Lectures on probability and statistics

I. Nagy (2020) - distant study (should be self-explanatory; if not, please, send comments or questions to nagy@fd.cvut.cz)

Contents

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 every 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 the studied 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: $x = \{4, 2, 3, 2, 2, 3, 3, 2, 4, 3, 3\}$

can be saved as

X_i	2	3	4
n_i	4	5	2

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**

data x_i	ordered data	ranks r_i
5,2,8,3,6	2,3,5,6,8	3,1,5,2,4

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* $x = [1, 2, 1, 1, 2, 2, 1, 1]$ *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* $1 \times \frac{5}{8} + 2 \times \frac{3}{8} = \frac{11}{8}$ *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** ζ_α, z_α
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

$$ord(x) = [1, 2, 2, 3, 4, 4, 5, 5, 6, 8].$$

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 and 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,$\cdots$
- **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 - a 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 $E \cdots$ "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

What is the probability that the randomly drawn card from a set of marriage cards will be an ace?

The total number of cards is 32. There are 4 aces. So the probability is

$$P = \frac{4}{32} = \frac{1}{8}$$

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 a box with 3 white and 5 black balls. We randomly draw one ball, and without returning it we draw the second one. What is the probability that the second will be white if we know that the first one was black?

For the second draw one black is missing. So we have 3 white and 4 black. The probability of white then will be

$$P = \frac{3}{7}$$

Remark

What will be the situation if we would return the first ball before drawing the second one?

Here we draw constantly form 3 white and 5 black. The condition has no effect.

2.3 Random variable

We have spoken about the notion of 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

$$X = [X_1, X_2, \cdots, X_n]$$

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

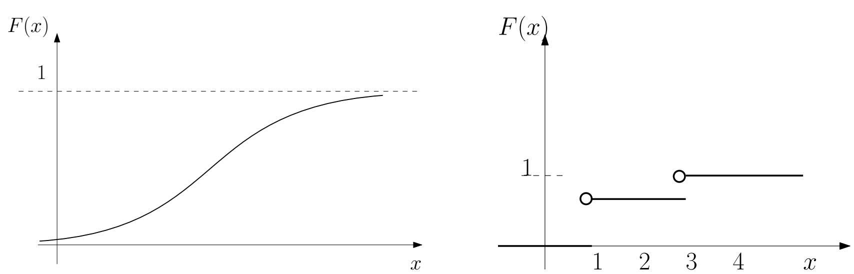
$$F_X(x) = P(X \leq x)$$

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 a probability that the value occurs in a given interval. Here, the intervals are $(-\infty,\ x\rangle$.

The following two pictures show examples of distribution functions for both the cases.



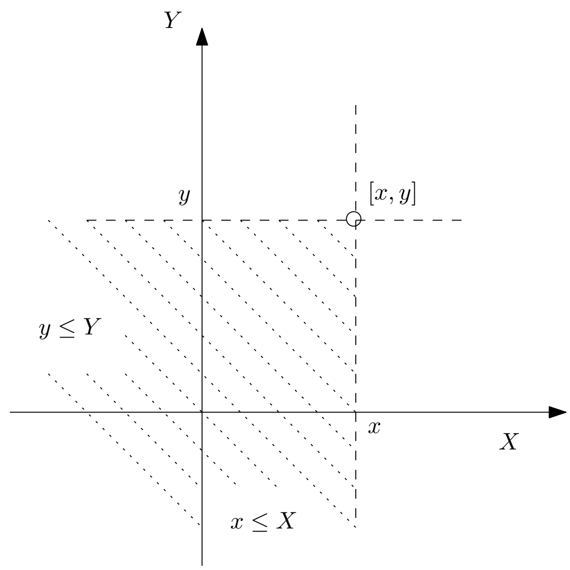
Continuous distribution function   Discrete distribution function

3.2 Distribution function of random vector

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

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

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 a 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 a 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

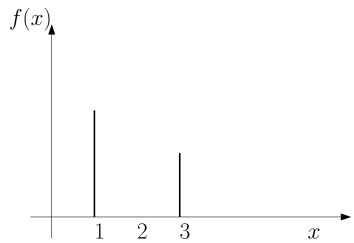
$$f_X(x) = P(X = x), \; x \in X$$

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).*

A general form of probability function in a graph is in the following picture



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 $p \in \langle 0, 1 \rangle$ 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

x	0	1
$f(x)$	$1-p$	p

3.3.2 Characteristics of discrete random variable

In difference to the probability function that gives a full stochastic description of 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

$$E[X] = \sum_X x_i f(x_i)$$

- variance, standard deviation

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

- quantile, critical value

$$\sum_{x_i \le \zeta_\alpha} f(x_i) = \alpha, \quad \sum_{x_i \ge z_\alpha} f(x_i) = \alpha$$

[1]It is so called Bernoulli distribution.

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

Example

Random variable X is defined through the following table

x	1	2	3	4	5	6
$f(x)$	0.2	0.1	0.1	0.3	0.2	0.1

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 ∞, 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 an integral form. Using this integral form we can easily derive the formula for probability of an interval (a, b)

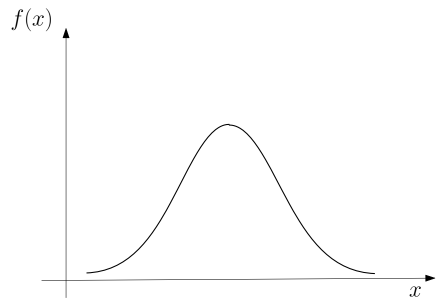
$$P(X \in (a, b)) = \int_a^b f(x)\,dx = F(b) - F(a)$$

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

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

Quantile ζ_α and critical value z_α

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

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

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

Example

[2] The distribution used in the picture as normal or Gaussian one.

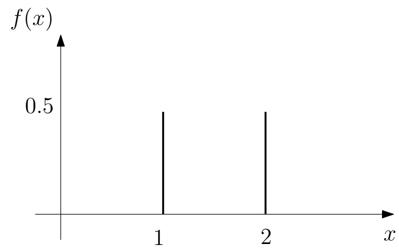
Depict the probability function of the experiment of flipping a regular coin with the assignment: "head" $\rightarrow$ 1, "tail" $\rightarrow$ 2.

Solution

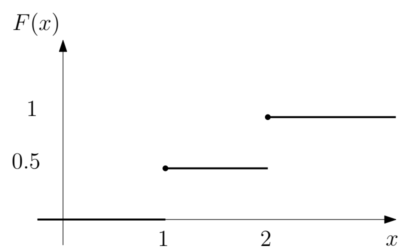
The probability functions can be expressed by a table

x	1	2
$f(x)$	0.5	0.5

Graphical form is a discrete function define only in the values of the random variable. It is as follows



Distribution function is a cumulative probability function and it has the form

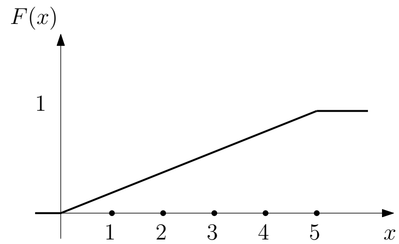


Example

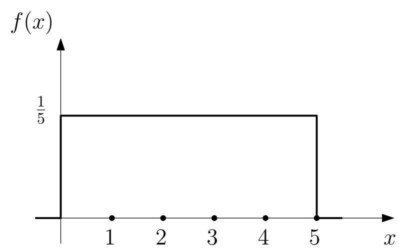
For the experiment waiting on a bus (see Example **??**) construct the distribution and density functions.
Let us recall: We are randomly coming to a bus station where the bus goes with exactly 5 minutes interval. Random variable is defined as a time interval we need to wait for the bus.

Solution

We are looking for the distribution function $F(x) = P(X \leq x)$. So precisely speaking, we are looking for probabilities that we will wait the time x or less. Waiting $x = 0$ has probability 0, waiting $x = 5$ has probability 1 (in the interval $\langle 0, 5\rangle$ the bus surely has to come). As we come randomly (each time instant is equally probable), the distribution function will be linear. So, it will have the form



Density function is a derivative of the distribution function (where the derivative exists). The derivative in the interval (0, 5) exists and it is equal to $\frac{1}{5}$. Outside this interval, i.e. in $(-\infty, 0) \cup (5, \infty)$ it is zero and in 0 and 5 we can define it as zero, too. Then it is



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

$$f_{X,Y}(x, y) = P\left(X = x \underbrace{,}_{\text{and}} Y = y\right)$$

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

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

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

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

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

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

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

$x \backslash y$	1	2
1	0.4	0.1
2	0.2	0.3

Determine covariance $C[X,Y]$.

First, we must 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

$$C[X, Y] = \sum_x \sum_y (x - E[X])(y - E[Y]) f(x, y)$$
$$= (1 - 1.5)(1 - 1.4) \cdot 0.4 + (1 - 1.5)(2 - 1.4) \cdot 0.1 +$$
$$= (2 - 1.5)(1 - 1.4) \cdot 0.2 + (2 - 1.5)(2 - 1.4) \cdot 0.3 = 0.1$$

3.5 Moments

Here, we introduce notions of general and central moments. They are computed with the use of distribution and as the distribution for discrete random variable (probability function) and continuous one (density function) are defined deferentially, also the moments have to be defined separately.

Overall comparison of data moments and population ones can be found in Appendix **??**

3.5.1 k-th general moments

– discrete random variable

$$m_k^{'} = \sum_X x_i^k f(x_i)$$

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

– 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.

– 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 with random variables

Example:

For two random variables X and Y with probability functions $f(x)$ and $f(y)$

x	1	2
$f(x)$	0.4	0.6

y	1	2	3
$f(y)$	0.3	0.5	0.2

construct joint $f(x, y)$ and conditional $f(x|y)$, $f(y|x)$ functions.

The variables X and Y are independent (a change in x does not influence probability of y) so the joint probability function is a product

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

which means a product entry by entry. The result is

$x \backslash y$	1	2	3
1	0.4·0.3	0.4·0.5	0.4·0.2
2	0.6·0.3	0.6·0.5	0.6·0.2

=

$x \backslash y$	1	2	3
1	0.12	0.2	0.08
2	0.18	0.3	0.12

The conditional distributions are

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

The marginals are given or they can be obtained by summing the joint (the first one is a sum of columns, the second one of rows).

$x\backslash y$	1	2	3	$f(x)$
1	0.12	0.2	0.08	0.4
2	0.18	0.3	0.12	0.6
$f(y)$	0.3	0.5	0.2	

The division is again entry-vise. So, $f(x|y)$ is

$x\backslash y$	1	2	3
1	$\frac{0.12}{0.3}$	$\frac{0.2}{0.5}$	$\frac{0.08}{0.2}$
2	$\frac{0.18}{0.3}$	$\frac{0.2}{0.5}$	$\frac{0.12}{0.2}$

=

$x\backslash y$	1	2	3
1	0.4	0.4	0.4
2	0.6	0.6	0.6

Remark

The sum over columns (i.e. for values of x) must be always equal to one. For given value of y, all possibilities for x are 1 or 2. And probability of "all" is one.

The columns of the conditional distribution are all the same. It is due to the independency of the variables. I.e. it holds $f(x|y) = f(x)$.

The second conditional distribution $f(y|x)$ is

$x\backslash y$	1	2	3
1	$\frac{0.12}{0.4}$	$\frac{0.2}{0.4}$	$\frac{0.08}{0.4}$
2	$\frac{0.18}{0.6}$	$\frac{0.3}{0.6}$	$\frac{0.12}{0.6}$

=

$x\backslash y$	1	2	3
1	0.3	0.5	0.2
2	0.3	0.5	0.2

Remark

Here the situation is similar, only transposed.

Example:

Two dependent discrete random variables.

Let us have random variables X and Y with the joint probability function $f(x, y)$

$x\backslash y$	1	2	3
1	0.1	0.2	0.1
2	0.3	0.1	0.2

Construct marginal and conditional probability functions.

Marginals $f(x)$ and $f(y)$ are sums of the joint one

$x\backslash y$	1	2	3	$f(x)$
1	0.1	0.2	0.1	0.4
2	0.3	0.1	0.2	0.6
$f(y)$	0.4	0.3	0.3	

Conditional probabilities are:

$f(x|y) = \frac{f(x,y)}{f(y)}$

$x \backslash y$	1	2	3
1	$\frac{1}{4}$	$\frac{2}{3}$	$\frac{1}{3}$
2	$\frac{3}{4}$	$\frac{1}{3}$	$\frac{2}{3}$

and $f(y|x) = \frac{f(x,y)}{f(x)}$

$x \backslash y$	1	2	3
1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
2	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{3}$

Remark 1

Conditional distribution is a function of the argument before the sign $|$*. So, e.g.* $f(y|x=1)$ *is the first row of the table above. The second row describes the distribution* $f(y|x=2)$*. That is, the whole table (e.g. the last one) is parameterized distribution* $f(y|x)$ *for* $x = 1, 2$*. The same holds for the last but one table. It is* $f(x|y)$ *for* $y = 1, 2, 3$*.*

Remark 2

Notice, that in difference to the previous example with independent variables, here joint distribution is not equal to the marginal one. I.e.

$$f(x,y) \neq f(x|y), \text{ for } y = 1, 2, 3$$

and

$$f(x,y) \neq f(y|x), \text{ for } x = 1, 2$$

Example:

Let us have two random variables X and Y with joint density function

$$f(x,y) = x + xy + \frac{y}{2}, \text{ for } x, y \in (0,1)$$

The marginal $f(x)$ is

$$f(x) = \int_0^1 f(x,y)\,dy = \frac{1}{6}(6x+1), \; x \in (0,1)$$

and $f(y)$

$$f(y) = \int_0^1 f(x,y)\,dx = \frac{1}{2}(2y+1), \; y \in (0,1)$$

Conditional $f(x|y)$

$$f(x|y) = \frac{f(x,y)}{(y)} = \frac{(2x+1)\,y + 2x}{2y+1}$$

and $f(y|x)$

$$f(y|x) = \frac{f(x,y)}{(x)} = \frac{(4x+2)\,y + 4x}{6+1}$$

As it holds $f(x,y) \neq f(x)\,f(y)$ or $f(x|y) \neq f(x)$ or $f(y|x) \neq f(y)$, the variables x and y are dependent.

The expectations are

$$E[X] = \int_0^1 x f(x)\,dx = \frac{5}{8}$$

$$E[Y] = \int_0^1 y f(y)\,dy = \frac{7}{12}$$

Variances

$$D[X] = \int_0^1 (x - E[X])^2 f(x)\,dx = \frac{13}{192}$$

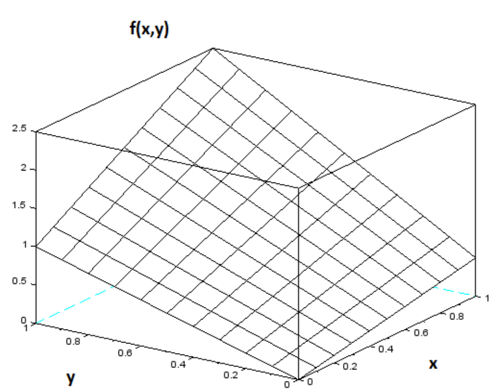
$$D[Y] = \int_0^1 (y - E[Y])^2 f(y)\,dy = \frac{11}{144}$$

Covariance

$$C[X,Y] = \int_0^1 \int_0^1 (x - E[X])\,(y - E[Y])\,f(x,y)\,dxdy = -\frac{1}{288}$$

(Solved in Maxima http://maxima.sourceforge.net)

The density function is plotted in the following picture



Example:

Determine distribution function $F(x)$ of random variable with density function

$$f(x) = \frac{1}{2}x, \text{ on } x \in (0,2)$$

For the distribution function on $x \in (0,1)$ it holds

$$F(x) = \int_0^x f(t)\,dt = \int_0^x \frac{1}{2}t\,dt = \frac{1}{2}\left[\frac{t}{2}\right]_0^x = \frac{1}{4}x^2$$

The whole distribution function for $x \in R$ it holds

$$F(x) = \begin{cases} 0 & \text{for } x \le 0 \\ \frac{1}{4}x^2 & \text{for } x \in (0,2) \\ 1 & \text{for } x \ge 2 \end{cases}$$

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 π.

Example: a car turns to left or right.

Probability function

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

Binomial distribution

n times independently 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

$$P(x;n,\pi) = \binom{n}{x}\pi^x (1-\pi)^{n-x}, \; x = 0,1,\cdots,n \tag{2}$$

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

$$P(x;\lambda) = \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!},\ x = 0, 1, 2, \cdots \tag{3}$$

It is a distribution of counts - e.g. number of cars passing a point of monitoring or number of customers entering a shop.

Geometric distribution

Geometrically distributed random variable counts the number of unsuccessful results of independently repeated Bernoulli experiments until the first success appears.

Example: A shooter shots at a small target. The probability of hitting is $p = 0.2$. What is the probability that the first hit will occur it the x-th shot.

Probability function

$$P(x, p) = p(1-p)^x$$

where p is the probability of success in a Bernoulli experiment.

It expresses the probability that the first success will precede x failures.

General discrete (categorical) distribution

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

Probability function

$$P(x; p_1, p_2, \cdots, p_n) = p_x,\ x \in \{x_1, x_2, \cdots x_n\} \tag{4}$$

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

Each value x of the random variable has its own probability p_x.

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

$$f\left(x; a, b\right) = \frac{1}{b - a},\ x \in \left(a, b\right),\ \text{and zero otherwise} \tag{5}$$

Has a rigid bounds. It expresses absolute uncertainty - all admissible values are equally probable.

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

$$f\left(x; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}},\ x \in R \tag{6}$$

Is generated by a large number of small independent random accidents.

Lognormal distribution

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

Probability density function

$$f\left(x; \mu, \sigma^2\right) = \frac{1}{x\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2}\frac{(\log(x)-\mu)^2}{\sigma^2}},\ x > 0 \tag{7}$$

It resembles normal distribution but it is for nonnegative variables.

Exponential distribution

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

Probability density function

$$f(x;\delta) = \frac{1}{\delta} \mathrm{e}^{-\frac{x}{\delta}},\ x > 0 \tag{8}$$

It is related to theory of reliability and queues.

4.3 Sample distributions

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

χ^2 distribution

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

Generation

$$\chi^2(n) = \sum_{i=1}^{n} \left(N_i(0,1)\right)^2 \tag{9}$$

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 when the variance of the tested variable is not known.

Generation

$$St(n) = \frac{N(0,1)}{\chi^2(n)/n} \tag{10}$$

where n is so called number of degrees of freedom.

