \( m \) and \( n \) are numbers of degrees of freedom.
5 Regression analysis

5.1 Linear regression

Describes linear dependence of explained variable $y$ on explanatory variable $x$. In geometrical view we have a set of points with coordinates $[x_i, y_i]$, $i = 1, 2, \ldots, N$ and we approximate these points by regression line. The approximation is to be optimal - the vertical distance of the regression line to individual points must be minimal. The situation is sketched in the figure.

Here, one point with subscript $i$ is described. This point has coordinates $[x_i, y_i]$. The corresponding point (the same $x$ coordinate) lying on the line is denoted $\hat{y}_i$ and is called prediction. The distance between $y_i$ and $\hat{y}_i$ denoted by $e_i$ is residuum. The line with the equation

$$y = b_1 x + b_0$$

has the position so that the sum of squares of all residua is minimal

$$\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} e_i^2 \rightarrow \min$$

(12)

5.1.1 Derivation for $b_0 = 0$

The derivation for the practical case - when it goes through origin - is very simple. Each point is described by its prediction (lies on the line) plus residuum $e_i$

$$y_i = b_1 x_i + e_i$$

As the line goes through the origin, the coefficient $b_0$ is zero. From it we have $e_i = y_i - b_1 x_i$. We substitute into the criterion (12)

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - b_1 x_i)^2 = \sum_{i=1}^{N} \left[ y_i^2 - 2b_1 y_i x_i + b_1^2 x_i^2 \right] = \sum_{i=1}^{N} y_i^2 - 2b_1 \sum_{i=1}^{N} y_i x_i + b_1^2 \sum_{i=1}^{N} x_i^2$$

26
\[
\begin{align*}
&= \sum_{i=1}^{N} y_i^2 - 2b_1 \sum_{i=1}^{N} y_i x_i + b_1^2 \sum_{i=1}^{N} x_i^2 = S'_y - 2b_1 S'_{xy} + b_1^2 S'_x
\end{align*}
\]

Derivative:
\[-2S'_{xy} + 2b_1 S'_{xx} = 0\]
from which
\[b_1 = \frac{S'_{xy}}{S'_{xx}}\]

5.1.2 General solution

In general case, when \(b_0 \neq 0\) is

Compute averages and second central moments

\[
\bar{x} = \sum_{i=1}^{N} x_i, \quad \bar{y} = \sum_{i=1}^{N} y_i
\]

\[
S_{xx} = \sum_{i=1}^{N} (x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^{N} (y_i - \bar{y})^2, \quad S_{xy} = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})
\]

and then we have:
- regression coefficients
\[b_1 = \frac{S_{xy}}{S_{xx}}, \quad b_0 = \bar{y} - b_1 \bar{x}\]
- correlation coefficient
\[r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}\]

Coefficients \(b_1\) and \(r\) give evidence about the quality of regression.

Show it.

5.1.3 Multivariate regression

Let the model equation is
\[y_i = b_{0,i} + b_{1,i} x_1 + b_{2,i} x_2 + \cdots + b_{n,i} x_{n,i} + e_i\]
where we denote \(x_i = [x_1, x_2, \cdots, x_{n,i}]'.\)
The collected data are \( y = [y_1, y_2, \cdots, y_N] \) and \( x = [x_1 x_2, \cdots, x_N] \) where \( x_i \) are column vectors. We construct matrices

\[
Y = y', \quad X = [1, x']
\]

which is

\[
Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \quad \text{and} \quad X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{n1} \\ 1 & x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1N} & x_{2N} & \cdots & x_{nN} \end{bmatrix}
\]

**Remark**

When we denote \( b = [b_0, b_1, \cdots, b_n]' \) we can write the matrix equation for all measured data

\[
Y = Xb + E
\]

where \( E \) is vector of residuals.

The general solution has the form

\[
b = (X'X)^{-1} X'Y
\]

Prediction for all data is

\[
\hat{y} = Xb
\]

and residual estimates are

\[
\hat{e}_i = y_i - \hat{y}_i.
\]

Variance of residuals is the estimate of model noise variance.

### 5.2 Nonlinear regression

#### 5.2.1 Polynomial regression

Equation

\[
y_i = b_0 + b_1 x_i + b_2 x_i^2 + \cdots + b_n x_i^n + e_i
\]

Estimation - via vector algorithm where in the rows in the matrix \( X \) are vectors \([1, x_i, x_i^2, \cdots, x_i^n]\).

