С.Н. Лапач, А.В. Чубенко, П.Н. Бабич Статистические методы в медико-биологических исследованиях с использованием Excel –2 изд. перераб. и доп. –К.: 2001, Морион. – 408с.

На английском языке представлен фрагмент книги: вторая глава.

Chapter 2. Law and randomness

Statistic is a collection of methods that enable us to make decisions in conditions of uncertainty.

Abraham Wald.

Studying different disciplines as at school as well as in institutions of higher education a determinacy of events and functions is suggested when every event is a consequence of the other and physic law are strict mathematical laws that present the dependence of one values upon the others. Though, everyday human activity constantly refutes this statement. Thus, verifying any physical laws (even at the level of laboratory classes at school or institutions it is revealed that every new experimental value determination provides different results. There are also values that are even indeterminate, since it is impossible, for example, to determine the number of passengers in the subway carriage or trolley bas at a definite period of time. The quantities whose exact value for every realization is not known are called random variates. Still, randomness does not mean that it is not possible to obtain and utilize in practice some process data. Thus, in the sick rate analysis one can trace a changing tendency of the exponent for 1000 people and draw conclusion about the appropriateness of measures, planning of the required number of medical institutions, medical stuff, and drugs. Despite the random character of some quantities certain laws used in actual practice can be deduced when studying them.

Theoretical basis of statistical methods is a probability theory.

2.1. Basic Concepts of Probability Theory

The subject of probability theory is studying probabilistic laws of mass homogeneous random events.

It serves as a theoretical basis of other disciplines (for example, mathematical statistics, and reliability theory).

2.1.1. Main Definitions

The trial is a realization of complex of some conditions. The result of a trial is an event. Each of probable trial results is called an elementary outcome.

Those outcomes, in which the event of interest is realized are called favorable outcome.

Classical definition of probability

Probability of *A* event is a relation of the amount of favorable elementary outcomes to their total number:

P(A)=m/n *(2.1)*

Properties subsequent upon the definition

1. Probability of a certain event (that is the event which under the given set of conditions will always succeed) is equal to 1: *P(A)=m/n=n/n=1.*

2. Probability of an impossible event (which will not succeed under the given set of conditions) is equal to 0: *P(A)=m/n=0/n=0.*

3. Probability of a random event (which either can succeed or not) is a positive number between 0 and 1:$1 \le p(a) \le 1$.

For example, probability of the even number in dicing equals to $\frac{1}{2}(0,5)$: total number of probabilities is 6 (1, 2, 3, 4, 5, 6), favorable of them are three (2, 4, 6).

Types of Random Events

Exclusive events — the occurrence of one of then excludes the occurrence of other in the same trial.

Event makes up a full group if in the result of a trail at list one of them is certain.

Events are equally probable when there are grounds to believe that neither of them is more probable then the others.

Combinatorial formulas are widely used to calculate probability.

Elements of Combinatorics

Permutations are combinations of the same *n* elements of an element different only in the order of their arrangement:

$P_n=n!$ (2.2)

The number of code combinations of figures 1, 2, 3, if every figure can be used in the code only once 1*2*3=6

In the event that combinations are formed not from all *n* elements, but only from *m* selected ones the formula for calculating their number appears as:

$$P^m_n=n!/(n–m)!$$

For example, the number of possible combinations for a combination lock with 3 digits (10 figures) is: $P^3_{10}=10!/(10–3)! = 720$.

To calculate the number of permutations in Excel the function PERMUT (number; number_selected). For the previous example PERMUT (10; 3).

Arrangements are combinations formed of *n* different taken *m* at a time (in every combination), which differs either in the order or in the composition of elements:

$$A^m_n=n(n-1)...(n-m+1) \qquad (2.3)$$

The number of signals, which can be formed of six signal flags of different color taken at 2, is calculated 6*5=30.

Combinations are collections formed of *n* different elements taken *m* at a time that differ at list in one element.

$$C^m_n=n!/(m!(n-m)!) \qquad (2.4)$$

The number of methods the group of two persons can be formed when selecting them from 10 candidates is: $C^2_{10}=10! /(2!8!)=45$.

In Excel to calculate the number of combinations, the function NUMCOMB (*n; m*) is used. Bear in mind that it is in the group of mathematical but not statistical functions (in contrast to PERMUTE()). The relationship between the given combinatorial formulas is:

$$A^m_n = P_n C^m_n \qquad (2.5)$$

The expression *n!* frequently used in combinatorial formulas is a factorial. It is calculated as a product o f all values from 1 to *n*. That is, *n!=1*2*3*4*5*6=720*. In this case, the factorial 0 is equal to one. That is *0!=1*. in Excel, to calculate a factorial the function FACTOR (number) is used. It belongs to mathematical functions.

Rule of Sum. If an object A can be selected by *m* methods and object *B* by *n* methods then either A or *B* can be selected by *m+n* methods.

Rule of Product. If an object A can be selected by *m* methods and after every such selection an object *B* can be selected by *n* methods then the pair of objects in the specified order A, *B* can be selected by *m*n* methods.

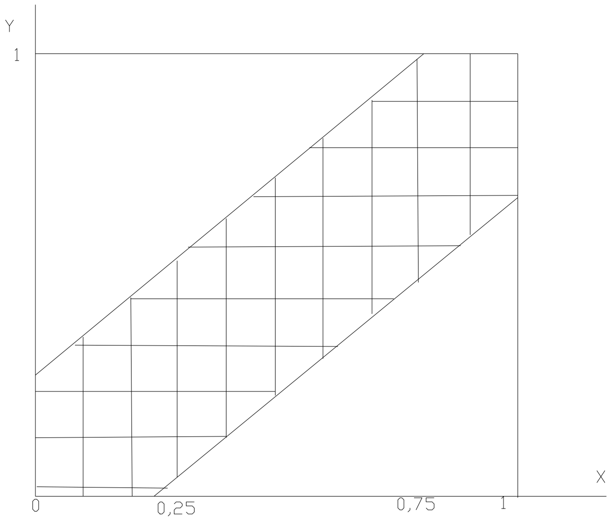
Statistical probability

The classical probability definition suggests the final number of elementary outcomes. In most cases it is impossible to present the outcome of a trial as a collection of elementary events. There are problems with indication of equal probability of elementary events. In such a case, theoretical definition is impossible. Since it is impossible to calculate the number of total and favorable possibilities. Thus, statistical probability definition is being introduced. According to this definition the probability is a relation of the number of trials (in which the event occurred) to the total number of conducted trials. This probability is also called *relative frequency.*

Geometrical probability is a probability of point entering in the area (segment, a part of plane), which is calculated as the relation of lengths and squares.

Usually to illustrate how the problem can be employed in calculating geometrical probability the following example is given. Two persons agreed to meet in the interval from 0 to 1 hour. In doing so, every of them comes at a random moment of time and waits another for fifteen minutes. What is the probability these two will meet? Let the time of arrival of one be *x* and another — y. They will meet if the condition |*x*-y| ≤ 0,25 h. is satisfied.

For the case *x* > y we will have the inequality *x*-0,25 ≥ y and for the case *x* < y – the inequality y ≥ *x*-0,25. Let's consider the situation in the geometrical representation (Fig. 2.1). All area of probable outcomes is represented by the rectangular limited by lines y = *x*-0, 25 and y = *x*+0,25. Thus, the probability of the meeting will be the relation of the cross-hatched square to the total rectangular square.



Pic. 2.1. Example of Geometrical probability

2.1.2. Sum and Product of Events

In this section we will consider main theorems of probability theory and conclusions from them without given proofs.

Concepts of Sum and Product of Events.

The sum *A*+*B* of two elements *A* and *B* is an event consisting of the occurrence of one of these events or two events together.

If *A*+*B* is exclusive, then *A*+*B* is an event consisting of the occurrence of one of the events.

Probability of the occurrence of one of exclusive events is equal to the sum of probabilities of these events:

$$P(A+B)=P(A)+P(B).$$

Consequence. The probability of the occurrence of one or several pairwise exclusive events is equal to the sum of probabilities of these events:

The sum of probabilities of the events, which make up exhaustive events, is equal to 1.

Complementary events are two singularly possible events, which make up exhaustive events.

Sum of probabilities of complementary events is equal to 1.

Product of the events A and *B* is the event *AB* that consists of joint occurrence of the events *A* and *B*.

Conditional Probability

Conditional probability $\mathrm{P}_\mathrm{A}(B)$ is a probability of the event *B* calculated provided the event A has already occurred.

$$\mathrm{P}_\mathrm{A}(B)=P(AB)/P(A) \qquad (2.6)$$

The probability of joint occurrence of two events is equal to the probability product of one of them into the conditional probability of another, calculated supposing the first event has already occurred:

$$P(AB)=P(A)\, P_A(B). \qquad (2.7)$$

Consequence. The probability of joint occurrence of several events is equal to the probability product of one of them into conditional probabilities of all others. The probability of every next event is calculated supposing all previous events have already occurred:

$$\mathrm{P}(\mathrm{A}_1\mathrm{A}_2\mathrm{A}_3.....\mathrm{A}_n)=\mathrm{P}(\mathrm{A}_1)\ \mathrm{P}_{\mathrm{A1}}(\mathrm{A}_2)\ \mathrm{P}_{\mathrm{A1A2}}(\mathrm{A}_3)\ ...\mathrm{P}_{\mathrm{A1A2A3.....An\text{-}1}}(\mathrm{A}_n) \qquad (2.8)$$

Events are called *independent* if the occurrence of one of them does not change the probability of the occurrence of another.

For independent events the probability of their product is

$$P(AB)=P(A)P(B) \qquad (2.9)$$

Several events are called independent in the aggregate if every two of them are independent and independent are every event and all possible products of others.

Probability of occurrence of at least one event

The probability of occurrence of at least one of the events A1, A2, A3,… An, independent in the aggregate, is equal to the difference between 1 and the product of probability of complementary events:

$$P(A) = 1 – q_1\ q_2\ q_3 ...\ q_n \qquad (2.10),$$

where $q_i = 1 – p_i$.

If the events have identical probability, then

$$P(A) = 1 – q^n \qquad (2.11)$$

Probability formula for the sum of exclusive events:

$$P(A+B)=P(A)+P(B)-P(AB) \qquad (2.12)$$

Composite probability formula

The probability of the event *A* which occurs supposing the occurrence of one of exclusive events B1, B2, B3,… Bn, that form exhaustive events, is equal to the sum of probability products of every of them into the respective conditional probability of the event A:

$$P(A) = P(B_1)P_{B1}(A) + P(B_2)P_{B2}(A) + ... + P(B_n)P_{Bn}(A) \qquad (2.13)$$

Bayes' formula (probabilities of hypotheses)

In case when the event *A* can occur as the result of the appearance of one of the exclusive events *Bi,* that form exhaustive events, it is possible to recalculate the conditional probability after the event A has already occurred:

$$P(B_k / A) = \frac{P(B_k)P(A/B_k)}{\sum_{i=1}^{n} P(B_i)P(A/B_i)} \qquad (2.14)$$

The events *Bi* are called hypotheses in this case. Thus, the probabilities of hypotheses are calculated that the event *A* has been the consequence of *Bi*.

2.2. Scales of measurements

> Measure all measurable and make all immeasurable measurable.
>
> *Galileo Galilei*

Statistical methods can be used to process only something measurable. In this connection we shall concern ourselves with the available scales of measurements. *Measurement* is an assignment of figures to subjects or events, based on a certain system of rules. The following conditions should be met for quantities representing the measurement results of the studied feature.

Identity

1. *Either A=B or A≠B*
2. *If A=B then B=A*

Transitivity

If A=B and B=C then A=C

Rank order

1. *If A>B then B<A*
2. *If A>B and B>C then A>C*

Additivity

1. *If A=B and C>0 then A+C>B*
2. *A+B=B+A*
3. *If A=B and C=D then A+C=B+D*
4. *(A+B)+C=A+(B+C)*

Depending on the possibility of fulfilling these conditions and also operations over the measured values (“equal”, ” not equal”, “more than”, “less than”, “addition”, “subtraction”, “

multiplication", and "division" there are the following *scales of measurements:*

- Size scale (of names);
- Order scale;
- Interval scale;
- Ratio scale.

Let's deal with the peculiarities of these scales.

Size scale (nominal)

No operations of comparison excepting "equal" and "not equal" are possible. Numeration and naming are used only to identify an object — house number, sportsman's T-shirt number, number of method of treatment and so on.

Order scale

Comparing objects by quantity "more than", "less than" is possible. Other operations are not possible. The example can be mineral hardness scale containing calibration minerals arranged in a row where every next mineral is harder than the previous one. In medicine, the example of comparing objects can be: severity of illness, "good", "satisfactory", "bad" condition, stage course of disease, etc. Only operations of comparison such as "more than", "less than", "equal" are possible. The values specified by different experts may not agree (scale shift).

Interval scale

In this scale not only comparison by quantity is possible but also specification "how much more" (i.e. operations of "addition" and "subtraction" are possible). The examples can be temperature scales (Celsius, Kelvin's, Fahrenheit's, and Reaumur).

Ratio scale

In this scale it is possible to find out "in how many times" (all operations are admissible "comparison", "addition", "subtraction", "multiplication" and "division"). Example – weight, length, etc. In such cases, there is a natural reference point.

In the process of developing of science and measuring instruments transition from one scale of measurement to another, more sophisticated becomes feasible. First thermometers, for example, measured temperature in the order scale ("moderate", "warm", "hot", etc.).

Sometimes they say about *discrete* and *continuous* scales of measurements. In the general case, *size scale* and *order scale* belong to discrete scales. In these scales there are no intermediate values, therefore they are often referred to as *no quantitative*.

It stands to reason that scale of measurement imposes limitations on statistical characteristics that can be calculated for a random variate measured in the specific scale and on processing methods that can be correctly applied to these characteristics. (Tables 2.1, 2.2). Generally, to process data measured in discrete scales, *non- parametrical methods are used.* Depending on the type of measuring scales of variables different statistical methods are used to study connections between them (Fig. 2.3)

Table 2.1

Possible operations in different scales of measurements

Name of scale	**Type of scale**	**Possible operations**
Size	Discrete	$= \neq$
Order	Discrete	$= \neq\ > <$
Interval	Continuous	$= \neq\ > < + -$
Ratio	Continuous	$= \neq\ > < + - / x$

Table 2.2

Statistical characteristics that can be calculated

Name of scale	**Statistical characteristics that can be calculated**
Size	Frequencies, modal class
Order[1]	Frequencies, mode, median, centiles, rank correlation
Interval	Frequencies, mode, median, centiles, rank correlation, average,

	variance
Ratio	All available

[1] In some works for such scales median quadratic deviation is used. It is calculated in the same manner as variance but median is used instead of average

Table 2.3[2]

Connection between scales of measurements and applied methods

Scale of measurement of influencing variables	**Scale of measurement of dependent variable**	**Applied methods**
Intervals or relations	Intervals or relations	Regression and correlation analysis
Time	Intervals or relations	Time series analysis
Names or order	Intervals or relations	Variance analysis [3]
Mixed	Intervals or relations	Covariance and regression analysis
Names or order	Names or order	Rank correlation analysis and analysis of contingency tables [4]
Names or Order	Intervals or relations	Discriminant analysis; cluster analysis; taxonomy

[2] taken from [1] with insignificant changes.

[3] if the number of factors is more than two regression analysis is preferable

[4] for multicellular tables regression analysis can be used

2.3 Random variates

Everything, that seems a random coincidence to you, is not such at all.

Stanislaw Lem “Sobyscha”

2.3.1. Common notions

To inquire into laws that manifest themselves through randomness, distribution laws of random variates and their numeric characteristics are put to studying.

Relative frequency is a concept of sick rate since it makes up the total number of patients for 1000, 10000 or 100000 of population. It should be borne in mind that relative frequency never tallies precisely with theoretical probability. For example, the relative probability of the “head” when tossing a coin is 0,5 (1/2). If tossing the coin you calculate the relative frequency it will become evident that it is not consistent with the theoretical probability. But according to Bernoulli theorem, given quite a big number of trials, the probability that the departure of relative frequency from theoretical probability will be as large as is wished, tends to 1. Hence it follows that relative frequency possesses the property of stability.

Random is a variate, which as the result of an experiment may assume previously unknown value. *Discrete random* is a variate that assumes individual values (for example, the number of the newly born).

Independent random variates. These are variates, which are the result of the independent random events. That is, of the events for which the occurrence of one event influences in no way the probability of the occurrence of the other.

2.3.2. Random Variate Distribution laws

Distribution law is a correspondence between the values of a random variate and probabilities of their realization. This variate can be specified as a table, formula or plot. For a discrete random variate it is usually specified as a distribution series (Table 2.4) or graphically as a distribution polygon (Table 2. 2). Let’s take the concrete example (Table 2.5).

Table 2.4

X	x_1	x_2	…	x_n
P	p_1	p_2	…	p_n

Attention. The sum in the line *P* should be one. Let’s construct the plot for this table (see Fig. 2.3).

Table 2.5

X	1	3	4	8
P	0,15	0,35	0,4	0,1

For continuous random variate, tabular specification is impossible. Thus, distribution functions are used for this purpose.

Distribution function is a function Fx which specifies the probability that the random variate *X* will assume the value less than *x* in a trial.

$$F(x)=P(X < x) \qquad (2.15)$$

Sometimes it is also called integral function.

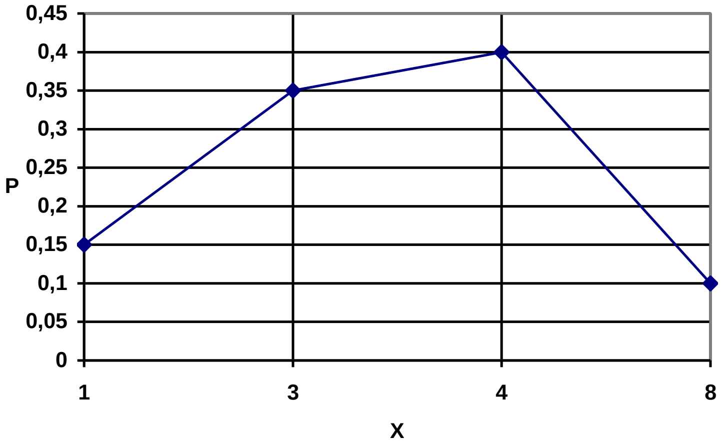


Рис. 2.2. Многоугольник распределения

Distribution function *F(x)* is a no decreasing function. That is, if a > *b* then *F(a)* ≥ *F(b).* Given *F(-∞)=0* and *F(+∞)=1.* For discrete random variate the distribution function is a discontinuous function (Fig. 2.3 distribution function for the distribution described in Table 2.5).

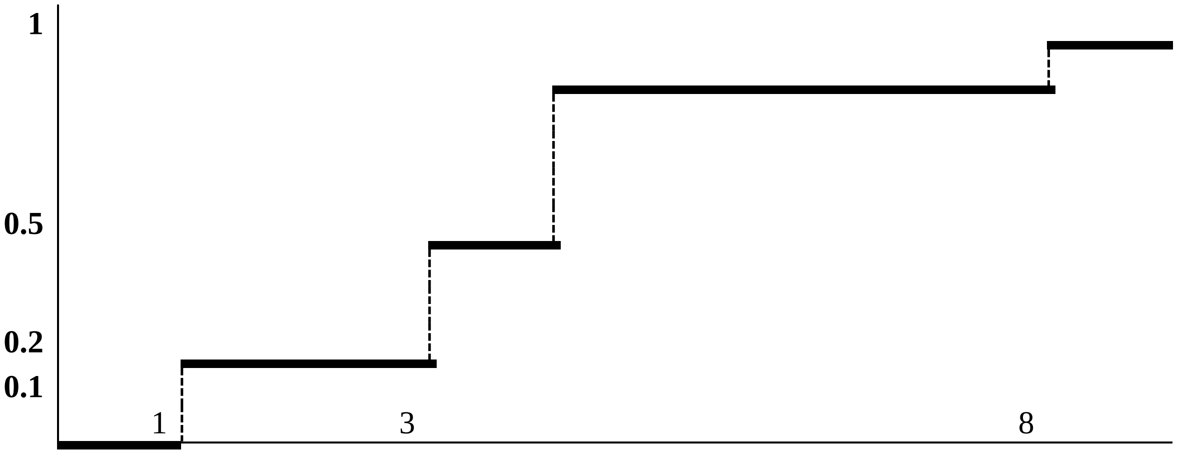


Fig. 2.3 Discrete random variate distribution function

In this case the probability the random variate will be included in the interval is defined from the formula.

$$\mathbf{P(a<X<b) = F(a) - F(b)} \qquad (2.16)$$

Distribution density is

$$f(x)=F`(x) \qquad (2.17)$$

Given $\int\limits_{-\infty}^{\infty} f(x)dx = 1$.

In this case the probability of being included into the interval (a, *b*) is defined by formula:

$$P(a < X < b) = \int\limits_{a}^{b} f(x)dx \qquad (2.18)$$

Expectation

For discrete random variate it is defined by formula:

$$M(X) = \sum_{i=1}^{\infty} x_i p_i , \qquad (2.19)$$

Where x_i is a random variate value, p_i is a probability of the occurrence of these values.

