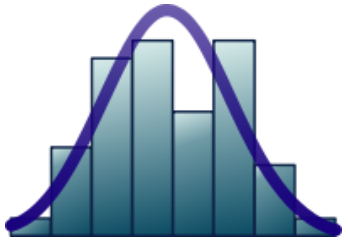
Methods of statistical analysis

Statistical analysis

Statistical analysis is basing on wide range of different techniques and methods for study of statistical distributions and their parameters (methods of mathematical statistics)



Mathematical statistics — science, which is developing mathematical methods for systematization and usage of statistical data for scientifical and practical conclusions.

Mathematical statistics — (also) a field of mathematics, developing methods for registration, description and analysis of the observation data, data from experiments to obtain in result a probability models of a massive random phenomena.

In many fields mathematical statistics uses the theory of probability, which allow us to estimate reliability and accuracy of the conclusions from limited statistical material (e.g. to estimate necessary sample size to obtain the results with required accuracy).

Frequentist probability

Frequentist probability - is the standard interpretation of probability, it determines the probability of an event as the limit of its relative frequency in a large number of studies.

Random experiment (test) is the implementation of a set of conditions that can be practically or mentally reproduce an arbitrarily large number of times.

Several examples of a random experiment: *flipping a coin or dice (dice), extracting one card from a shuffled deck.*

Probabilities

Phenomena that occur as a result of random experiment, called elementary outcomes (events). It is believed that during the random experiment implemented only one of the elementary events.

If you toss a coin once, the outcome can be considered as Heads (**H**) and Tails (**T**).
If the random experiment considered tossing a coin three times, then the elementary outcomes are the following:

HHH, HHT, HTH, THH, HTT, THT, TTH, TTT.

Probabilities II

The set of all elementary outcomes random experiment is called the space of elementary events. Lets denote space of elementary events with letter **Ω**, i-th elementary outcome will be **ω_i**.

If the space of elementary events contains n elementary outcomes, then

$$\Omega = (\omega_1, \omega_2, ..., \omega_n)$$

For example, the experiment consists of tossing a coin three times. The sample space is

Ω =(HHH, HHT, ..., TTT).

An event could be for example "2 times out of 3, a head was tossed", and corresponds to 3 sample points. The probability of such an event is therefore 3/8

If a random experiment - throwing the dice, then **Ω =(1,2,3,4,5,6).**

Probabilities: definitions

Elementary event or just an event is a subset of the space of elementary events. However, this event may be set of several elementary events.

In probability theory the weight associated with each event is normalized so that the total weight of the entire space of outcomes were 1:

$$\sum_i W(x_i) = 1; \int w(x)dx = 1$$

These weights – the "intuitive" likelihood that a particular event will occur.

If the event is not elementary outcome, but a collection of these, then the probability of an event is the sum of all the weights associated with elementary outcomes belonging to the event.

The overall probability is also normalized to **1**:

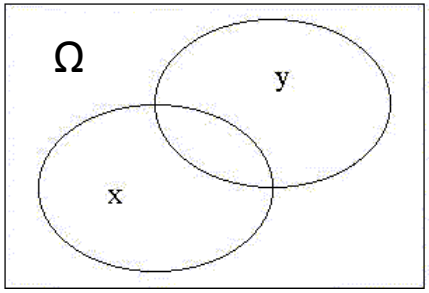
$$\sum_j P(x_j) = 1; \int p(x)dx = 1$$

Probabilities: definitions II

Two events are mutually exclusive if their overlap is zero.

Two events are complementary to each other if their union gives the full space of elementary events, and their overlap - zero.

Events conveniently represented in a pattern called a Venn diagram.
The right figure the space of elementary events **Ω** is depicted as a rectangle, and a set of elementary outcomes favorable events **X** and **Y**, are enclosed in ellipses.



Outcomes themselves on the Venn diagram are not shown, and the information about the relationship between their sets contained in the borders of the location areas.
We see that the events are overlapping area, then they are neither mutually exclusive nor complementary.

Probabilities: Exercise

Lottery. Need to guess the 5 digits of 36. This is the biggest win, 4-medium, 3-small. Find the probability of the large, medium and small winnings.