F distribution

It is used in confidence intervals and testing of hypotheses about a ratio of two variances, e.g. F-test and Analysis of variance.

Generation

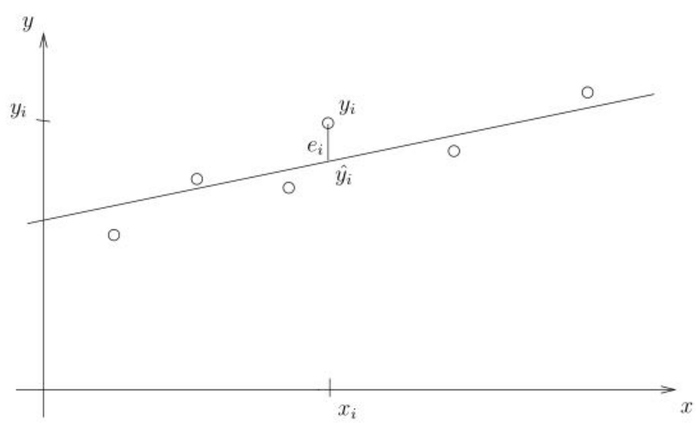
$$F\left(m,n\right)=\frac{\chi_{1}^{2}\left(m\right)/m}{\chi_{2}^{2}\left(n\right)/n} \tag{11}$$

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 a geometrical view, we have a set of points with coordinates $\left[x_i, y_i\right],\ i=1,2,\cdots,N$ and we approximate these points by a 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 $\left[x_i, y_i\right]$. 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

$$y=b_1x+b_0$$

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

$$\sum_{i=1}^{N}\left(y_i-\hat{y}_i\right)^2=\sum_{i=1}^{N}e_i^2\rightarrow \min \tag{12}$$

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 though 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)

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

$$= \sum_{i=1}^{N} y_i^2 - 2b_1 \sum_{i=1}^{N} y_i x_i + b_1^2 \sum_{i=1}^{N} x_i^2 = S_y^{'} - 2b_1 S_{xy}^{'} + b_1^2 S_x^{'}$$

Derivative:

$$-2S_{xy}^{'} + 2b_1 S_{xx}^{'} = 0$$

from which

$$b_1 = \frac{S_{xy}^{'}}{S_x^{'}}$$

5.1.2 General solution

In a 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 the regression.

Show it.

5.1.3 Multivariate regression

Let the model equation is

$$y_i = b_{0,i} + b_{1,i}x_1 + b_2 x_{2,i} + \cdots + b_n x_{n,i} + e_i$$

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

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

which is

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

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 a vector of residuals.

The general solution has the form

$$b = \left(X'X\right)^{-1} X'Y$$

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 $\left[1, x_i, x_i^2, \cdots, x_i^n\right]$.

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 Population and data sample

Population is a set from which we select values, **sample** is what we have selected. Population is fixed, sampling is performed in random, so each sample differs form other samples.

Let us start with two examples:

Example 1

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

person	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
age	27	28	42	35	27	33	56	37	27	44	59	42	38	35	29

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 the average age requested based on those present.

Let the sample (those who are randomly present) is

person	2	3	6	11	15
age	28	42	33	59	29

The estimate of the average age of all (expectation of the age) is

$$\frac{28+42+33+59+29}{5} = 37.27$$

True expectation (average of all) of all is 38.20

We can see that

1. *The sample average is not too far from the real expectation (and the longer the sample will be, the more precise the is).*

2. *If the length of the sample will be 15 (all are present), the estimate will be precise.*

Here:

The **population** are ages of all the persons working at the department.

The **sample** are the ages of those five present.

It is clear that the population is still the same, however, the sample changes day by day. Specifically, the average of the population (expectation) is constant, while the average (sample average) of the sample changes from sample to sample (sampling is random). Also it is clear, that the differences between sample averages will be smaller than the difference between values of the population. The longer the sample is the closer is the sample average to the expectation.

Important: The population and its characteristics are constant, the sample and its characteristics are random (they depend on the sample performed)

Remark

The probability function of the population mentioned is

x	27	28	29	33	35	37	38	42	44	56	59
$f(x)$	$\frac{3}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$

. It can be used for generation. I.e. the age 27 will be generated (or chosen to the sample) with probability $\frac{3}{10}$,the age 28 with probability $\frac{1}{15}$ etc.

Example 2

We measure speeds of passing cars. Here, the sample are measured speeds. But what is the population?

Here, the situation is more complex then in the previous example. The population can be characterized as "all possible speeds that were, are and will be measured" but it is too vague. Instead, we will view the population just like random variable that produces values randomly but according to some probabilistic rules. This random variable is described by its density function (in a discrete case it was probability function). Then the population can be viewed either like all the possible generated values (each with its probability of generating) or better just as the random variable.

As the the probability density is fixed then also the properties and the characteristics of the population will be constant.

The sample (a given number of measured speeds) is random. That is why again, the sample itself as well as its characteristics (e.g. sample average) are random.

Specifically, let the speeds are distributed normally with the expectation $\mu = 84$ km/h and standard deviation $\sigma = 21$ km/h. Denoting speeds by x the population is given by the density function

$$f(x) = \frac{1}{\sqrt{2\pi 21^2}} \exp\left\{-\frac{1}{2}\left(\frac{x-84}{21}\right)^2\right\}$$

Let us generate 10 samples each with 5 measurements. The result is in the table

samples	x_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
1	75.9	100.1	86.4	97.7	119.7	93.3	51.0	80.6	92.1	77.3
2	110.4	70.7	92.5	85.2	95.6	85.5	107.4	74.4	63.5	67.8
3	107.6	73.6	89.7	90.7	76.5	71.7	61.9	63.8	34.1	96.5
4	64.8	109.0	132.1	73.2	112.7	116.5	78.6	121.7	64.8	98.1
5	69.7	99.1	18.3	73.7	51.9	62.5	49.0	119.0	61.7	76.4
s. average	85.68	90.5	83.8	84.1	91.28	85.9	69.58	91.9	63.24	8.22

Average of sample averages: 82.92

Standard deviation of sample averages: 9.38

We can see:

(1) The sample averages vary less than the data (e.g. in the first table row). They are not far form the expectation 84.

(2) The standard deviation of data is 21 which is more than standard deviation of the sample averages which is 9.38.

(3) The average of the sample averages 82.92 is close to the expectation 84. With the growing number of sample averages their average converges to the value of expectation.

Remark

All these observations will further be formulated as assertions or definitions.

6.1 Population and its parameters

Population is given by the experiment we monitor and investigate. It can be e.g. speeds of passing cars, seriousness of traffic accident or intensities of the traffic flow in arms of a controlled crossroads.

Notice: The population concerns the experiment generally. Measuring the speeds - no specific speed is given, yet. We can imagine it either as the experiment or the set of all possible values that can be measured, including probabilities of measuring these values.

The description of population is provided by its distribution. This distribution id generally unknown (we can hardly define the probabilities of a car going with a speed in specified interval, say 60-65 km/h; it is not possible to analyze all possible drivers that can go through a specified point, the possibilities of their cars etc.) However, often we select some type of distribution (frequently normal) and say that only its parameters are unknown. We say, the distribution is known up to some of its parameters. The general task then is to estimate these parameters.

Let the **population** is given by the density function

$$f(x, \theta)$$

where θ are the **unknown parameters** (e.g. expectation).

The task is to construct the **point estimate** $\hat{\theta}_t$ of the parameter θ based on the set of measured data $\{x_i\}_{i=1}^t$.

6.2 Random sample

The notion of random sample is the most difficult from the whole statistics.

As we already have said, the sample is used for estimation of the properties of the population, specifically of the unknown parameters of the distribution by which the population is described.

Important !!!

It is necessary to distinguish between a set of measured data, which we will call **realization of a sample** and **sample** itself as a possibly measured set of values when we repeat our choice. To explain this rather vague assertion we return to the Example 2 from the beginning of this section. Here, we have made a serial of 10 choices, each with 5 sampled data. As the choices have been random, all the choices were different. At the bottom of the table we listed also averages from the choices. Naturally, they were also different. From it we can see, that the random choice and its characteristics (here average) are random. Now, similarly as we characterized random variable as a variable whose values differ in each measurement, we can define the sample as a vector random variables on which we can measure its vector values - realizations of sample.

In this way we define:

Sample is a vector of independent and equally distributed random variables.

Realization of sample is a realization of data sample as random vector.

Remarks

1. Independence in the definition of a sample expresses the fact, that the measured values must be representative - they must represent the whole population. And this will be guaranteed if the choice will be performed independently.
2. Equally distributed from the definition says that the data must be measured on the same experiment. E.g. if I want to analyze traffic during the day I must not include data measured at night. Or, if I analyze speeds of cars in a free movement I must exclude all data measured when the cars were stacked in a queue.
3. Realization of a sample is a single data vector which is used for further actions as e.g. parameter estimation. In practice, no repetition of data choices is used. The possibility of repetitive choices is used only for theoretical considerations.
4. To stress randomness of a sample it is often called **random sample**. Also realization of a sample is often called **data sample**.

6.2.1 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 $D[X] = \sigma^2$. On this random variable, we have a sample

$[X_1, X_2, \cdots, X_N]$ of the length N, where $X_1, X_2, \cdots, X_N$ are independent and equally distributed (with the same expectation μ and variance σ^2) random variables. We can define the following characteristics:

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, define earlier.

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. They have he following properties:

Expectation of sample average

$$E\left[\bar{X}\right] = \mu$$

Proof

$$E\left[\bar{X}\right] = \int_{-\infty}^{\infty} \bar{X} f(x,\theta)\, dx = \int_{-\infty}^{\infty} \frac{X_1 + X_2 + \cdots + X_N}{N} f(x,\theta)\, dx =$$
$$= \frac{1}{N}\left(\int_{-\infty}^{\infty} X_1 f(x)\, dx + \int_{-\infty}^{\infty} X_2 f(x)\, dx + \cdots + \int_{-\infty}^{\infty} X_N f(x)\, dx\right) dx =$$
$$= \frac{1}{N}\left(\underbrace{\mu + \mu + \cdots + \mu}_{N\,\text{times}}\right) = \mu$$

as the choice is taken from the same random variable with expectation μ.

Variance of sample average

$$D\left[\bar{X}\right] = \frac{\sigma^2}{N}$$

Proof

$$D\left[\bar{X}\right] = D\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \frac{1}{N^2}\sum_{i=1}^{N} D\left[X_i\right] = \frac{1}{N^2}\sum_{i=1}^{N} \sigma^2 = \frac{\sigma^2}{N}$$

as $D\left[aX\right] = a^2 D\left[X\right]$ and random variables in the sample are independent (definition).

These properties show the following:

1. If we take infinitely many sample realizations, compute sample averages of them and make an average of these sample averages (which makes the expectation $E\left[\bar{X}\right]$), we obtain precisely the expectation of the population.

2. The more sample realizations we take the more accurately we obtain the value of the population expectation when averaging their sample averages.

Conclusion: Estimation of population expectation by replacing it with the sample average has a good sense.

6.3 Point estimation

In the previous paragraph, we have shown that sample average is a good estimator of population expectation. Now, we are going to generalize this assertion.

We have random variable $X \sim f(x, \theta)$ which is known, up to the parameter θ. We want to construct 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). To this end we introduce the notion of statistics:

Statistics T is a function of random sample whose value can be regarded as estimate of the parameter θ.

Remark

As we have seen, sample average is a good approximation of the parameter μ (expectation)

$$\bar{x} \to E[X]$$

Similarly for other parameters (variance, proportion, correlation coefficient etc.) there can be found functions of random sample, that approximate their true value. Such function is called statistics, and should have some properties (that guarantee it estimates just this parameter). The statistics should be:

– **unbiased**

$$E[T] = \theta$$

– **consistent** = unbiased and

$$\lim_{N \to \infty} D[T] = 0$$

– **efficient** (for T_1, T_2 unbiased)

$$T_1 \text{ is more efficient then } T_2 \text{ if } D[T_1] > D[T_2]$$

Example

Sample average is an unbiased and consistent statistics with respect to the parameter $\theta = \mu$.

Proof

$E\left[\bar{X}\right] = \mu$; $\lim_{N\to\infty} D\left[\bar{X}\right] = \lim_{N\to\infty} \frac{\sigma^2}{N} = 0$

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

Solution

If $N_1 < N_2$ then $\frac{\sigma^2}{N_1} > \frac{\sigma^2}{N_2}$ and the second statistics is more efficient than the first one.

The notion of statistics is the basic one in the whole subject of the theory of Statistics. Its detailed treatise can be found (and it is recommended to read) in the Appendix **??**.

Now, we can formally define the notion of point estimate:

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

Example

A good statistics for estimation of population expectation μ is the sample average. Thus, sample average $\bar{x}$ is a point estimate of μ.

Notice

For estimation we use average of the sample realization $\bar{x}$ which is a number and not average of the sample $\bar{X}$ which is random variable. For estimation we use only a single realization of the sample. No repetition of sampling is performed.

Practical example

Let us return to the Example 2 from the beginning of this Section 6. Here we had the population with expectation 84 and standard deviation 21. We took 10 samples to show theoretical properties of sample average (here we looked at the sample a average theoretically as a random variable).

Now, we take only the first sample realization

$$x = [75.9,\ 110.4,\ 107.6,\ 64.8,\ 69.7]\,.$$

This sample realization will be used for estimation of the expectation. The estimate $\hat{\mu}$ will be the value of the sample average that has been computed

$$\hat{\mu} = 85.68$$

If we use the same data generator (normal distribution wit the expectation 84 and standard deviation 21) and generate longer samples, we obtain

sample length	10	20	50	100	1000	10000
point estimate	87.08	85.45	81.40	88.36	83.07	84.01

Here, we can see, that the estimates are almost true only for samples longer than 1000. The reason is that the variance of the population is rather large.

7 Statistical inference

In the previous Chapter, we spoke about point estimation of unknown parameters of population. We have said that the point estimate is constructed from a single sample realization taken from the population under investigation. Also we know, that we could have taken another sample realization (samples can theoretically be repeated). And as the sampling is random, the new sample realization would differ from this we have done and also the statistics (e.g. sample averages) would differ. We have said, that they are random variables (random variable gives different values; statistics based on different sample realizations gives also different values). So it, as a random variable, is described by its distribution $f(T)$, where T is the statistics.

Very important summarization !!!

We have **population** X that is described by its distribution $f(x,\theta)$ which is known up to the parameter θ.

We define **statistics** $T(x,\theta)$ (for estimation of θ)

Finally, we have the **distribution of the statistics** T, which can be derived as a transformation of $f(x,\theta)$ according to the function $T = T(x,\theta)$. This transformation is complex and we will not discuss it here.

The distribution $f(x,\theta)$ points at data from the population.

The distribution $f(T,\theta)$ points at the parameter θ - the value of T is a point estimate of θ:

$$T(x) = \hat{\theta}$$

As we are now interested in parameters, all further derivations will concern the distribution of the statistics $f(T,\theta)$.

Example

Let us consider population described by normal distribution with variance 1 and unknown expectation μ. We have taken a sample realization $x = [x_1, x_2, \cdots, x_{50}]$ of the length 50. We want to determine point estimate of μ and the distribution of the chosen statistics T.

The first step is to choose a statistics suitable for estimation of μ. We choose sample average

$$T = \bar{x} = \frac{1}{50}\sum_{i=1}^{50} x_i$$

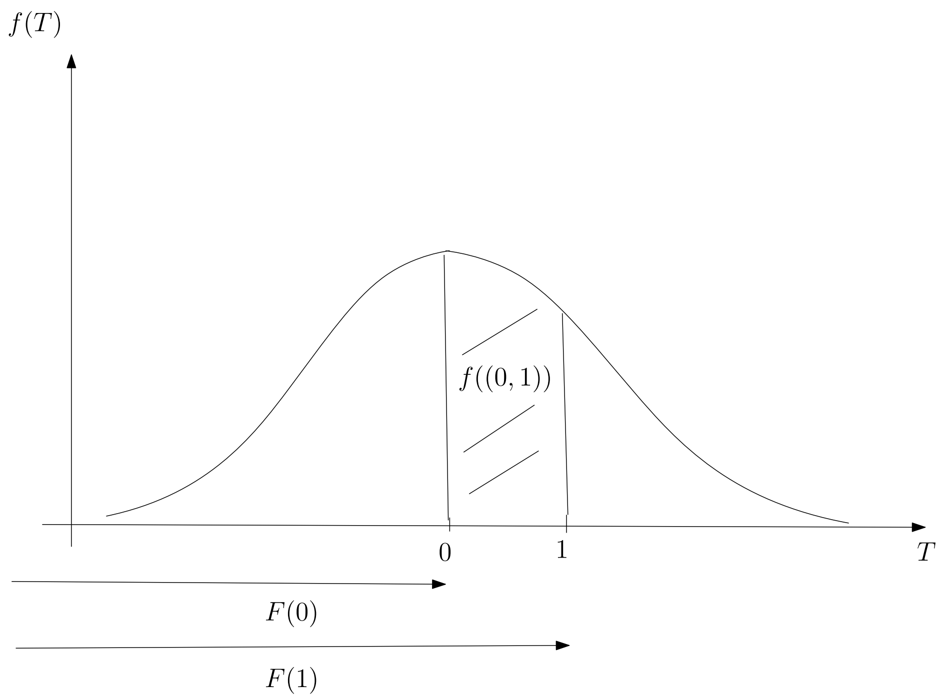
We know its suitability - it is unbiased and consistent and its efficiency is given by the length of the sample realization (we suppose that 50 is enough).

The point estimate is given by the value of the statistics.

The distribution of the statistics $T = \bar{X}$ *(here the statistics is viewed as random variable) will be normal as the population is (this result is not easy to derive, but it can be found on web, e.g. https://online.stat.psu.edu/stat414/node/172/) with the same expectation* $E[T] = \mu$ *as of the population and variance equal to* $D[T] = \frac{\sigma^2}{50}$ *(it has been derived in the previous Chapter).*

On the density function of the statistics, we can determine probabilities of point estimates of intervals. E.g. for $f(T)$ with standard normal distribution ($\mu = 0, \sigma^2 = 1$) the probability of $T \in (0,1)$ will be

$$P(T \in (0,1)) = F(1) - F(0) = 0.84 - 0.5 = 0.34$$



7.1 Confidence intervals

α-confidence interval CI is an interval, in which the true value of the estimated parameter lies with the probability $1 - \alpha$.

It can be **both sided** (as in the picture) where the borders are $1 - z_{\frac{\alpha}{2}}$ and $z_{\frac{\alpha}{2}}$ or

left-sided, with borders $1 - z_{\alpha}$ and ∞ or

right-sided, with borders $-\infty$ and z_α,

where z_α is the α-critical value of the statistics distribution.