For the example presented in Table 2.5, expectation is calculated in the following manner: *M(X)=1*0,15+3*0,35+4*0,4+8*0,1=3,6*

For continuous random variates the formula (2.19) appears a s follows:

$$M(x) = \int xf(x)dx \qquad (2.20)$$

Probabilistic Meaning of Expectation

Let *n* trials have been conducted in which the random variate *X* assumed the following values (Table 2.6).

Table 2.6

Value *X*	x_1	x_2	…	x_k
Number	m_1	m_2	…	m_k

The total number of trials is m$_1$+m$_2$+…+ m$_k$=*n*. to calculate the average value taken by the random variate in the process of trials we will have $\overline{X} = \frac{x_1 m_1 + x_2 m_2 + ... + x_k m_k}{n}$

Herefrom: $\overline{X} = \frac{x_1 m_1}{n} + \frac{x_2 m_2}{n} + ... + \frac{x_k m_k}{n}$

Relations m$_i$/*n* represent relative frequencies of the occurrence of the value x$_i$. When approaching the number of trial to infinity, frequencies are approaching to the probability of events (analyzed further).

Then the formula assumes the form: $\overline{X} \approx x_1 p_1 + x_2 p_2 + ... + x_k p_k = M(X)$

Hence it follows that given a big number of trials expectation is approximately equal to the average arithmetic value of the random variate.

Expectation is also called *distribution center*. If we fancy that the random variate values are coordinates of the location of masses (for example, on the column) and probabilities are proper values of masses then coordinates of the center of masses are calculated from the formula:

$x_c = \sum_{i=1}^{k} x_i p_i \Big/ \sum_{i=1}^{k} p_i$. Поскольку $\sum_{i=1}^{k} x_i p_i = M(X)$, а $\sum_{i=1}^{k} p_i = 1$, то x$_c$ = M(X).

Let's have a look at Fig 2. 4 where the example from the table 2.5 is given. Assume that it is column (horizontal line), in which weights are distributed (vertical line). The center of gravity will be in the point with the value 3,6 that corresponds to the expectation value.

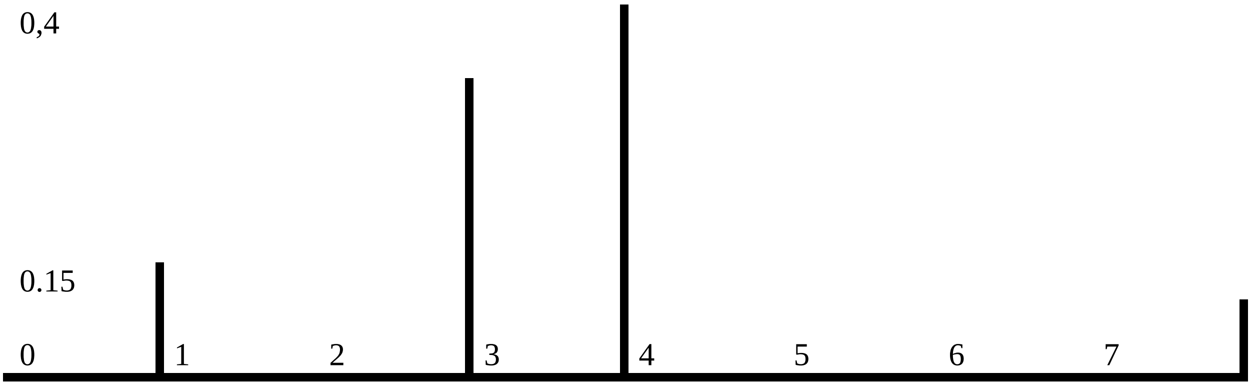


Fig 2.4. Mathematical Expectation as a Center of Gravity

Mathematical expectation properties.

1. Expectation of the constant is equal to the constant itself. $M(C)=C$

2. Constant factor can be removed from under the expectation sign $M(CX)=CM(X)$.

3. Expectation of the sum of random variates is equal to the sum of expectations of summands:

$M(X_1+X_2+\ldots+X_k)=M(X_1)+M(X_2)+\ldots+M(X_k)$

4. Expectation of the product of independent random variates is equal to the product of expectations of these variates. $M(X_1*X_2*\ldots*X_k)= M(X_1)*M(X_2)*\ldots*M(X_k)$

Expectation of the number of the occurrence of the event A in *n* independent trial is equal to the number of trials into the probability of the occurrence of the event in one trial $M(X)=np$.

One expectation is not enough to characterize a random variate as it specifies only the center grouping location but not dispersion.

Centered random variate *X-M(X)* and its expectation

Variance is defined as the expectation of the centered random variate square:

$$\mathbf{D(X)=M((X–M(X))^2)} \qquad (2.21)$$

It is usually calculated by formula:

$$\mathbf{D(X)=M(X^2)–(M(X))^2} \qquad (2.22)$$

Standard deviation is defined in the following way:

$$\partial(\mathrm{X})=\sqrt{\mathrm{D(X)}} \qquad (2.23)$$

Quantile is a solution relative to *x* of the equation:

$$F(x)=p \qquad (2.24)$$

Where *p* is a set probability.

The following specific cases of the quantiles, which have their proper names, are distinguished. For example, quartile is a value of the exponent that divides the half of a sample into two equal parts (there are two of them). Q_1 or lower quartile is a value fro which the following condition is met: a quarter of observations is less then this quartile. Q_3 or upper quartile is less than the quarter of observations. That is, the median and quartiles divide rank series into 4 equal parts. The value Q_3-Q_1 is called interquartile width. The value $(Q_3$-$Q_1)/2$

called semiinterquartile width is of a very frequent use. It is a median of absolute deviations from the average quartile *$(Q_3+Q_1/2)$*.

Deciles are respective values, which divide rank series into ten equal in volume parts (centiles — 100). Under *p* — percent quantile is understood an exponent value which does not exceed *p*% of observations.

There are QUARTILE and PERCENTILE functions in Excel. They allow you to determine some quantiles.

The function QUARTILE has the format QUARTILE (array; parameter).

Array is a cell interval or cell array and parameter determines which value should be defined. Thus, QUARTILE (array; 1) returns Q_1, or lower quartile, and QUARTILE (array; 3) returns $Q_{3,}$ or upper quartile. By means of the rest of parameters the values, which can be obtained also by other functions are transferred. QUARTILE (array; 0)=MIN (array), that is the minimum value; QUARTILE (array; 4)=MAX (array), that is maximum value, and QUARTILE (array; 2)=MEDIAN (array).

For example, QUARTILE *({0,2; 0,6; 1,1; 3; 4,3; 5; 5,4; 7}; 3)=5,1*. It means that the quarter of all observations is more than 5,1.

The function PERCENTILE ahs the format PERCENTILE (array; percentile) where PERCENTILE is *p*-percent quantile expressed in fractions (values from 0 to 1).

For example, PERCENTILE *({0,2; 0,6; 1,1; 3; 4,3; 5; 5,4; 7}; 0,9)=5,88.* I t means that 90% of all observations are less than 5,88. and PERCENTILE *({0,2; 0,6; 1,1; 3; 4,3; 5; 5,4; 7}; 0,8)=5,24.*

Interaction among distributions

The interactions among principle distribution laws, to which the present work is devoted, are presented in Fig. 2.5.

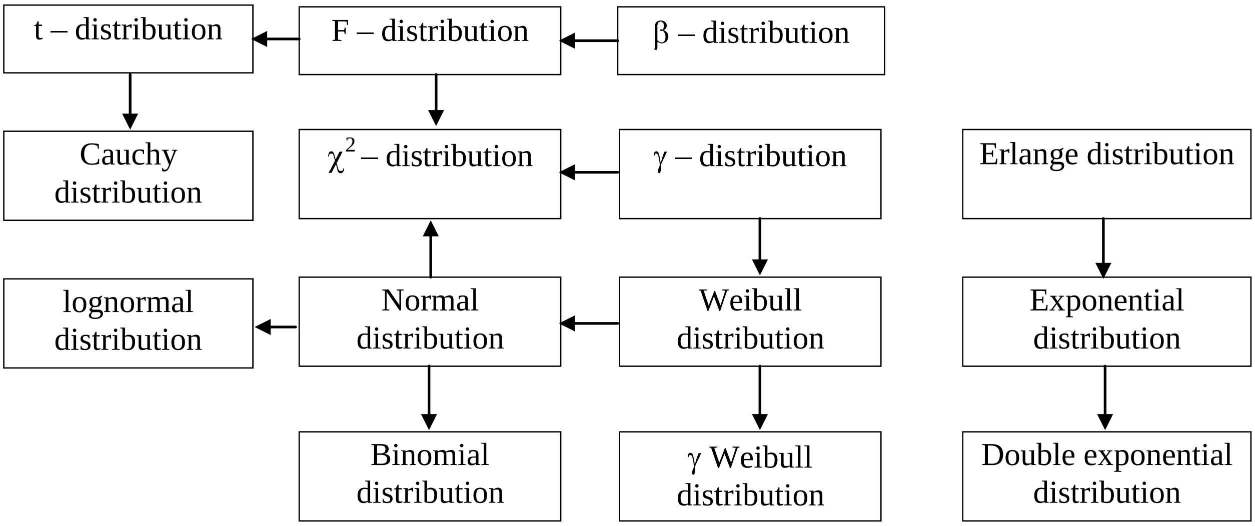


Fig. 2.5. Interaction of Continuous Distributions

2.3.3 Laws of Large Numbers

Probability theory limiting theorems

The fact that in certain conditions random variates behave in a randomless manner allows predicting the result of random phenomena, practically with the complete definiteness. Limiting theorems of probability theories deal with conditions and laws of such situations.

Chebyshev inequality

Let there be a random value *X* with the expectation *M* and dispersion D. no matter what the positive α may be the probability that the random variate will deviate from its expectation no less than for α, is less than D/α ^{2}:

$$P(|X-M|\ge\alpha)\le\frac{D}{\alpha^2} \qquad (2.25)$$

The law of large numbers (Chebyshev theorem). The sampling method is based on the inferences from Chebyshev theorem. The theorem states: if X_1, X_2, X_3, … , X_n are pairwise independent random variates, given their dispersion are evenly limited (do not exceed a certain constant value, than for as small as is wished primarily prescribed number ε the probability of fulfilling the condition: $\left|\frac{X_1+X_2+...+X_n}{n}-\frac{M(X_1)+M(X_2)+...+M(X_n)}{n}\right|<\varepsilon$ will be as much close to 1 as is wished, given quite a big number of random variates *n*.

Otherwise:

$$\lim_{n\to\infty} P(\left|\frac{X_1+X_2+...+X_n}{n}-\frac{M(X_1)+M(X_2)+...+M(X_n)}{n}\right|<\varepsilon)=1 \qquad (2.26)$$

For a specific case when all independent random variates have the same expectation *m*, the indicated expression will assume the following form:

$$\lim_{n\to\infty} P(\left|\frac{X_1+X_2+...+X_n}{n}-M\right|<\varepsilon)=1\,. \qquad (2.27)$$

The essence of Chebyshev theory is that the arithmetic average of the large number of random variates, whose dispersions are evenly limited, loses its random character and converges in probability to expectation.

Inferences from Chebyshev theory are widely used in practice by the virtue of the fact that increasing the number of quantity measurement the accuracy of its determination goes up. Unfortunately it is forgotten that the theorem contains the conditions under which a phenomenon will take place. They are:

- Pairwise independence of random variates (experiments);
- Variances are evenly limited;
- The same expectation.

Chebyshev theorem, Bernoulli theorem and Chebyshev in a quality are often called the *law of large numbers*.

Generalized Chebyshev theorem

If independent random variates X_1, X_2, … , X_n have expectations M_1, M_2, … , M_n and variances D_1, D_2, … , D_n if all variances are limited at the top by one value then increasing *n* the arithmetic average of values of the quantities X_1, X_2, … , X_n — converges to the arithmetic average of their expectations (the difference between one and another converges in probability to 0:

$$P(\left|\frac{\sum_{i=1}^{n} X_i}{n}-\frac{\sum_{i=1}^{n} M_i}{n}\right|<\varepsilon)>1-\delta \qquad (2.28)$$

Corollaries of the Law of Large Numbers

Bernoulli Theorem. It should be borne in mind that relative frequency never tallies precisely with theoretical probability. For example, the relative probability of the “head” when tossing a coin is 0,5 (1/2). If tossing the coin you calculate the relative frequency it will become evident that it is not consistent with the theoretical probability.

The theorem states that given quite a big number of trials, the probability that the departure of relative frequency from theoretical probability will be as small as is wished, tends to 1: $\lim\limits_{n\to\infty} P(\left|{}^{m}\!/_{n} - p\right| < \varepsilon) = 1.$

Hence it follows that relative frequency possesses the property of stability. It is well to bear in mind that in this theorem we are dealing with the convergence in probability. That is, increasing the number of experiment ad infinitum, the difference between relative frequency and probability will not necessarily be as small as it is wished. For some *n* this condition may be not fulfilled. The convergence in probability means that probability of such event is infinitesimal.

Poisson’s Theorem. If performing *n* independent trials, the probability of the occurrence of the event A in i-th trial is equal to p_i. Then, increasing the number of trials, the frequency of the event A converges in probability to the arithmetic average of probabilities p_i.

It is of high practical importance since in fact nominally equal conditions of trials are being slightly changed.

Central limiting theorem is a set of theorems relating to limiting distribution laws of the sum of random variates. The most important is Lyapunov’s theorem. In accordance with this theorem the distribution law of the sum of random variates tends to the normal law provided the unlimited growth of the number of random variates and fulfillment of the following conditions: all quantities have end expectations and variances and none of the quantities differs markedly from the rest.

This conditions mean that the effect of separate summands on the sum should be evenly small, that is there are no predominant components influencing the variants or expectations.

2.3.4. Sampling Method

As a rule, conducting researches the population of values of the studied quantity is not available (that is we do not have all possible values at our disposal). Thus, we are not able to measure height, weight or arterial blood pressure of all inhabitants of the earth (theoretically, it is possible, but we will be in charge of it and who will pay for this?) so we are bound to employ data samples in work. Using sampling method has a number of advantages compared to a continuous examination:

- It requires less expenses of all resources (financial, time, human, material and others);
- Due to less expenses you can conduct researches more often, more rapidly and consequently, their results will correspond to the current moment;
- Sampling methods can be carried out in place of a continuous examination where the latter is impossible;
- Reduction of subjective errors, as the result of decreasing the number of participants and employing more efficient stuff;

Though, there are some problems when using a sample method:

- Provision of a homogeneous and representative sample. If it is not done, the results of the sample examination cannot be distributed over the whole general population;
- Estimate of accuracy of the received results;
- Methods of forming a sample;
- Correspondence of methods of data processing to their peculiarities.

The formulas to define sample sizes are given in 7.10.

In detail, the sampling method can be found in [10, 29]. In our work, describing any methods we focus the attention of the reader on these problems.

2.4. Some Distribution Laws

2.4.1. Even Distribution (discrete)

Even distribution is a distribution for which the probability of every value of the random variate is identical. That is:

$$P(X = x) = \frac{1}{N} \qquad (2.29)$$

where N is a number of possible values of the random variate.

2.4.2. Binomial Distribution

It is a distribution of probabilities of the occurrence of m events in n independent trials, given a constant probability of the event in every trial is p. the probability of a possible number of events is defined by Bernoulli distribution formula:

$$\mathrm{P_n(X = m) = C_n^m p^m q^n}, \qquad (2.30)$$

where p is a probability of the occurrence of the event in every trial, m is an expected number of events, n is a total number of trials, $p=1-p$,

$$\mathrm{C_n^m = \frac{n!}{m!(n-m)!}} \qquad (2.31).$$

Conditional short designation of the binomial distribution law appears as follows: *X ~ B(n, p).*

Binomial distribution can be specified as a series, presented in Table 2.7.

Table 2.7

X=m	0	1	…	k	…	N
$\mathrm{P_n}(m)$	$\mathrm{q^n}$	$\mathrm{p^1q^{n-1}}$	…	$C_n^k = n!/k!(n-k)!$	…	$\mathrm{p^n}$

Binomial distribution expectation is *np* and variance — *npq*. Given a large number of trials binomial distributional approaches very narrowly to the normal (Fig. 2.6). As is seen, the distribution form with the height n is approaching to the normal, but the less is p, the slower. This fact is proved in Moivre-Laplace local theorem (it relates to limit theorem). It follows from here that given large n the probability that in n trials the event will occur m times can be defined from the formula.

$$P_n(m) = \frac{1}{\sqrt{npq}}\varphi(x), \qquad (2.31)$$

where $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, $x = \frac{m - np}{\sqrt{npq}}$.

In Excel there are two functions for binomial distribution: BINOMDIST and CRITBINOM.

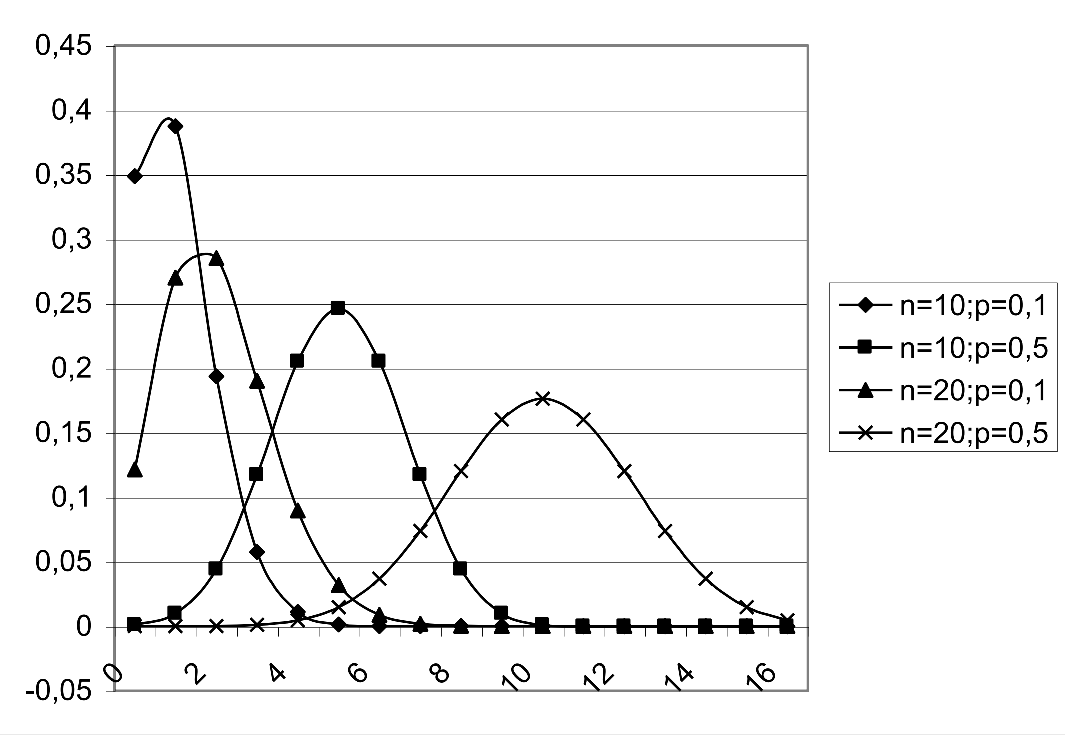


*Fig. 2.6. Probability Function for Binomial Random Variate given different **p** and **n**.*

The function BINOMDIST has the following format: BINOMDIST (*num_suc; num_trials; prob_suc; item*). *Num_suc* is a number of trials with a successful outcome, *num_trials* is a total number of trials, *prob_suc* is a probability of successful outcome in one trial, *item* is a type of the returned function (given TRUE - the integral distribution function value returns, that is the probability that the number of successful outcome will not be less then the value *num_suc* and FALSE – distribution function returns, that is a probability that the number of successful outcomes is exactly the same as the value *num_suc*). (Fig. 2.7)

If we need to calculate the probability that 6 of 10 newly born children will be boys then BINOMDIST (6; 10; 0,51; FALSE)=0,213022. But if we need to calculate the probability that no less than 6 of 10 newly born children will be boys then BINOMDIST (6; 10; 0,51;

TRUE)=0,811227.