Number of combinations of n by k is the binomial coefficient:

$$\binom{n}{k} = C_n^k = \frac{n!}{k!(n-k)!}$$

Probabilities: Exercise II

Lotto game contains 50 kegs. We choose 5 kegs. What event is more likely:

а) all kegs will be from one dozen

б) each keg will be from different dozen

Number of combinations of n by k is the binomial coefficient:

$$\binom{n}{k} = C_n^k = \frac{n!}{k!(n-k)!}$$

Probabilities: Probability density function

Probability density function (PDF) f(x), or density of the continuous random variable - is a function that describes the relative likelihood for this random variable to take a predetermined value.

$$P(x \in [x, x+dx]) = f(x)dx$$

$$P(x \in [a,b]) = \int_a^b f(x)dx$$

Function f(x) itself is not a probability.
It is always normalized:

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

Cumulative density function (CDF) – distribution function:

$$F(x) = \int_{-\infty}^{x} f(x')dx'$$

Expectation value:
(mean of the random value)

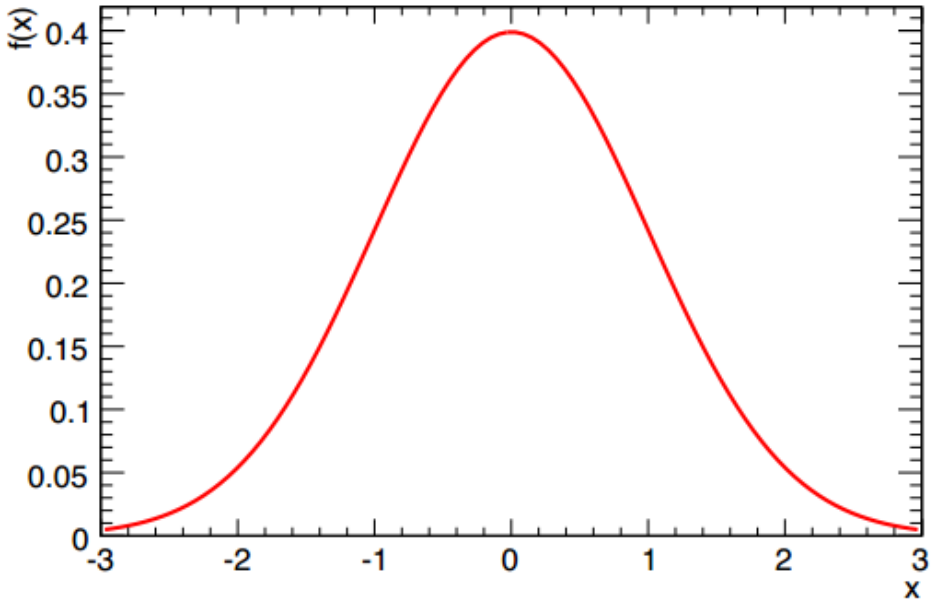
$$M[X] = \mu = \int_{-\infty}^{+\infty} x \cdot f(x)dx$$

The variance:
(measure of spread of the random variable, i.e. its deviation from expectation)

$$D[X] = M[X^2] - (M[X])^2$$

Probabilities: Probability density function II

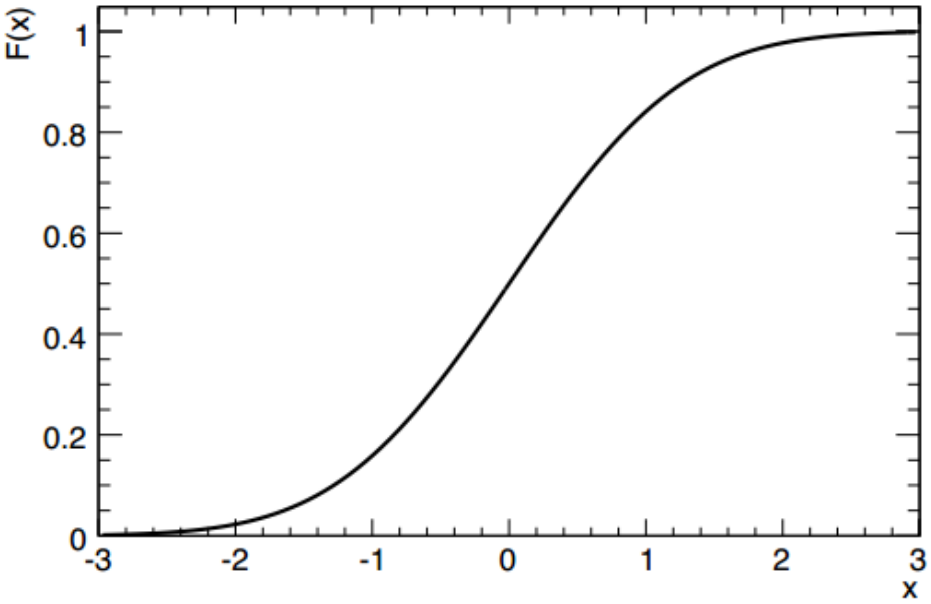
PDF



Alternatively, we define as the partial density of the cumulative:

$$f(x) = \frac{\partial F(x)}{\partial x}$$

CDF



The same expression, but for the connection of total and differential cross sections :

$$f(E) = \frac{1}{\sigma}\frac{\partial \sigma}{\partial E}$$

Probabilities: Probability density function III

Nuisance parameter – parameter of a probability density, but not a measurement

Example:

1) Poisson distribution $$P(\{N\}) = \frac{e^{-\mu}\mu^{N}}{N!}$$

discrete

we use it for event counting

μ – (nuisance) parameter of the distribution

2) Gauss distribution

continuous

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

μ, σ – (nuisance) parameters of the distribution

We use it for systematic errors

Probability density function: Exercise

Random variable X is given by the distribution function

$$F(x) = \begin{cases} 0, & x \leq 1 \\ \frac{1}{2}\left(x^2 - x\right), & 1 < x \leq 2 \\ 1, & x > 2 \end{cases}$$

Find the probability density function, mean and variance of the random variable X. Plot distribution function and density.

Distribution of several values

Joint probability distribution function of two variables in the general case can be written as f (x, y). Its normalization condition

$$\int\int f(x,y)dxdy = 1$$

Limiting distribution of x - the projection of this function to the variable x:

$$g(x) = \int f(x,y)dy$$

The conditional distribution of x for a given y0 – it is a cut for a given y0:

$$f(x \mid y_0)\frac{f(x,y_0)}{\int f(x,y_0)dx}$$

Two variables are independent if their joint probability distribution function can be expressed as follows:

$$p(x,y) = f(x)\cdot g(y)$$

Distribution of several values II

In this case (case of independent variables) boundary and the conditional distributions are equal.
The expectation of any function g of x and y:

$$E[g(x,y)] = \int\int g(x,y)\, f(x,y) dxdy$$

In particular the mean value and variance, for instance, the variable x are equal to:

$$\bar{x} = \mu_x = E(x) = \int x \int f(x,y) dydx$$

$$\sigma_{x^2} = E(x-\bar{x})^2 = \int x^2 \int f(x,y) dydx - \bar{x}^2$$

Covariance

Covariance (correlation moment, covariance moment) - a measure of the linear dependence of two random variables.
Covariance of two variables x and y is defined as follows:

$$\operatorname{cov}(x, y) = E[(x - \bar{x})(y - \bar{y})] = E(xy) - E(x) \cdot \bar{y} - E(y) \cdot \bar{x} + \bar{x} \cdot \bar{y} =$$
$$= E(xy) - E(x) \cdot E(y) = \operatorname{cov}(y, x)$$

Thus, it is obvious, that:

$$\operatorname{cov}(x, x) = \sigma_x^2, \ \operatorname{cov}(y, y) = \sigma_y^2$$

Covariance matrix - a generalization of covariance for vectors of random variables.
By definition it is:

$$V = \begin{pmatrix} V_{xx} & V_{xy} \\ V_{yx} & V_{yy} \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \operatorname{cov}(x, y) \\ \operatorname{cov}(x, y) & \sigma_y^2 \end{pmatrix}$$

Correlation

Correlation - a statistical relationship between two or more random variables. Correlation refers to any of a broad class of statistical relationships involving dependence. Changes of values for one or more of quantities are accompanied by systematic variation of value of other variable/variables. Mathematical measure of correlation between two random variables is the coefficient of correlation (correlation factor).

Correlation factor - a dimensionless quantity:

$$\rho_{xy} = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y} \qquad (1)$$

Main property: $|\rho_{xy}| \leq 1$

Lets prove: the variance of any quantity is always positive, for example, a linear combination ax + y:

$$V(ax+y) = E([ax+y-(a\bar{x}+\bar{y})]^2) = E([ax+y]^2) - (a\bar{x}+\bar{y})^2 =$$
$$= a^2\overline{x^2} + \overline{y^2} + 2a\overline{xy} - a^2(\bar{x})^2 - (\bar{y})^2 - 2a\bar{x}\bar{y}$$