#### 5.2.2 Exponential regression

Equation

\[
y_i = \exp\{b_0 + b_1 x_i + e_i\}
\]

By taking logarithm we obtain

\[
\ln\{y_i\} = b_0 + b_1 x_i + e_i
\]

and the estimation is performed with \( \tilde{y} = \ln\{y\} \) and \( x \).
6 Data sample and point estimates

6.1 Data sample

In this lecture we will work with parametric density functions

\[ f(x, \theta) \]

which means, we know the density function up to the unknown parameter \( \theta \).

**Example**

We measure speeds of passing cars. We assume normal distribution of the speeds (discuss the assumption) with known variance, which is given by experimentally detected percent of too cautious drivers and those who like to exceed the reasonable speed. However, the expectation which depends on the quality of the road is not known. It is the unknown parameter of the distribution.

The estimation of the parameter is performed from a measured data sample. This sample is relatively small with respect to the size of the whole population, so the estimate is not accurate. We need to know the estimate and its possible error.

**Example**

15 persons work at a department of certain institute. They can work at home, but 5 of them must be present. Those who should be present are chosen each day randomly. The age of the persons is

<table>
<thead>
<tr>
<th>person</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>27</td>
<td>28</td>
<td>42</td>
<td>35</td>
<td>27</td>
<td>33</td>
<td>56</td>
<td>37</td>
<td>27</td>
<td>44</td>
<td>59</td>
<td>42</td>
<td>38</td>
<td>35</td>
<td>29</td>
</tr>
</tbody>
</table>

One day, suddenly, we are asked what is the real average age (expectation) of the people employed in the department. As only 5 of them is present, we have to estimate.

With the five present we compute the average age of all persons as the average those five. The result is following:

True expectation (average of all) of all is 38.2

Randomly chosen (present):

<table>
<thead>
<tr>
<th>person</th>
<th>2</th>
<th>3</th>
<th>6</th>
<th>11</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>28</td>
<td>42</td>
<td>33</td>
<td>59</td>
<td>29</td>
</tr>
</tbody>
</table>

Estimated expectation (average from the five) is 37.27

This is the best we can do in our situation.

**Remarks**
1. The sample average is not too far from the real expectation and the longer is the sample, the more precise the result will be (if we choose all, the result will be correct).

2. The minimal age in population is 27, the maximal is 59. The range is 32 and the half-range is 16. The deviation of the estimate is $38.2 - 37.27 = 0.93$, which is much less than the half-range.

3. The sample is not very lucky - one item of the sample is the oldest person ($11 \rightarrow 59$) from all. And still the estimate is very good.

Now let us generalize and justify the results of the example.

**Population** is the basic set of numbers from which we sample. It is random variable $X$ with density $f(x, \theta)$ and the numbers are realizations of this random variable.

**Realization of random sample** is a vector of measured data (numbers, realizations of random variable).

**Random sample** is a vector of equally distributed and independent random variables.

Explanation to random sample:

What we do, is measuring a sample of data - realizations of random variable $X$. They must be independent (why?) and taken from the same random variable (equally distributed) which is to be investigated. However, only theoretically, the sampling can be repeated and as it is random, we obtain another realization of random sample. This theoretical difference (which can be derived from the knowledge of $f(x, \theta)$) can serve as information about the possible error which can be done in estimation.

To use the information about variability in different samples, we define random sample as a vector of random variables. The realizations of the first variable are data on the first position of all possibly taken random samples. The second one corresponds to the data on second position of samples and so on.

In this way the random sample is a vector of random variables and realization of the sample is vector of realizations of these random variables.

In reality, indeed, only one realization of the random sample is taken and used for estimation.

### 6.1.1 Characteristics of random sample

Let us have random variable $X \sim f(x)$ (for now with no parameter) with expectation $E[X] = \mu$ and variance $D[X] = \sigma^2$. From this rv we are going to take a sample $[X_1, X_2, \cdots, X_N]$ of the length $N$.

**Sample average**

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$
Sample variance
\[ S^2 = \frac{1}{N - 1} \sum_{i=1}^{N} (x_i - \bar{x})^2 \]
where the sample variance differs in denominator from the variance, defined earlier (why?).

The characteristic are computed not from numbers, but from random variables. Thus, they are not numbers but again random variables for which we can compute their expectations and variance. We show here the characteristic of the random average.