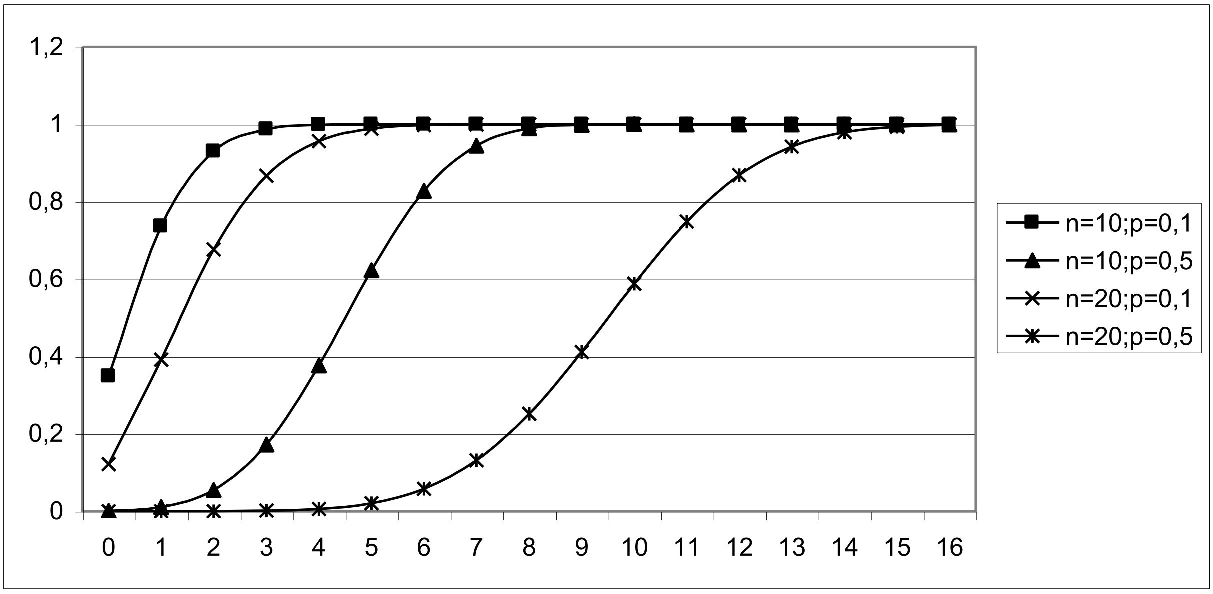


*Fig. 2.7. Integral Distribution Function for Binomial Random Variate Given Different **p** and **n***

The function CRITBINOM returns the critical value, that is the least value for which integral binominal distribution is more than or equal to the specified *item*. It has the following format: CRITBINOM (num_trials; prob_suc; alpha). Here, alpha is a significance level (section 3.1).

For example, CRITBINOM (100; 0,5; 0,05)=42. It means that 42 is a minimum value for which the probability of the occurrence of no less 42 successful outcomes out of 100 is more than or equal to 0,05 and indeed, BINOMDIST (41; 100; 0,5; TRUE)=0,044313; and BINOMDIST (42; 100; 0,5; TRUE)=0,066605.

2.4.3 Geometrical distribution

This distribution represents the probability of the successful outcome after the first *k* outcomes were unsuccessful. Formally, in the short form it is designated as *X ~ NB(1, p)*. The probability of such event is defined from the formula:

$$P(X=k)=p(1-p)^k \qquad (2.33)$$

Plots are shown on Fig. 2.8.

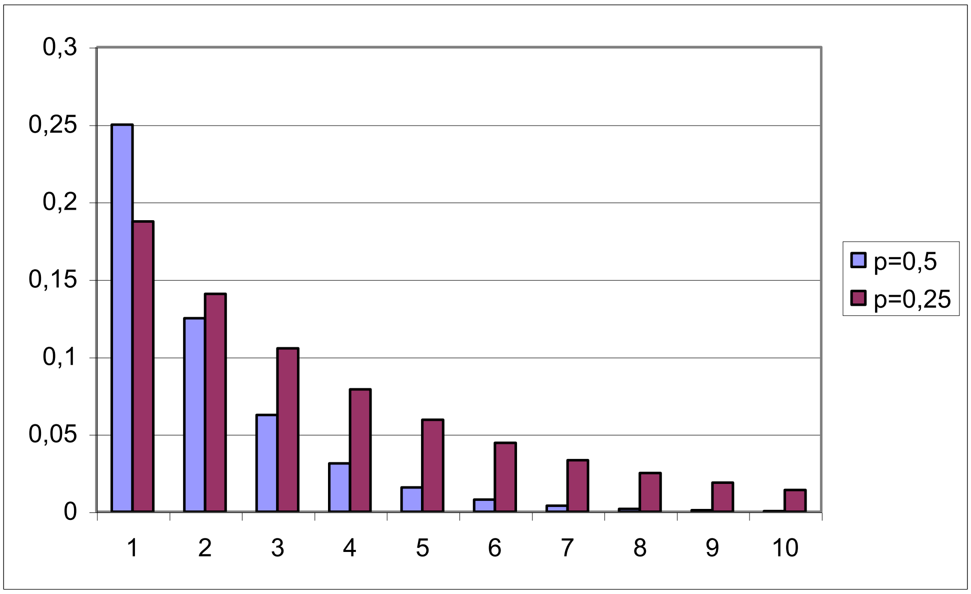


Fig. 2.8. Geometrical distribution for different ***p***

2.4.4. Hypergeometrical distribution

Hyper geometrical distribution density is specified by the following probability quantity.

$$P(X=k)=\frac{C_k^M C_{n-k}^{N-M}}{C\frac{N}{n}} \qquad (2.34)$$

Furthermore, the following conditions should be fulfilled $N \geq M$, $n \geq k$, $M \geq k$. Formally, in the short form, hypergeometrical distribution is designated as *X ~ H(M, N, n).* The plot of the function for several values of parameters is given in Fig. 2.9. Hypergeometrical distribution is represented by the following urn model. Let *N* balls be in the urn, *M* of them is white. From the urn, *n* balls are taken. The probability that *k* of them are white is defined by formula 2.32. To calculate the values of hypergeometrical distribution there is the function in Excel – HYPERGEOMETR(*k*; *n*; *M*; *N*). For example, there are 100 syringes in a box. 20 of them have been used. 10 syringes have been removed from the box at random. What is the probability that 5 of them have been used? HYPERGEOMETR(5; 10; 20; 100)=0,0215.

2.4.5 Pascal Distribution

Generalization of geometrical distribution. It is also called negative binomial distribution. In the formal short form it is designated as

X ~ NB(k, p). it represents the probability of the presence of *k* negative results before *m*-th positive ones.

$$P(X=k)=C_{m+k-1}^{k}p^{m}(1-p)^{k} \qquad (2.35)$$

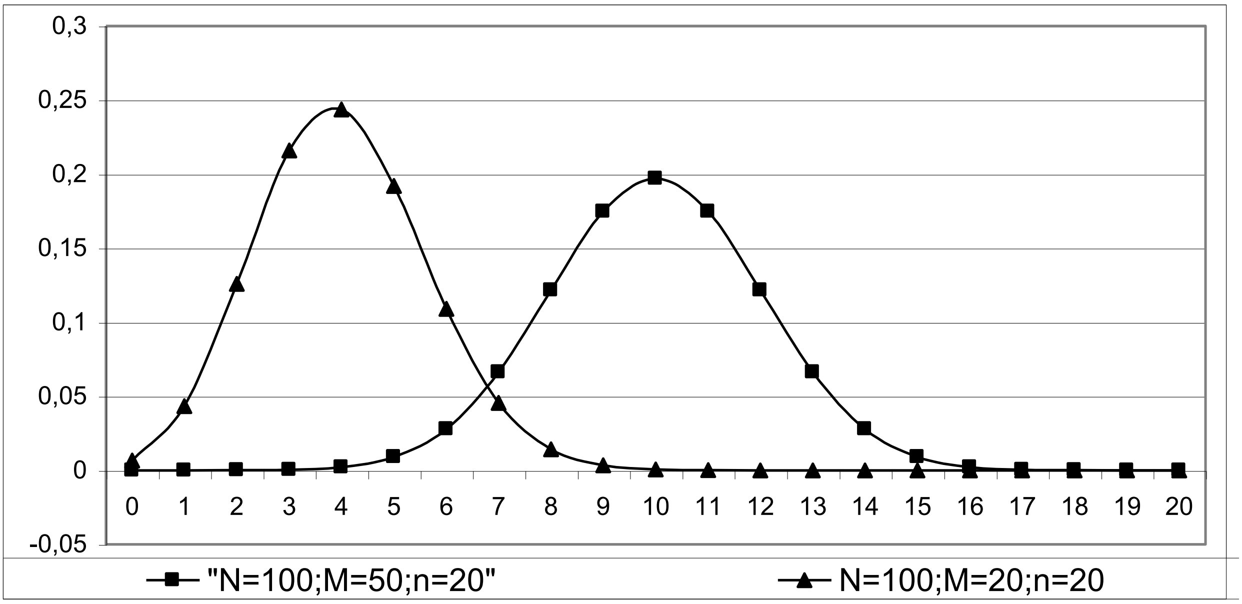


Fig. 2.9. Hypergeometrical distribution

In the specific case when *m*=1 it is reduce to geometrical probability distribution *X ~ NB(1, p)*. The plots of Pascal distribution are given on Fig. 2.10.

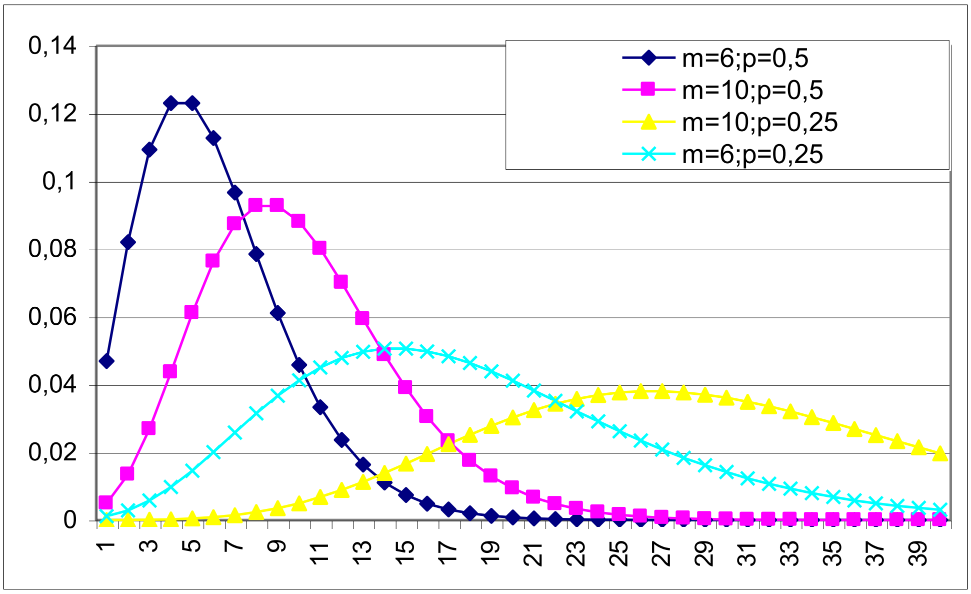


Fig. 2.10. Pascal distribution

To calculate the values of probability corresponding to the formula 2.33 there is the function in Excel — NEGBINOMDIST(*k*; *m*; *p*;). For example, to study the drug effect on behavioral reactions we need to select rats with certain qualities. The presence of qualities is determined by means of the tentative test. The probability of this quality being in a rat is 0,6. We need twenty rats to carry out an experiment. The probability that 20 rats are to be culled before we have 20 rats with proper qualities is determined by NEGBINOMDIST(20; 20; 0,6)=0,0277.

2.4.6. Poisson distribution (distribution law of rare events)

In case when *p* or *p* is small, increasing the number of trials, binomial distribution tends to Poisson distribution, which is defined by formula:

$$P_m = \frac{\lambda^m e^{-\lambda}}{m!}, \qquad (2.36)$$

where $\lambda = np$ is an intensity. It is considered to be constant, independent of the number *n*. Both variance and expectation in Poisson distribution are $np=\lambda$.

Poisson distribution is widely used to model the systems of mass service (telephone central offices, commercial institutions, banks etc. Poisson distribution probability functions for different λ are given in Fig. 2.11.

To calculate the values of probability corresponding to Poisson distribution there is a function in Excel — POISSON (*cumulative; intensity; item*). *Cumulative* is a number of events for which we calculate; *intensity* is a value of the parameter λ; if *item*=FALSE then the returned value is equal to the probability that there will be events from 0 to the number determined by the parameter *cumulative* (integral function), if *item*=TRUE then the returned value is equal to the probability that there are as many events as determined by the parameter *cumulative*. For example, the number of the unsuccessful outcomes of some surgical operations in a hospital per annum is 5 on the average. Then, the probability that there will be 4 failures this year is calculated in the following way POISSON(4; 5; FALSE)=0,1755 and the probability that there can be 4 or less unsuccessful outcomes is calculated POISSON(4; 5; TRUE)=0.44049.

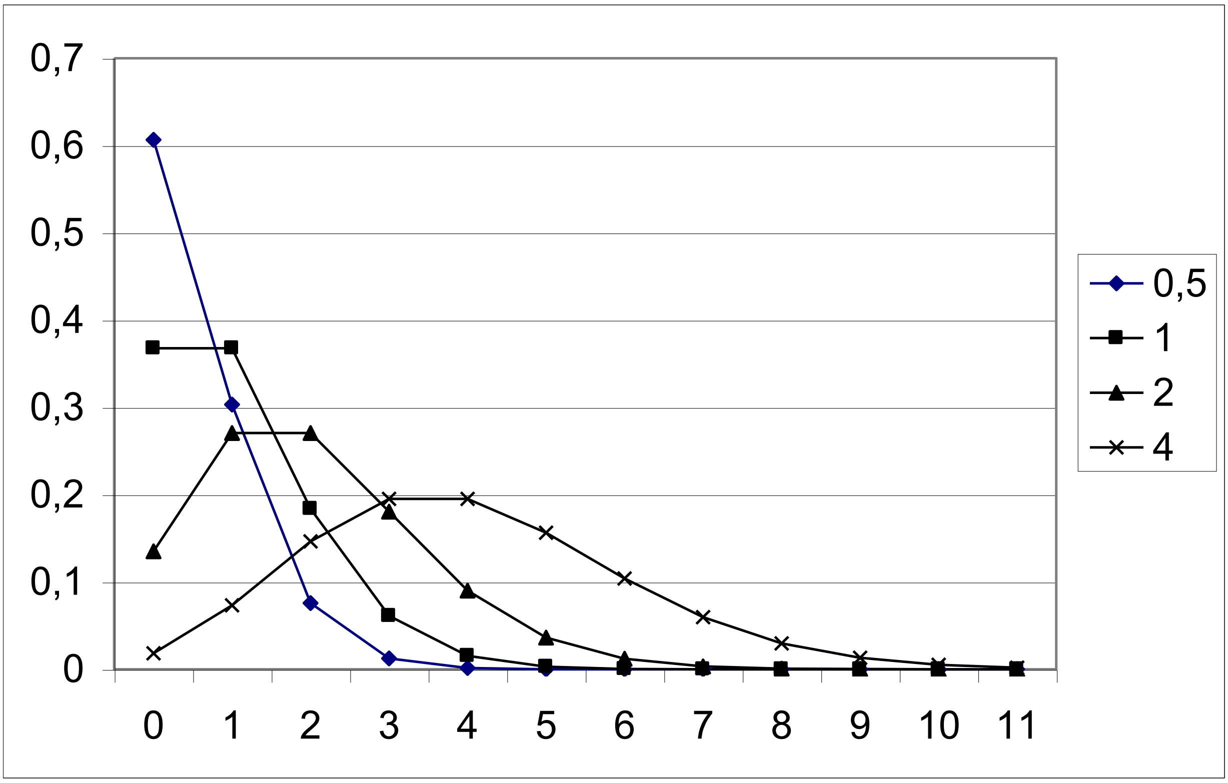


Fig. 2.11. Poisson Density probability distribution functions (polygons of distribution) for different λ

2.4.7. Pareto distribution

Density distribution is defined by formula 2.37. function distribution is defined by formula 2.38.

$$f(x) = \frac{a}{b}\left(\frac{b}{x}\right)^{a+1} \qquad (2.37)$$

$$F(x) = \begin{cases} 0, npu: x \le b \\ 1 - \left(\frac{b}{x}\right)^{a}, npu: 0 < a, 0 < b < x \end{cases} \qquad (2.38)$$

Density plots for some values of parameters are given on picture 2.12. This distribution is good at representing population income distribution and market distribution among individual firms. Besides, the influence force of individual sectors on a certain process is usually distributed according to Pareto. That is, the relatively small number of factors of general number explains, in greater part, almost the whole process conduct.

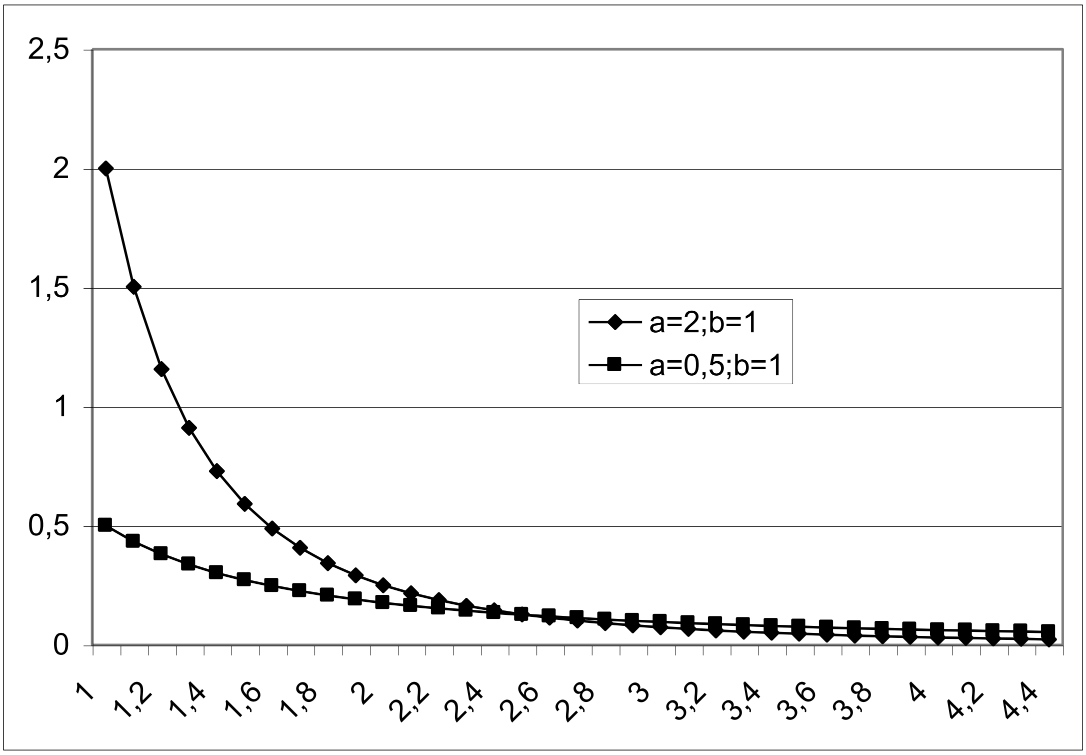


Fig. 2.12. Pareto density distribution plots

Sales volume distribution of the most frequently sold drugs in Ukraine (1997) is presented on Fig. 2.13. Sick rate distribution in the regions of Ukraine and many other random variates has a very similar form.

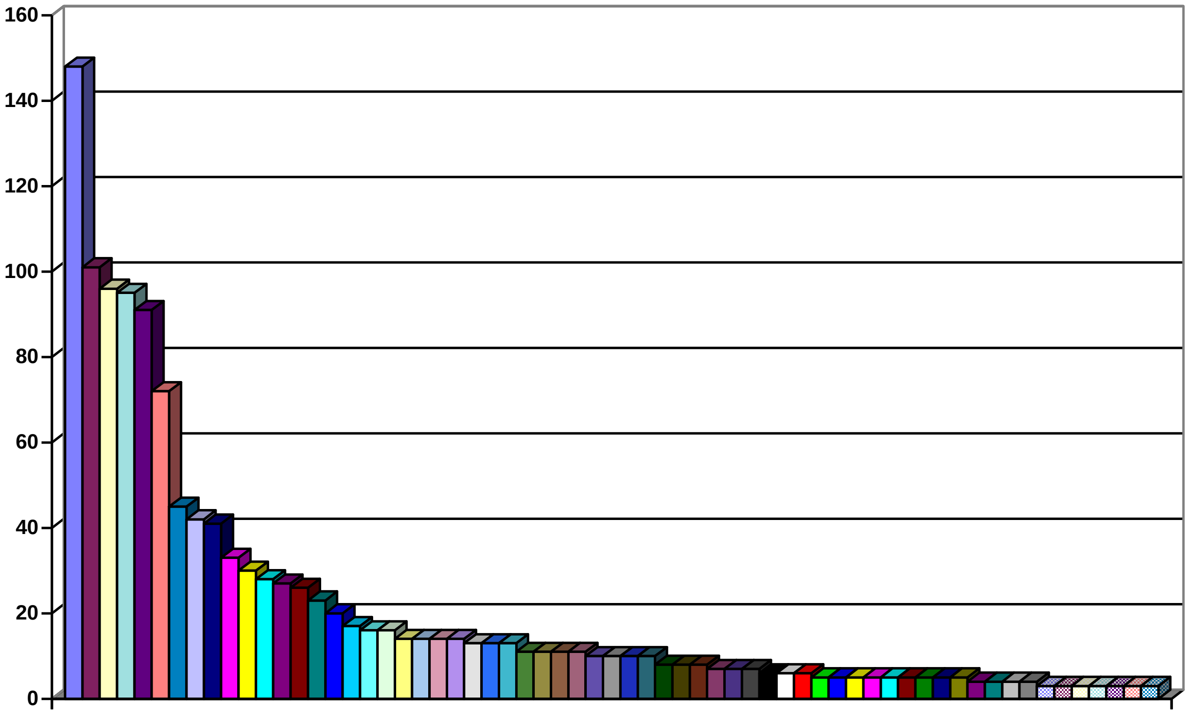


Fig. 2.13. Distribution of the most frequently sold drugs

2.4.8. Continuous distribution

In the short form it is designated as *X* ~ U(a, *b*). Density of probability and distribution function is determined by formula.

$$f(x) = \frac{1}{b-a}, -\infty < a < x < b < \infty \qquad (2.39)$$

$$F(x) = \frac{x-a}{b-a}, -\infty < a < x < b < \infty \qquad (2.40)$$

Plots of both functions are presented in Fig. 2.14.