Correlation II

So we have a quadratic equation for a:

$$0 \le a^2\sigma_x^2 + \sigma_y^2 + 2a\,\mathrm{cov}(x,y)$$

Solving this equation, it must have a maximum of one root, thus its discriminant is negative:

$$D = 4\,\mathrm{cov}(x,y)^2 - 4\sigma_x^2\sigma_y^2 \le 0$$

$$\mathrm{cov}(x,y)^2 - \sigma_x^2\sigma_y^2 \le 0$$

$$|\,\mathrm{cov}(x,y)\,| \le |\,\sigma_x\sigma_y\,| \Rightarrow |\,\rho_{xy}\,| \le 1$$

proved

Quadratic equation V(ax + y) has a root when ax + y - a constant, i.e. when there is a linear combination of x and y, in which case ρ_xy = ± 1.
Covariance (and hence, the correlation factor) of two independent variables equal to 0:

$$\mathrm{cov}(x,y) = \int\int (x-\bar{x})(y-\bar{y})\,f(x)g(y)dxdy =$$

$$= (\int (x-\bar{x})\,f(x)dx)\cdot(\int (y-\bar{y})\,f(y)dy) =$$

$$= (\bar{x}-\bar{x})(\bar{y}-\bar{y}) = 0$$

Correlation III

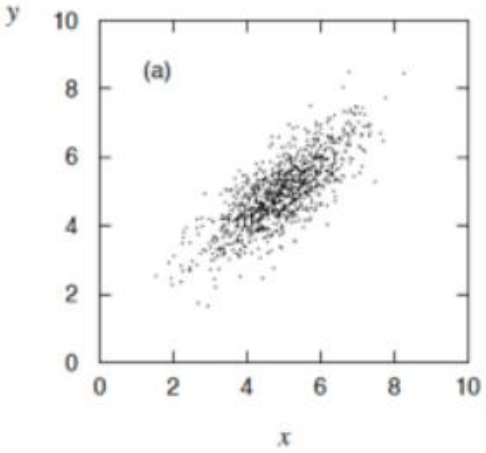
Consequence of zero correlation of independent variables is also the fact that the variance of the sum of two (or more) independent variables is simply the sum of the variances of these values:

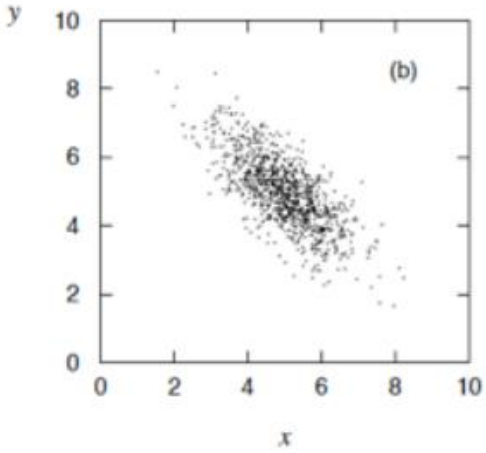
$$V(x+y) = \sigma_x^2 + \sigma_y^2 + 2\,\mathrm{cov}(x,y) = \sigma_x^2 + \sigma_y^2$$

Reverse - not equivalent! If the correlation of two variables is equal to zero, they are not necessarily independent