Remark

We will compute confidence intervals in the software Statext. This concise explanation is just for us to know what we are doing.

7.2 Test of hypotheses

Another method of parameter analysis, based on the confidence interval, is testing of hypotheses.

Again we have a population X described by distribution $f(x, \theta)$ which is known up to the unknown parameter θ. We formulate two hypotheses

H0: the parameter is equal to θ_0; it is $\theta = \theta_0$.

HA: it is not equal to θ_0;it is $\theta < \theta_0$ or $\theta > \theta_0$.

Remark

H0 is called zero hypothesis, and it usually expresses the current state of affairs. HA is called alternative hypothesis and it denies H0. The denial can have three forms

H0: $\theta = \theta_0$; HA: (i) $\theta \neq \theta_0$, (ii) $\theta > \theta_0$, (iii) $\theta < \theta_0$. Accordingly, the test can be (i) both-sided, (ii) right-sided or (iii) left-sided.

The principle of testing hypotheses is as follows:

The hypothesis H0 is assumed to be valid. HA tries to deny it, based on the measured sample realization $x = [x_1, x_2, \cdots, x_N]$. Using this sample realization, we construct confidence interval for zero hypothesis H0. If the sample realization comes for distribution according to H0 then majority of its items (specifically $(1-\alpha) \cdot 100\%$ of them) should be within this confidence interval. If it is not true, the H0 is rejected.

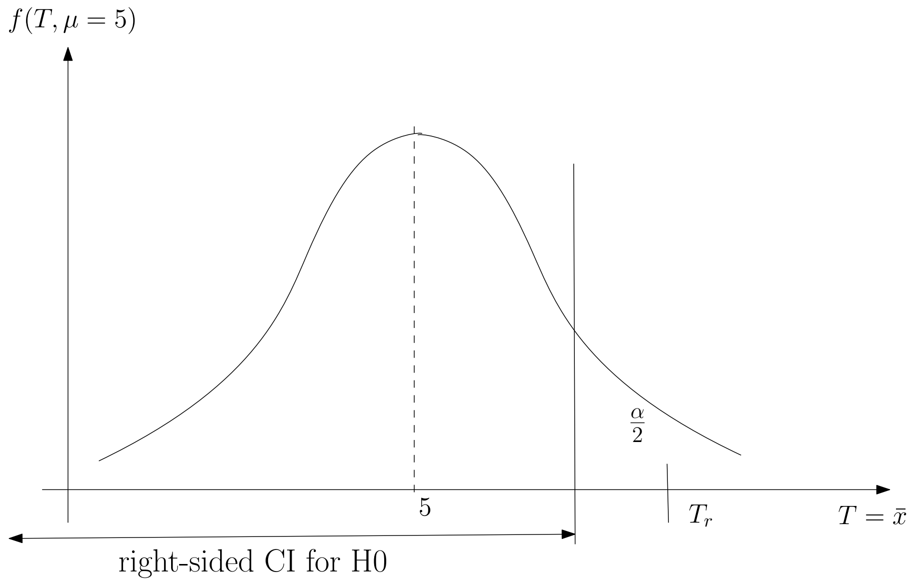
Example

Let the population is normal with the unknown expectation μ and known variance 3. Let we have

H0: $\mu = 5$; HA: $\mu > 5$.

Test H0 with a sample realization of the length 10;

The statistics for estimation μ is sample average $\bar{x}$. Its distribution will be again normal with expectation $\mu = 5$ and variance $\sigma^2 = \frac{3}{10} = 0.3$. As the alternative hypothesis HA says: $\mu > 5$, the confidence interval for H0 will be right-sided (so that HA is the opposite). The situation is sketched in the following picture



If the realized statistics T_r (which is computed from the sample realization) is as in the picture - outside the CI for H0, the hypothesis H0 is rejected. If it lies in the CI, we say that evidence is not sufficient for rejecting H0 (however, HA is not acknowledged).

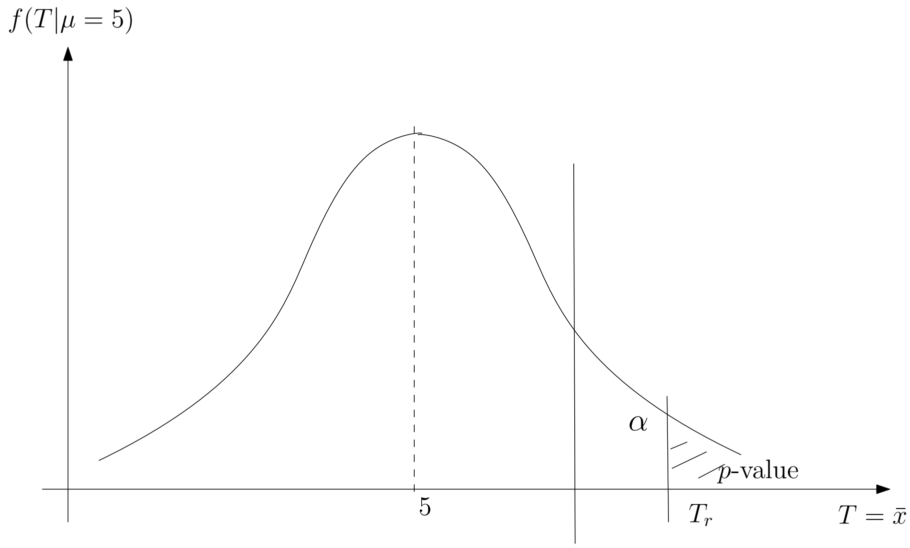
Here, the CI is called region of acceptance, its complement is denoted W and called critical region. So, if

$T_r \in W$ we reject H0, and if

$T_r \notin W$ we do not reject H0.

This principle rejects or does not reject. The power of rejection is not apparent. To show this, we introduce so called p-value.

p-**value** is the probability that in the future we will obtain realized statistics that are greater or equal then the one we have already obtained. It can be visible on the picture. It is the area under the probability function greater than T_r. For better orientation we will show it in the following picture



p-value is the most frequently used way of presentation of statistical testing. Its detailed exposition is given in Appendix **??**.

Remark

We will test hypotheses in the software Statext. This concise explanation is just for us to know what we are doing.

7.3 Important tests with one sample

7.3.1 Parametric tests (normality required)

- expectation (known $\times$ 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.*

- 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 are convinced that the ratio is higher.*

- 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.*

7.3.2 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.*

7.3.3 Tests of distribution type

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

- 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

- 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 weekdays and weekends. Test if they are unif*

7.4 Important tests with two samples

7.4.1 Parametric

- two expectations (independent × 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.

  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.*

- two proportions
  Ex: *Ratio of drivers violating rules in town is greater than outside. Test it.*

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

7.4.2 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.*

- 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.*

- 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.*

7.5 Important tests with more samples

7.5.1 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.*

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

Auxiliary tests to anova

- Bartlett - test of equality of more variances

- Scheffé - detects different samples

7.5.2 Nonparametric

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

- 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.*

7.6 Important 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?*

- 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.*

- Spearman test: nonparametric Pearson. Works with ranks.
  – H0: rvs are independent
  – test is both sided

- 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 their pay (low, normal, high). Test if the pay depends on the area.*

7.7 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 mus be any relation. This is expressed in **regression coefficient**

   $$\rho = \frac{C[X,Y]}{\sqrt{D[X]\,D[Y]}} \quad \longleftrightarrow \quad r = \frac{S_{xy}}{\sqrt{S_x S_y}}$$

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

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

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

   Pearson t-test has H0: $\rho = 0$, HA: $\rho \neq 0$; both sided test with Student distribution.

   H0: x and y are uncorrelated - regression does not have sense. To be able to use regression, H0 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

   $$y_i - \bar{y} = \underbrace{(\hat{y}_i - \bar{y})}_{\text{expl.}} + \underbrace{(y_i - \hat{y}_i)}_{\text{unexpl.}}$$

   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 H0: $F = 0$ is nothing explained and the regression does not have sense. The test is right-sided. Regression has sense, it H0 is rejected.

4. Test of independence of residuals

   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). H0: 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. Relative prediction error of residuals RPE

   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

$$RPE = \frac{var(e)}{var(y)}$$

   which is variance of prediction error e_i relative to variance of dependent variable y_i.

8 Work with Statext

8.1 Introduction to Statext

The statistical software Statext can be found on web

https://www.statext.com/

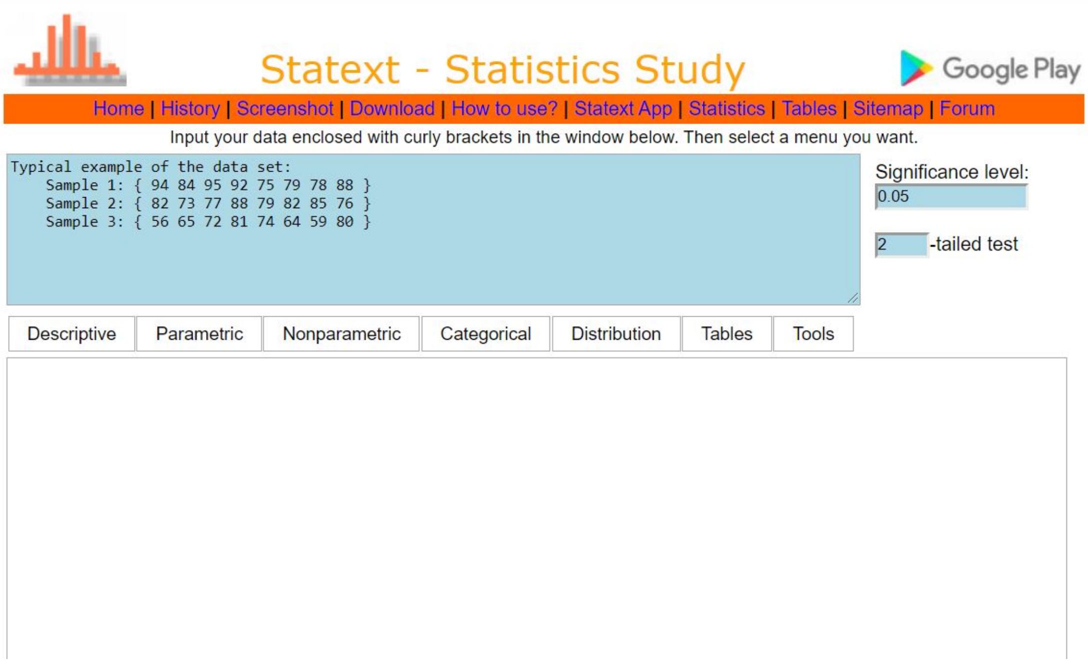
This is the main page of the Statext project. The download can be found on the folder Download.

However, there is much more here.

You can work on-line, here. But this on-line version is a bit limited in the menu possibilities. So, off-line work is preferred. No installation is necessary. Just download the file Statext.zip, unzip it and run Statext.exe.

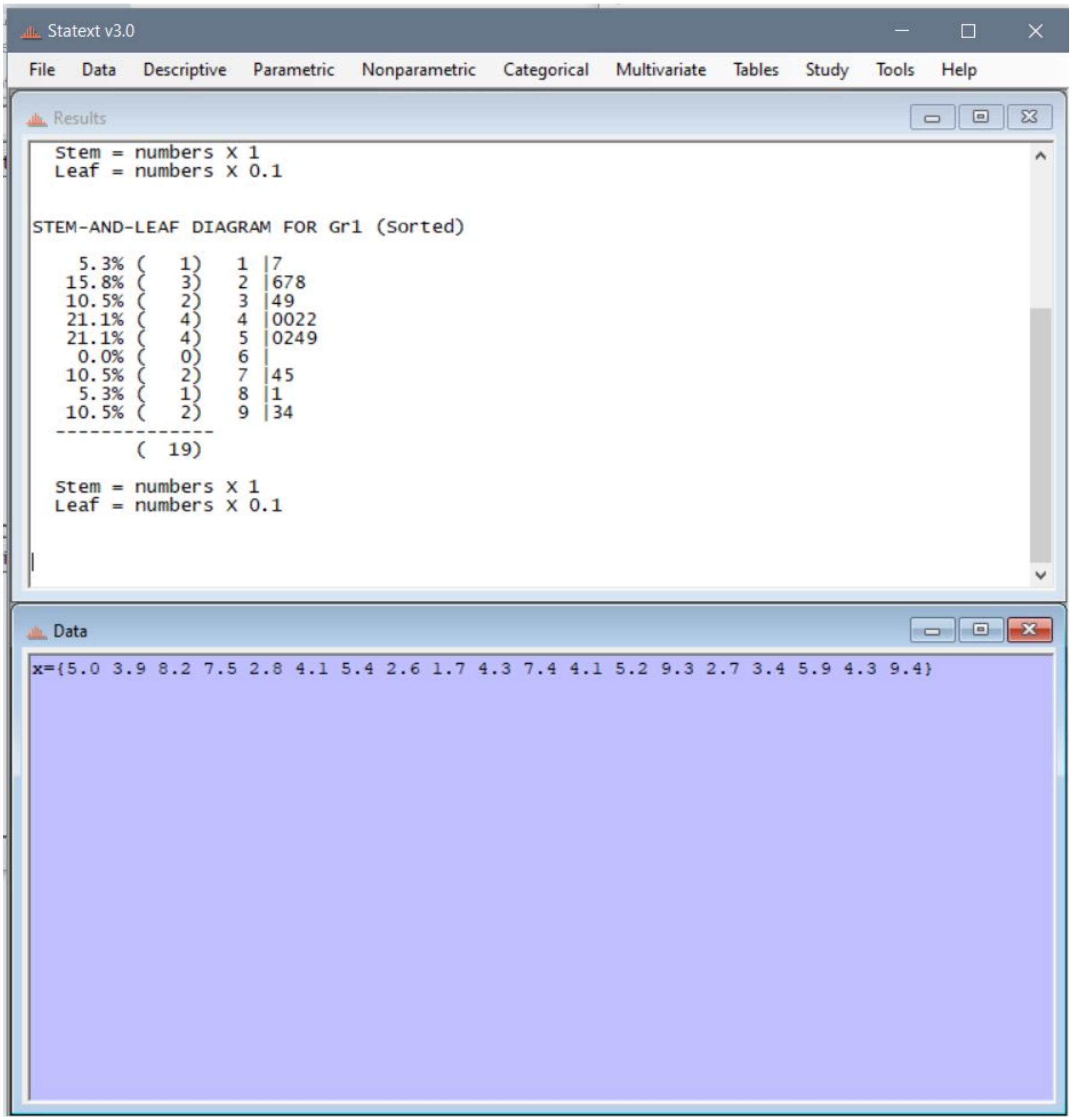
In the Statext homepage you can also find many useful things on the red bar. E.g. list of all sub-menu (in How to use?) or brief introduction to basic statistical notions (in Statistics).

The picture of Statext homepage is here



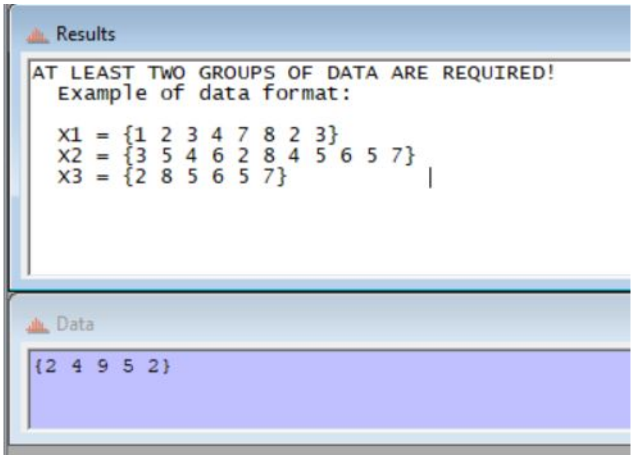
Here in the blue (upper) window, the data can be set (closed in curly brackets), in the menu (Descriptive, Parametric, etc.) the task is selected and the result (including the procedure of computation) appears in the white (lower) window. The result is mainly given by the p-value.

The picture of the off-line Statext version is similar

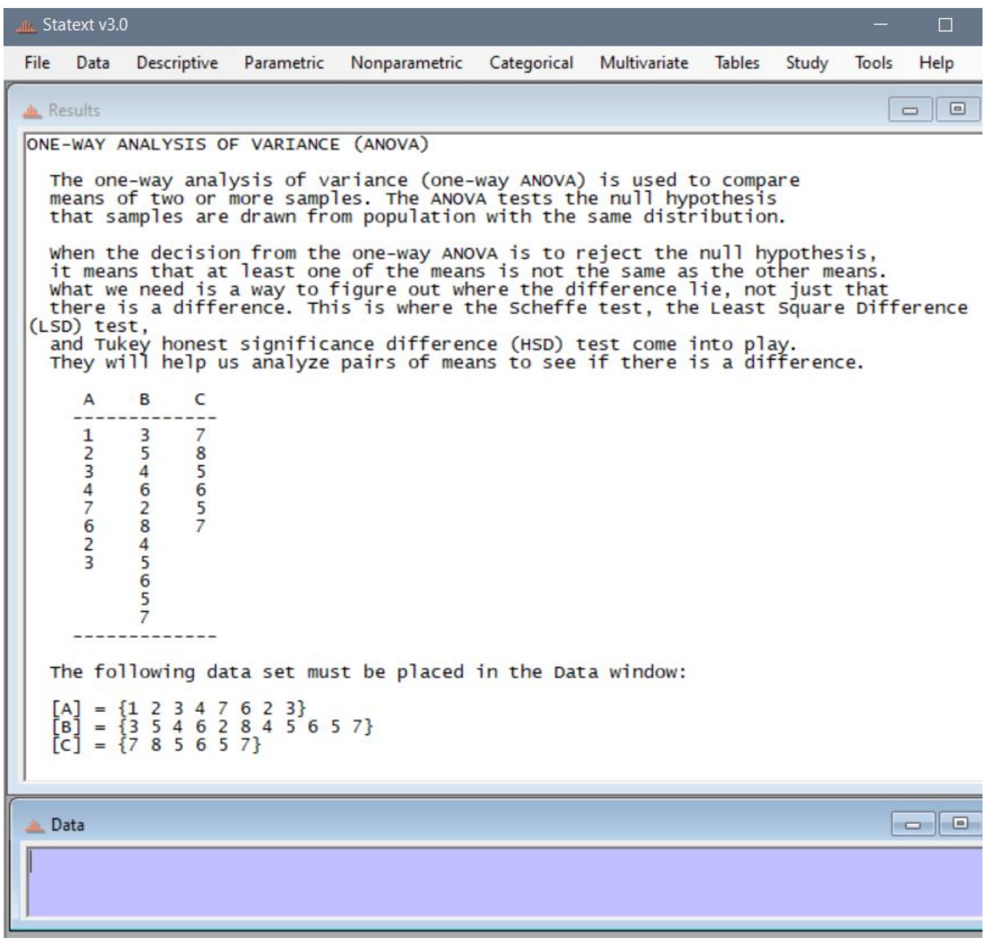


The blue (here lower) window obtains data, the white (here upper) window sets the results of the task selected in the menu (at the top).