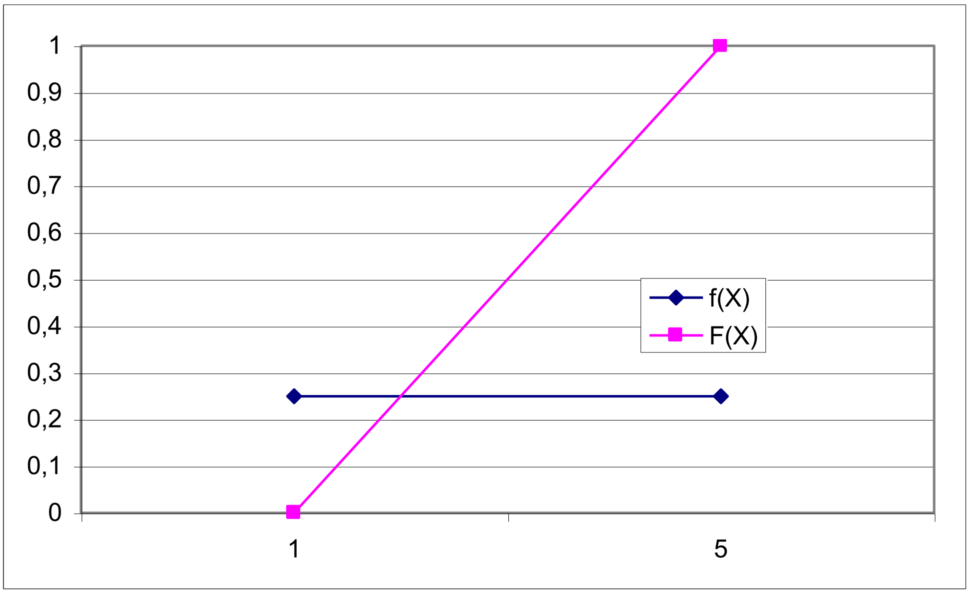


Fig. 2.14. Probability density and distribution function for continuous distribution law

2.4.9. Normal (Gauss) distribution law

For normal law, distribution density appears as:

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}, \qquad (2.41)$$

where *m* is an expectation and σ is a standard deviation (σ^2 — variance). In the short form, normal distribution is designated as *X* ~ *N(m, σ^2)*.this distribution law has found wide use in probability theory and mathematical statistics. Standard normal distribution is a distribution with zero expectation and unit variance whose density has the following form:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \qquad (2.42)$$

Standard normal distribution density probability appears as in Fig. 2.15, its distribution function is presented in Fig. 2.16.

Changing expectation does not change the curve but only moves it along *X*-axis. The curve form is changed when changing variance Fig. 2.17.

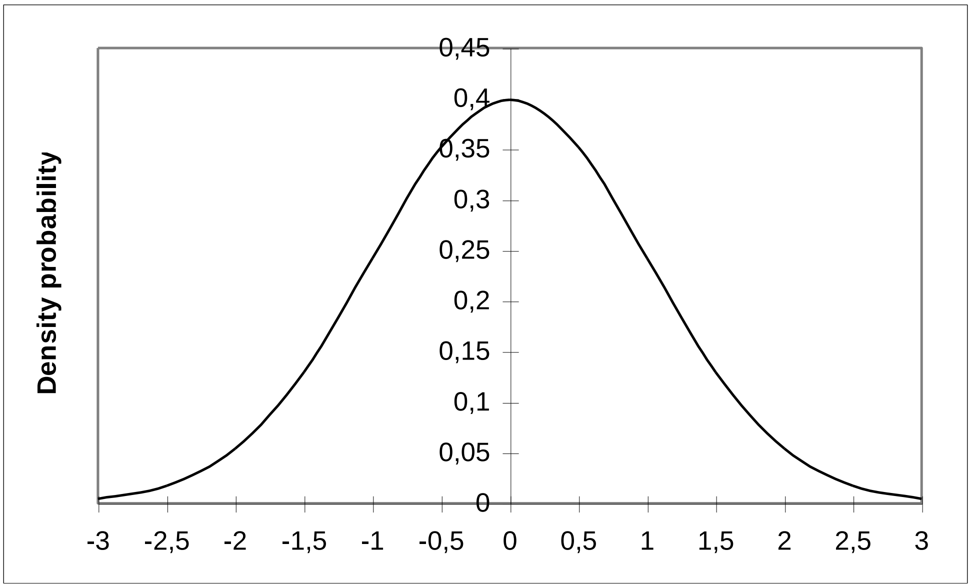


Fig. 2.15. Density probability of standard normal random variate

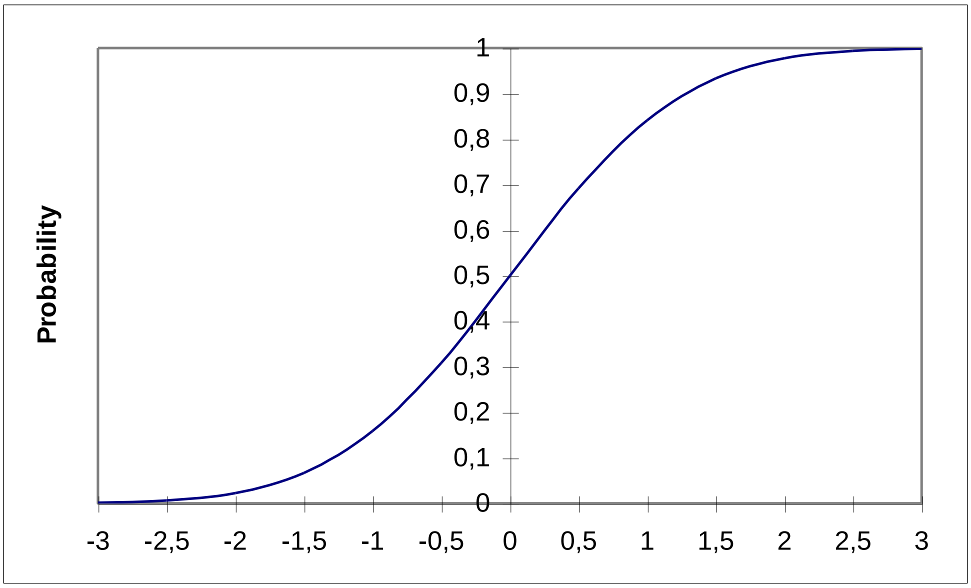


Fig. 2.16. Distribution function of standard normal random variate

From the figure it is evident that the bigger is the variance, the more steep and stretched is the curve (and vice versa).

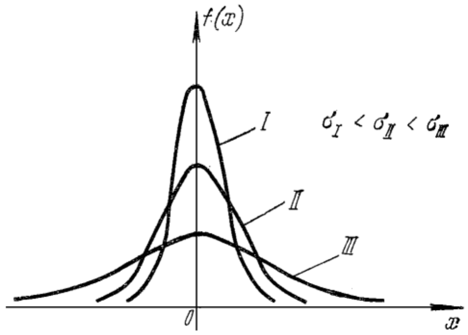


Fig. 2.17. Changing probability density form according to variance

Almost all parametric statistics is based on normal distribution law. It has to do with the fact that most part of distributions used to test statistical hypotheses (Fisher's, Student's) are transformations of normal distribution law.

The main peculiarity of normal distribution law is that all other distribution laws tend to it, when fulfilling certain laws. It follows from the *central limit theorem* (See 2.3.3).

There are 5 functions available in Excel to calculate normal distribution. NORMSDIST(*x*) returns the standard normal cumulative distribution according to value *x*. For example, NORMSDIST(0)=0,5 (Fig. 2.16). NORMSINV(probability) calculates *x* corresponding to the specified probability value, that is NORMSINV(0,5)=0. Functions NORMDIST(*x*; mean; st_dev) and NORMSINV(probability; mean; st_dev) are alanologos to the those mentioned before but they are meant for freeform normal distribution with the parameters μ=mean and σ=st_dev. The function NORMALIZATION(*x*; mean; st_dev) returns the normalized *x*, that is the value that should have had *x* if had been distributed according to normal distribution law with the parameters μ=mean and σ=st_dev.

2.4.10. Student's distribution

Student's distribution is a distribution of a random variate:

$$t_n = \frac{x_0}{\sqrt{\sum_{i=1}^{n} x_i^2 \Big/ n}}, \qquad (2.43)$$

where random variates x_i have standard normal distribution. Density distribution appears as:

$$f(x)=\frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{\pi n}}(1+\frac{x^2}{n})^{-(n+1)/2} \qquad (2.44)$$

Student’s distribution expectation is 0 and variance — *n*/(*n*-2). Probability density and Student’s distribution function with the number of degrees 1 are represented in Fig. 2.18 and 2.19 respectively.

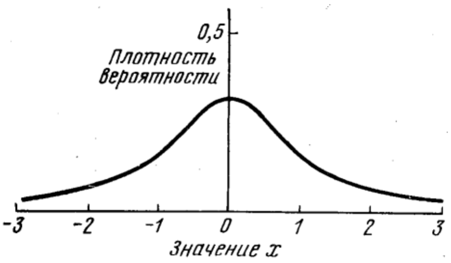


Fig 2.18. Density probability of Student’s distribution

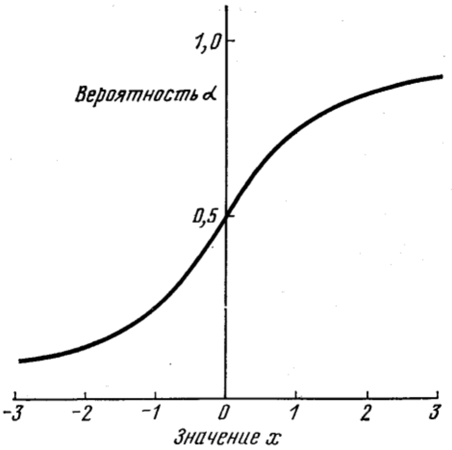


Fig 2.19. Distribution function of Student’s random variate

Let’s consider Student’s density distribution function and normal distribution function (Fig. 2.20). It is obvious that this function as close similarity to the normal density distribution, but in students distribution probabilities on tails are bigger.

There are 2 functions in Excel that can be used as alternatives to tables of Student’s distribution. This is the function STUDIST(*x*; degrees_freedom; item). Here *x* is a value that requires the distribution to be calculated, *degrees_freedom* is a number of distribution degrees of freedom. If item=1 then the value for one-side distribution returns. Given item=2 the value for two-side distribution returns. The function STUDISTINV(probability; degrees_freedom) allows calculating Student's distribution value which corresponds to the specified probability and the number of degrees_freedom. After that, this value can be used for comparison with the experimentally calculated value to test different statistical hypotheses.

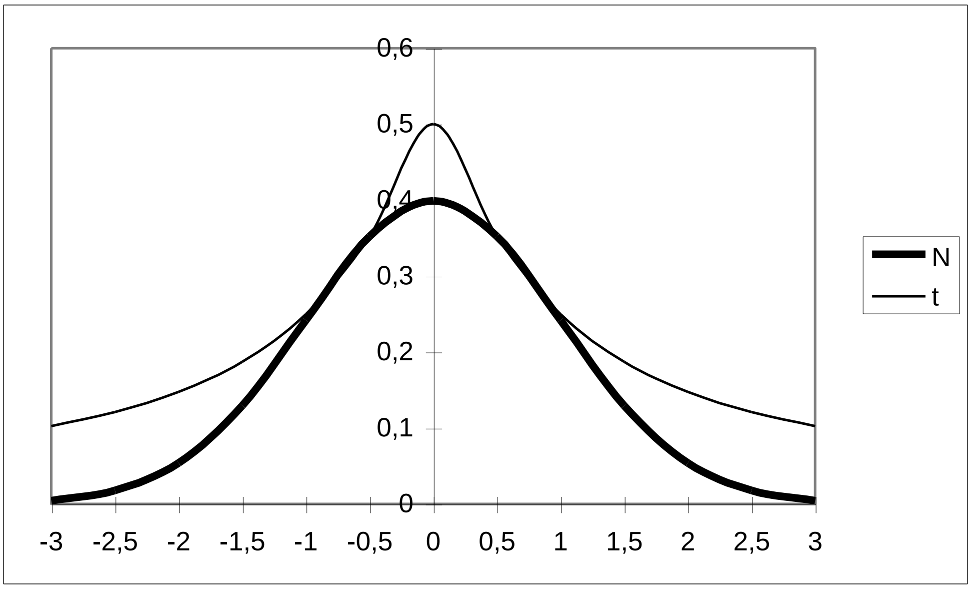


*Fig. 2.20. Probability density of Student's (t-distribution and normal distribution (*N*)*

2.4.11. Fisher's distribution

Fisher's distribution is attributed to the following random variate:

$$F_{n,m} = \frac{\sum_{i=1}^{m} x_i^2 / m}{\sum_{j=1}^{n} y_j^2 / n} \tag{2.45}$$

Random variates x_i and y_j are distributed according to normal standard distribution. Density distribution function, therewith, is defined by formula:

$$f(x) = \frac{\Gamma(\frac{\nu_1 + \nu_2}{2}) \nu^{\nu_1/2}{}_1 \nu^{\nu_2/2}{}_2 x^{(\nu_1/2)-1}}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})(\nu_1 + \nu_2 x)^{(\nu_1+\nu_2)/2}} \tag{2.46}$$

Fisher's density distribution functions and its corresponding distribution function are presented in Fig. 2.21, 2.22.

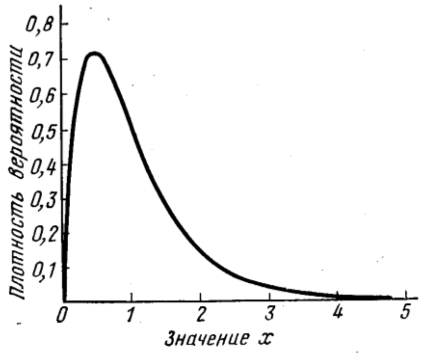


Fig. 2.21. Density probability of Fisher's distribution for degrees of freedom v_1=4, v_2=40

Fisher's density distribution functions with different degrees of freedom are given in Fig. 2.23.

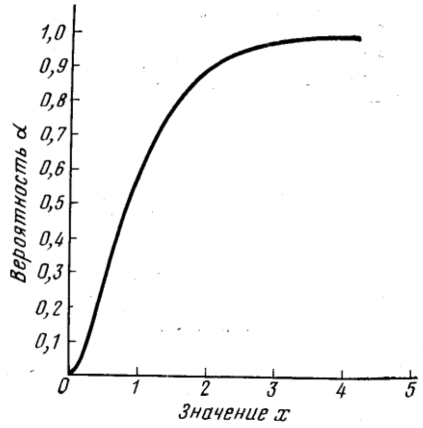


Fig. 2.22. Fisher's distribution function for degrees of freedom v_1=4, v_2=40.

In Excel there is the function FDIST(*x*; degrees_freedom1; degrees_freedom2) which calculates the value of the integral distribution function for the specified number of degrees of freedom *x*. there is also the function FDISTINV(probability; degrees_freedom1; degrees_freedom2) which returns the value *x* for the specified numbers of degrees of freedom and probability. It can also be used instead of tables.

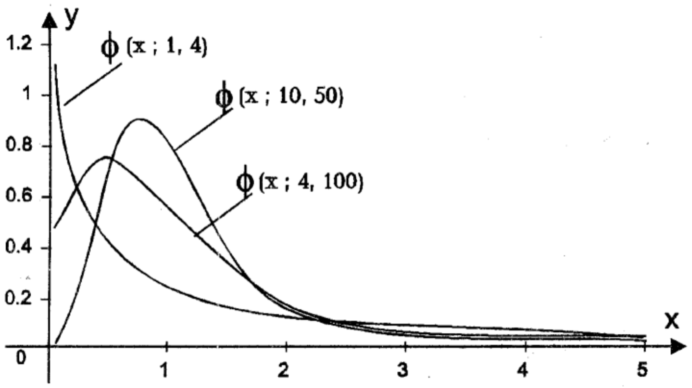


Fig. 2.23. Fisher's density distribution functions with different degrees of freedom

2.4.12. Pearson's chi-square distribution

Distribution χ^2 is attributed to the random variate which makes up the sum of squares of random variates, every of which is distributed according to the normal distribution law. χ^2 density distribution appears as:

$$f(x) = \frac{1}{2^{n/2}} \frac{1}{\Gamma(n/2)} x^{n/2-1} e^{x/2} \qquad (2.46),$$

where Γ() is a gamma function: $\Gamma(z) = \int_0^\infty x^{z-1} e^x dx, z > 0$. Mathematical expectation of χ^2 is *n* and variants is 2n. The plots of density probability function and χ^2 distribution functions are given in Fig. 2.24, 2.25.

The difference between densities of distribution of this function for different degrees of freedom is shown in Fig. 2.26.

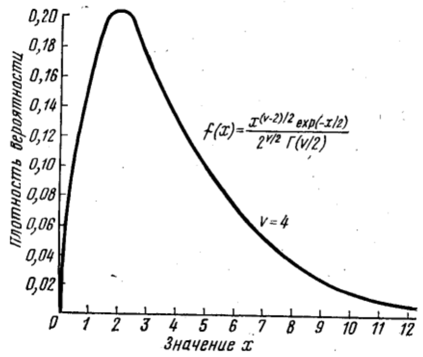


Fig. 2.24. Probability density of distribution χ^2

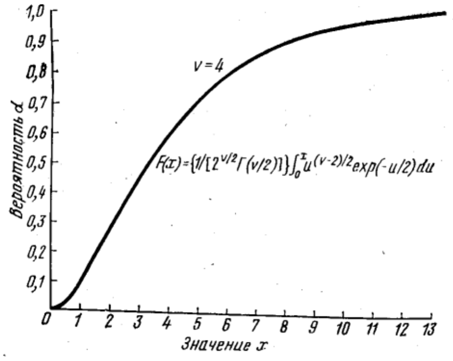


Fig. 2.25. Distribution function χ^2

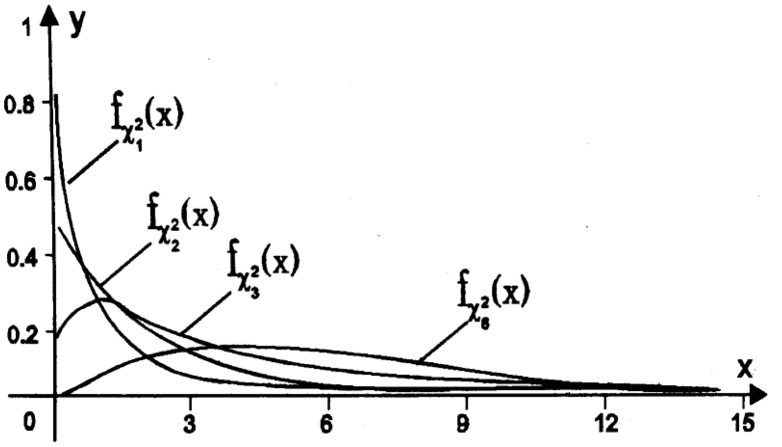


Fig. 2.26. Distribution density χ^2 for different degrees of freedom

To provide for calculations with the use of χ^2 , there is the function in Excel –

CHI2DISTR(*x*; degrees_freedom), which returns the probability for one-side χ^2 distribution for the specified *x* and degrees of freedom. There is also the function CHI2INV(probability; degrees_freedom), which delivers the value corresponding to the specified probability and the number of degrees of freedom. For example, CHI2DISTR(3;4)=0,557825 and CHI2INV(0,5;4)=3,356695 (Fig.2.25)

2.4.13. Lognormal distribution

It represents the random variate whose logarithm has the normal distribution. Density distribution is defined by formula:

$$f(x) = \frac{1}{2\pi\sigma^2}\frac{1}{x}e^{-\frac{(\ln x - a)^2}{2\sigma^2}} \qquad (2.48)$$

The plots of density for different values of a parameter are presented in Fig. 2.27. There are two functions available in Excel LOGNORMDISTR(*x*; mean; st_dev) and LOGNORINV(probability; mean; st_dev). Parameters are analogous to the parameters of the functions for normal distribution. For example, (Fig.2.27), LOGNORMDISTR(0,51; 0; 1)= 0,250364 and LOGNORINV(0,25; 0; 1)=0,509416

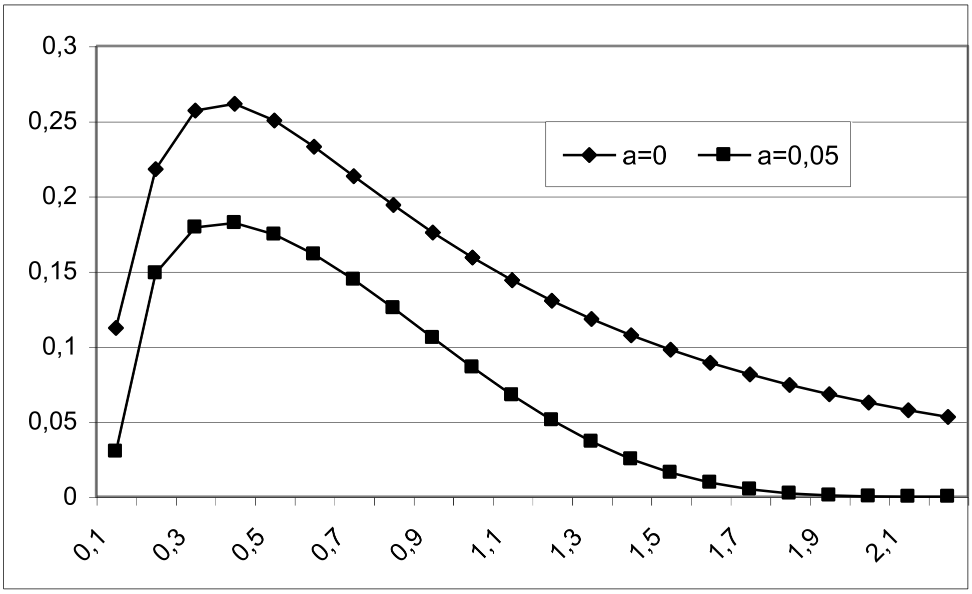


Fig.2.27 Plots of lognormal distribution density

2.4.14 Exponential distribution

Quite a lot random variates are distributed according to the exponential [distribution] law, for example, time intervals between the calls of ambulance.