Correlation: Examples

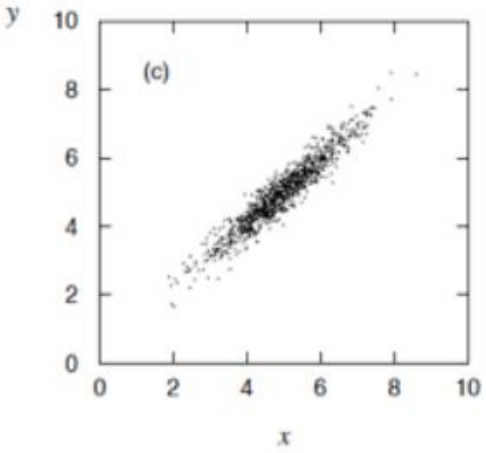
$\rho = 0.75$

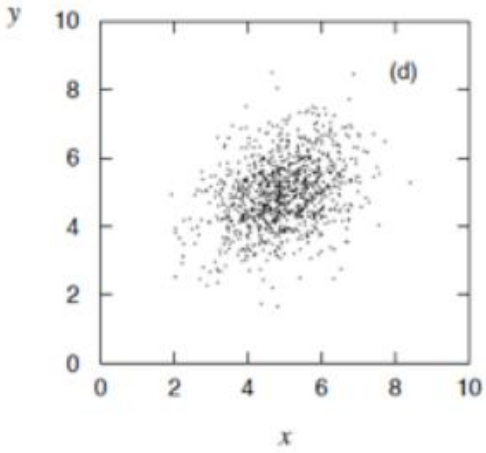




$\rho = -0.75$

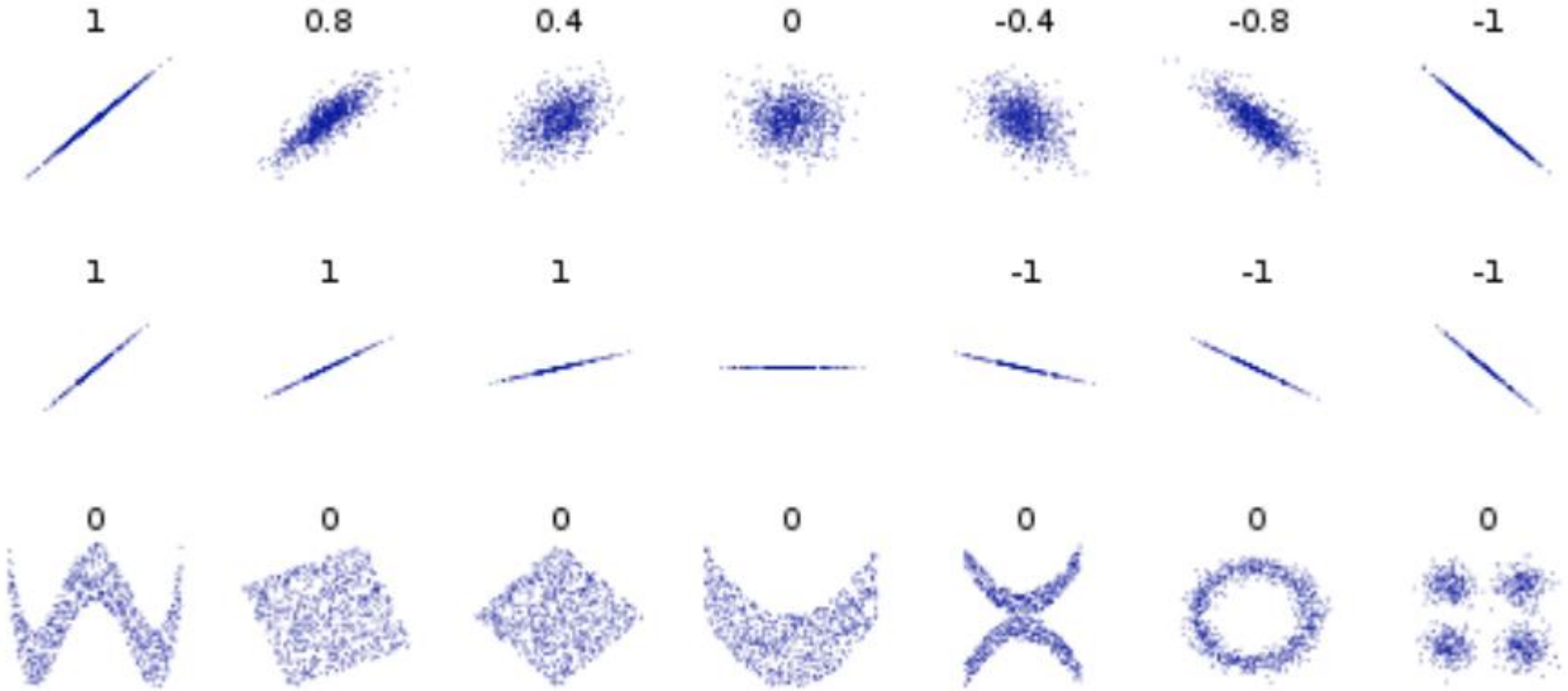
$\rho = 0.95$





$\rho = 0.25$

Correlation: Examples II



Correlation: Exercise

Plot a two-dimensional histogram of the variables x and y, a regression line construct (using LSM) and calculate for them the correlation of (Formula 1) in cases where :

1) x, y – two pseudorandom variables
2) x – pseudorandom variable, y=-5*x+0.6
3) x, z – pseudorandom variables, y=z*x

Correlation: Exercise II

Question: How will change the correlation coefficient if we reduce /increase all variables by the same number?

Correlation: Addition

Because the evaluation of the correlation coefficient is calculated on the final sample, and therefore may deviate from its general values, you must check the significance of the correlation coefficient. Verification is performed using the t-test:

$$t = \frac{R_{x,y}\sqrt{n-2}}{\sqrt{1-R_{x,y}^2}}$$

Random variable **t** follows the t-distribution and the t-distribution table must find the critical value of the criterion ($t_{cr.\alpha}$) for a given level of significance **α**. If calculated by the above formula t modulo will be less than $t_{cr.\alpha}$, the dependence between random variables X and Y is not. Otherwise, the experimental data do not contradict the hypothesis about the dependence of random variables.