**Expectation of sample average**
\[
E [\bar{X}] = \int_{-\infty}^{\infty} \bar{X} f(x, \theta) \, dx = \int_{-\infty}^{\infty} \frac{X_1 + X_2 + \cdots + X_N}{N} f(x, \theta) \, dx = \frac{1}{N} \left( \int_{-\infty}^{\infty} X_1 f(x) \, dx + \int_{-\infty}^{\infty} X_2 f(x) \, dx + \cdots + \int_{-\infty}^{\infty} X_N f(x) \, dx \right) dx = \frac{1}{N} \left( \mu + \mu + \cdots + \mu \right)_{N\text{times}} = \mu
\]
as the choice is taken from the same random variable with expectation \( \mu \).

**Variance of sample average**
\[
D [\bar{X}] = D \left[ \frac{1}{N} \sum_{i=1}^{N} X_i \right] = \frac{1}{N^2} \sum_{i=1}^{N} D [X_i] = \frac{1}{N^2} \sum_{i=1}^{N} \sigma^2 = \frac{\sigma^2}{N}
\]
as \( D[aX] = a^2 D[X] \) and random variables in the sample are independent (definition).

### 6.2 Point estimation

We have random variable \( X \sim f(x, \theta) \) which is known, up to the parameter \( \theta \). Point estimate \( \hat{\theta} \) of this parameter, based on the realization of sample \( x = [x_1, x_2, \cdots, x_N] \) (vector of measured data).

**Statistics** \( T \) is a function of random sample whose value can be regarded as estimate of the parameter \( \theta \).

**Remark**
As we have seen, sample average is a good approximation of the parameter \( \mu \) (expectation)
\[ \bar{x} \to E [X] \]
Similarly for other parameters (variance, proportion, correlation coefficient etc.) can be found functions of random sample, that approximate their true value. Such function is called statistics, and should have some properties. They should be

- unbiased
  \[ E[T] = \theta \]

- consistent = unbiased and
  \[ \lim_{N \to \infty} D[T] = 0 \]

- efficient (for \( T_1, T_2 \) unbiased)
  \( T_1 \) is more efficient then \( T_2 \) if \( D[T_1] > D[T_2] \)

**Example**
Sample average is unbiased and consistent statistics with respect to the parameter \( \theta = \mu \).

Prove

We have two samples: first of the length \( N_1 \) and second \( N_2 \). Which sample average is more efficient with respect to parameter \( \mu \).

| Point estimate \( \hat{\theta} \) of parameter \( \theta \) is the value of the statistics corresponding with this parameter (should be unbiased, consistent and sufficiently efficient). |

**Example**
Sample average \( \bar{x} \) is a point estimate of the parameter \( \mu \) (expectation).

7 Statistical inference

7.1 Confidence intervals

**Example**
Ages of persons from certain firm

<table>
<thead>
<tr>
<th>person</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>27</td>
<td>28</td>
<td>42</td>
<td>35</td>
<td>27</td>
<td>33</td>
<td>56</td>
<td>37</td>
<td>27</td>
<td>44</td>
<td>59</td>
<td>42</td>
<td>38</td>
<td>35</td>
<td>29</td>
</tr>
</tbody>
</table>

Estimate the average age from a sample of length 5.

True average age: average from all.

Estimate: average from randomly selected five.

1st sample: 44, 27, 59, 29, 56., average: 43.0

2nd sample: 56, 27, 59, 42, 42., average: 45.2
3rd sample: 29. 28. 38. 59. 27., average: 36.2

For many samples the averages form density function. The area under the curve on same interval is a probability, that realization will fall into this interval.

Here:

$$f(\bar{x}) \sim N(\mu, \frac{\sigma^2}{n})$$

Random variable is the statistics, realizations are point estimates of the parameter. Interval is the area, where the point estimates occur with the mentioned probability.

If the probability is defined as $1-\alpha$ ($\alpha$ being the prob. level), we speak of the **Confidence interval**

![Confidence Interval Diagram]

### 7.2 Test of hypotheses

Similar approach is **Testing of hypothesis**. We have hypotheses

- $H_0$: the parameter is equal to $\cdots$; its estimates come from $f_0(\theta)$.
- $H_A$: it is not $\cdots$; estimates come from another distribution

Solution:

Construct CI according to $H_0$. Compute point estimate $\hat{\theta}$. If $\hat{\theta} \in CI$, do not reject $H_0$. If $\hat{\theta} \notin CI$, reject $H_0$, but you do not acknowledge $H_A$ (it is not expressed as single distribution).