In the short form it is designated as *X~P(1,λ)*

Density distribution for exponential distribution appears as:

$$p(x,\lambda)=\lambda e^{-\lambda x}. \qquad (2.48)$$

Distribution function for exponential distribution appears as:

$$F(x,\lambda)=\begin{Bmatrix} 1-e^{-\lambda x}, \forall x \geq 0 \\ 0, \forall x < 0 \end{Bmatrix} \qquad (2.49)$$

Expectation for exponential law is 1/λ and variance — $1/\lambda^2$. Distribution density and density distribution are given in Fig. 2.28 and 2.29

To find values for exponential distribution there is the function in Excel EXPDISTR(*x*; λ; item). If item=FALSE density distribution value is calculated. If item=TRUE integral distribution function is calculated.

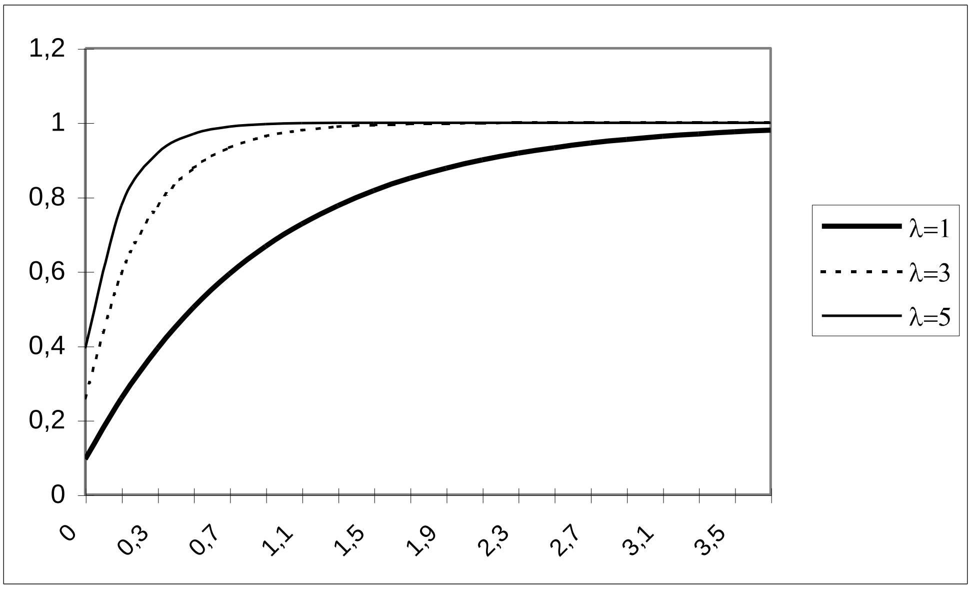


Fig. 2.28 Distribution function of exponential random variate for different values of λ

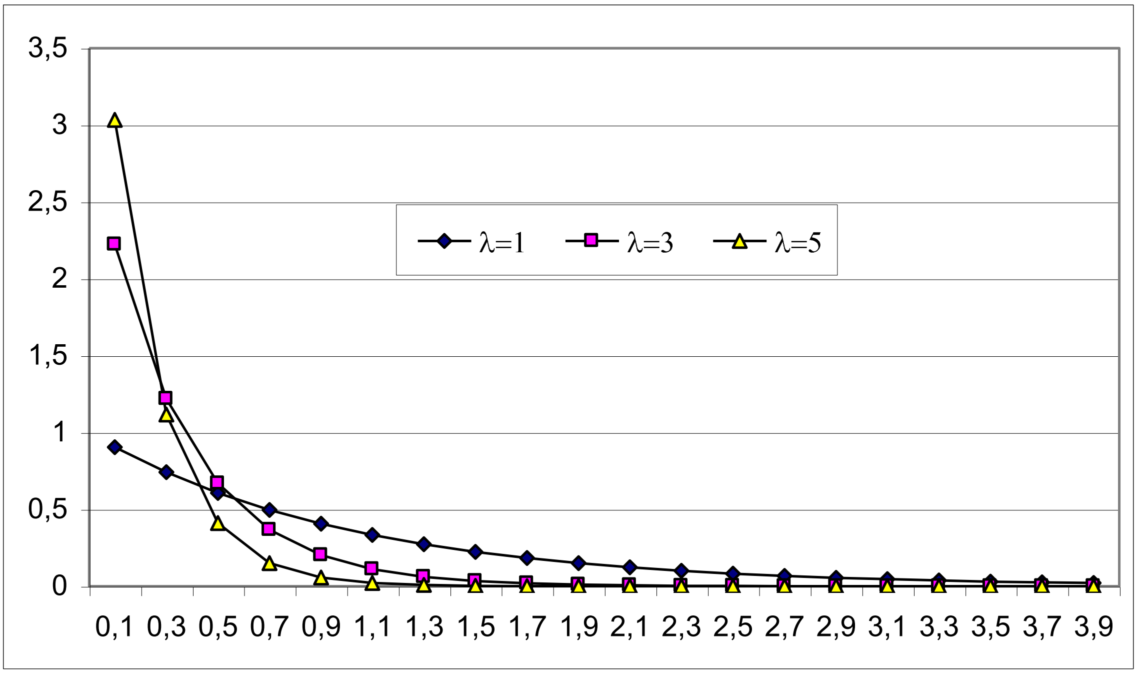


Fig. 2.29. Density of exponential distribution given different λ

2.4.15. Two-side exponential distribution

It represents one of the generalizations of exponential distribution whose density distribution is defined by formula

$$f(x) = \begin{cases} \frac{\lambda}{2} e^{\lambda x}, \text{для} : x \le 0 \\ \frac{\lambda}{2} e^{-\lambda x}, \text{для} : x > 0 \end{cases} \qquad (2.51)$$

The plots of density distribution for different values of λ are given in Fig.2.30. Such random variates as average “life” endurance of technical equipment, probability of dying in infancy and others are distributed in conformity with this distribution law.

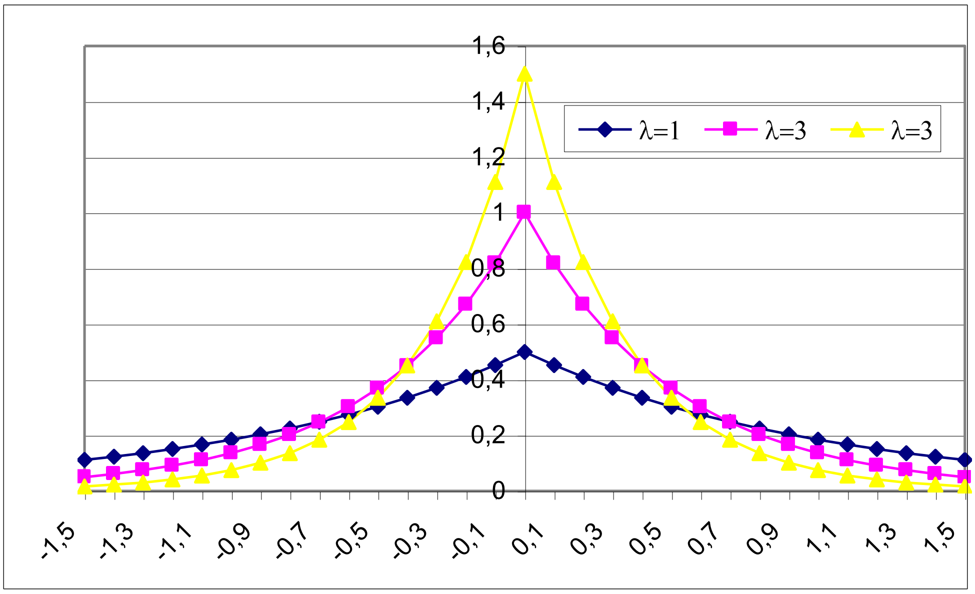


Fig. 2.30 Density distribution of random variate with two-side exponential distribution law

2.4.16. Gamma distribution

Density distribution is defined by formula:

$$f(x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)}, \alpha, \beta > 0; 0 < x < \infty \qquad (2.51)$$

Given β=1 it is called standard gamma distribution. Given α=1 it is reduced to exponential distribution, given α is positive and whole – reduced to Erlange distribution. The plots of density distribution for different values α and β are presented in Fig. 2.31 and distribution functions are given in Fig. 2.32. It is used in the theory of mass service systems.

To make calculations connected with gamma distribution there are functions in Excel GAMMADISTR(*x*; β; α; item) and GAMMAINV(probability; β; α). GAMMADISTR given item=TRUE returns interval distribution function and given item=FALSE – density distribution. GAMMAINV is used to calculate gamma distribution value, which corresponds to the specified probability. For example, GAMMADISTR(B233; 0,5; 3; FALSE)=0,161313816 (Fig.2.31). GAMMADISTR(B233; 0,5; 3; TRUE)=0,6826892 (Fig.2.32). And GAMMAINV(0,6; 0,5; 3)=1,062489901 (Fig. 2.31).

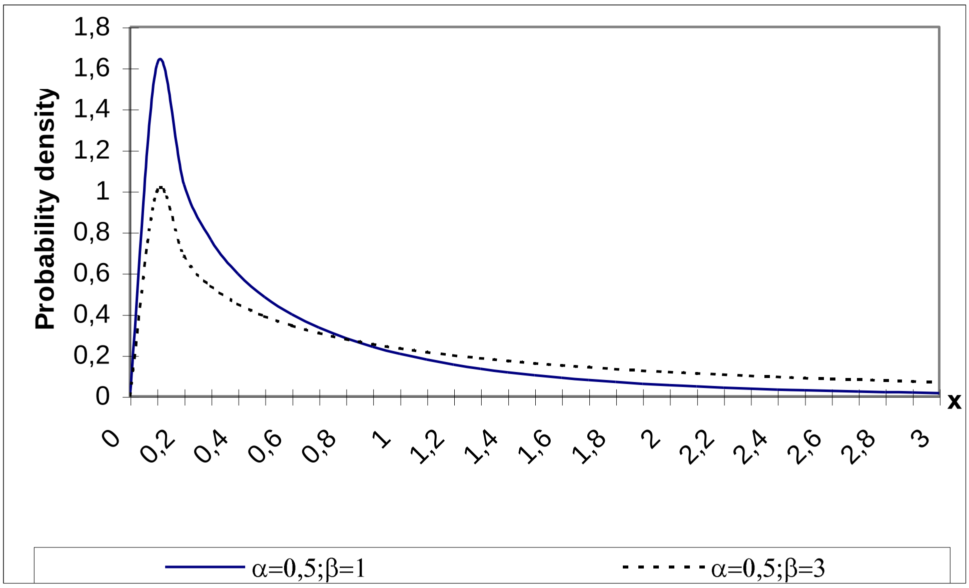


Fig. 2.31 Probability density of gamma distribution

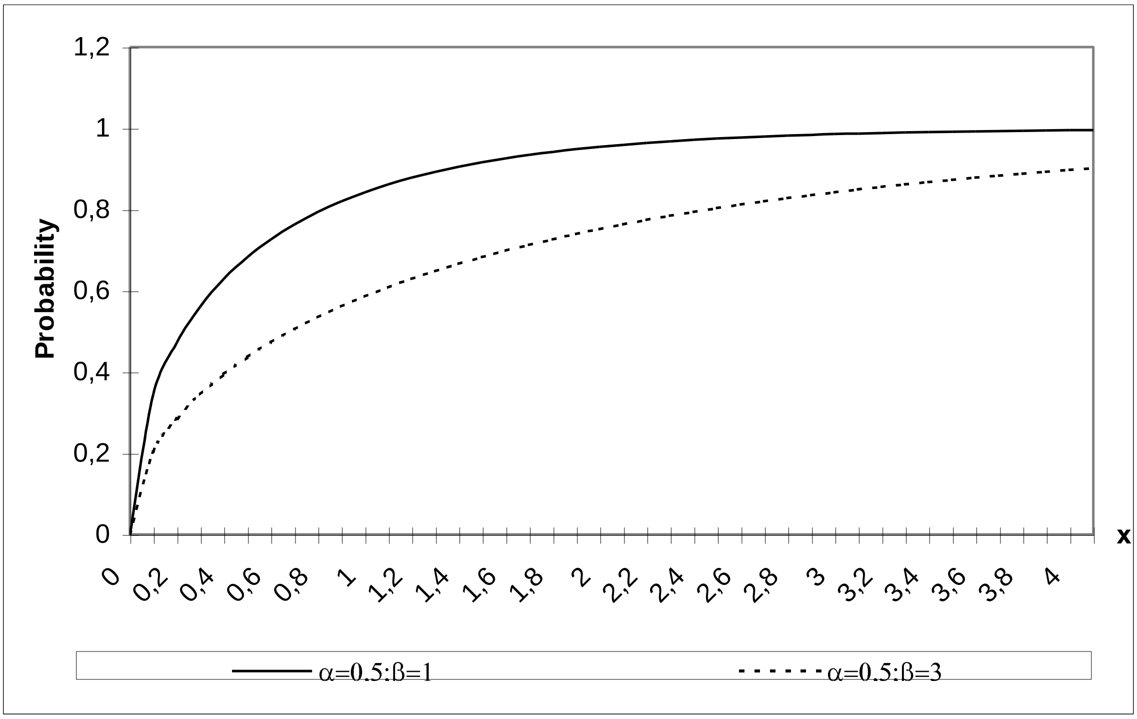


Fig. 2.32 Plot of gamma distribution function

2.4.17. Beta distribution

It is of vital theoretical importance. Density distribution is defined by formula:

$$f(x) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)x^{\alpha-1}(1-x)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}, \forall : 0 < x < 1; 0 < \alpha, \beta \\ 0, \text{во всех остальных случаях} \end{cases} \qquad (2.53),$$

where $\Gamma(z) = \int\limits_0^\infty x^{z-1} e^x dx, z > 0$. The plots of density probability and distribution function are given in Fig. 2.33 and 2.34. Frequently used to present the changes of a certain random variate in fractions (percents).

There are functions available in Excel — BETADISTR(*x*; α; β; A; *B*) and BETAINV(probability; α; β; A; *B*). A and *B* set the change interval of a variable. For standard beta distribution A=0 and *B*=1. BETAINV(probability; α; β; A; *B*) delivers *x* of distribution function corresponding to the specified probability.

For example, BETADISTR(0,1; 1; 3; 0; 1)=0,271 (Fig. 2.34) and BETAINV(0,271; 1;3;0; 1)=0,099999905.

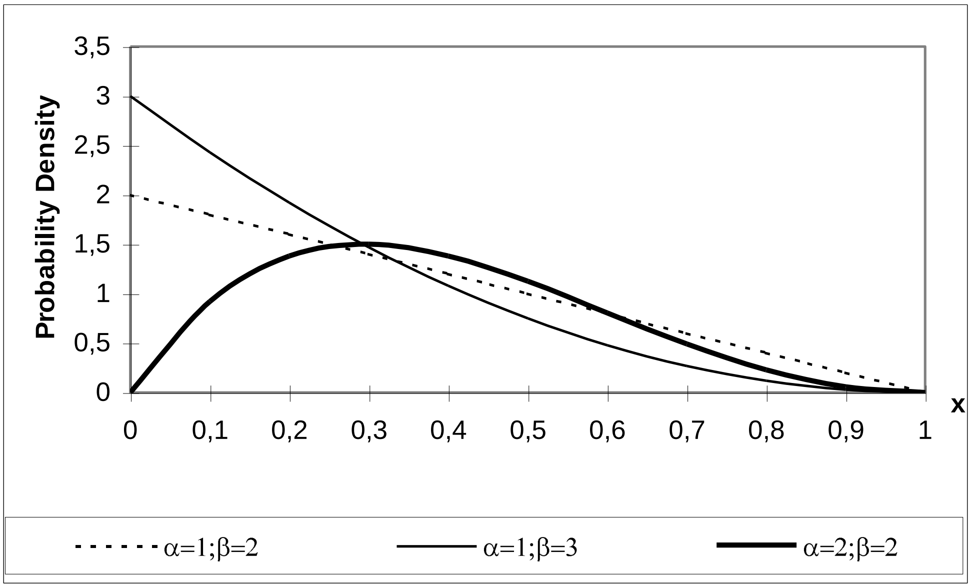


Fig. 2.33 Density of beta distribution probability

2.4.18. Weibull-gamma distribution

Represents random variate with probability density:

$$f(x) = \begin{cases} \dfrac{bkdx^{b-1}}{(x^b + d)^{k+1}}, x \geq 0; b, d, k > 0 \\ 0, x < 0 \end{cases} \qquad (2.54)$$

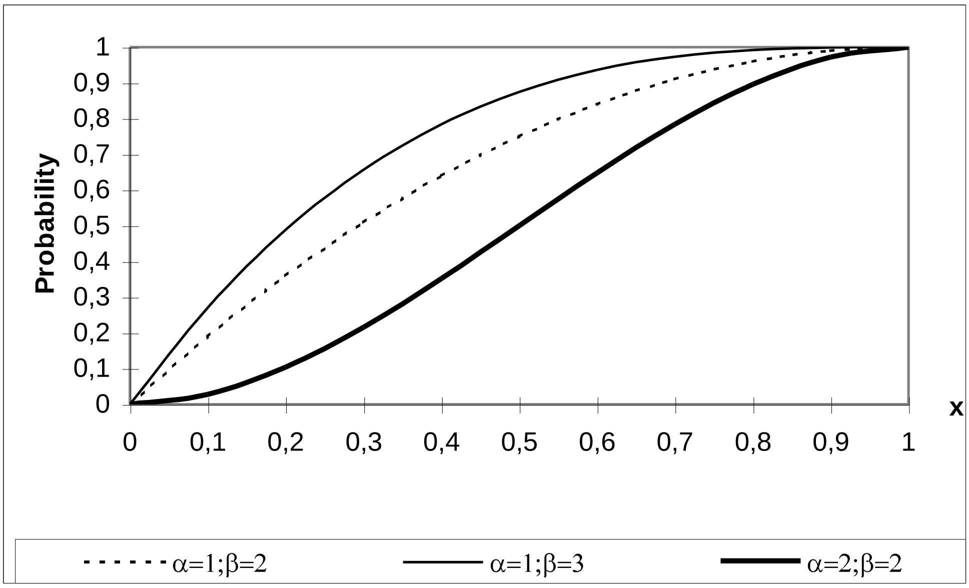


Fig. 2.34 Beta distribution function for different values of parameters

Described by three parameters. The plots of density probability for several combinations of parameters are given in Fig. 2.35.