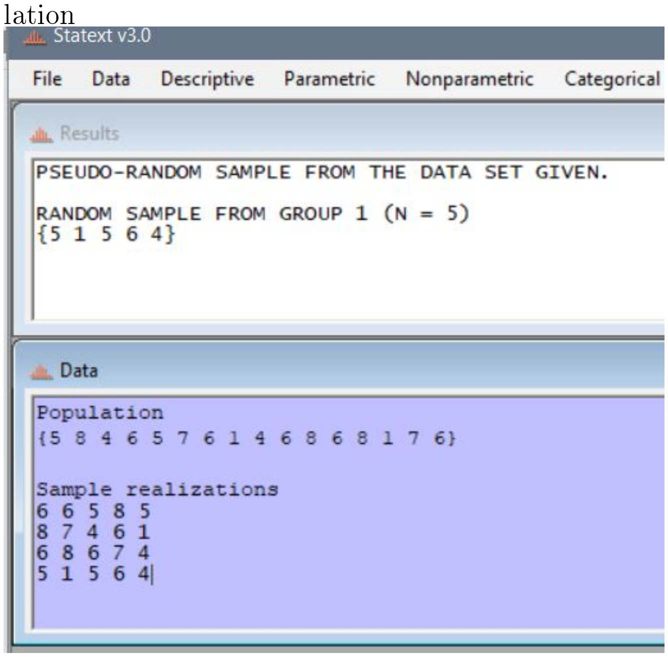
If you do not know the format of the data for a specific task, just set some, run the task and Statext will advice you. E.g. for ANOVA, we get



If you want help to some task, run it without any data



Another possibility that Statext allows is copying block of data from the result window into the data one. For example: We want to generate sample realizations form a population



Here we have set the population (in {}) and call Data|Random Sample.. We obtain result it the result window. We can take it and copy to the data window (without {} it will be just a text). If we want to use the sample realizations for further computation as data, we need to close them into {}.

Further possibilities can be got by experimenting. What we are mainly interested in, are the menus solving directly our statistical tasks. They are

Parametric, Nonparametric and Categorical

In the following text, we will look at them in more details.

Remark

Sometimes it happens that after running the task the results window stays rolled up and nothing can be seen. It must be rolled down manually.

9 Examples solved in Statext

Statext version 3.3 is used.

For us, compulsory are only those tasks that are supported by Statext. The others were listed only for your information.

The following examples should be the main practical outcome of the subject. The theory is important just to understand the examples.

9.1 Regression analysis

9.1.1 Example

In a factory, dependence of the overall costs y (in thousands of Kč) on the production x has been investigated. The following data have been measured

$x = \{532\ 297\ 378\ 121\ 519\ 613\ 592\ 497\}$

$y = \{\ 48\ 32\ 42\ 27\ 45\ 51\ 53\ 48\}$

a) Using linear regression estimate the costs for the production of 1000 products

b) For which production the costs would be equal to \$ 100 000.

Results

line: $y = 0.053x + 19.51$

cost_1000 = 73.02

prod_100 = 1504.09

Statext

The data can be directly copied from the pdf.

Call: Parametric|Simple regression|Linear...

In the result window you can find not only coefficients of regression line, but also the analysis of the regression performed.

The second task must be computed manually from the estimated regression line $Y = 19.510024 + 0.053514X$ and express X. The precise parameters are at the top of the Result window.

Notice !!!

After running the task from the menu, a window appears. Here the characteristics of the data set are computed and can be used. If you have your own characteristics, you can set them here replacing the computed ones.

9.1.2 Example

A harmful substance leaked into the container with water. Neutralizing agent has been applied and the concentration of the harmful substance has been measured at time instants x. The measured concentrations y are

$x = \{5\ 12\ 20\ 26\ 29\ 38\ 65\ 126\}$

$y = \{19\ 17\ 18\ 17\ 17\ 15\ 14\ 7\}$

Compute the correlation coefficient of linear regression and conclude about its suitability. If suitable, estimate when the concentration will be zero.

Results

Correlation coefficient r = -0.9832531

Line $y = -0.095x + 19.305$

Zero concentration will be at x = 203.58

Statext

The same as in Example 14.1.2.

Call: Parametric|Simple regression|Linear...

9.1.3 Example

At certain process we have measured the data

$x = \{5\ 12\ 20\ 26\ 29\ 38\ 40\ 45\}$

$y = \{9\ 7\ 12\ 12\ 27\ 35\ 44\ 76\}$

Perform the polynomial regression of the order 3 and the exponential regression. Using p-value of the regression decide which type of regression is better.

Results

$pv_{lin} = 0.0057$

$pv_{exp} = 0.0025$

Polynomial is better.

Statext

Call: Parametric|Simple regression|Linear...

Parametric|Simple regression|Exponential...

Better is that one which has smaller p-value.

9.1.4 Example

At certain process we have measured the data

$x1 = \{15\ 12\ 11\ 9\ 9\ 8\ 5\ 3\}$

$x2 = \{3\ 9\ 5\ 11\ 28\ 14\ 32\ 58\}$

$y = \{9\ 7\ 22\ 12\ 27\ 31\ 44\ 36\}$

Perform multivariate linear regression and test its suitability using p-value from the ANOVA table.

Results

pv = 0.0415

On the level 0.05 is the regression OK.

Statext

Call: Parametric|Multiple regression|Linear...

9.2 Confidence intervals

9.2.1 Example

Assume, that the height of children in the age 10 has normal distribution with the variance 38. Determine the interval α-I, in which the true height will be if we have measured the data sample of the length 12 with the average 127.3. Compute on the level $\alpha = 0.01$.

Results

CI = (122.72, 131.88)

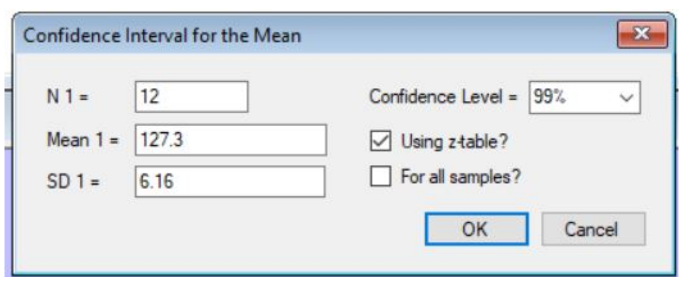
Statext

Here, instead of data, we have computed characteristics of average 127.3. The variance is given as known (z-values will be used).

Call: Parametric|Confidence interval for the mean...

Remark: some data, no matter which, must be set in the data window.

The window that appears must be set like this:



The value 6.16 is square root of 127.3. Confidence level is set as 0.99 (it is 1-0.01). Using z-table must be checked, as the value of variance (here standard deviation) is known, not computed from data.

9.2.2 Example

Assume, that the height of children in the age 10 has normal distribution. Determine the interval α-I, in which the true height will be if we have measured the data sample of the length 12 with the average 127.3 and variance 38. Test on the level $\alpha = 0.01$.

Compare with the previous result and justify.

Results

CI = (121.77, 132.83)

Statext

The same as the previous Example 14.2.1

Call: Parametric|Confidence interval for the mean...

The difference is that the variance now is computed from data, not known exactly. So, Using the z-table must stay unchecked.

9.2.3 Example

To learn the accuracy of a method for measuring the volume of manganese in the steel, we performed independent measurements of several variances. We would like to know the border for which it holds that only 5% of possibly measured variances will be greater than this border. The measured values are

$x = \{4.3\ 2.9\ 5.1\ 3.3\ 2.7\ 4.8\ 3.6\}$

Results

The border of variance is 3.2

Statext

Call: Parametric|Confidence interval for standard deviation...

The border will be given by the upper border of right-sided confidence interval. As the intervals offered are only both-sided, we have to set 10% (i.e. 0.1) instead of 5%.

In the results window we must find interval for variance.

9.3 Parametric tests

9.3.1 Example

At the motorway with recommended speed 80 km/h we monitored the speeds of passing cars and obtained data presented in the following table of values and frequencies

x	0-78	79	80	81	82	83-∞
n	543	32	45	15	8	2

Test the hypothesis H0 that only 3% of drivers exceed the speed 80 km/h. Test on the level 0.05.

Results

$pv = 0.096$

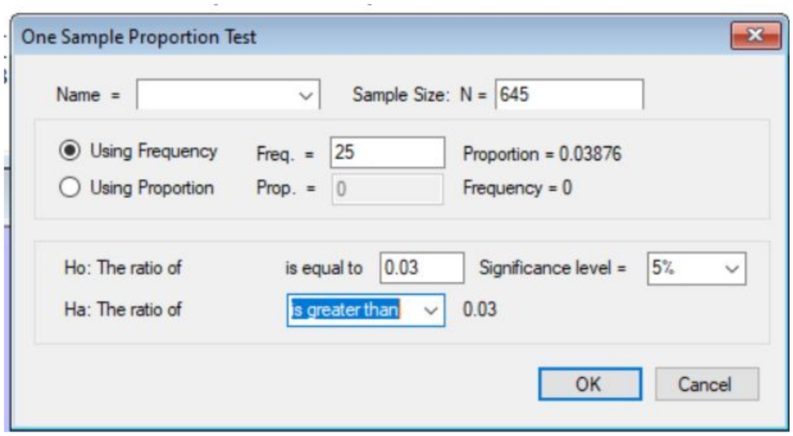
We do not reject.

Statext

Exceed: 25, All: 645

Call: Categorical|One sample proportion test...

The window to be filled in is here



9.3.2 Example

From a set of steel rods with equal nominal length 6.2 cm, we have selected random choice with the lengths x. The producer guarantees that the standard deviation of the lengths is 0.8 cm. At the significance level 0.05 test the assertion of the producer that the produced rods have the nominal length if the measured lengths are

$x = \{6.2\ 7.5\ 6.9\ 8.9\ 6.4\ 7.1\}$

Compare the result with the situation when the variance is not known.

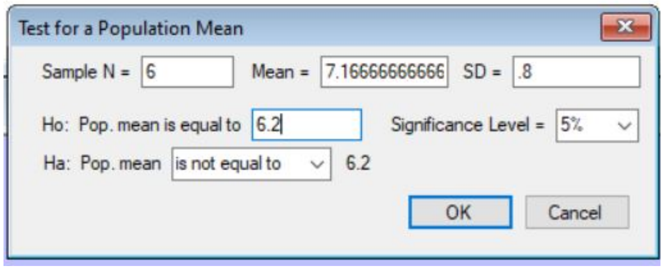
Results

$pv_z = 0.008, \quad pv_t = 0.059$

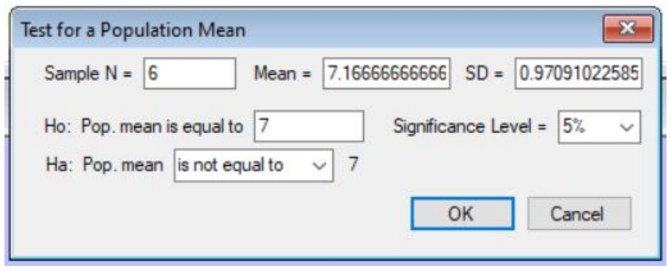
Statext

Call: Parametric|Test for a population mean...

– for known variance



– for unknown variance



Here in the item SD= we leave the computed value.

9.3.3 Example

The accuracy of setting of certain machine can be verified according to the variance of its products. If the variance is greater then the level 28, it is necessary to perform new setting. A data sample has been measured

$x = \{102\ 113\ 108\ 119\ 114\ 102\ 115\ 119\ 99\ 117\ 108\ 101\}$

On the level 0.05 test if it is necessary to set the machine.

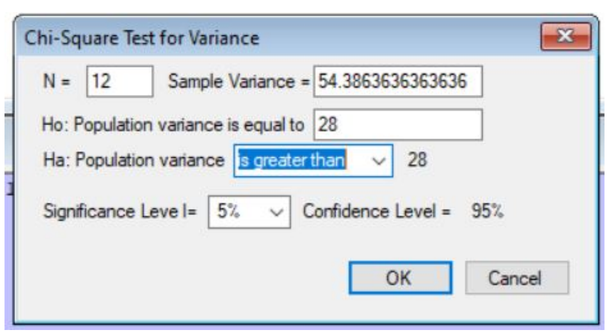
Results

pv = 0.03

On the level 0.05 reject H0: the machine is OK

Statext

Call: Parametric|Chi-square test of variance...



9.3.4 Example

Solidity of materials is verified by two methods A and B. The same material has been subdued testing by both methods. The results are

$A = \{20.1\ 19.6\ 20.0\ 19.9\ 20.1\}$

$B = \{20.9\ 20.1\ 20.6\ 20.5\ 20.7\ 20.5\}$

On the level 0.05 test equality of both methods if the variability of methods is assumed to be equal.

Results

pv = 0.0027

Equality is rejected.

Statext

Call: Parametric|Test for two population means|t-test (indep.samp.)

9.3.5 Example

We are going to test if the tire removal on left and right sides of the front wheels of cars is equal. The measured values are

$xL = \{1.8\ 1.0\ 2.2\ 0.9\ 1.5\}$

$xP = \{1.5\ 1.1\ 2.0\ 1.1\ 1.4\}$

Test at the level 0.05.

Results

pv= 0.55, are the same.

Statext

Call: Parametric|Test for two population means|Paired t-test

9.3.6 Example

At the motorway with recommended speed 80 km/h speeds of passing cars have been monitored the in the direction to the town (xT) and from the town (xF). The data measured are

$xT = \{95\ 88\ 71\ 82\ 69\ 75\ 78\ 67\ 77\ 82\ 79\}$

$xF = \{81\ 69\ 75\ 91\ 77\ 76\ 88\ 68\ 91\ 74\ 92\}$

At the level 0.05 test the hypothesis H0: To the town the cars go faster.

Results

1. *test of variances: pv = 0.3 - variances are equal.*

2. *test of speeds: pv = 0.322 - to town are higher is not rejected*

Statext

Call: Parametric|Test for two population means|t-test (indep.samp.)

To is the first sample, From is the second sample.

H0 says: To > From. HA: To<From → HA ... is less than

Now, in the results table we must find result for "equal variances".

The variance can be tested in: Parametric|Barlett's test for variance...

9.3.7 Example

During a check of the front lights of cars we have measured the data xL (left light) and xR (right light).

xR= {-3 5 16 9 -8 -2 23 5 -6 -3}

xL= {-5 -12 22 -3 -9 1 -1 2 -13 -5}

The values are distances (in cm) above (positive) and below (negative) of the real level with respect to the optimal level. At the level 0.1 test if

a) the light levels at each car are the same,

b) the left lights are higher then right.

Results

a) pv = 0.075; On the level 0.1 we reject equality.

b) pv = 0.03; Left are higher is rejected.

Statext

Call: Parametric|Test for two population means|Paired t-test

a) H0 ... is not equal

b) H0 ... is greater than

because: R is first, L is second. H0: R<L, HA: R>L

so HA says: R (first) is greater than L (second).

9.3.8 Example

At a crossroads, we have written down numbers of cars going straight (S) turning to left (L) and right (R). The measured data are $xS = 62$, xL =39 and $xR = 46$. On the level 0.1 test assertion that the ratio of cars

a) going straight is equal to those that are turning,

b) going straight exceed those turning.

Results

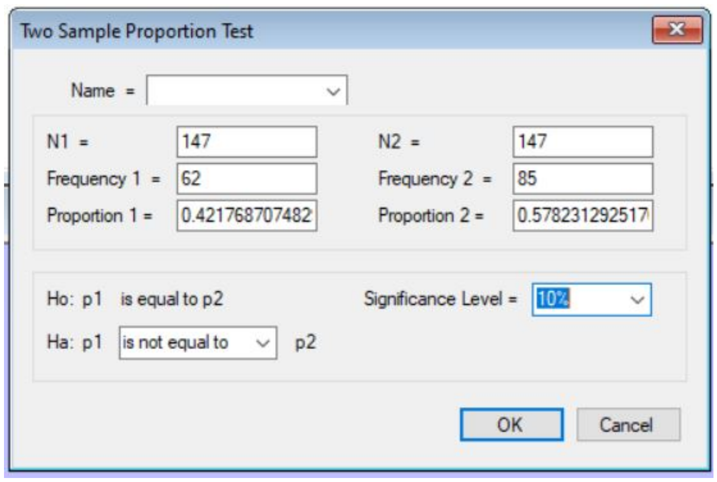
a) pv = 0.0073 - equality is rejected

b) pv = 0.0036 - "straight is more" is rejected

Statext

Call: Categorical|Two sample proportion test...

a)



b) The same, but HA: $\cdots$ is less than

9.4 ANOVA

9.4.1 Example

We monitor three machines. Randomly, we measure their productions per hour

$x1=$ {53 55 49 58 52 61 56 55}

$x2=$ {49 56 52 45 51 56 44 51}

$x3=$ {52 53 52 54 55 53 53 52}

At the level 0.05 test the equality of their production.

Results

pv= 0.054

Statext

Call: Parametric|Analysis of variance|One-way anova...

9.4.2 Example

For one month in the years 1999 2000 2001 2002 2003, we monitored number of accidents at five crossroads. The results are as follows

1999: {3 5 2 1 3}

2000: {6 2 5 3 4}

2001: {3 2 1 1 2}

2002: {4 1 1 2 2}

2003: {4 2 5 5 6}

At the level 0.01 test hypothesis H0: The average number of accidents is equal at all monitored crossroads.

Results

One-way anova pv = 0.0207

Two-ways anova Equality in crossroads pv = 0.0195

Equality in time pv = 0.275

Statext

Call: Parametric|Analysis of variance|One-way anova...

Call: Parametric|Analysis of variance|Two-way anova (without...)...

9.4.3 Example

A factory produces some products whose weight must be constant. For the production it uses four machines. A sample of products has been taken from all machines to test equality of the product weights. The measured values are

x1={39.4 34.8 35.6 35.1 35.8}

x2={34.4 34.2 35.1 31.1 32.5 33.8}

x3={30.2 35.1 34.2 36.3 30.8 35.6 35.2}

x4={39.1 34.3 38.6 34.5 36.4 36.1}

Test the equality of the product weights on all four machines.