α or p-level - a measure inversely proportional to the reliability of results. For example, p-level = .05 (т.е. 1/20) shows, that there is a 5% probability that the sample found in the relationship between the variables is the only feature of this random sample.

Many studies p-value of .05 is considered an acceptable level boundary errors.

Correlation: Addition II

t- distribution

$$f_t(y) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{y^2}{n}\right)^{-\frac{n+1}{2}}$$

Number of degrees of freedom (n - 2)	α = 0.1	α = 0.05	α = 0.02	α = 0.01	α = 0.002	α = 0.001
1	6.314	12.706	31.821	63.657	318.31	636.62
2	2.920	4.303	6.965	9.925	22.327	31.598
3	2.353	3.182	4.541	5.841	10.214	12.924
4	2.132	2.776	3.747	4.604	7.173	8.610
5	2.015	2.571	3.365	4.032	5.893	6.869
6	1.943	2.447	3.143	3.707	5.208	5.959
7	1.895	2.365	2.998	3.499	4.785	5.408
8	1.860	2.306	2.896	3.355	4.501	5.041
9	1.833	2.262	2.821	3.250	4.297	4.781
10	1.812	2.228	2.764	3.169	4.144	4.587
…	…	…	…	…	…	…
∞	1.645	1.960	2.326	2.576	3.090	3.291

Least square method

If some physical quantity depends on another value, then this dependence can be investigated by measuring y for different values of x. The measurements obtained by a number of values:

$$x_1, x_2, ..., x_i, \ ... \ , x_n;$$

$$y_1, y_2, ..., y_i, \ ... \ , y_n.$$

According to this experiment can be plotted as y = ƒ (x). The resulting curve gives you the opportunity to judge a function ƒ (x).
However, the constant coefficients, which are included in this function remain unknown. Allows to determine their method of least squares.
Experimental points usually do not lie exactly on the curve.

Least squares method requires that the sum of squared deviations of the experimental points from the curve, i.e. $\mathbf{[y_i - ƒ(x_i)]^2}$ to be minimal.
In practice, this method is most frequently (and more simply) used in the case of the linear dependency, i.e. when:

$$\mathbf{y = kx} \quad \text{или} \quad \mathbf{y = a + bx.}$$

Least square method II

Lets consider dependence **y = ax + b** .

The challenge is that by having a set of values xi, yi find the best values of a and b.

Form the quadratic form ϕ, equal to the sum of squares of deviations xi, yi from the line:

$$\phi = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

and we can find values a and b , when ϕ is minimal

$$\frac{\partial \phi}{\partial b} = -2\sum (y_i - ax_i - b) = 0; \quad \frac{\partial \phi}{\partial a} = -2\sum x_i (y_i - ax_i - b) = 0$$

Simultaneous solution of these equations gives:

$$\begin{cases} a = \dfrac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ b = \dfrac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n} \end{cases}$$

Least square method: Exercise

Question. Experimental data on the values of the variables x and y are shown in the table.

	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
x_i	0	1	2	4	5
y_i	2,1	2,4	2,6	2,8	3,0

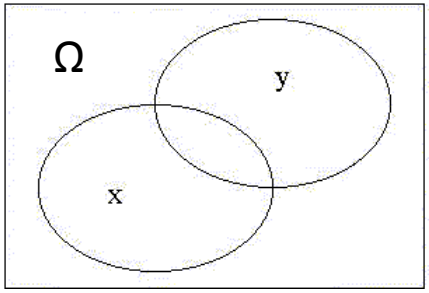
Using the least squares method, you have to approximate the data by a linear dependence y = ax + b (find parameters a and b).

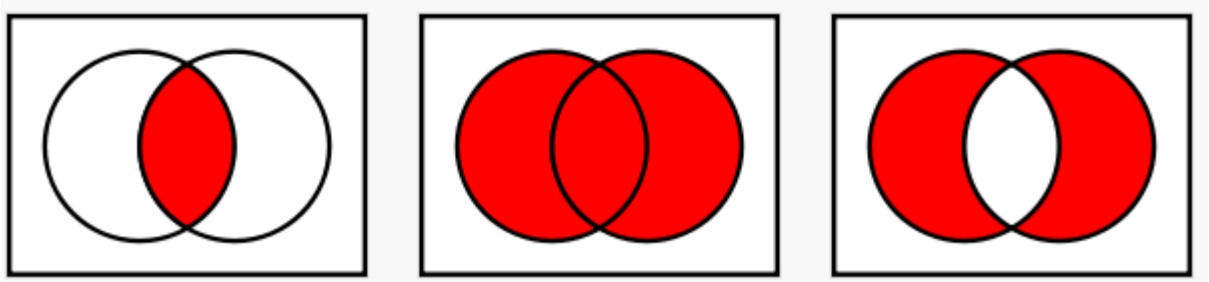
Probabilities: defenitions - repetition

Two events are mutually exclusive if their overlap is zero. **A∩B=0**

Two events are complementary to each other if their union gives the full space of elementary events, and their overlap - zero. **AUB=Ω, A∩B=0**