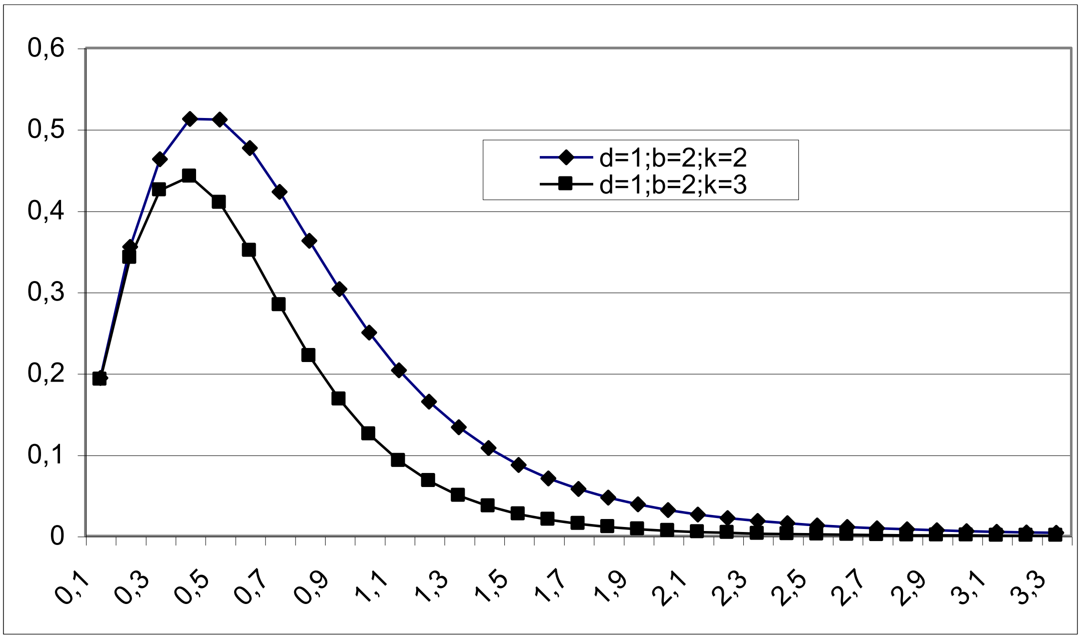


Fig. 2.35 Probability density of Weibull-gamma distribution

2.4.19. Weibull distribution

Quite widely used distribution with the density defined by formula:

$$f(x)=\frac{n}{x}\left(\frac{x-\mu}{a}\right)^{n-1}e^{\left(-\frac{x-\mu}{a}\right)^{n}},x\geq\mu \qquad (2.55)$$

Given *m*=0 and a=1 it is called Weibull standard distribution for which density function appears as:

$$f(x) = \begin{cases} 0, npu: x \le 0 \\ nx^{n-1}\lambda e^{-\lambda x^{n}}, npu: x > 0 \end{cases} \qquad (2.56)$$

Weibull distribution function appears as:

$$F(X) = 1 - e^{-x^{n}} . x \ge 0 \qquad (2.57)$$

Given *m*=0, we will have the so-called two-parameter Weibull distribution and given *n*=1 - distribution turns in exponential. So-called survival function is based on Weibull distribution.

$$F(X) = e^{-x^{n}}, x \ge 0 \qquad (2.58)$$

Distribution densities of Weibull standard distribution are presented in Fig. 2.36

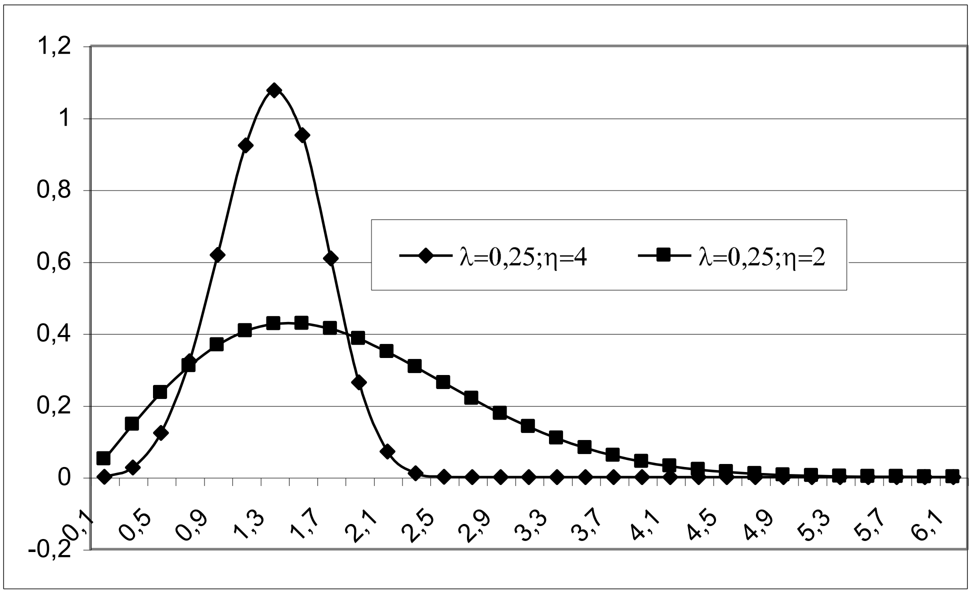


Fig. 2.36 Probability density of Weibull distribution

2.4.20. Cauchy distribution

Distribution density is defined by formula:

$$f(x) = \frac{k}{\pi(k^{2} + (x-m)^{2})}, k > 0; -\infty < x < \infty \qquad (2.59)$$

The plots of density probability for some values of parameters are given in Fig. 2.37

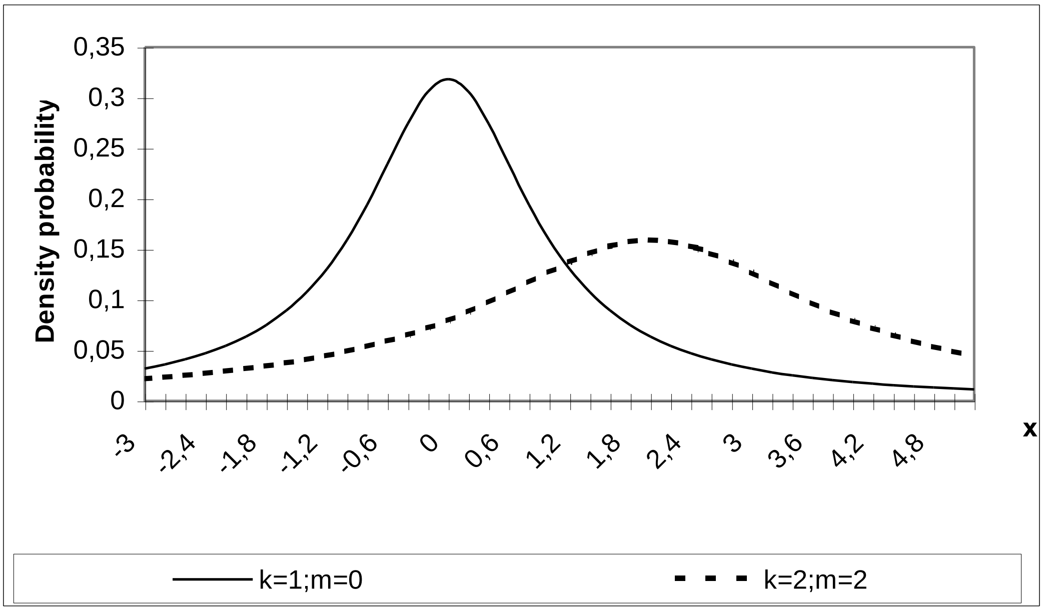


Fig. 2.37 Density probability of Cauchy distribution

2.4.21. Erlange distribution

Extensively used in the theory of mass service systems. Conventional sign is *X*~*P*(*n*, λ). Density distribution:

$$f(x) = \begin{cases} 0, npu: x \le 0 \\ x^{n-1}\lambda^{n}\dfrac{e^{-\lambda x}}{(n-1)!}, npu: x > 0 \end{cases} \qquad (2.60)$$

Given *n*=1 it turns into the exponential distribution. The plots of density of Erlange distribution for some values of parameters are given in Fig. 2.38

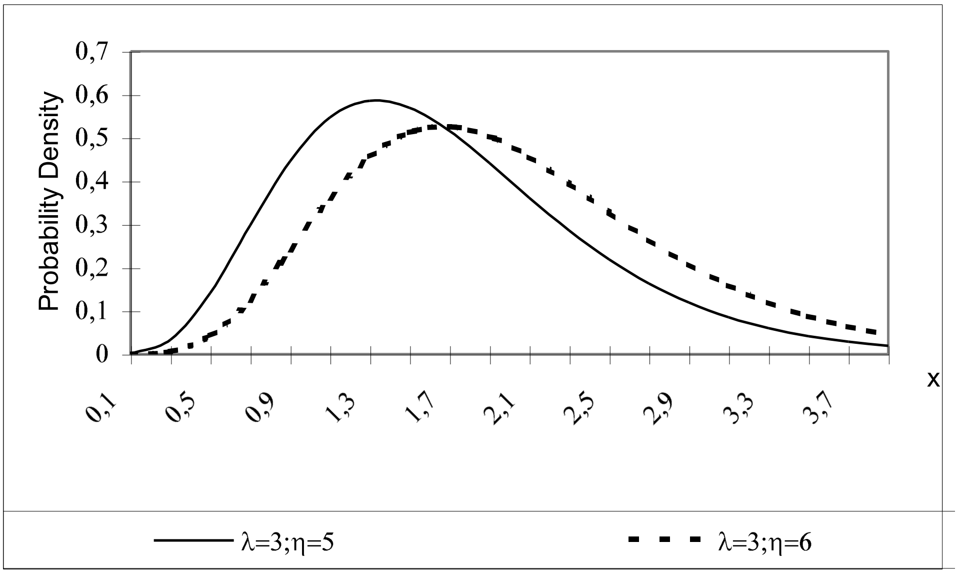


Fig. 2.38 Density probability of Erlange distribution

The following density formula can often be encountered $f_k(x) = \frac{\lambda(\lambda x)^k}{k!} e^{-\lambda x}$.

In this case it is called Erlange distribution of *k*-th order. In mass service systems this distribution law is convenient because any degree of consequence can be set according to the order (*k* value): from total absence given *k*=0 to functional dependence between periods of requests given *k*=∞.

2.5. Empirical distribution law

In most cases, solving real-world problems distribution law and its parameters are not known. Thus, to determine the type of a distribution law, one should fulfill a set of actions on the analysis of the received original data.

Histogram

It is used to represent an interval series: intervals are plotted on the abscissa and rectangles with the height equal to the frequency (sales volume) or relative frequency (sales rates from the total volume) of the corresponding interval are constructed on these segments (Table 2.8, Fig. 2.39)

Table 2.8

Observed values of the random variate

Random variate change interval	**Realization frequency of the variate entering this interval**
0-12	4
12-24	14
24-36	32
36-48	48
48-60	44
60-72	33
72-84	15
84-96	5

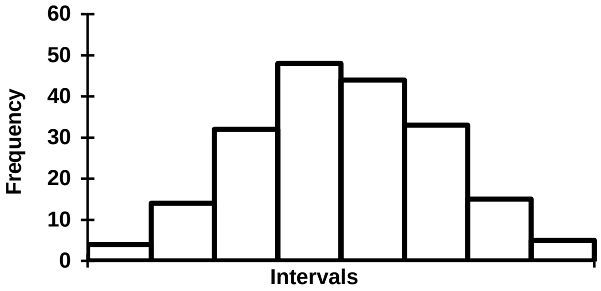


Fig. 2.39 Histogram

If on *X* axis set the interval width equal to 1 and on Y axis — the interval width equal to one observation, then the histogram area will be equal to the number of observations for frequencies and 1 for relative frequencies.

Polygon of frequencies

Histogram analog for discrete distribution. Values of magnitudes are plotted on the abscissa and frequencies or relative frequencies are laid off as an ordinate. Obtained points are joined with a polygonal line. Polygon can be constructed for a continuous distribution and in doing so, the middles of the upper segments of rectangles in the histogram are joined with a polygonal line. (Fig. 2.40)

Constructing histogram, the choice of intervals is of a paramount importance since the distribution form will depend on this choice. There are design formulas to calculate an interval size but as a rule, it is supposed that there should be 12—15 intervals and each interval should include no less than 5—6 realizations of magnitude (in the general case, intervals can be of a different size).

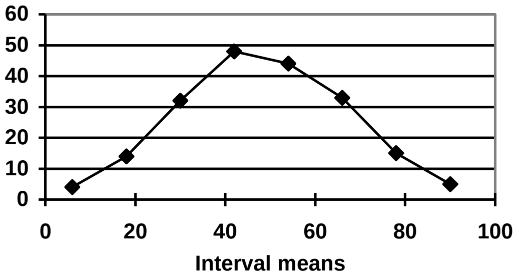


Fig. 2.40 Polygon of frequencies for continuous distribution

Cumulative curve

Intervals are plotted on the abscissa and on the ordinate is plotted the number or part of the elements that have smaller or equal value as compared to the specified number. In Fig. 2.41 the cumulative curve is given for the sample presented in the histogram (Fig. 2.39)

As can be seen from the Fig. 2.41, increasing the sample size to infinity the histogram turns into the plot of density distribution and the curve — into the plot of distribution function.

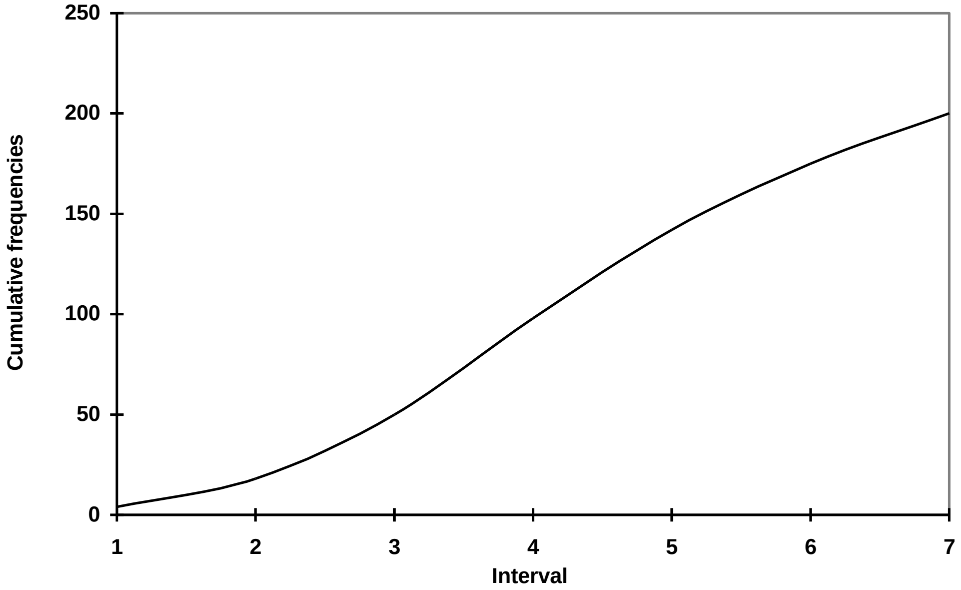


Fig. 2.41 Cumulative curve

Determining random variate distribution law

In classical mathematical statistics much attention is attributed to the determination of type and parameters of random variate distribution law. There are various criteria and software to test respective hypotheses. In practical investigations this problem requires great efforts and its solution is problematic for many reasons: different criteria produce different results; change of intervals may change conclusions; empirical distribution can be a mixture of different distributions or can be “loaded” as compared to theoretical distribution. For an experimenter there is a need to take up this problem only in the following cases:

- The methods he/she employs as prerequisites require specific distribution law;
- The problem under solution (for example, imitation modeling) requires knowledge of a type and parameters of distribution law.

In the first case, it is usually restricted to general simple checks, sufficient to make a decision (or non-parametric methods are used since they do not require the knowledge of distribution law). In the second case, empirical distribution law is approximated with Pearson's functions.

Approximating distribution by application of Pearson's functions, density distribution appears as:

$$\frac{1}{f(x)}\frac{df}{dx} = -\frac{a+x}{c_0 + c_1 x + c_2 x^2} \qquad (2.61)$$

If set the origin of coordinates in the point which corresponds to average value and designate as:

$$d=2(5b_2 - 6b_1 - 9), \qquad (2.62)$$

then, constants can be calculated by such formulas:

$$a = \frac{\sqrt{m_2}(b_2+3)\sqrt{b_1}}{d} \qquad (2.63)$$

$$c_0=m_2(4b_2-3b_1)/d \qquad (2.64)$$

$$c_1=a \qquad (2.65)$$

$$c_2=(2b_2-3b_1-6) \qquad (2.66)$$

Here $\sqrt{b_1} = {m_3}/{(\sqrt{m_2})^3}$ (2.67) — asymmetry, $b_2 = {m_4}/{m_2^2}$ (2.68) — эксцесс.

m_i are central distribution moments. They are calculated from the initial distribution moments by the following general formula:

$$m_k = m_k' - C_k^1 m_1' m_{k-1}' + C_k^2 m_1'^2 m_{k-2}' - ... + (-1)^k m_1'^k \qquad (2.69)$$

For the moments that are of interest to us these formulas assume the form:

$$m_2 = m_2' - (m_1')^2 \qquad (2.70)$$

$$m_3 = m_3' - 3m_2' m_1' + 2(m_1')^3 \qquad (2.71)$$

$$m_4 = m_4' - 4m_1' m_2' + 6m_1'^2 m_2' - 3(m_1')^4 \qquad (2.72)$$

Initial moments are defined in the following manner:

$$m'_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \quad (2.72)$$

The initial moment of the first order is the average and the central moment of the second order is the variance.

After solving differential equation we will have the formula for density distribution of the given empirical distribution. According to the combination of parameters, Pearson divided them in 12 groups that are designated in Roman figures, for example distribution of the type XI.

You may find the examples of solution in [16].

In many cases it will suffice to make sure whether the random variate distribution law is normal. There are rigorous criteria of this check but they usually require no less than 50-100 values for reliable conclusions. Because of this, we can use the simplified check.

It is suggested that the random variate is distributed according to the normal law if the following conditions, that are the consequences from normal distribution law, can be met. To perform a check absolute average deviation is found:

$$\Delta_{abs} = \frac{\sum_{i=1}^{N} \left| X_i - \overline{X} \right|}{N}. \qquad (2.74)$$

Then, the fulfillment of the following conditions is tried:

- The number of positive and negative deviations from the mean is approximately equal.
- The half (or a little more) of the deviations from the mean on the absolute value is less than the absolute average deviation $\sigma^2 < \Delta_{abs.}$
- None of the deviations exceeds average absolute deviation in more than 3 … 4 times: $\varepsilon_{i\ max} <$ 3 … 4Δ_{abs}.

Other possible methods of checking are also suggested. For example, it will suffice to try the fulfillment of the following condition:

$$\left| \frac{\Delta_{abs}}{S} - 0{,}7979 \right| < \frac{0{,}4}{\sqrt{N}}. \qquad (2.74)$$

For those who do not want to calculate absolute average deviation, there is the following set of conditions to try.

- Almost all deviations (99,7%) from the mean are less than three sigma : $\varepsilon_i < 3\ \sigma$.
- Two third of deviations (68,3%) is less than σ.
- Half of deviations are less than $0{,}625\sigma$.

If all these conditions are met it may be considered that the normal distribution hypothesis does not contradict the available data.

In the specialized literature there are recommendations on the actions that should be undertaken for the distribution to become normal:

- Functional transformations;
- Censoring in sampling.

As the transformations it is usually proposed to calculate a logarithm, arc sine, square and cube roots from the original magnitudes. Using these transformations, it should be borne in mind:

- The transformed magnitude should have a natural sense. For example, if the estimated values differ in tens of times, the use of taking logs (or transformation of the type of the square root for values that are areas) is intuitively thought as being natural.
- Conclusions you will make on the basis of the further analysis relate not to initial values but to their functions.

When censoring, the least and the largest values are usually rejected. It is well to bear in mind that in doing so “emasculation” of the initial problem, its idealization is taking place. According to modern ideas, normal distributions with “heavy tails” are in wide use and in contrast to the standard normal distribution the frequency of the occurrence of maximum and minimum values for them is appreciably high.

It is worth remembering that statistics is based on the sampling method when not the whole population is studied but only its certain sample. The truth and validity of this approach is based on the set of theorems. One of the most important theorems is Chebyshev theorem. It contains the conditions to be met for its use:

1) pairwise independence of random variates;
2) identical mathematical expectation in these random variates;
3) Even (uniform) limitation of variances of these random variates.

If at least one of the listed conditions is not met conclusions will be invalidated.

2.6. Random variate characteristics

> A common man in Society is like the center of gravity in a physical body; considering this central point we come to understand all phenomena of equilibrium and motion
>
> Adolf Ketle

In statistics, analyzing data the following problems may be come across:

1. ***How much*** data to select and ***in what way*** they should be selected.

2. ***Rightfulness*** of the distribution of conclusions that were made on the basis of sample data over the whole population

3. ***Selecting optimal methods*** of estimation.

4. Selecting methods of data ***generalization***, ***classification*** and ***presentation***.

The researcher ought to always remember about these problems, as he/she is responsible for the results of investigations and conclusions that can affect welfare, health and life of many people.

2.6.1. Properties of parameter estimates

Parameter estimates should conform to the following requirements.

Unbaisedness. It means that conduction very big number of trials with samples of the identical size, the average value of every sample tends to the true value of the population. Baisedness is usually caused by the presence of a systematic error.

Consistency. Increasing the size of a sample, the estimate should tend to the value of the corresponding parameter of the population with the probability tending to 1.

Efficiency. The selected estimate should have minimum variance for the sample of the same size.

Sufficiency. The estimate should contain all necessary data and require no additional information.

When working out estimates certain prerequisites are proposed. Thus, as a rule estimates comply with the aforesaid requirements only when fulfilling this prerequisites. It is well to bear in mind when using estimates.

To estimate parameters different methods are made use of. The maximum likelihood method ranks first among all others. It is employed when distribution law is known. The essence of the maximum likelihood method is that estimates should be equal to values with which a sample has maximum probability to occur.

Characteristics of univariate distribution consist of:

1. Measures of location (median; mean; mode; and others).
2. Measure of scattering (range, coefficient of variation, variance, standard deviation).
3. Measures of form (asymmetry, excess, methods of third and fourth order).

2.6.2. Arithmetic mean (sample)

$$\overline{X} = \frac{\sum_{i=1}^{N} x_i}{N} \qquad (2.76).$$

Properties of the mean (sample)

- The sum of deviations from the mean is 0.
- If all sample values add or subtract, multiply or divide in the same number, the mean value will change accordingly.
- As the number of measurements grows, the accuracy of estimations increases and the mean approaches to the expectation but only when there are no systematic errors and observations are independent.