Results

pv = 0.036

Statext

Call: Parametric|Analysis of variance|One-way anova...

9.5 Nonparametric tests

9.5.1 Example

At a crossroads we have written down numbers of passing cars. The lengths of monitoring were $d = \{15\ 10\ 20\ 35\ 10\ 50\}$ and the measured numbers $x = \{71\ 56\ 98\ 121\ 44\ 271\}$. At the level 0.05 test the hypothesis that the cars go uniformly (the same number per time unit).

Results

pv = 0.002

Remark: E=70.82 47.21 94.43 165.26 47.21 236.07

Statext

First, you must compute the expected frequencies $E = d\,(\sum x_i / \sum d_i)$. Then copy $x = \{\cdots\}$ and $E = \{\cdots\}$ into the data window and then

Call: Categorical|Chi-square Goodness-of-Fit test...

Remark: Homogeneity test compares if the two samples come from the same distribution. It is not the same as Goodness-of-Fit test.

9.5.2 Example

The following data are frequencies (f) of incidents at certain big factory at time intervals (i)

i: 8-10h. 10-12h. 12-13h. 13-17h.

f: 2 7 1 16

At the level 0.05 test the hypothesis that the accidents occur uniformly.

Results

pv= 0.13

Remark: E=5.78 5.78 2.88 11.56

Statext

The same as previous. Only d must be constructed as lengths of intervals in i. I.e. d ={2 2 1 4}. Then $E = d\,(\sum f_i / \sum d_i)$. Copy f and E to data window and

Call: Categorical|Chi-square Goodness-of-Fit test...

9.5.3 Example

A connection between color of eyes and hair has been investigated. In a collected data sample we obtained the following frequencies

eyes\hair	light	brown	dark
blue	90	75	55
gray	96	136	88
brown	108	135	119

At the level 0.05 test the hypothesis that the color of eyes and hair are independent.

Results

pv= 0.017

Statext

Copy the data from the table into the data window and close the rows into the curly brackets, like this:

blue {90 75 55}
gray {96 136 88}
brown {108 135 119}

Then

Call: Categorical|Chi-square independence test...

9.5.4 Example

Two operators O1 and O2 alternate regularly in production of certain articles. The produced articles are checked for quality (either 1 or 2). The following data have been measured

O1: 1 2 1 1 2 2 2 1 2 1 1 1 2 1 2 2 2 1 2 1 2

O2: 2 2 1 2 1 1 2 2 2 1 2 2 1 2 1 2 2 2 1 1 2

At the level 0.05 test the assertion that both the operators are with respect to the production quality independent.

Results

pv= 0.78

(the result listed is p-value, not G test)

Statext

First the frequency table T must be constructed (frequencies of individual configurations of [O1,O2]. It is T = {{3 7}{5 6}} (the matrix is set in this form). Then

Call: Categorical|Chi-square independence test...

9.5.5 Example

Two doctors recommend curing a cold with two different methods. The results (number of days of the treatment) are $x1$ and $x2$. Test equality of the methods.

$x1$={5 8 7 8 4 5 5 6 9 3 5 8 6 8 7 5 8 5 7 5 6 8 4 7 7 5 6}

$x2$={3 4 9 5 4 9 9 8 3 3 5 3 6 4 5 6 2 2 3 4 2 3}

Results

pv=0.005

Statext

Call: Nonparametric|Two samples (unpaired)|Mann-Whitney U test...

9.5.6 Example

Eight sportsmen in a certain sports club were tested with respect to their performance. All of them threw a javelin once and then they were subdued to intensive training. Then they threw once more. The measured lengths were $x1$ and $x2$. The hypothesis is that one day of training is not enough to improve their performance. Test on the level 0.05.

x1={68 81 69 72 66 91 98 89 75 68 69 75 72 83 88 79 88 76 81 85}

x2={79 62 70 75 68 81 85 94 71 62 81 70 74 85 82 91 85 82 83 73}

Results

pv = 0.68

Statext

Call: Nonparametric|Two samples (paired)|Wilcoxon signed-rank test...

9.5.7 Example

Test if mice and stags have equally long front legs. The measured values are

x1={135 123 3.1 2.5 98 124 131 3.4 2.8 128 154 135 2.9 137 2.7 3.0 3.2 131 2.8 148}

x2={136 121 2.9 2.6 101 121 130 3.5 2.9 126 162 141 2.8 142 2.9 2.8 3.0 132 3.1 151}

Results

pv=0.31

Statext

Call: Nonparametric|Two samples (paired)|Wilcoxon signed-rank test...

9.5.8 Example

Tree inspectors are to evaluate functionality of five fast food stands. Each inspector evaluates each stand. The result is at the table (rows correspond to inspectors, columns columns to stands). Evaluation marks are 1,2,..,10. The mark 10 is the best one. Test if the quality of the stands is equal.

Table

{{10 8 3 9 7}

{8 7 5 9 10}

{8 9 5 7 6}}

Results

pv= 0.155

Statext

Call: Nonparametric|Block design|Friedman test...

(Data sets are for subjects)

9.5.9 Example

Let X denote the length (in centimeters), of a certain fish species. We obtained the data set

X={5.0 3.9 8.2 7.5 2.8 4.1 5.4 2.6 1.7 4.3 7.4 4.1 5.2 9.3 2.7 3.4 5.9 4.3 9.4 8.2 4.8 3.3 4.7 5.3 4.2 4.0}.

Can we conclude that the median length of the fish species differs significantly from 4.1 centimeters?

Results

pv = 0.052

Statext

Call: Nonparametric|One sample|Wilcoxon signed-rank test...

Remark: Do not forget to insert 4.1 into the H0: population window

10 Examples to the textbook on Statistics

10.1 Examples to probability

10.1.1 Example

We throw a dice. What is the probability that we obtain an odd number if somebody watched the result and reveals us that the result is a) greater than 3; b) greater than 4.

Solution

The example can be demonstrated in a picture. The basic situation with a dice has 6 equally possible outcomes

1	2	3	4	5	6

The condition a) excludes 1, 2, 3. In the next picture they are denoted by X

X	X	X			
1	2	3	4	5	6

Now, all possible are 4, 5, 6 from which only 5 is odd. So, we have

$$P1 = \frac{1}{3}$$

For the condition b), the excluded are 1, 2, 3, 4 and positive (odd) is 5

X	X	X	X		
1	2	3	4	5	6

Then, we have

$$P2 = \frac{1}{2}$$

Notice

The original probability (of odd number) is $P = \frac{1}{2}$. For the first case a) it holds $P1 \neq P$ which means that the events are dependent. In the case b) it is $P2 = P$, so the events are independent. From the example we can see that independence means that in the new sample space (created by the condition) the ratio of positive and negative results is the same as in the original sample space - odd/even are 1/1. In the case of dependent events, the ratio is unbalanced - here odd/even are 1/2. The result of odd number is less probable so the condition brings an information for us which can be used in guessing the result.

10.1.2 Example

Consider an example of throwing two dices. What is the probability that: a) The sum on both dices will be greater than ten? b) We get an even sum if we know that there was not 6 at the first dice? c) We get an even sum if we know that there was not 6 neither on the first nor on the second dice?

Solution

a) We would like to use the 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 are couples of numbers that fell at first and second dice. The situation can be depicted as follows (vertical axis represents toss on the first dice and horizontal one the second - inside the table are sums of tosses on both dices)

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	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

As all the results denoted inside the table are equally possible, the classical definition of probability can now be used. 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

$$P\left(x > 10\right) = \frac{3}{36} = \frac{1}{18}.$$

b) The above table for the condition no 6 on the first dice looks looks like this

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11

From it the number of positive results is 15 (all entries with even sums) and possible results are 30 (all entries). Thus the probability is

$$P\left(no\,6\,on\,first\,dice\right) = \frac{15}{30} = \frac{1}{2}$$

which is the same result as without the condition.

c) With no 6 on both dices the table is

	1	2	3	4	5
1	2	3	4	5	6
2	3	4	5	6	7
3	4	5	6	7	8
4	5	6	7	8	9
5	6	7	8	9	10

Here, positive results are 13 and possible ones ore 25. Then the probability is

$$P\,(no\,6\,on\,both) = \frac{13}{25}$$

Here the condition influences the result. Even sum is more probable.

Remark

The reason why at b) the probability is the same as without the condition can be seen. The condition b) leads to omitting the results 7 8 9 10 11 12. Here, the number of even and odd is the same. In difference, in the case c) we omit 7 8 9 10 11 12 11 10 9 8 7 where even results are 5 and odd ones are 6. It means that more even results stays in the game and thus even is more probable.

10.1.3 Example

We have a box with five white (w) two red (r) and three black (b) balls. Randomly we draw one ball and without returning it we draw the second one.

1. What is the probability, that the first ball will be white?
   Solution: According to the classical definition of probability we look for positive and all possible results[3]. There are five positive results (five white balls) and ten possible ones (all balls in the box). So the probability is

$$P = \frac{5}{10} = 0.5$$

2. What is the probability that the second drawn ball will be white if the first was a) white; b) black?
   Solution a) white: If the first ball was white then before the second draw we have four white and altogether nine balls. So, denoting $1w$ as the first draw white and $2w$ the second one white, we have

$$P\,(2w|1w) = \frac{4}{9}$$

   Solution b) black: Here before the second draw we have five white and nine balls altogether. Then

$$P\,(2w|1b) = \frac{5}{9}$$

[3] The conditions for classical definition are met: there are a finite number of possible results and each result is equally probable.

3. What is the probability that the second drawn ball will be white?
Solution: Here, we do not know which ball has been drawn as the first. So, we must take into account all possibilities. What we know in analogy to the previous examples are marginal and conditional probabilities (whit the analogous denotation) $P(1w) = \frac{1}{2}$; $P(1r) = \frac{1}{5}$; $P(1b) = \frac{3}{10}$; $P(2w|1w) = \frac{4}{9}$; $P(2w|1r) = \frac{5}{9}$ and $P(2w|1b) = \frac{5}{9}$. Then the probability $P(2w)$ is given as a union of all possible first draws

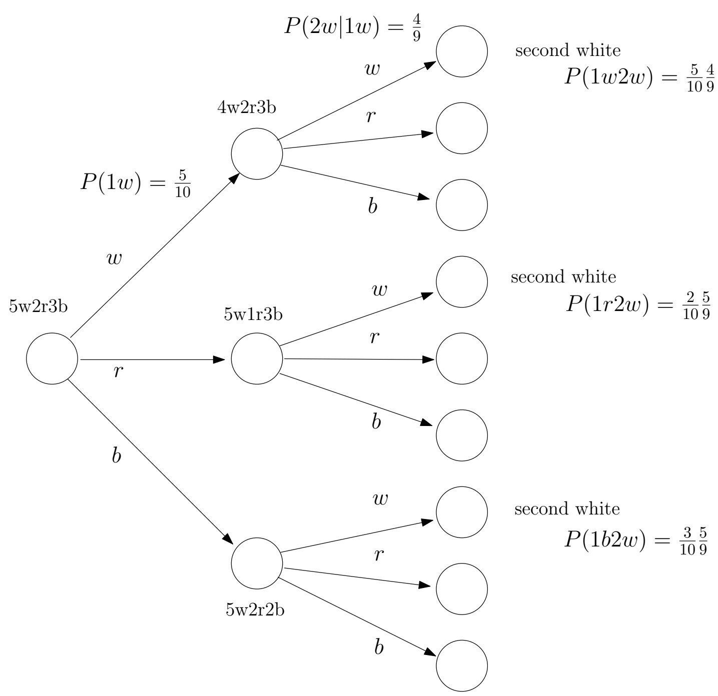
$$P(2w) = P(2w, 1w) + P(2w, 1r) + P(2w, 1b) =$$

$$= P(2w|1w) P(1w) + P(2w|1r) P(1r) + P(2w|1b) P(1b)$$

where $P(\cdot, \cdot)$ denotes joint probability on which we applied the chain rule $P(A, B) = P(A|B) P(b)$. Now, all these probabilities are known and can be substituted

$$P(2w) = \frac{4}{9}\frac{1}{2} + \frac{5}{9}\frac{1}{5} + \frac{5}{9}\frac{3}{10} = \frac{1}{2}$$

The situation can be easily represented graphically



The nodes represent states (numbers if individual balls in the box). The leftmost one denotes the initial state (5white,2red,3black). The arrows are actions (colors of the drawn balls). Up white, right red and down black. The second column of nodes are states after the first draw (always one corresponding color is missing). The probabilities are computed on the base if the input node and are given by number of positive colors divided by the number of all colors. The rightmost column are the final states. Their probabilities are given by the product of probabilities on th path from the beginning to this node.

To solve the problem - probability of second white - we must select the nodes where the last drawn ball was white. They are the first, fourth and seventh one. The final probability is sum of probabilities of these selected nodes.

10.1.4 Example

Let the sample space be given by sex of residents of certain village x and their age y. By $x = 1$ we denote female (girl) and by $x = 2$ mail (boy). The age has values $y = 1$ for age less then 18 and $y = 2$ otherwise. We obtained the following population

sex	1	1	1	2	2	1	2	1	1	1	1	2	2	1	2	2	1	1	1	2	2
age	2	2	1	1	1	2	2	2	1	2	2	2	1	1	2	2	2	2	2	1	2

1. What is the probability that a randomly chosen person will be male?
   Solution: The question concerns only x - the first row of the data-table. Mail is 2 and there are nine entries equal to 2. The total number of persons (entries of row in the table) is 21. So the probability of male is

   $$P = \frac{9}{21}$$

2. What is the probability that a randomly chosen person will be a male if we know that the age is less than 18?
   Solution: The condition is age < 18 which is what we know. So, from the above table we select only columns which have 1 in their second entry. We obtain

sex	1	2	2	1	2	1	2
age	1	1	1	1	1	1	1

   and what remains in the first row is: three times 1 and four times 2; altogether seven columns. That is the probability of a boy is

   $$P = \frac{4}{7}$$

3. What is the probability that a randomly chosen person will be a boy (i.e. a male with the age < 18).
   Solution: Here, we are guessing both the sex as well as the age. So, from the table, we have to select all positive columns $\begin{matrix} 2 \\ 1 \end{matrix}$. There are four of them. The total number is 21, so the probability is

   $$P = \frac{4}{21}$$

10.1.5 Example

Let us have a buss station with a bus coming exactly in five minutes interval. We come randomly at the bust station and want to go by the bus. The result of the experiment is defined as a time we must wait. What is the probability of waiting a) one minute; b) less than 1 minute; c) more than one minute.

Solution

a) Evidently, the probability of waiting exactly a given time is zero. In the interval $\langle 0,\, 5\rangle$ is uncountably infinitely many time instants which all are candidates for a possible waiting time. The positive result is only one. So, one divided by infinity gives zero.

b) The probability of maximum time of waiting is proportional to the length of this interval. No waiting has probability 0 (see the case a)) and maximum waiting 5 min minutes has probability 1. If we denote X as maximum time of waiting and x some real number, then we have

$$P\left(X \leq x\right) = \frac{x}{5},\ \text{ for } x \in \langle 0,\, 5\rangle$$

and zero otherwise.

So, waiting less than 1 minute has the probability $\frac{1}{5}$.

Remark

If we take two time instants $x_1 \leq x_2$,both within the interval $\langle 0,\, 5\rangle$ then

$$P\left(X \leq x_1\right) = \frac{x_1}{5},\ P\left(X \leq x_2\right) = \frac{x_2}{5}$$

and the probability of waiting more than x_1 and less than x_2 is

$$P\left(X \in \langle x_1,\, x_2\rangle\right) = \frac{x_2 - x_1}{5}$$

From it we can also see that if $x_1 = x_2$ then the probability is zero.

c) Probability of waiting more than 1 minute can be determined on the basis of the last formula. It is

$$P\left(X > 1\right) = \frac{5-1}{5} = \frac{4}{5}.$$

10.2 Examples to random variable

10.2.1 Example

Let us consider an experiment with flipping two coins. The results are "both heads" (H), "both tails" (T) and "different" (D). Introduce random variable X, construct its probability function $f\left(x\right)$ and compute the expected flip $E\left[X\right]$.

Solution

Let us assign $H \to 1$, $T \to 2$ and $D = 3$. (The assignment is arbitrary, but in the following computations it must be preserved)

Then

$$f(1) = P(H) = \frac{1}{4},\ f(2) = P(T) = \frac{1}{4},\ f(3) = P(D) = \frac{1}{2}$$

So, the probability function is

x	1	2	3
$f(x)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$

The expectation is

$$E[X] = 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{2} = 2.25$$

10.2.2 Example

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

x	1	2	3	4	5
$f(x)$	p	$2p$	$0.3+p$	0.2	p

Solution

First we determine $p : p + 2p + (0.3 + p) + 0.2 + p = 5p + 0.5 = 1$

$\to p = 0.1$. So the table will be

x	1	2	3	4	5
$f(x)$	0.1	0.2	0.4	0.2	0.1

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

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

10.2.3 Example

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

Solution

$$P(X > 3) = 1 - P(X \le 3) =$$

$$= 1 - \exp\{-4\} \frac{4^0}{0!} - \exp\{-4\} \frac{4^1}{1!} - \exp\{-4\} \frac{4^2}{2!} - \exp\{-4\} \frac{4^3}{3!} = 0.566$$

10.2.4 Example

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

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

Solution

– expectation

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

where $y = ax; \quad dy = a \cdot dx$.

– 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 $y = ax; \quad dy = a \cdot dx$ and $x = \frac{y}{a}$.

– distribution function

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

$$= 1 - \exp\{-ax\}, \ x \geq 0 \ \text{othewise} \ 0.$$

– 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

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

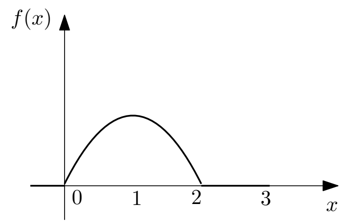
10.2.5 Example

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

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

Solution

The density function is in the following picture



Expectation

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

$$= \frac{3}{4}\left[\frac{2}{3}x^3 - \frac{1}{4}x^4\right]_0^2 = \frac{3}{4}\left(\frac{2}{3}8 - \frac{1}{4}16\right) = 4-3=1$$

which also follows directly from the graph.