Events conveniently represented in a pattern called a Venn diagram.
The right figure the space of elementary events **Ω** is depicted as a rectangle, and a set of elementary outcomes favorable events **X** and **Y**, are enclosed in ellipses.





A∩B **AUB** **AΔB**

Main theorems: Conditional probability

A, B – events from space Ω

$$P(A \Delta B)=P(A \text{ или } B)=P(A)+P(B)-P(A \text{ и } B)$$
$$P(A \text{ и } B)=P(A \cap B)=P(AB)$$

Probability $P(A|B)$ – conditional probability of the event A if the event B satisfied:

$$P(AB)=P(A|B)\cdot P(B)=P(B|A)\cdot P(A)$$

From this line we can write down the Bayes theorem (conditional probability):

$$P(A|B)=P(B|A)\cdot P(A)/P(B)=P(AB)/P(B)$$

if A_1,A_2,..., A_n - comprehensive and exceptional sets (that means that any event belongs to one and only one A_i), the marginal probability of event B can be overwritten as:

$$P(B)=\sum_j P(A_j B)=\sum_j P(B|A_j)\cdot P(A_j)$$

Main theorems: Conditional probability II

Thus, Bayes theorem could be overwritten as:

$$P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{\sum_j P(B \mid A_j) \cdot P(A_j)}$$

Case of independent events. Events A and B - independent if P (A | B) = P (A), ie P (A) is not affected by an event B is executed or not.
If we have such case, then:

$$P(AB)=P(A)P(B)$$

и

$$P(B|A)=P(B)$$

Case of mutually exclusive (incompatible) events. Events A and B - mutually exclusive if their overlap is zero. **A∩B=0**
If we have such case, then:

$$P(A|B)=P(A\cap B)/P(B)=0$$

Conditional probability: Exercise

Question. Urn contain 3 white and 3 black balls. Twice removed from the urn one ball without returning them back. Find the probability of a white ball in the second test (event B), if the first test was extracted black ball (event A).
Solution. After the first test in an urn in 5 balls, 3 of them white. Seeking the conditional probability **P(B|A)=3/5**.

We can obtain the same result using Bayes formula:

$$P(B \mid A) = \frac{P(AB)}{P(A)}$$

The probability of a white ball in the first test is
P(A)=3/6=1/2.
We find the probability P (AB) that appears in the first Test black ball, and the second - white. Total number of outcomes - the co-occurrence of two balls, no matter what color, equal to the number of placements $C_6^2 = 6 \cdot 5 = 30$
From this number of outcomes favorable to the event AB 3 * 3 = 9 outcomes.
Thus, **P(AB)=9/30=3/10**.
So the conditional probability equals

$$P(B \mid A) = \frac{P(AB)}{P(A)} = \frac{3/10}{1/2} = \frac{3}{5}$$

Results are identical!

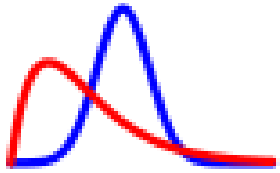
Conditional probability: Exercise II

Five students pull on the exam five tickets, one of which is very easy. What are the chances for someone who is third, get a this easy ticket?

Bayesian(conditional) probability

Bayesian probability — an interpretation of the concept of probability used in Bayesian theory. Probability is defined as the degree of confidence in the truth of a proposition. To determine the degree of confidence in the truth of a proposition if new information in the Bayesian theory is used Bayes' theorem.

Bayesian probability is opposed to the frequentist in which the probability is determined by the relative frequency of occurrence of a random event for sufficiently long observation.



Bayesian probability II

In general, Bayesian methods are characterized by the following concepts and procedures:

1) Usage a random variable to model all the sources of error in the statistical model. It includes not only the true sources of randomness , but also uncertainty due to lack of information.

2) Consistent application of Bayes' formula : when more data arrives after the calculation of the posterior distribution , the latter becomes the next primary.

3) Frequentist probability hypothesis is a statement (which must be either true or false) , so that the frequency is equal to the probability of the hypothesis of either 1 or 0 . In Bayesian statistics, the hypothesis can be assigned a probability that is other than 0 or 1 , if the true value is not determined.