Now, all we have said holds as both-sided, left-tailed and right-tailed. p-value
8 Tests with one sample

8.0.1 Parametric tests (normality required)

- expectation (known × unknown variance) - test of true average
  Ex: A company declares that its production is more than 150 products per day. Somebody opposes and says that it is less.
  Prg: z_test, t_test

- proportion - test of a part from the whole
  Ex: City manager says that only 5% of drivers exceed the permitted speed at certain street. Police is convinced that the ratio is higher.
  Prg: prop_test

- variance - test of variability of a variable
  Ex: Quality of production is given by the dispersion of weight of products is. If it is higher then a given level, the machines must be adjusted. Test, if the machines are OK or it is necessary to tune them.
  Prg: var_test

8.0.2 Nonparametric tests (normality is not required)

- Wilcoxon test: tests median of rv from one sample
  - H₀: median is equal to the assumed value
  - test is all sided
  Ex: Compare caloric intake measured at 11 selected women with the recommended value 7725 kJ.
  Prg: wilcoxon_test

8.0.3 Tests of distribution type

- w/s test of normality (statistics = range / std)
  - H₀: rv is normal
  Prg: norm_test

- Kolmogorov-Smirnov test: tests given distribution. It is based on comparison of assumed and empirical DF.
  - H₀: rv has assumed distribution
  - right sided test with special crit. vals
  Prg: ks_test_cont, ks_test_disc, ks_test_2 (comp.of two distr.)

- Chi-square test of homogeneity: test of distribution type. It compares observed and expected frequencies.
  - H₀: rv has the assumed distribution
right sided test  
Ex: We have measured number of accidents for week days and weekends. Test if they are uniformly distributed.  
Prg: chisquare_test

9 Tests with two samples

9.0.1 Parametric

• two expectations (independent \times paired samples)  
Ex (indep): Company A claims that its production is greater than that of B. Assistant of company B denies it. Test. ... how to determine the side.  
Prg: t_test_2s, t_test_2n  
Ex (paired): Uniformity of tire removal at the front wheels of cars of a specific mark has been investigated. The producer of the cars proclaims uniformity. Test it.  
Prg: t_test_2p

• two proportions  
Ex: Ratio of drivers violating rules in town is greater than outside. Test it.  
Prg: prop_test_2

• two variances  
Ex: Variability of weights of products from company A is greater than those, from company B. Test it.  
Prg: var_test_2

9.0.2 Nonparametric

• Mann-Whitney test: tests equality of two medians (independent samples)  
  - H0: the medians are equal  
  - both sided test  
Ex: Marks form math were checked at two classes of secondary school. 5 marks from the first and 8 marks from the second class were recorded. Compare the classes.  
Prg: mannwhit_test

• Wilcoxon: tests two medians (paired samples)  
  - H0: medians are equal  
  - all sided test  
Ex: At a secondary school an improvement of students in math was checked. In the 1st class eight students were selected and their marks recorded. In the 2nd class the marks of the same students were recorded again. Test, if the results of individual
students are improved.
Prg: wilcoxon_test
• McNemar: tests improvement after some action. Data are yes/no - two by two table of frequencies.
  – H0: no improvement
  – right sided
Ex: 22 selected people were tested for cold (yes/no). Then, they received some drug and after a week they were tested again. Test the effectiveness of the drug.
Prg: mcnemar_test

10 Tests with more samples

10.0.1 Parametric

• Analysis of variance: tests equality of several expectations
  – H0: expectations are equal
  – right sided test
Ex: Test if the power of engine of vehicles of five marks is the same.
Prg: anova_1

• Anova with two factors: tests equality in columns and rows.
Ex: Five cars are tested by three drivers. Test the cars and the drivers.
Prg: anova_2

Auxiliary tests to anova

• Bartlett - test of equality of more variances
Prg: bartlett_test

• Scheffé - detects different samples
Prg: scheffe_test

10.0.2 Nonparametric

• Kruskal-Wallis: nonparametric anova.
  – H0: medians are equal
  – right sided test
Ex: as for anova1
Prg: kruskal_test

• Friedman - block test of equality of medians
  – H0: medians are equal
- test is right sided
Ex. 5 shops are rated by 3 inspectors (each shop is rated by each inspector; inspectors are factors of no interest = block). Evaluate quality of the shops. 
Prg: friedman_test