Variance

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

$$= \frac{3}{4}\left[-\frac{1}{5}x^5 + x^4 - \frac{5}{3}x^3 + x^2\right]_0^2 = \frac{1}{5}$$

Remark

Computation of 0.05*-quantil:*

$$\int_0^{\zeta_{0.05}} \frac{3}{4}\left(2x-x^2\right)dx = 0.05$$

$$\frac{3}{4}\left[x^2-\frac{1}{3}x^3\right]_0^{\zeta_{0.05}} = 0.05 \quad \to \quad \zeta_{0.05}^2 - \frac{1}{3}\zeta_{0.05}^3 = \frac{4}{3}0.05$$

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

It cannot be solved analytically.

10.2.6 Example

Random variable X is defined by the formula

$$f(x) = p(1-p)^{x-1}, \; x = 1, 2, \cdots$$

Find α-quantile ζ_α and α-critical value z_α of this distribution with $\alpha = 0.05$.

Solution

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

$$\ln(1-\alpha) = n \ln(1-p)$$

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

The result is not an integer. Its roundup is an approximation of the quantile.

The critical value can be obtained by setting $\alpha = 0.95$.

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

10.2.7 Examples

Construct distribution function of random variable X with the density function

$$f(x) = \frac{3 - |x^2 - 4x + 3|}{8}, \qquad \text{pro } x \in (0,\, 4).$$

Solution

The density function is defined piece-wise on several intervals. First we need to solve quadratic equation

$$x^2 - 4x + 3 = 0$$

and get rid of the absolute value. The quadratic function can be written as $(x-1)(x-3)$ and the solution is $x_1 = 1$ and $x_2 = 3$. From it we have the table of the signs of the quadratic function inside the absolute value

x	$(0,1)$	$(1,3)$	$(3,4)$
sign	$+$	$-$	$+$

Now, we can write the density function for individual intervals:

$x \in (0,1) \qquad f(x) = \frac{-x^2+4x}{8},$

$x \in (1,3) \qquad f(x) = \frac{x^2-4x+6}{8},$

$x \in (3,4) \qquad f(x) = \frac{-x^2+4x}{8}.$

For the distribution function it applies

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

Notice: For each $x \in R$ we need to compute the integral on $(-\infty, x)$.

For $x \in (-\infty, 0)$ the integral is zero, i.e. $I_1(0) = 0$.

For $x \in (0,1)$ we compute

$$I_2(x) = \int_{-\infty}^{x} f(t)\,dt = I_1(0) + \int_0^x \frac{-t^2+4t}{8}dt =$$

$$= 0 + \frac{1}{8}\left[-\frac{1}{3}t^3 + 2t^2\right]_0^x = -\frac{1}{24}x^3 + \frac{1}{4}x^2$$

and $I_2(1) = \frac{1}{8}\left(-\frac{1}{3}+2\right) = \frac{5}{24}$

For $x \in (1,3)$

$$I_3(x) = \int_{-\infty}^{x} f(t)\,dt = I_2(1) + \int_1^x \frac{t^2-4t+6}{8}dt =$$

$$= \frac{5}{24} + \frac{1}{8}\left[\frac{1}{3}t^3 - 2t^2 + 6t\right]_1^x = \frac{1}{8}\left(\frac{1}{3}x^3 - 2x^2 + 6x - \left(\frac{1}{3} - 2 + 6\right)\right) =$$

$$= \frac{1}{24}x^3 - \frac{1}{4}x^2 + \frac{3}{4}x - \frac{1}{3}$$

and $I_3(3) = \frac{19}{24}$.

For $x \in (3,4)$

$$I_4 = I_3 + \int_3^x \frac{-t^2+4t}{8}dt = \frac{19}{24} + \frac{1}{8}\left[-\frac{1}{3}t^3 + 2t^2\right]_3^x =$$

$$= \frac{1}{8}\left(\frac{19}{3} - \frac{1}{3}x^3 + 2x^2 - \left(-\frac{1}{3}\cdot 27 + 2\cdot 9\right)\right) = -\frac{1}{24}x^3 + \frac{1}{4}x^2 - \frac{1}{3}$$

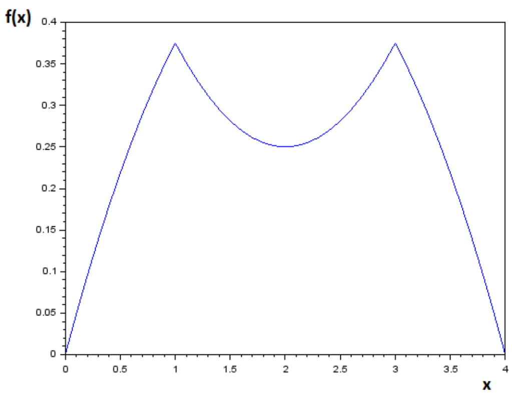
$I_4(4) = 1$ (check of correctness)

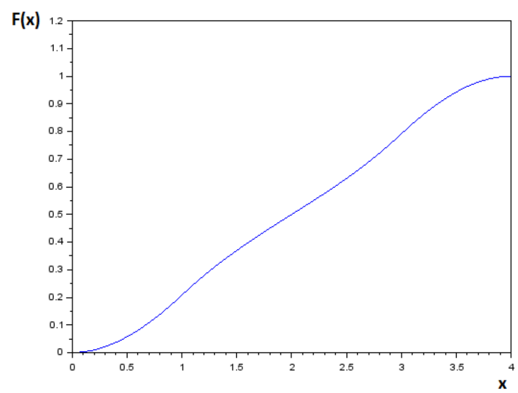
For $x\,(4,\infty)$ the integral is equal to 1.

Altogether, we can write

$$F(x) = \begin{cases} 0 & \text{for } \ x \in (-\infty, 0) \\ -\frac{1}{24}x^3 + \frac{1}{4}x^2 & \text{for } \ x \in (0, 1) \\ \frac{1}{24}x^3 - \frac{1}{4}x^2 + \frac{3}{4}x - \frac{1}{3} & \text{for } \ x \in (1, 3) \\ -\frac{1}{24}x^3 + \frac{1}{4}x^2 - \frac{1}{3} & \text{for } \ x \in (3, 4) \\ 1 & \text{for } \ x \in (4, \infty) \end{cases}$$

Density and distribution functions are in the following pictures





10.2.8 Example

For the density function

$$f(x, y) = \frac{4}{5}\left(x^2 + xy + 2y^2\right), \ x, y \in (0, 1)$$

determine marginal and conditional distributions and decide if the random variables are independent.

Solution

Marginals are

$$f(x) = \frac{4}{5}\int_0^1 \left(x^2 + xy + 2y^2\right) dy = \frac{2}{15}\left(6x^2 + 3x + 4\right)$$

$$f(y) = \frac{4}{5}\int_0^1 \left(x^2 + xy + 2y^2\right) dx = \frac{2}{15}\left(12y^2 + 3y + 2\right)$$

Conditional distributions

$$f(x|y) = \frac{f(x,y)}{f(y)} = \frac{6\left(x^2+xy+2y^2\right)}{12y^2+3y+2}$$

$$f(y|x) = \frac{f(x,y)}{f(y)} = \frac{6\left(x^2+xy+2y^2\right)}{6x^2+3x+4}$$

As evidently $f(x,y) \neq f(x|y)$ and also $f(x,y) \neq f(y|x)$, the variables are not independent.

Expectations

$$E[X] = \int_0^1 xf(x)\,dx = \frac{3}{5}$$

$$E[Y] = \int_0^1 yf(y)\,dy = \frac{2}{3}$$

Variances

$$D[X] = \int_0^1 (x-E[X])^2 f(x)\,dx = \frac{7}{90}$$

$$D[Y] = \int_0^1 (y-E[Y])^2 f(y)\,dy = \frac{29}{450}$$

Covariance

$$C[X,Y] = \int_0^1 \int_0^1 (x-E[X])(y-E[Y]) f(x,y)\,dxdy = -\frac{1}{90}$$

10.2.9 Example

We have a random variable X with density function

$$f(x) = a^2 x \exp(-ax), \quad x \in (0,\infty)$$

Compute expectation $E[X]$ and variance $D[X]$.

Solution

Expectation

$$E[X] = \int_0^\infty xf(x)\,dx = a^2 \int_0^\infty x^2 \exp(-ax)\,dx = (*)$$

The integral is solved by twice per partes:

$$\int_0^\infty x^2 \exp(-ax)\,dx = \begin{vmatrix} u = x^2 & v' = \exp(-ax) \\ u' = 2x & v = \frac{-1}{a}\exp(-ax) \end{vmatrix} =$$

$$\underbrace{\left[\frac{-x^2}{a}\exp(-ax)\right]_0^\infty}_{=0 \text{ (L'Hospital)}} - \int_0^\infty \frac{-2x}{a}\exp(-ax)\,dx =$$

$$= \frac{2}{a}\int_0^\infty x\exp(-ax)\,dx = \begin{vmatrix} u = x & v' = \exp(-ax) \\ u' = 1 & v = \frac{-1}{a}\exp(-ax) \end{vmatrix} =$$

$$= \frac{2}{a}\left\{\underbrace{\left[\frac{-x}{a}\exp(-ax)\right]_0^\infty}_{0 \text{ (L'Hospital)}} - \int_0^\infty \frac{-1}{a}\exp(-ax)\,dx\right\} =$$

$$= \frac{2}{a^2}\int_0^\infty \exp(-ax)\,dx = \frac{2}{a^2}\left[\frac{-1}{a}\exp(-ax)\right]_0^\infty = \frac{2}{a^3}$$

$$(*) = a^2\frac{2}{a^3} = \frac{2}{a}.$$

Variance

It is solved similarly by three times per partes with the result

$$D[X] = \tfrac{2}{a^2}$$

Remark

The result can be obtained also by Maxima (http://maxima.sourceforge.net/).

10.3 Examples to regression analysis

10.3.1 Example

A certain firm monitored its yearly profit. The data collected are in the table

year	1995	2000	2003	2008	2011	2015	2016	2017	2018	2019
profit (mil $)	55	50	50	53	45	60	61	67	65	66

They are interested about the prediction of the profit for the year 2030. To this end perform linear and cubic polynomial regression, select the better one and perform the prediction. Compare it with the prediction from the other regression.

Solution

As the numbers denoting the years are unnecessarily too big (or computation) we introduce the following transformation

$$x = \text{year} - 1900, \text{ and } y = \text{profit}$$

and with it we have data

x	95	100	103	108	111	115	116	117	118	119
y	55	50	50	53	45	60	61	67	65	66

The solution can be found with Statext:

Linear regression

Parametric|Simple regression|Linear...

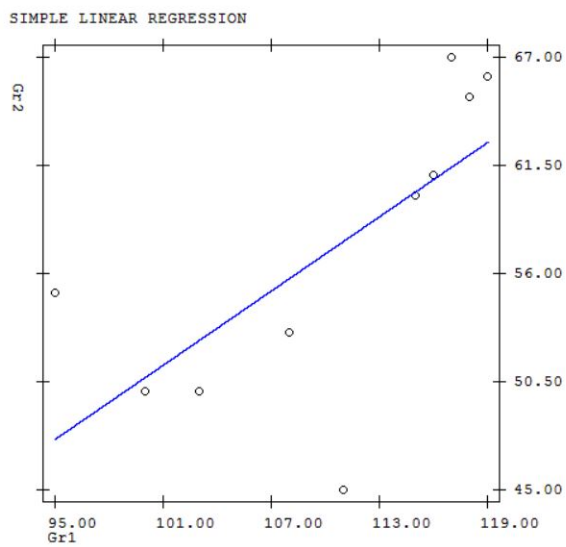
The equation is

$$y_p = 0.63x - 12.13$$

p-value = 0.028

prediction $yp = 0.63 \cdot 130 - 12.13 = 69.77$

The resulting regression is in the following picture



Cubic polynomial regression

Parametric|Simple regression|Cubic...

Equation

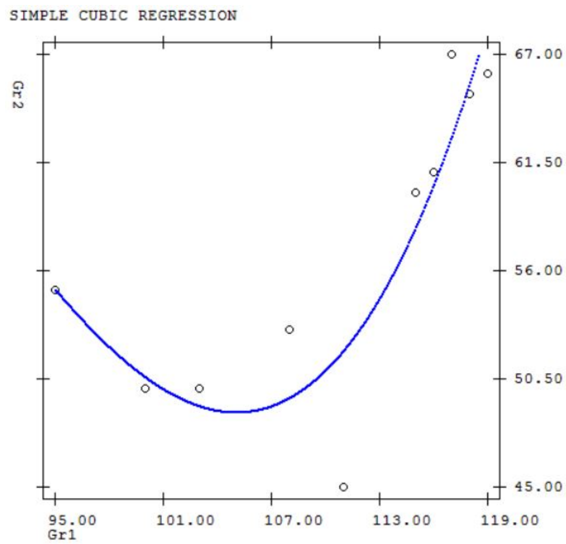
$$y_p = 0.0016x^3 - 0.437x^2 + 37.688x - 982.634$$

p-value = 0.0096

prediction $y_p = 0.0016 \cdot 130 - 0.437 \cdot 130^2 + 37.688 \cdot 130 - 982.634 = 46.71$

Conclusion: The cubic regression is better (it has smaller p-value). The prediction from it is $y_p = 46.71$. However, both the predictions differ considerably. So, we would like to try other regression - 5th order polynomial. As it is not supported by Statext, we will continue with formulas.

The regression for cubic case is



10.3.2 Polynomial regression of the 5-th order

As a basis for our demonstration, we will consider a general multi-regression model in the form

$$y_t = x_t \cdot \theta + e_t$$

where

$$x_t = [1,\, x_1,\, x_2, \cdots, x_n]_t \;\text{ and }\; \theta = [b_0, b_1, b_2, \cdots, b_n]'$$

Remark

The product $1 \cdot b_0$ produces a constant term of the model.

Now, for $t = 1, 2, \cdots, N$ we can write individual prediction models one below another and to construct a matrix form of the model for all data

$$Y = X\theta + E$$

where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{1;1} & x_{2;1} & \cdots & x_{n;1} \\ 1 & x_{1;2} & x_{2;2} & \cdots & x_{n;2} \\ \cdots & \cdots & \cdots\cdots\cdots & \cdots & \cdots\cdots \\ 1 & x_{1;N} & x_{2;N} & \cdots & x_{n;N} \end{bmatrix}, \quad E = \begin{bmatrix} e_1 \\ e_2 \\ \cdots \\ e_N \end{bmatrix}$$

Then for the optimal estimate $\hat{\theta}$ of the parameter θ it holds

$$\hat{\theta} = \left(X'X\right)^{-1} X'Y$$

Derivation of the formula is very simple. The sum of prediction errors can be written as $E'E = \sum_{t=1}^{N} e_i^2$ *and substituting the model we have*

$$E'E = (Y - X\theta)'(Y - X\theta) = \theta' X' X\theta - 2\theta' X'Y + Y'Y$$

Le look for minimum, so we differentiate and lay the derivative equal to zero[4]

$$2X'X\theta - 2X'Y = 0 \quad \rightarrow \quad \theta = \left(X'X\right)^{-1} X'Y$$

that is the result which was to be proved.

Now, to our problem. In the case of 5-th order polynomial regression, the model is

$$y_t = b_0 + b_1\theta + b_2\theta^2 + b_3\theta^3 + b_4\theta^4 + b_5\theta^5 + e_t$$

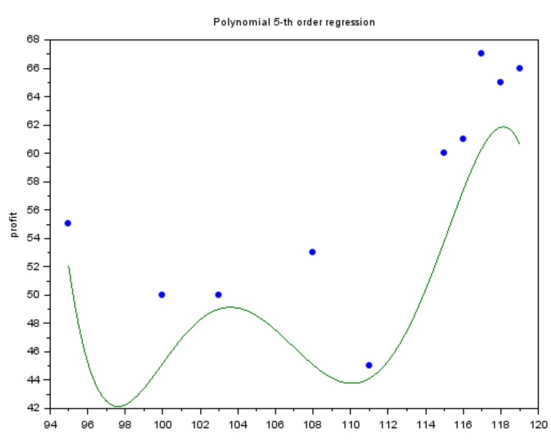
For the measured data

x	95	100	103	108	111	115	116	117	118	119
y	55	50	50	53	45	60	61	67	65	66

we have

$$Y = \begin{bmatrix} 55 \\ 50 \\ 50 \\ 53 \\ 45 \\ 60 \\ 61 \\ 67 \\ 65 \\ 67 \end{bmatrix}, \quad X = \begin{bmatrix} 1,\ 95,\ 95^2,\ 95^3,\ 95^4,\ 95^5 \\ 1,\ 100,\ 100^2,\ 100^3,\ 100^4,\ 100^5 \\ 1,\ 103,\ 103^2,\ 103^3,\ 103^4,\ 103^5 \\ 1,\ 108,\ 108^2,\ 108^3,\ 108^4,\ 108^5 \\ 1,\ 111,\ 111^2,\ 111^3,\ 111^4,\ 111^5 \\ 1,\ 115,\ 115^2,\ 115^3,\ 115^4,\ 115^5 \\ 1,\ 116,\ 116^2,\ 116^3,\ 116^4,\ 116^5 \\ 1,\ 117,\ 117^2,\ 117^3,\ 117^4,\ 117^5 \\ 1,\ 118,\ 118^2,\ 118^3,\ 118^4,\ 118^5 \\ 1,\ 119,\ 119^2,\ 119^3,\ 119^4,\ 119^5 \end{bmatrix}$$

The result is shown in the picture



[4]We differentiate according to θ which is a vector. We must preserve the rules for differentiation of matrices.

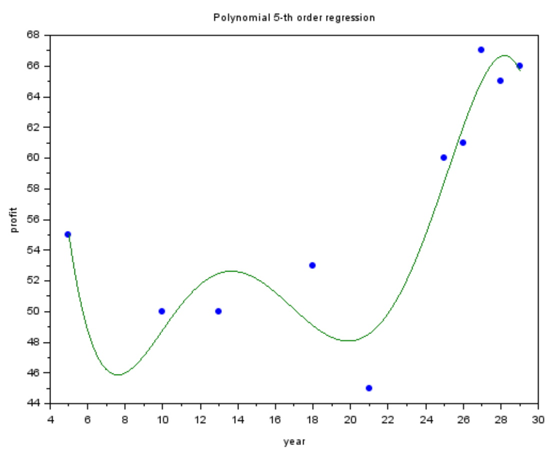
where we immediately can see that something is wrong. The whole prediction lies below the data points. From the estimated parameters

$$\hat{\theta} = [3290679.2 - 154287.122888.85 - 27.000410.1259682 - 0.0002347]$$

we can see, that the regression curve will be numerically sensitive. Compare the largest and the smallest regression coefficient. The reason can be in still relatively big values of x. We try to transform more $x_{new} = x_{old} - 90$. This gives the x values

x	5	10	13	18	21	25	26	27	28	29

Now the picture looks like this

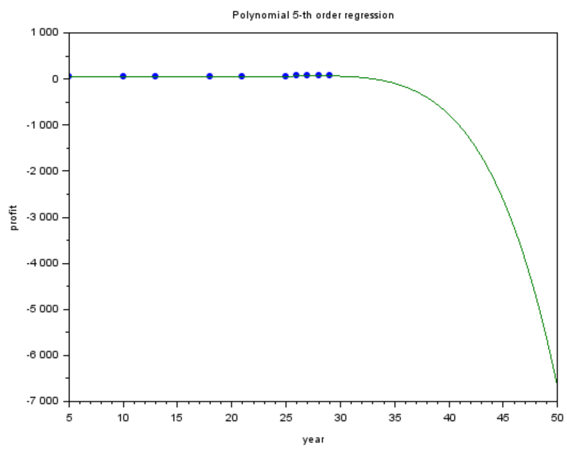


which is much more better.