- The mean of the sum of two samples is equal to the sum of their means (analogously for difference):

$$\overline{X+Y}=\overline{X}+\overline{Y}. \qquad (2.77)$$

- If the series of observation consists of *k* groups the arithmetic mean of the whole series is equal to the weighed grouping mean. In this case, the weights are sizes of groups[1]:

$$\overline{X}=\sum_{i=1}^{k}X_i n_i / \sum_{i=1}^{k} n_i , \qquad (2.78)$$

where *n* is the size of i-th group, X_i is the mean of the i-th group.

Some discouraging remarks.

- The mean does not necessarily signify "typical". For example, the average income is not typical at all.
- The mean does not coincide with the expectation. With the exception of normal distribution, the arithmetic mean is not even and unbiased estimate of the expectation with the least variants. Moreover, even for normal distribution it is impossible to set the estimate which will be closer to the expectation (though without some effective properties of the mean).

2.6.3. Geometric mean (sample)

Geometric mean is used if:

- The variable changes in time with constant ratio between its measurements $\frac{X_{i-1}}{X_i}=\frac{X_{i+1}}{X_i}=\text{const}$ (for example, the growth of the number of bacteria, capital surplus, operational costs and others).
- Certain values in the sample are at a long distance from each other (for instance, differ for the order). Geometric mean is calculated from the formula:

$$\overline{X}_G=\sqrt[n]{\prod_{i=1}^{n}X_i} . \qquad (2.79)$$

[1] Such formula is usually used in cases when we deal with grouped data, for example, intervals of income (from and to) and that part of the population which have these incomes. Then the middle of the interval is though to be the mean and we can make calculations of the mean for the whole sample.

2.6.4. Harmonic mean

In a number of cases (for example, average life expectancy estimation, average speed determination) harmonic mean is used:

$$\overline{X}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}} \cdot \qquad (2.80)$$

2.6.5. Mode (modal value)

It is a value which is observed for the most number of times (the most probable quantity). For the interval variation series, it is calculated by formula:

$$\text{Mo} = \text{X}_{\text{Mo}} + \text{h}\ (\text{m}_{\text{Mo}} - \text{m}_{\text{Mo-1}})/(2\text{m}_{\text{Mo}} - \text{m}_{\text{Mo+1}} - \text{m}_{\text{Mo-1}}), \qquad (2.81)$$

where X_{Mo} is the origin of the modal interval (the one that has the biggest frequency); *h* is the quantity of the modal interval; m_{Mo} is the frequency of the modal interval; $m_{\text{Mo -1}}$ is the frequency of the interval prior to the modal one; $m_{\text{Mo +1}}$ is the frequency of the interval following after the modal one.

Remember that the mode is not applicable if distribution is multimodal (multiapexed).

2.6.6. Median (sampling).

It is a value which divides ranked variation series into two equal groups. Variation series is ranked.

If the number of the series members is odd, the median is a series value located in the middle, that is the element with the number (*N*+1)/2. .If the number of the series member is even, the median is equal to the mean of the series members with the numbers *N*/2 and *N*/2+1. For example, 4; 5; 6,7; 8; 12 the value 6,7 will be the median for the interval variation series, the median is calculated by formula:

$$\text{Me} = \text{X}_{\text{Me}} + \text{h}\ (\Sigma \text{m}_{\text{X}}/2 - \text{m}^{\text{max}}{}_{\text{X}})/\ \text{m}_{\text{m}}, \qquad (2.81)$$

where X_{Me} is the origin of the median interval; m_{X} are frequencies on all intervals; $m_X^{\max}$ is the frequency cumulated to the origin of the median interval; m_{m} is the median interval frequency.

Median interval is an interval which includes the median value.

Properties of median

The sum of absolute values of the variants of deviations from the median multiplied in the corresponding frequencies is less than from any other value[2]:

$$\Sigma|x - Me|\, m_x < \underset{\forall x}{\Sigma}|x - a|\, m_x. \qquad (2.83)$$

- The median is not influenced by the change of variation series extreme values if only the smaller of the median remains smaller and the larger continues to remain larger than the median[3]

2.6.7. Variation exponents

Variation range

$$\boldsymbol{R_B = X_{max} - X_{min}} \qquad (2.84)$$

Unreliable since it is influenced by extreme values. It does not change whatever the changes of the variation series may be if these changes do not concern extreme values.

Empirical variants

$$S^2 = \sum_{i=1}^{N}(X_i - \overline{X})^2 \Big/ (N-1) \qquad (2.85)$$

Properties of variance

- Variance of constant quantity is 0.
- If increase or decrease all results by common number, the variance will not change.
- If change all results in *K* times the variance will change in K^2 times[4].
- $S^2 = X^2 - (\overline{X})^2$ (2.85).

The square root of the variance presents standard deviation.

[2] This property of a median can be used to select, for example, the location of city transport stops, fuel stations, etc.

[3] In connection with this property the median is used instead of the mean if variation series extreme values differ radically from others.

[4] The fact that in time of inflation the poor become poorer and the rich – richer is just the variance property in terms of statistic.

Some remarks on the issue about *N* or *N*-1.

One of the most frequently encountered question is: by what should the sum of squares be divided in the variance formula, by *N* or *N*-1?

Generally, the difference between these estimates is not great but it can be important when the expectation is known. The unbiased estimate of variance is:

$$S^2 = \sum_{i=1}^{N}(X_i - M(X))^2 \Big/ N \qquad (2.86)$$

But the sum of squares in the formula numerator is minimal only when expectation is equal to the mean. Yet, this is not so at all. So, Bessel correction is needed. It brings about *N*-1 in the denominator. In practice, it is often done in the following way: if the expectation is known or its estimate is obtained from another sample, the formula with *N* is used. In case, when the mean average estimate is obtained from the same sample as the variance *N*-1 is used.

Coefficient of variation

$$V = \frac{S}{\overline{X}}100\% \quad (2.87)$$

In case when the frequency polygon of the variation series has no considerable skewness and all series are positive members are positive then V < 30%. If the coefficient of variation is more than 100% it means that data inhomogeneous.

2.6.8. Confidence interval

Confidence interval is an interval which can be asserted to contain the unknown value of θ parameter with the primarily specified probability P=1-α:

$$P[\theta_1 < \theta < \theta_2] = 1 - \alpha, \qquad (2.88)$$

where 1–α is the confidence probability, α — significance level.

Properties

- Increasing the number of measurements the accuracy improves. It holds true only when there are no systematic errors and observations are independent.

- Increasing reliability with the fixed sample leads to increasing the confidence interval and decreasing the accuracy.

If we increase the number of measurements, the estimate of parameter becomes more accurate and confidence interval decreases. It does not relate to the situations when measurements are dependent or there are no systematic errors in them. In such cases increasing the number of measurements, the accuracy is likely to decrease instead of getting higher. Let's take up a simple example.

It's known that the time averaging indicated on the watches of a big number of randomly taken people provides sufficiently exact time. Assume, in the sample we used to perform the averaging, most people (just before the check) set the time on their watches in keeping with the watches of one of the respondents. In this case measurements will not be independent and the mean will shift. As the consequence the calculated value of the confidence interval will not represent the fact. Conducting the analogous survey on the day when the shift between summer and winter time takes place, the measurements will contain the systematic error because most people will not reset their watches.

Lets given example of the inconsistency between confidence interval and actual value. The measurements of the earth where performed by means of the ground equipment (in the process of the development of scientific and technological advance with more and more accuracy). But when the measurements were carried out with the help of the artificial satellites of the earth it turned out that earth dimensions were outside the province of the confidence intervals. There were systematic errors in the design procedure and they could not be undone by increasing the number of measurement.

Confidence interval does not mean the probability of the value of the estimated parameter to enter into the limits of certain boundaries, but the fact that if we take sufficient number of samples in 100 *x p*% of cases, the parameter will be in the specified interval.

Significance level is usually selected within the interval from 0,01 to 0,05 where 0,05 are usual requirements of reliability, 0,01 — excess requirements, 0,001 — very high requirements, 0,1 — reduced requirements of reliability.

Confidence interval for the mean

Defined by formula:

$$\left[\overline{X} - t_{n,p}\frac{S}{\sqrt{n}}, \overline{X} + t_{n,p}\frac{S}{\sqrt{n}}\right], \qquad (2.90)$$

where S is a standard deviation, *n* is the number of trials, $t_{n,p}$ is the tabular value of students distribution within the number of degrees of freedom *n* and confidence probability *p*.

This formula is applied when the variance is not known and its estimate on experimental data is used. If the variance is known, the other formula is used (not given here since the probability of the occurrence of such situations is infinitesimal).

Confidence interval for standard deviation

$$\mathrm{P}[\sqrt{\mathrm{n}}\mathrm{S}/\chi_2 < \sigma < \sqrt{\mathrm{n}}\mathrm{S}/\chi_1] < 1 - \alpha\ , \qquad (2.91)$$

where χ_1 and χ_2 are quantiles of distribution χ^2 (chi-square).

Quantiles χ_1 and χ_2 have (*n*-1) degrees of freedom each and significance level 1-α/2 and α/2 respectively.

Notes.

- When several parameters are determined by one sample, the confidence intervals will be larger than the expected ones. It has to do with the fact that statistics used to create confidence intervals of parameters (for example, the mean or variance) are not independent.
- Given formulas for confidence intervals have been deduced on the basis of normal data distribution.

2.6.9. Notions about parametric, non-parametric and robust statistics

It is common knowledge that testing a hypotheses one should lean upon a certain collection of suggestions that are used to develop formulas necessary for this very test. Besides, along with other there are always suggestions about the sample distribution law. It is understood that the failure to fulfill these prerequisites makes the application of the corresponding methods incorrect.

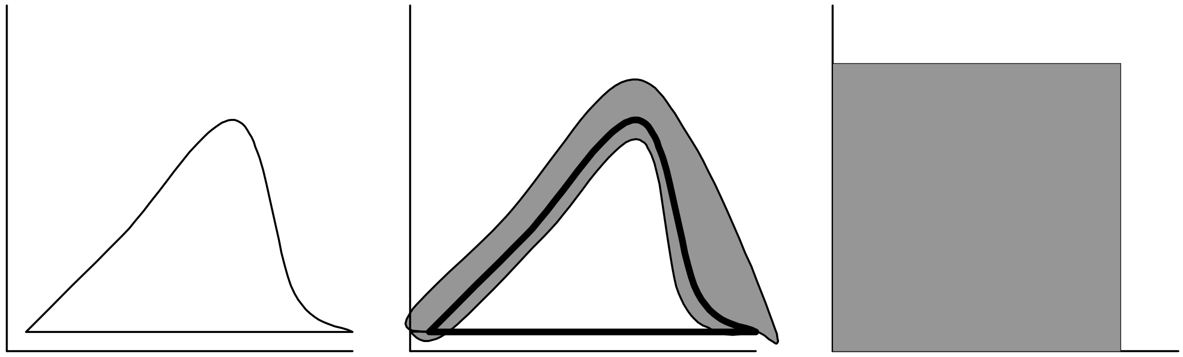
Parametric methods suggest specific distribution with the definite parameters (Fig. 2.42a). Practically, all traditional statistical tests and procedures relate to this group. They are usually well-defined and thoroughly studied. The problem is that in the overwhelming majority of practical problems, the prerequisites of normality either are not fulfilled or their fulfillment is impossible.

Robust methods also suggest specific distribution but allow for the deviation from it (Fig. 2.42b). The form and quantity of these deviations depend on the concrete methods. For all types of problems, the developed robust methods are available.

Non-parametric methods do not make specific suggestions about distribution law. They give only most common notions (Fig. 2.42c). For example, the sample has continuous distribution law or both samples have one distribution law. Usually, they are strictly substantiated. For many problems especially for multivariate ones, there are no proper non-parametric methods.

Data analysis is usually represented by some heuristical procedure, developed to solve specific problems. Their validity is leant upon logic and computing experiment with specially designed artificial data set.

The selection of methods is carried out according to the purpose of the research and peculiarities of the available data.



a b c

Fig. 2.42. Comparison of prerequisites of different methods

2.7. Examples of calculations and plotting.

The learning itself gives very general guidelines, if the are not specified by experience.

Francis Beacon

2.7.1. Calculating the average and median, comparing their stability

Let's consider the example of the average value and the median calculations. In Table 2.9 the prices of a drug in different trading organizations are given.

Table 2.9

Organization	**A**	***B***	**C**	**D**	**E**	**F**	**G**	**H**
Price	100	110	115	125	140	145	145	150

Average price value is (100 + 110 + 115 + 125 + 140 + 145 + 145 + 150)/8=128,75.

Ranking series {100; 110; 115; 125; 140; 145; 145; 150}

Price median (125 + 140)/2=132,5 (the number of members of series is even).

The location of the average (square) and median (triangle) in the data series is shown in Fig. 2.43.

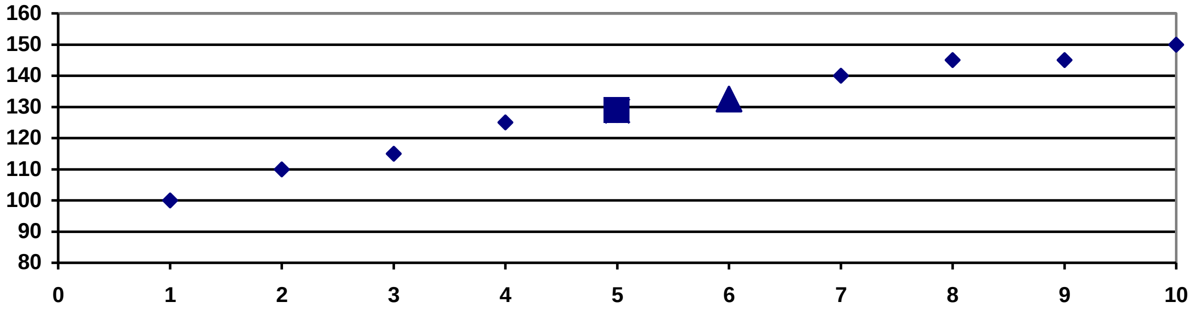


Fig. 2.43. Location of the average and median

Assume we have prominent data (Table 2.10).

Table 2.10

Organization	**A**	***B***	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Price	100	110	115	125	140	145	145	150	450

Average price value is (100 + 110 + 115 + 125 + 140 + 145 + 145 + 150 + 230)/9=128,75.

Ranking series {100; 110; 115; 125; 140; 145; 145; 150; 230}

Price median 140 (the number of members of series is odd).

The location of the average (square) and median (triangle) in the data series is shown in Fig. 2.44.

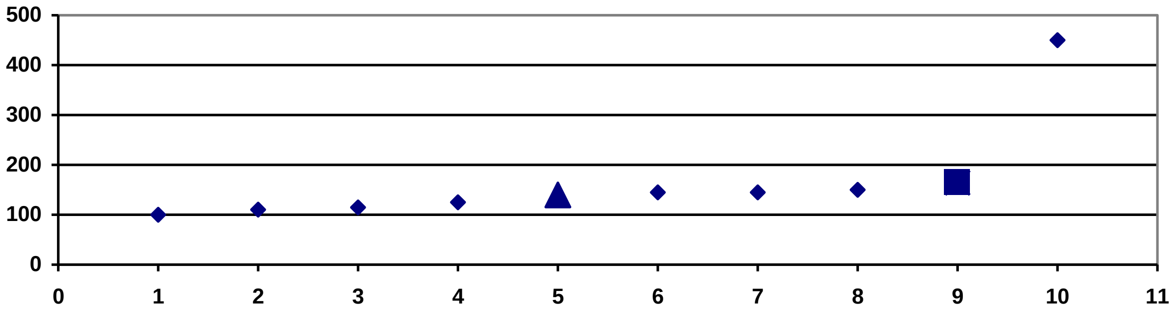


Fig. 2.44. Location of the average and median

This example shows that in the presence of prominent data or the data differ considerably the median is not more stable estimate then the average value.

2.7.2. Calculating the average, mode and median for grouped data

This problem is typical for medical statistics and marketing. Assume, there is a Table 2.11, which needs the average value, mode and median of the patients' age to be calculated.

Table 2.11

Age	20—29	30—39	40—49	50—59	60—69
Number of patients	45	36	175	361	825
Cumulated frequencies	45	81	256	617	1442

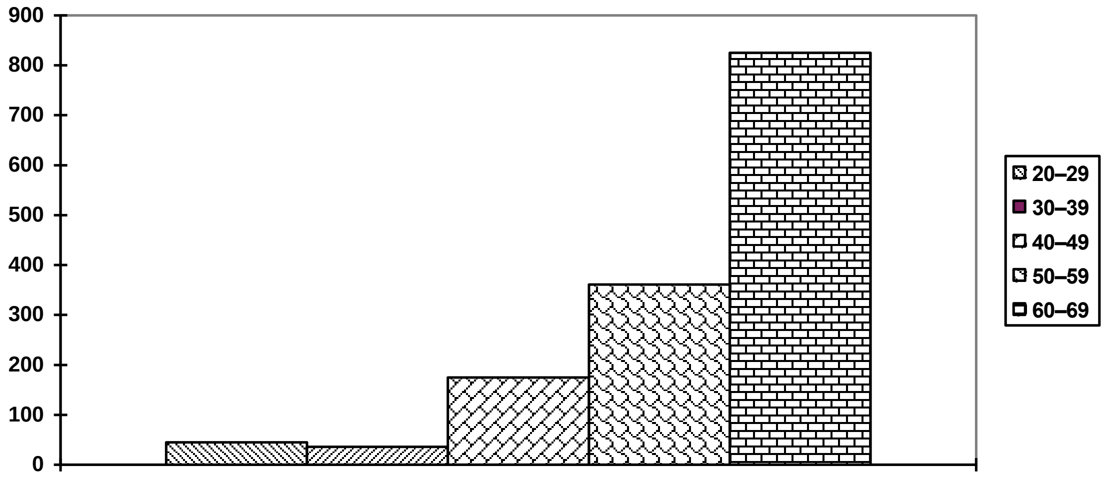


Рис. 2.45. Гистограмма для данных таблицы 2.11

Average age value of a patient.

X_{avg}=(24,5*45+34,5*36+44,5*175+54,5*361+64,5*825)/(45+36+175+361+825)=58,57.

Median is calculated by formula: $Me = X_{Me} + h\ (\Sigma m_X /2 - m^{max}{}_X)/\ m_m$.

X_{Me}=60 is the beginning of the interval which contains the median. Since the sum of frequencies 1442 the half (the median divides into two) is 721. In the beginning of the interval 60-69 only 617 is cumulated. Thus, it is clear that the median is in the interval.

The width of the median interval is h=69-59. 59 is subtracted as 60 is a part of the interval.

The total cumulated frequency is divided into $\Sigma m_X /2$ =1442/2

The frequency cumulated by the beginning of the median interval is $\text{м}^{max}{}_X$=617

The frequency of the median interval is m_m=825.

Hence, the median is M_e=60+10*(1442/2-617)/825=61,26.

The mode is calculated by formula: $Mo = X_{Mo} + h\ (m_{Mo} - m_{Mo-1})/(2m_{Mo} - m_{Mo+1} - m_{Mo-1})$.

The beginning of the modal interval is X_{mo}=60. Modal interval is an interval which has the biggest frequency.

The width of the modal interval is h=69-59

The frequency of the modal interval is M_{Mo}=825.

The frequency of the interval prior to the modal is m_{Mo-1}=361 and m_{Mo+1}=0.

Since it is the frequency of the interval following after the model, the modal interval is the last.

Mode is X_{Mo}=60+10*(825-361)/(2*825-0-361)=65

2.7.3. Calculation variation exponent

Let’s consider table 2.12 and calculate variation exponents of its data.

Variation range is R=91,2-70,1=21,1

Average=76,32

Variance S^2=((75,7-76,32) ^{2} + (70,1-76,32) ^{2} + (91,2-76,32) ^{2} + (70,7-76,32) ^{2} + (71,4-76,32) ^{2} + (78,8-76,32) ^{2})/(6-1)=((-0,62) ^{2}+(-6,22) ^{2}+(14,88) ^{2}+(-5,62) ^{2} +(-4,92)^2 + (2,48)^2) / 6 = (0,3844+38,6884+221,4144+31,5844+24,2064+6,1504)/5=64,13.

Standard deviation — 8,008.

Coefficient of variation V=(8,008/76,32)·100%=10,49%

2.7.4. Calculating confidence intervals

Using data from table 2.12 ad calculated values of the average and standard deviation

Let’s construct confidence intervals for them.

Table 2.12

Patient	A	*B*	C	D	E	F
Blood hemoglobin contents, mc mole	75,7	70,1	91,2	70,7	71,4	78,8

Confidence intervals for average:

$\left[\overline{X}-t_{n,p}\frac{S}{\sqrt{n}},\overline{X}+t_{n,p}\frac{S}{\sqrt{n}}\right]$=[76,32-2,45*8,008/2,45;76,32-2,45*8,008/2,45] =[68,312;84,328].

In the numerator 2,45 is a tabular value of Student’s t-test with a number of degrees of freedom 6 and confidence probability 0,95. In the denominator 2,45 is a square root of 6.

Confidence interval for standard deviation:

$$[\sqrt{n}S/\chi_2<\sigma<\sqrt{n}S/\chi_1]<1-\alpha\cdot$$

If we set significance level 0,5, critical distribution value of chi-square with the significance level 1-0,05/2 and the number of degrees of freedom (5-1) χ_2=0,831. Critical chi-square distribution value with the significance level 0,05/2 and the number of degrees of freedom (5-1) χ_2=12,832.