11 Tests of independence

- Gamma coefficient: test of association of two discrete rvs. It compares prediction from marginal and conditional pf.
- result: how many times the prediction from cond. pf is better than from marginal.
Ex: We measure speed and consumption on driven cars. Is there a relation between these two variables?
Prg: gamma

- Pearson test: tests independence of two rvs. It tests correlation coefficient. (parametric test)
- H0: rvs are independent
- test is both sided
Ex: Test the data x and y if they are suitable for linear regression.
Prg: pearson_test

- H0: rvs are independent
- test is both sided
Prg: spearman_test

- Chi-square test of independence: test if independence of two rvs. Compares observed and expected frequencies. Based on the definition of independence $f(x, y) = f(x)f(y)$.
- H0: rvs are independent
- test is right sided.
Ex: We asked 200 people from three different areas about they pay (low, normal, high). Test if the pay depends on the area.
Prg: chisquare_test_i

12 Validation in regression analysis

Regression can be viewed as approximation of dependence of y on x from data sample by some curve - linear, exponential, polynomial etc. However, not each data sample must be convenient for such approximation. Here we will discuss this question.

1. Draw xy-graph: ideal, good, possible and no good regression.
2. Pearson $t$-test of correlation coefficient

For approximation of a relation between $x$ and $y$ there must be any relation. This is expressed in the **regression coefficient**

$$
\rho = \frac{C[X,Y]}{\sqrt{D[X]D[Y]}} \quad \leftrightarrow \quad r = \frac{S_{xy}}{\sqrt{S_xS_y}}
$$

where $C$ is covariance, $D$ are variances, $S$ are sums

$$
S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) \quad S_x = \sum (x_i - \bar{x})^2 \quad S_y = \sum (y_i - \bar{y})^2
$$

The true property of random variables is expressed in population regression coefficient $\rho$. Its true value is estimated from sample by the statistics $r$ (sample regression coefficient).

Pearson $t$-test has $H_0$: $\rho = 0$, $H_A$: $\rho \neq 0$; both sided test with Student distribution.

$H_0$: $x$ and $y$ are uncorrelated - regression does not have sense. To be able to use regression, $H_0$ has to be rejected.

Prg: **pearson_test**

3. Fisher $F$-test of explained and unexplained variance

In regression, we have data and predictions of data which lie on the regression line. If we want to characterize data $\{y_i\}_{i=1}^N$ without regression, we can compute the average value $\bar{y}$. Then, for a selected $x_i$ we have the value $y_i$ and its prediction $\hat{y}_i$.

Now, the deviation of $y_i$ from $\bar{y}$ can be decomposed as

$$
y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)
$$

where

- $y_i - \bar{y}$ is the error in measurement without taking into account the regression (overall error),
- $\hat{y}_i - \bar{y}$ is a deviation from the average explained by regression (explained error),
- $y_i - \hat{y}_i$ is a deviation of the measured point from the regression line - if regression is precise, all points should lie no the line (unexplained error).

Taking variances, we obtain explained $S_r$ (regression) and unexplained $S_e$ (residual) variances. The statistics is defined as $F = \frac{S_r}{S_e}$ with $F$ distribution. For $H_0$: $F = 0$ is nothing explained and the regression does not have sense. The test is right-sided. Regression has sense, it $H_0$ is rejected.

4. Test of independence of residuals

38
Residuals are deviations of the data from regression line. For correct regression
the residuals must be independent. If not, the relations between them could be
used to construct better regression line.

The test has the statistics
\[
z = \frac{2b - (n - 2)}{\sqrt{n - 1}} \sim N(0, 1)
\]
where \(b\) is number of sequences (deviations from median with the same sign). \(H_0:\)
is independence (for \(z = 0\)).

5. Test for auto-correlation of residuals
It is a similar test to the previous one. We test if a current residuum \(e_i\) can be estimated from the previous one \(e_{i-1}\). We estimate the dynamical regression
\[
e_i = ae_{i-1} + b + \epsilon_i
\]
If \(|a| < 0.3\) and \(k \to 0\), the regression is OK.

6. Standard error of residuals \(SE\)
Residuals \(e_i = y_i - \hat{y}_i\) are errors of approximation of data with regression curve. The smaller the errors are, the better approximation. The standard error is defined as
\[
SE = \frac{var(e)}{var(y)}
\]
which is variance of prediction error \(e_i\) relative to variance of dependent variable \(y_i\).

13 Revision