The regression parameters, now, are better

$$\hat{\theta} = [210.37, -64.68,\ 9.39, -0.62,\ 0.019, -0.00022]$$

the approximation is not good. And the prediction is visible from the picture with prolonged x to the value 50 (which corresponds to the year 2030)



We can conclude: the higher order of the polynomial used in regression does not need to lead to better quality.

The result is, that the 5-th order regression is of no use.

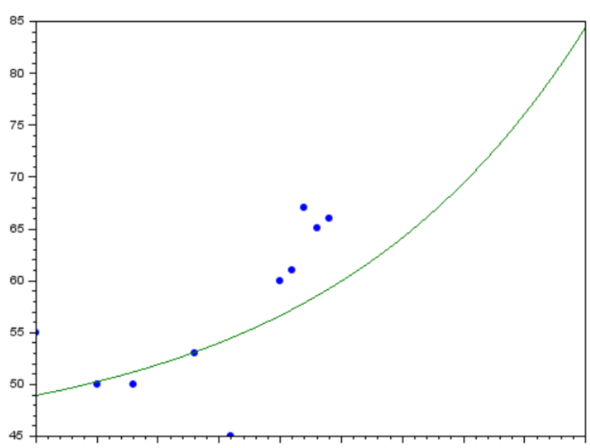
10.3.3 Exponential regression

When we look at the data, we can try to improve the prediction using the exponential regression. This can be taken from Statext, however, it is a nonlinear one using the prediction equation

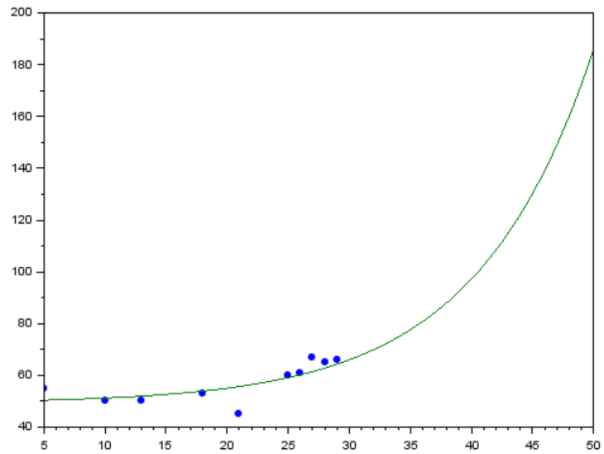
$$y_t = a + b_2 \exp\left(b_1 x\right)$$

with the estimated parameters $a = 43.9$, $b_0 = 0.062$ and $b_1 = 0.046$.

the graph of the regression (with the prolonged x axis to the point prediction - year 2030) is here



which also is not ideal. The reason is the extremely small value at he point $x = 21$. If we take it as an outlier and omit it, we get



and we can say that ow the regression is well. However, the prediction is very uncertain. It relays on the exponential course of the data which is not fully acknowledged by the data themselves. All in all we can state that this prediction is to long ahead to be reliable.

In the end, we try the standard linearized exponential regression with the prediction equation

$$y_t = b_0 \exp(b_1 x) \quad \rightarrow \quad \ln(y_t) = \ln(b_0) + b_1 x$$

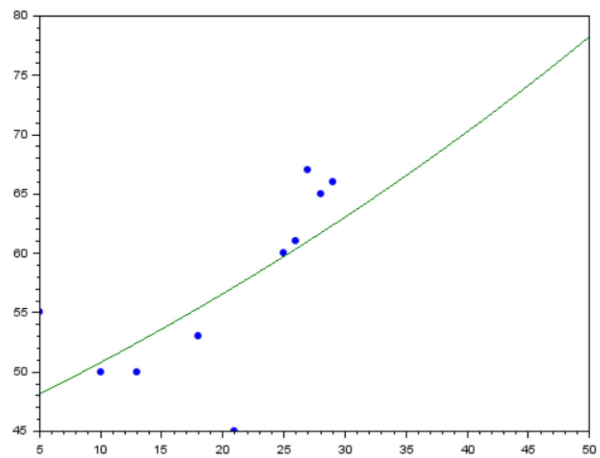
It can be computed similarly as the 5th order polynomial regression. We construct vector Y and matrix X

$$Y = \begin{bmatrix} \ln(y_1) \\ \ln(y_2) \\ \cdots \\ \ln(y_{10}) \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_N \end{bmatrix}$$

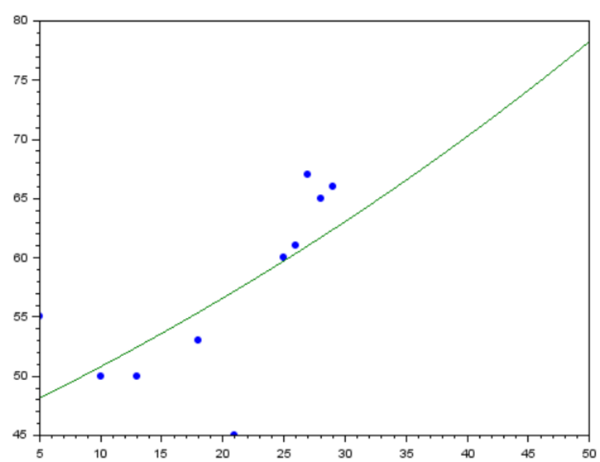
and with the formula $\hat{\theta} = (X'X)^{-1} X'Y$ we get the parameters

$\ln(b_0) = 3.82 \quad \rightarrow \quad b_0 = 45.6$ and $b_1 = 0.01$

and the graph



and this is not bad. After omitting the outlier $[21, 45]$ we get



which is approximately the same as with it. So the solution can be considered to be relatively stable.

The p-value is 0.031 which in comparison with the 3-th order polynomial regression computed at the beginning of this example shows that the polynomial one is better and so it is an absolute winner.

10.4 Examples to confidence intervals

10.4.1 Example

From a long-time monitoring, it is known, that the standard deviation of speeds of cars going in a road with restriction to 40 km/h is 6.9 km/h. We have measured the speeds and get the data

$$x = [41, 45, 38, 43, 35, 37, 42, 61, 37, 40, 42, 44, 38, 45, 39]$$

We would like to construct 0.05-confidence interval for the true speed expectation.

Solution

In the Statext the task is very simple.

We choose: Parametric|Confidence interval for the mean... and check the square at Using z-table (as we know the variance).

We obtain

$$I_{0.05} = (38.31,\ 45.29)$$

As the speed 40 km/h is within the interval, we can say that we do not reject the hypothesis that the cars go with this speed.

Now, for those who are interested we will show how the interval is derived.

We have the population x of speeds for which we can assume that it is normal and in addition we assume its real standard deviation to be $\sigma = 6.9$ km/h. That is $f(x, \mu) = N_x(\mu, \sigma^2)$.

For estimation of the expectation we choose the statistics in the form of sample average $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$. From the theory we know, that the sample average from normal data is also normal with the expectation μ and variance $\frac{\sigma^2}{N}$, where μ and σ^2 are characteristics of the population. So, $f(\bar{x}, \mu) = N_{\bar{x}}\left(\mu, \frac{\sigma^2}{N}\right)$.

Now, we want to find interval $I_{0.05}$=(I_L, I_U) for which it holds

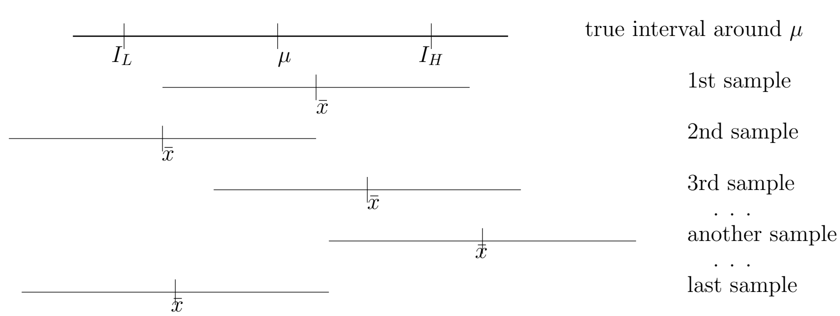
$$P(\bar{x} \in I_{0.05}) = 0.05 \tag{13}$$

Remark

According to the definition of confidence interval, which is

$$P(\mu \in I_{0.05}) = 0.05$$

it could seem that what we have written is wrong. However, it is necessary to realize that both the probabilities are equal. It can be seen from the following picture



Here, on the upper line we have denoted expectation μ and around it some interval $I = (I_L, I_U)$, let it be the confidence interval. On the lines below it we denoted several samples and their sample averages $\bar{x}$. Around each sample average we draw the same interval I. Now, For the first three of them and the last one it holds that the sample average belongs to the interval around μ and also that μ also belongs to the interval around the sample average. The assertion is reciprocal. For the fourth line represents a sample for which, again in reciprocal way, it holds that μ is not form I around μ and $\mu$is not in the interval around $\bar{x}$. So, if there are 95% of samples within the confidence interval then th probability of μ being in the interval is 0.95.

Now, for the outside of the interval (13), it can be written[5]

$$P(\bar{x} < I_L) = \frac{\alpha}{2} \text{ and } P(\bar{x} > I_U) = \frac{\alpha}{2} \tag{14}$$

where we generally denoted the confidence level 0.05 by α and I_L, I_B are interval borders.

To evaluate these probabilities (without using computer) we can use tables of the normalized normal distribution where we can find quantiles and critical values of the random variable. For α-quantile it holds $P(z < \zeta_\alpha) = \alpha$where z is normalized random variable and ζ_α is the quantile. For α-critical value $P(z > z_a) = \alpha$with z_a the critical value. The normalization is

$$z = \frac{x - \mu}{\sigma} \text{ and } \bar{z} = \frac{\bar{x} - \mu}{\sigma}\sqrt{N}$$

where x is non-normalized random with characteristics μ and σ and $\bar{x}$ is sample average of x whose characteristics are μ and $\frac{\sigma}{\sqrt{N}}$. If we switch to normalized variables we can write (14) like this

$$P\left(\bar{z} < \zeta_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2}, \quad P\left(\bar{z} > z_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$$

where $\bar{z}$ is normalized sample average.

Substituting the above formulas into the argument of the probabilities we obtain

$$\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{N}}} < \zeta_{\frac{\alpha}{2}} \text{ and } \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{N}}} > z_{\frac{\alpha}{2}} \tag{15}$$

and from it

$$\bar{x} - \zeta_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{N}} < \mu \text{ and } \mu < \bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{N}}$$

and because it holds $\zeta_{\frac{\alpha}{2}} = -z_{\frac{\alpha}{2}}$ it holds (for the outside of confidence interval)

$$\underbrace{\bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{N}}}_{I_U} < \mu \text{ and } \mu < \underbrace{\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{N}}}_{I_L}$$

[5]We consider symmetrical interval in the sense that we demand so that the probability to the left and right of the interval be one half of 0.05.

So, the interval is

$$I = \bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}} = \left(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}},\ \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}} \right)$$

For our example we get

Sample average $\bar{x} = 41.8$, $\sigma = 6.9$ (given as known), $N = 15$, $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$ and

$$I_{0.5} = (38.308,\ 45.292)$$

which is the same as form Statext.

10.5 Examples to tests of hypothesis

10.5.1 Example

From a long-time monitoring, it is known, that the standard deviation of speeds of cars going in a road with restriction to 40 km/h is 6.9 km/h. We have measured the speeds and get the data

$$x = [41, 45, 38, 43, 35, 37, 42, 61, 37, 40, 42, 44, 38, 45, 39]$$

Test H0: the average speed of cars does not exceed 40 km/h.

Solution

The answer can be provided by Statext

We choose: Parametric|Test for a population mean...

Here we set SD = 6.9, H0 Pop. mean is equal to 40, Ha: is greater than

and we get (for Normal distribution) pv= 0.156 which means, that H0 is not rejected.

But how we can come to this result?

The population is assumed to have normal distribution with standard deviation $\sigma = 6.9$ and unknown expectation μ. H0 claims that the expectation is 40 km/h[6], so HA: the expectation is greater than 40 km/h.

So, according to the H0 ($\mu_0 = 40$km/h) we have the distribution

$$f(x|H0) = N_x\left(40, 6.9^2\right)$$

[6]In the example we have H0: speed is nor greater than $\cdots$. However, for H0 we need to have a fixed number as a value of the parameter. So, we set it to $\mu = 40$ and the formulation "is not greater$\cdots$" we use for determining of the direction of the test. With respect to H0, the hypothesis HA says the opposite "is greater $\cdots$", which means right-sided test.

However, in testing, we are nit interested of x but of μ which is represented by its point estimate $\mu = \bar{x}$ (sample average) for which the distribution is

$$f\left(\bar{x}|H0\right) = N_x\left(40, \frac{6.9^2}{15}\right)$$

where 15 $=N$. So, if the data come from the distribution according H0, their sample average $\bar{x}$ should mostly lie within the confidence interval. For it, we have (see the previous Example 15)

$$\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{N}}} < \zeta_{\frac{\alpha}{2}} \text{ and } \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{N}}} > z_{\frac{\alpha}{2}}$$

where, now, we assume $\mu = \mu_0$ according to H0. So, we can use this formula as follows

$$\bar{z} < -z_{\frac{\alpha}{2}} \text{ and } \bar{z} > z_{\frac{\alpha}{2}}$$

where $\bar{z} = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{N}}}$ is normalized sample average with respect to H0 and we used the property $\zeta_\alpha = -z_\alpha$. The above inequalities specify the area outside the confidence interval (the probability is α). So, for $\bar{z}$ to be within the confidence interval, it must hold

$$\bar{z} \in \left(-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}\right)$$

The conclusion is:

If $\bar{z} \in \left(-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}\right)$ then we admit that the sample average we have calculated from the measured sample comes from the distribution according to H0 and thus, we do not reject that $\mu \leq \mu_0$.

If $\bar{z} \notin \left(-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}\right)$ then if H0 is valid only for $\alpha \times 100\%$ samples we would get this result. As α is very small probability we consider events with probability α as almost impossible and that is why we reject H0, that is we say that $\mu \neq \mu_0$.

Now, to our example:

First of all we have right-sided test, so the interval will be $(-\infty, z_\alpha) = (-\infty, 1.645)$ for $\alpha = 0.05$.

The statistics (sample average) is $\bar{x} = 41.8$.

The normalized statistics is

$$\bar{z} = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{N}}} = \frac{41.8-40}{\frac{6.9}{\sqrt{15}}} = 1.01$$

As $\bar{z} \in (-\infty, 1.645)$ we do not reject H0.

This is the decision about rejecting. If we would like to know more - how strongly we reject / do not reject, we need to evaluate p-value.

P-value

The definition of p-value (for right-sided interval) is

$$pv = P\left(\bar{z} > \bar{z}_r | H0\right) \tag{16}$$

What does it mean. $\bar{z}$ is normalized statistics (as random variable), $\bar{z}_r$ is realized statistics (computed value of normalized statistics from measured sample) and H0 in the condition means, that the normalization is performed using characteristics according to H0, i.e.

$$\bar{z} = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{N}$$

where μ_0 from H0 is used.

The probability in (16) can be computed as

$$pv = 1 - F\left(\bar{z}_r\right)$$

where F is a distribution function of the distribution of the normalized statistics (here it will be $N\left(0,1\right)$).

Our realized statistics is $\bar{z}_r = 1.01$ (see above) and using a computer or from tables we have

$$pv = 0.156$$

which is the same result as from Statext.

11 Questions for knowledge validation

11.1 Variables and data

1. What is a data file?
2. Explain the ways of storing data in a plain form and as values and frequencies.
3. Having data 8 6 2 8 9 7 5 4 - determine their ranks.
4. Which characteristic of data file express their level and which ones describe their variability?
5. Determine median of the data file 2 4 6 1 7 3 9 .
6. Compute average and variance of data file 3.15 2.22 10.45 3.57 1.58 6.9 5.13 -0.75 2.24 6.58 4.65 6.52 3.79 4.95 1.87 3.87 4.22 3.68 6.3 3.64 ?
7. What is the mode of data file 2 5 4 6 4 2 4 2 6 2 2 2 6 4 2?
8. What is the difference between general bar graph and histogram?

11.2 Probability and random variable

1. What is random experiment?
2. What is an event?
3. What data produces the experiment of flipping a coin?
4. What data produces random variable describing the experiment of flipping a coin?
5. Which are the three important properties of the probability?
6. Define classical definition of probability.
7. Define statistical definition of probability.
8. What is the major difference between classical and statistical definition of probability.
9. After ten flips of coin we obtained 3 heads and 7 tails. We concluded that the probability of head is 0.3. Which definition of probability we have used?
10. We inspected a coin and concluded that it is not damaged. So we determined the probability of head is 0.5. Which definition of probability we have used?

11. What is the definition of conditional probability?

12. Consider an experiment of tossing a dice. What is the probability of even number if we know, that the number that really fell is less than 4.
Hint: Choose only from those that could have fallen.

13. Consider an experiment of drawing colored balls from a box. The drawn ball is not returned back. Are the draws independent?
Hint: Is the probability of drawing specified color all the same during the draws?

14. We have an experiment of flipping two coins. The results are: "both heads", "both tails" and "different sides". Are the probabilities of these result equal?

15. A natural definition of independence x and y is $P(x|y) = P(x)$ - the knowledge of y does not influence the probability of x. Using the definition of conditional probability, derive the formula $f(x,y) = f(x)\,f(y)$.