Then, the confidence interval appears as (2,45*8,008/12,832; 2,45*8,008/0,831)=(1,53; 23,61).

As you can see confidence intervals are big enough. It is because of big variance (data spread). To provide for the specified probability (0,95) of the average and standard deviation to enter into the assigned interval, the latter should be increased. That is what we may witness.

2.7.5. Determining parameters using electronic worksheet Excel

Most of the described parameters and characteristics (with the exception of the grouped data) can be calculated through the agency of the electronic worksheet Excel. In table 2.13 not all functions relative to distribution laws are given, since every chapter includes only those functions that are primarily used in it. For more information, see subject index at the end of the book.

Table 2.13

Description of some functions

Name of option	Name of function	Option description
Average	AVERAGE	List of values or cell name interval
Median	MEDIAN	List of values or cell name interval
Mode	MODA	List of values or cell name interval
Variance	VAR	List of values or cell name interval
Standard deviation	STDEV	List of values or cell name interval
Geometrical mean	GEOMEAN	List of values or cell name interval
Half-width of confidence interval	CONFIDENCE	(alfa, mean square; number of trials) alfa – significance level, usually 0,5.
Harmonic mean	HARMIN	List of values or cell name interval

Excess	**EXCESS**	**List of values or cell name interval**
Asymmetry	**SKEW**	**List of values or cell name interval**
Absolute deviation average from the mean	**MEANDEV**	**List of values or cell name interval**
Standard deviation by general population	**STDEVPA**	**List of values or cell name interval. Data should be general population.**

Lets consider the example of the use of main functions. The screencopy with the example is shown in Fig. 2.46 in the cells from C2 to H2, the source data are located. In the cells from G5 to G9 and H9, the functions to calculate necessary parameters are given (Table 2.14).

Table 2.14

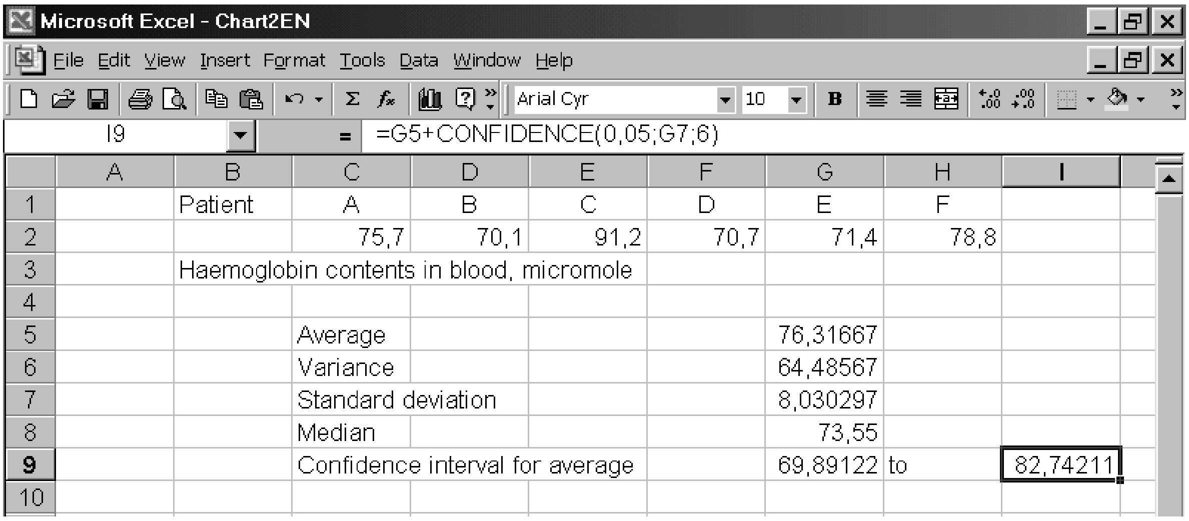
Cell name	**Contents**
G5	=MEANVAL(C2:H2)
G6	=VAR(C2:H2)
G7	=STDEV(C2:H2)
G8	=MEDIAN(C2:H2)
G9	=G5-CONFIDENCE(0,05;G7;COUNT(C2:H2))
H9	=G5+CONFIDENCE(0,05; G7; 6)

You can see that in H9 in the function CONFIDENCE(0,05; G7; 6) are the number of trials is obviously indicated while in G9 in the similar function CONFIDENCE(0,05; G7; COUNT(C2:H2). To determine the number of trials the function count (C2:H2) is used. Its parameters are the rage of the used cell and the result is the number.

Assume, we have data to construct a histogram (Fig. 2.47). First, we should set the intervals in which frequencies of the occurrence of a random variate will be counted. In this case we selected identical intervals with four units in length. The boundaries of the intervals are located in the column G (Fig. 2.48).

That is we have the intervals (0,24):(25,28) and so on. After that we enter the formula =FREQUENCY(B3:F17; G3:G17). Its first parameter represents the field with source data, the second one represents the fields with the right interval boundaries. After entering the formula the only thing to stretch the cell H3 to H12. These cells contain corresponding frequencies of the intervals. The result is shown in Fig. 2.49.

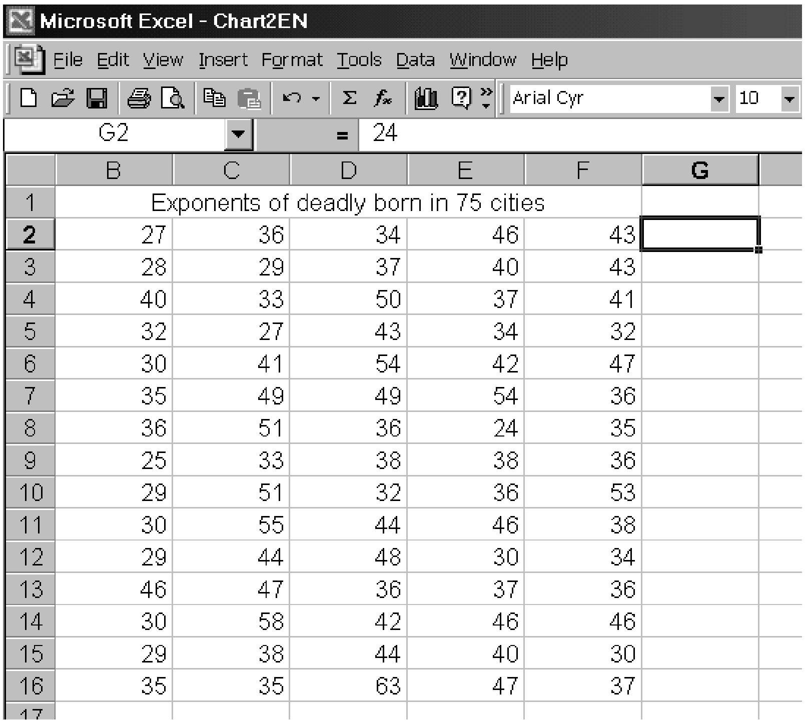
It will not take much effort to construct graphic presentation on the frequency column. It is possible to construct a histogram using functions available in Data analysis. For this purpose we select successively **Tools**, **Data analysis** in the menu. In the opened window we select **Histogram** (Fig. 2.50). After that the box appears on which source data should be set to construct a histogram (Fig. 2.51).


Microsoft Excel - Chart2EN
File Edit View Insert Format Tools Data Window Help
Arial Cyr
10

I9 = =G5+CONFIDENCE(0,05;G7;6)

	A	B	C	D	E	F	G	H	I
1		Patient	A	B	C	D	E	F	
2			75,7	70,1	91,2	70,7	71,4	78,8	
3		Haemoglobin contents in blood, micromole							
4									
5			Average				76,31667		
6			Variance				64,48567		
7			Standard deviation				8,030297		
8			Median				73,55		
9			Confidence interval for average				69,89122	to	82,74211
10									

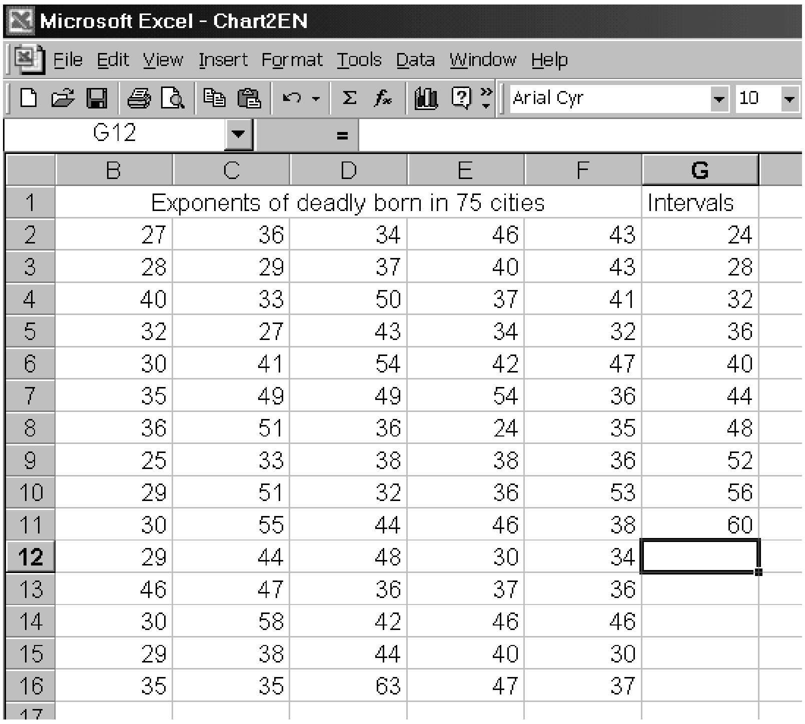
Fig. 2.46. Counting and construction histogram using Excel


Microsoft Excel - Chart2EN
File Edit View Insert Format Tools Data Window Help
Arial Cyr
10

G2 = 24

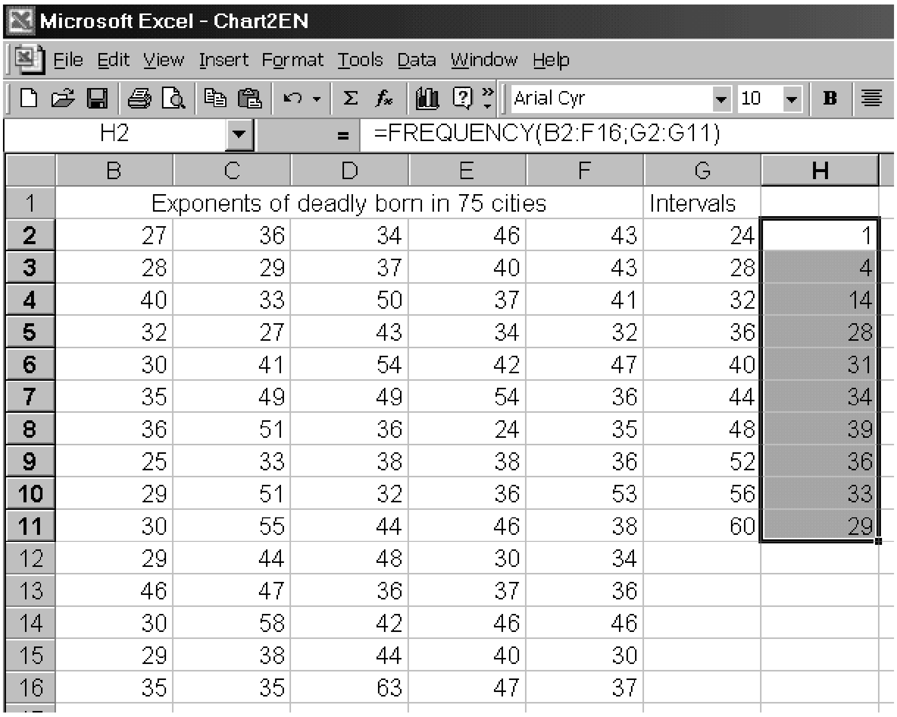
	B	C	D	E	F	G
1	Exponents of deadly born in 75 cities					
2	27	36	34	46	43	
3	28	29	37	40	43	
4	40	33	50	37	41	
5	32	27	43	34	32	
6	30	41	54	42	47	
7	35	49	49	54	36	
8	36	51	36	24	35	
9	25	33	38	38	36	
10	29	51	32	36	53	
11	30	55	44	46	38	
12	29	44	48	30	34	
13	46	47	36	37	36	
14	30	58	42	46	46	
15	29	38	44	40	30	
16	35	35	63	47	37	
17						

Fig. 2.47. Source data (exponents of the dead-born in 75 cities)


Microsoft Excel - Chart2EN
File Edit View Insert Format Tools Data Window Help
Arial Cyr
10
G12
=

	B	C	D	E	F	G
1	Exponents of deadly born in 75 cities					Intervals
2	27	36	34	46	43	24
3	28	29	37	40	43	28
4	40	33	50	37	41	32
5	32	27	43	34	32	36
6	30	41	54	42	47	40
7	35	49	49	54	36	44
8	36	51	36	24	35	48
9	25	33	38	38	36	52
10	29	51	32	36	53	56
11	30	55	44	46	38	60
12	29	44	48	30	34	
13	46	47	36	37	36	
14	30	58	42	46	46	
15	29	38	44	40	30	
16	35	35	63	47	37	

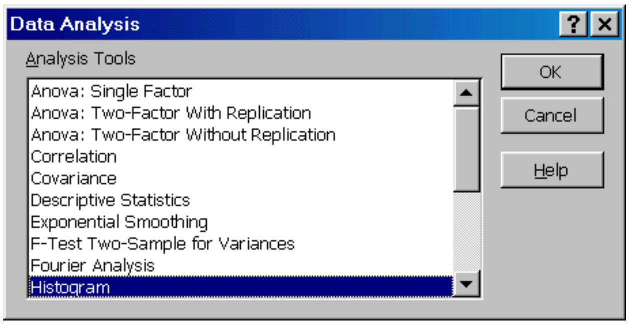
Fig. 2.48. Set boundaries of intervals (exponents of the dead-born in 75 cities)


Microsoft Excel - Chart2EN
File Edit View Insert Format Tools Data Window Help
Arial Cyr
10
B
H2
=

=FREQUENCY(B2:F16;G2:G11)

	B	C	D	E	F	G	H
1	Exponents of deadly born in 75 cities					Intervals	
2	27	36	34	46	43	24	1
3	28	29	37	40	43	28	4
4	40	33	50	37	41	32	14
5	32	27	43	34	32	36	28
6	30	41	54	42	47	40	31
7	35	49	49	54	36	44	34
8	36	51	36	24	35	48	39
9	25	33	38	38	36	52	36
10	29	51	32	36	53	56	33
11	30	55	44	46	38	60	29
12	29	44	48	30	34		
13	46	47	36	37	36		
14	30	58	42	46	46		
15	29	38	44	40	30		
16	35	35	63	47	37		

Fig. 2.49. Location of frequencies (exponents of the dead-born in 75 cities)

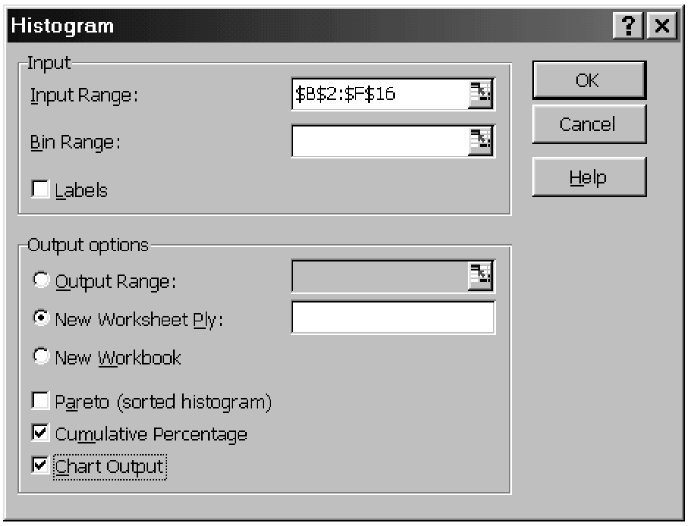


Data Analysis

Analysis Tools

Anova: Single Factor
Anova: Two-Factor With Replication
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram

Fig. 2.50. Function selection box



Histogram

Input

Input Range: $B$2:$F$16

Bin Range:

□ Labels

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

□ Pareto (sorted histogram)

☑ Cumulative Percentage

☑ Chart Output

Fig. 2.51. Box of setting new data for histograms

The options of the dialog box histogram have the following purposes:

Input range the reference to the cell range with source data should be set in this box. Source data should represent the list of values, but not frequencies.

Bin range enters in the field the cell range and optional collection of limiting values which determine the segments (the bins). These values should be entered in increasing order. The amount of data falling between the current beginning of the segment and adjacent one (if there is such) is calculated in Microsoft Excel. The values on the lower boundary are switched on while the values of the upper boundary are not. If the bin range was not entered, the collection of segment evenly distributed between minimum and maximum data values would be created automatically.

Labels. A label is set if the first line or first column of the input intervals contain titles. If there are no titles they will be assigned to output data range automatically.

Output range. A reference to the upper left cell of the output range is entered. The size of the latter is determined automatically and the message will appear on the screen if the output range can overlap on the source data.

New worksheet. A switcher is installed to open a new sheet in the book and insert the analysis results starting from the cell A1.

New workbook. To open a new book and insert analysis results in the cell A1, a switcher is installed on the first sheet in this book. If you want to have a chart this is a must-to-do.

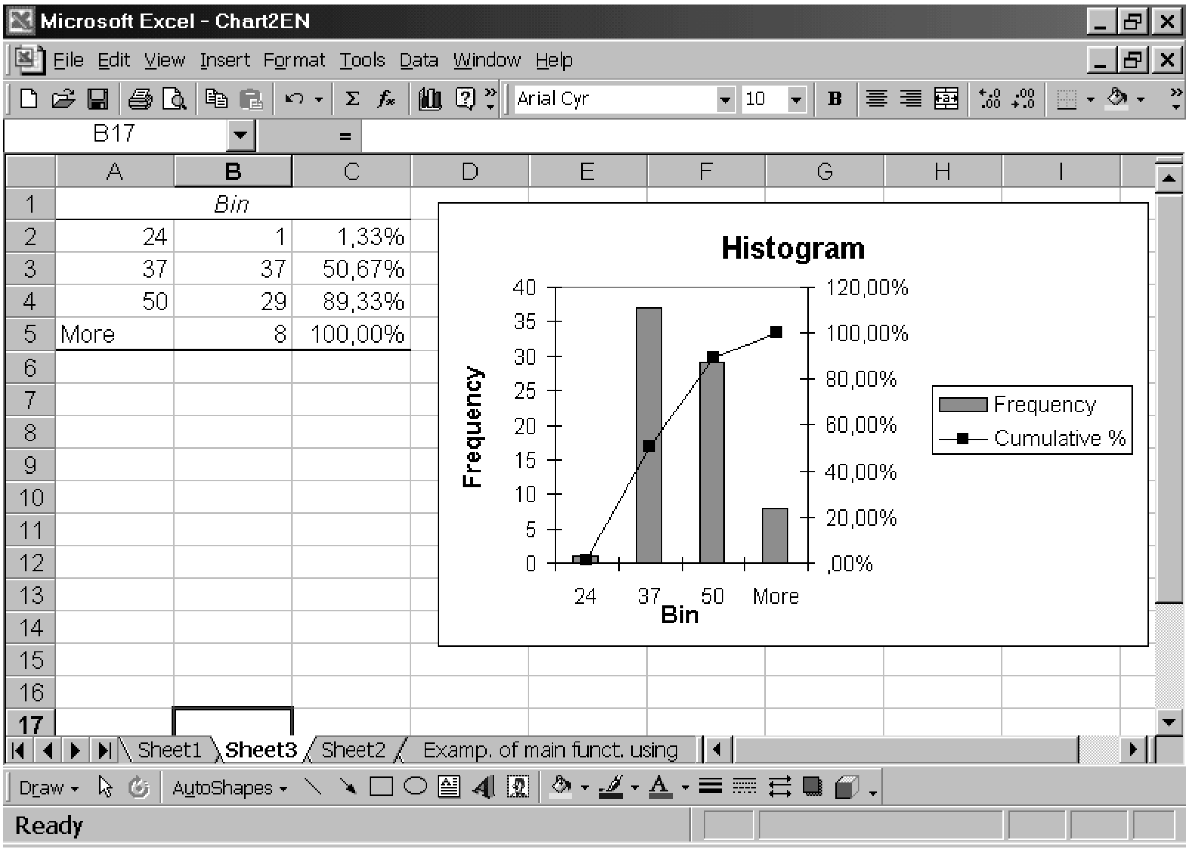
Pareto (sorted histogram). Selecting this possibility the data are represented in decreasing order of frequency. In mathematical statistics this type of histogram is not used.

Cumulative percentage. Values are calculated and the cumulated frequency chart is plotted.

Chart output. A check box is selected for automatic creation of the built-in chart.

ATTENTION! If you want to construct a chart it is necessary to assign New Book.

The results of the histogram (frequency table and chart are given in Fig 2.52.



A	B	C
	Bin	
24	1	1,33%
37	37	50,67%
50	29	89,33%
More	8	100,00%

Fig. 2.52. Results of Histogram creations