Frequentist/Bayesian probability

Frequentist point of view: probabilities describe the outcomes of experiments. Models are unknown parameters. The probabilities are given as a function of the model parameters.

Bayesian extension: probabilities are also used to describe the "degrees of confidence" in the model parameters → Nuisanse parameters itself can have probabilities assigned to them

Definitions for the Bayesian theory

And again:

$$P(B \mid A) = \frac{P(A/B)P(B)}{P(A)}$$

Each part of the formula has its own name:

- Prior: P(B), where B – theory (usually, parameters)
- Likelihood: P(A|B), where A – measurement
- Posterior: P(B|A) – analysis result

- P(A) – no special name

Normalization is often done by integrating a posterior on all theories :
P(B|A)+P(~B|A)=1

Definitions for the Bayesian theory: Exercise

Assume that the mixed school has 60% of boys and 40% of girls among the students. Girls wear pants or skirts, all the boys wear pants.
The observer sees from afar (random) student: everything he sees - a student in the pants.
What is the probability that the student - a girl?

Definitions for the Bayesian theory: Exercise II

(Poincaré problem) In gambling club half of the players are honest, half - sharpers. Probability of drawing a king from the deck is 1/8. Sharper probability of drawing king from the deck is 1. Man sitting in front of you pulls a player from the deck of the king the first time. What is the probability that you are playing with sharper?

Probabilities in high energy physics

- Probability: predicts the number of events for this theory and the experimental setup of the

- But we want to know what observation can say about theory

- Frequentist approach gives observation probability to each theory (there is no probability for theory)

- Bayesian approach: assigns probability (degree of confidence) for theories

- High-energy physics: can use both approaches (although there is some preference to the frequentist approach, in particular for discovery)

Confidence levels&intervals

Confidence Interval - the term used in mathematical statistics for interval (as opposed to point) evaluating statistical parameters, preferably with a small volume of sample.
Confidence interval is called, which covers the unknown parameter with a given reliability.

Confidence Interval tells about theory parameters
Confidence Level: associated probability

The difference in the confidence levels in the Frequentist / Bayesian interpretation

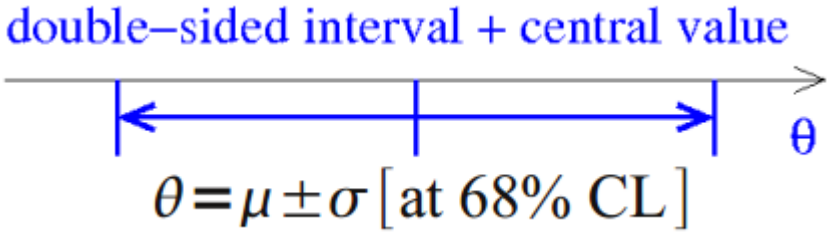
Frequentist **CL~P(набл|θ)**

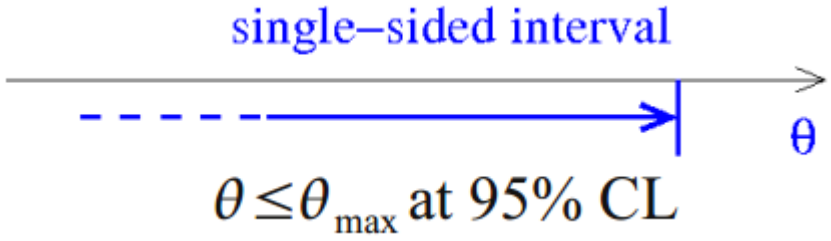
$$\mathbb{P}(L \leqslant \theta \leqslant U) = p$$

Bayesian: **CL~(θ |набл)**

$$\mathbb{P}(L \leqslant \theta \leqslant U | X) = p$$

Double-sided – measurement of 2 borders **(CL=68%)**
Single-sided measurement - limit **(CL=95%)**





Useful methods: Bootstrap

Bootstrap — practical computer method for determining the statistics of probability distributions based on the generation of multiple samples Monte Carlo based on the available sample. Allows you to quickly and easily evaluate a variety of statistics (confidence intervals, variance, correlation, and so on) for complex models.

Bootstrap - practice of estimator properties (e.g. , its variance) by measuring these properties sampling of approximating distribution. One standard choice for the approximated distribution - the empirical distribution of the observed data . In the case where a set of observations can be assumed independently and identically distributed , it may be carried out by constructing a series of resampling observed data (same size and the observed data set), each of which is obtained by a random sampling with replacement of the original dataset.