11.3 Description of random variable and vector

1. What is the difference between random experiment and random variable.

2. What types of random variable do you know?

3. Can random variable have negative values?

4. What is the merit of random vector in comparison with random variable?

5. What are realizations of a random vector?

6. Define distribution function.

7. What are the basic properties of a distribution function?

8. Determine the probability $P(a,b)$, $b > a$ using the distribution function.

9. What is the probability $P(X = 5)$ for a continuous random variable X equal to?

10. Is any distribution a continuous function?

11. Define probability function of discrete random variable.

12. Define density function of continuous random variable.

13. What are the basic properties of any probability or density function.

14. What assertion is correct: all values of any probability function are (i) non-zero, (ii) positive, (iii) non-negative.

15. What assertion is correct: all values of any density function are (i) non-zero, (ii) positive, (iii) non-negative.

16. How the probability $P(a,b)$, $b > a$ can be computed using probability or density function?

17. What is the definition of expectation for discrete random variable?

18. What is the definition of expectation for continuous random variable?

19. Define α-quantile and α-critical value of continuous random variable.

20. How can you find median of continuous random variable.

21. How can you find mode of continuous random variable.

22. We have a function $y = kx$. Determine the constant k so that this function would be a density function of the interval $x \in (0,\, 5)$.

23. The probability function of random variable X is given by the table

x	1	2	3	3	5
$f(x)$	0.2	0.1	0.3	0.2	0.2

Determine the distribution function.

24. Write the probability functions describing the experiments (i) "flipping a fair coin" and (ii) "flipping some unfair coin".

25. The random variable X describes the following experiment "time of waiting for a bus with an interval 5 min if your coming to the bus station is random". Write the density and distribution functions of this random variable.

26. We have two random variables X and Y with probability functions $f(x) = 2 - 2x$ and $f(y) = 2y$, both on $x \in (0,\, 1)$ and zero otherwise. Write the joint density function $f(x,y)$.

27. A random vector $[X,\, Y]$ has joint density function

$$f(x,y) = k \exp\left(x^2 + 2y^2\right)$$

for $x, y \in (-\infty,\, \infty)$. k in a normalization constant. Are these random variables independent?

28. A random vector $[X,\, Y]$ has joint density function $f(x,y) = 1$ on $x \in (0,\, 1)$ and $y \in (0,\, 1)$. Determine the probability $P([X,Y] \in (0,\, 0.1) \times (0,\, .2))$.

29. A random vector $[X,\, Y]$ has joint probability function given by the table

x/y	1	2	3
1	0.1	0.05	0.3
2	0.35	0.1	k

Determine the value of k.

30. A random vector $[X, Y]$ has joint probability function given by the table

x/y	1	2	3
1	0.2	0.1	0.1
2	0.3	0.1	0.2

Determine the marginal $f(y)$ and the conditional $f(x|y)$.

31. A random variable X has probability function $f(x) = p^x (1-p)^{1-x}$, $x = 0, 1$ and $p \in (0, 1)$.
a) What is the probability that in the next sampled value of X will be 1?
b) What is the expected number of results $x = 1$ in 100 experiments?

11.4 Important distributions

1. What is the definition of Bernoulli distribution?

2. What is the meaning of the parameter p in Bernoulli distribution?

3. The probability that a newborn will be a boy is 0.52. What is the probability that in a family with 5 children there will be 2 boys and three girls?

4. Describe the experiment connected with a binomial random variable.

5. For several years we have measured accidents in five points of a large traffic region. The results were

point of measurement	1	2	3	4	5
numb. of accidents	38	147	51	223	197

On the basis of the measured data determine the probability function describing these accidents.

6. We are flipping a fair coin. What is the probability, that the "head" appears at the third flip for the first time.

7. One percent of bits transmitted through a digital transmission are received in error. Bits are transmitted until the first error. Let X denote the number of bits transmitted until the first error. What distribution describes random variable X?

8. Compare the supports of Normal distributions $N(0, 1)$ and $N(10, 3)$, where the denotation is $N(\mu, \sigma^2)$.

9. A fixed distance 10 meters is repeatedly measured. We define two random variables X - the value of the measurement and Y - the error of measurement from the true value. What is the distribution of these random variable and what they differ in?

10. What are the main assumptions of a uniform distribution.
11. We have a random variable with uniform distribution on the interval (3, 9) . What is the probability that a realization of X will be within the interval (5, 7).

11.5 Regression analysis

1. Write regression line constructed only for two measured points: $x_1 = [1, 1]$, $x_2 = [2, 5]$. What will be the correlation coefficient?
2. The regression line is computed for three data points
$$x_1 = [1, 2] \, , \, x_2 = [2, 2] \, , \, x_3 = [3, 5]$$
Compute the criterion as a sum of squares of residuals.
3. We investigate a profit (y) in dependence on invested money (x). We obtained linear regression $y = 0.21x - 100$. What will be the profit if we invest 5000 (Kč)?
4. We investigate a profit (y) in dependence on invested money (x). We obtained linear regression $y = 0.21x - 100$. How much we need to invest to have the profit equal to 1000 (Kč)?
5. We have exponential regression $y = b_0 \exp(b_1 x)$. Perform its linearization.
6. Write a quadratic and cubic regression curves.
7. The polynomial regression has coefficients $b_0 = 2.1$, $b_1 = 0.6$, b_2=1.2 and $b_3 = 0.1$. Write the value of the prediction at $x = 2$.

11.6 Population and data sample

1. What are the differences between population and random sample?
2. What is the difference between random sample and sample realization?
3. Is an average of random sample a number or random variable?
4. Is an average of sample realization a number or random variable?
5. We measure speeds of passing cars. What is the population, what random sample and what sample realization?
6. We throw ten times a dice. What is the population, what random sample and what sample realization?
7. We monitor speeds of cars on a certain point of a motorway. We assume that the speeds are normally distributed with the variance 5.83. What parameter we need to estimate?

8. What is the difference between parameter and its point estimate?
9. What is the statistics in stochastic estimation?
10. What are the most important characteristics of random sample?
11. What are the expectation and variance of sample average equal to?
12. What is the meaning of the formula expressing the expectation of sample average?
13. What is the meaning of the formula expressing the variance of sample average?
14. Which properties has the sample average.
15. When comparing efficiency of two sample averages with different lengths of sample, which one would be better (has higher efficiency)?

11.7 Statistical inference

1. What statistics is suitable for estimation of expectation?
2. What statistics is suitable for estimation of variance?
3. What statistics is suitable for estimation of proportion?
4. What statistics is suitable for testing of independence?
5. A sample of the length n is taken from normal population with expectation μ and variance σ^2. What is the distribution of sample average?
6. We have a population $f(x)$. We want to compute a confidence interval for its unknown parameter μ. We chose the statistics $\bar{x}$ (sample average). Which distribution is used: $f(x)$ or $f(\bar{x})$?
7. What is the difference between both-sided, left-sided and right-sided interval?
8. What are zero and alternative hypotheses?
9. What are region of acceptance and critical region?
10. What is realized statistics T_r?
11. Explain the difference between both-sided, left-sided and right-sided test.
12. What is the connection between confidence interval and test of hypothesis.
13. What is the conclusion of a test if the realized statistics lies in the critical region?
14. What is the conclusion of a test if the p-value is smaller then the confidence level?

11.8 Important tests

1. What is the difference between parametric and nonparametric test.
2. Which are the parametric tests with one sample?
3. Which are the parametric tests with two samples?
4. Which are the parametric tests with more samples?
5. Which are the nonparametric tests with one sample?
6. Which are the nonparametric tests with two samples?
7. Which are the nonparametric tests with more samples?
8. Which samples of independence do you know?
9. Which are the tests of a distribution?

11.9 Validation in regression analysis

1. How can you evaluate the quality of regression analysis according to the xy-graph?
2. What is the statistics for Pearson t-test?
3. Is a regression analysis suitable if the p-value of the Pearson test is very small?
4. Is a regression analysis suitable if the p-value of the F-test is very small?
5. Is a regression analysis suitable if the sum of squares of residuals is very small?
6. Should residuals be independent or dependent?

12 Answers to the questions

Please, try to answer the questions first of all on the basis of the textbook. These answers should only acknowledge your answers if you think they are correct.

12.1 Variables and data

1. Data file is a set of measured data.
2. Plain form stores data in the order how they are measured. Alternative way is to store only different values and the numbers of their occurrence.
3. The ranks are 6.5, 4, 1, 6.5, 8, 5, 3, 2. For the same values, the average of their order is taken.
4. Level of data is given by expectation, mode or median, variability corresponds to variance, standard deviation.
5. Median is 4.
6. Average is 4.228, variance 5.80748.
7. Mode is 2.
8. General graph indicates values, histogram frequencies.

12.2 Probability and random variable

1. It is any action leading to some of correctly defined result. Repeating the action leads accidentally to different results.
2. Event is a set of results. It can be also specified verbally. E.g. at a dice, we can say "even number" and it determines the set $\{2, 4, 6\}$.
3. They are "Head" and "Tail".
4. They are usually 0 and 1. They also can be 1 and 2 or any other two numbers.
5. They are that probability is (i) non-negative $(P(A) \geq 0)$, (ii) less or equal to 1 $(P(A) \leq 1)$, (iii) σ-additive (for events A, B such that $A \cap B = \emptyset$ it holds $P(A \cup B) = P(A) + P(B)$.

6. $P=\frac{m}{n}$, where m is number of all possible positive results and n is number of all possible results.

7. $P=\frac{M}{N}$, where M is number of all performed positive results and N is number of all performed results.

8. Classical definition takes into account possible results while statistical one speaks about performed results. The former is theoretical and its result is constant, the latter practical and its results a bit vary with a specific setof experiments.

9. Here we have used the statistical definition.

10. This relates to the classical definition.

11. It is denoted $P(A|B)$ and it is computed plainly as a probability of A but on the set of results which meet the condition. With a dice $P(even| < 5)$ is a probability of even on a set $\{1,2,3,4\}$ which is $2/4 = 0.5$.
    The definition is
    $$P(A|B) = \frac{P(A,B)}{P(B)}$$

12. $A = "even" = (2,4,6)\,,\ B = " < 4" = \{1,2,3\}\,.$ The set B has 3 elements, one of them is even: $P = 1/3$.

13. They are not independent as the probability depends on which ball was previously drown.

14. They are not. “different sides” has twice much possibilities.

15. $P(x|y) = P(x,y)/P(y) \underbrace{=}_{def.} P(x)\,.$ Multiplication by $P(y)$ gives the result.

12.3 Description of random variable and vector

1. Random variable corresponds to random variable. However, values of random variable (which correspond to results of experiment) must be numeric. If results are not numeric (red, yellow, green), we must assign them numbers (e.g. 1,2,3).

2. Discrete and continuous.

3. Yes, it can.

4. There are two things: (i) common treatment of several random variables, (ii) mutual connection between random variables.

5. They are vectors of numbers.

6. $F_X(x) = P(X \le x)$, X is random variable, x is a number.

7. Zero for $x \to -\infty$, one for $x \to \infty$ and non-decreasing on the whole support.
8. $P(x \in (a, b)) = F(b) - F(a)$.
9. Probability of each single number is zero for continuous random variable.
10. No, discrete ones are only piece-wise continuous.
11. It is a discrete function with values equal to individual values of random variable.
12. It is a derivative of the distribution function.
13. Non-negative values and sum (integral) equal to one.
14. (iii) is correct.
15. (iii) is correct.
16. $P(a, b) = \int_a^b f(x)\, dx$
17. $E[X] = \sum_{x_i \in X} x_i f(x_i)$
18. $E[X] = \int_{-\infty}^{\infty} x f(x)\, dx$
19. $\int_{-\infty}^{\zeta_\alpha} f(x)\, dx = \alpha$, $\int_{z_\alpha}^{\infty} f(x)\, dx = \alpha$
20. Median $\tilde{x} : \int_{-\infty}^{\tilde{x}} f(x) = 0.5$.
21. Mode $\hat{x} = \arg\max (f(x))$
22. $k = \frac{1}{5}$ (integral of $f(x)$ must be 1)
23. Distribution function is a cumulative sum of $f(x)$
24. (i) $f(x) = \begin{cases} 0.5 & \text{for } x = 0 \\ 0.5 & \text{for } x = 1 \end{cases}$, (ii) $f(x) = \begin{cases} 1 - p & \text{for } x = 0 \\ p & \text{for } x = 1 \end{cases}$
25. The density function is a constant $\frac{1}{5}$ from $x = 0$ to $x = 5$. The distribution function is: zero on $x \in (-\infty, 0)$, $F(x) = \frac{1}{5}x$ for $x(0, 5)$ and one for $x \in (0, \infty)$.
26. Joint is a product (they are independent).
27. Yes, they are. $f(x) = k \exp\left(x^2\right) \exp\left(2y^2\right) = f(x) f(y)$.
28. $P = 0.1 \cdot 0.2 = 0.02$. It is the volume of a prism with edges 0.1, 0.2 and 1.
29. $k = 1 - sum(\text{of the rest}) = 0.1$
30. Marginal (sum over columns) 0.5, 0.2, 0.3; Conditional (joint columns divided by marginal) $\begin{matrix} \frac{2}{5} & \frac{1}{2} & \frac{1}{3} \\ \frac{3}{5} & \frac{1}{2} & \frac{2}{3} \end{matrix}$.
31. $f(x = 1) = p^1 (1 - p)^{1-1} = p$; Expected number is $p \times 100$ (p in one experiment, $p \times 100$ in 100 experiments).

12.4 Important distributions

1. Random variable with two outcomes 0 and 1, where 1 has probability p (that is constant). E.g. Selection of a product hat can be either good (1) or defective (0), if $p \times 100\%$ of products is good.

2. p is a probability of the result $x = 1$.

3. Binomial distribution $f(x = 2) = \binom{5}{2} 0.52^2 0.48^{5-2} = 0.299$.

4. We consider a Bernoulli trial with $P(x = 1) = p$. We perform n experiments a want to know with probability k of them will be 1. !! p must stay constant during the experiments) !!

5. It is he same table where the second row is divided by the sum of all its entries. 0.058, 0.224, 0.078, 0.340, 0.300.

6. Geometrical distribution $f(x = 3) = 0.5 \cdot (1 - 0.5)^2 = 0.125$.

7. It is the geometrical distribution with $p = 0.01$.

8. They are both the same $(-\infty, \infty)$.

9. $f(x) = N\left(10, \sigma^2\right), f(y) = N\left(0, \sigma^2\right)$ where σ^2 is the variance of measurements.

10. For $X \sim U(a, b)$ it holds a) all values in the range (a, b) are equally probable and b) all values outside this interval are impossible.

11. $f(x)$ on $(3, 9)$ is $\frac{1}{6}$. So, $P(x \in (5, 7)) = \frac{1}{6}(7 - 5) = \frac{1}{3}$.

12.5 Regression analysis

1. $y = 4x - 3$ (line going through the points); $r = 0$ (residuals are zero)

2. The regression line is (use Statext) $y = 1.5x$. The residuals -0.5, 1, -0.5 ($= \cdots y - yp$, where yp is prediction - value on the line). Sum of squares is $0.25 + 1 + 0.25 = 1.5$.

3. It will be $0.21 \cdot 5000 - 100 = 950$.

4. It i the solution of he equation $1000 = 0.21x - 100$; $x = 4285.7$.

5. $\ln(y) = \ln(b_0) + b_1 x$

6. Quadratic: $y = b_0 + b_1 x + b_2 x^2$; cubic: $y = b_0 + b_1 x + b_2 x^2 + b_3 x^3$.

7. It is $y = 2.1 + 0.6 \cdot 2 + 1.2 \cdot 2^2 + 0.1 \cdot 2^3 = 8.9$.

12.6 Population and data sample

1. Very roughly: population is big and constant, sample is relatively small and it changes when taking a new one.
2. Random sample is a vector of equally distributed and independent random variables and sample realization is a vector on measured numbers.
3. It is random variable (average of realization of sample is a number).
4. It is a number.
5. Population are speeds of all cars in the world. Random sample are speeds of a given number of cars that will potentially be measured (it is a definition of experiment - measure speeds of n cars), sample realization are really measured speeds.
6. Population are numbers $\{1, 2, 3, 4, 5, 6\}$ with equal probabilities. Random sample is a vector of 10 entries - each entry is prepared for a value obtained on the thrown dice. Sample realization is a vector with numbers that really were obtained on the dice.
7. Expectation is not mentioned so it must be estimated.
8. Parameter is an unknown constant, estimate is a number computed from measured data that is close to the true value of the parameter.
9. Statistics is a function of random sample which when a sample realization is inserted gives a value near the estimated value of the parameter.
10. They are sample average and sample variance.
11. $E\left[\bar{X}\right] = \mu$ and $D\left[\bar{X}\right] = \frac{\sigma^2}{n}$
12. It says that if we take very many sample realizations, from each we compute sample average and then we take average from these averages we get practically precise value of the population expectation.
13. It expresses the fact that the larger the sample is the higher is the precision of the sample average as an estimate of the expectation.
14. It is unbiased, consistent and the larger is the sample length the higher is its efficiency.
15. The one which is computed from larger sample.

12.7 Statistical inference

1. It is the sample average.
2. It is the sample variance.
3. It is the sample proportion.
4. It is the correlation coefficient.
5. It is normal with expectation μ and variance $\frac{\sigma^2}{n}$.
6. For computation of confidence interval we must use distribution of the statistics $f(\bar{x})$.
7. For both-sided interval we use $\frac{\alpha}{2}$ on both sides; for left-sided respectively right-sided interval we use α on the left respectively right side.
8. Zero H0 is currently valid, alternative HA rejects H0.
9. Region of acceptance is equal to the confidence interval, critical region is its complement.
10. T_t is the value of the statistics with the inserted sample realization.
11. HA for both-sided test says "is not equal to"; for left sided it says "is less than" and for right-sided "is greater than".
12. Confidence interval is equal to the region of acceptance.
13. We reject H0.
14. We reject H0.

12.8 Validation in regression analysis

1. The closer the data points are to the regression curve, the better is the regression.
2. It tests if x and y are uncorrelated. If yes, the regression is not possible.
3. Yes, it is. (x and y are not uncorrelated)
4. Yes, it is.
5. Yes, it is.
6. Yes, they should.

References

[1] Morris H. DeGroot: Probability and Statistics. Addison-Wesley Publising Company, Reading, Massachusetts, 1984.

[2] Athanasios Papoulis, S. U. Pillai: Probability, Random Variables and Stochastic Processes. Mc Graw Hill, 2002.

[3] T. T. Soong: Fundamentals of Probability and Statistics for Engineers. John Wiley & Sons, Ltd, 2004