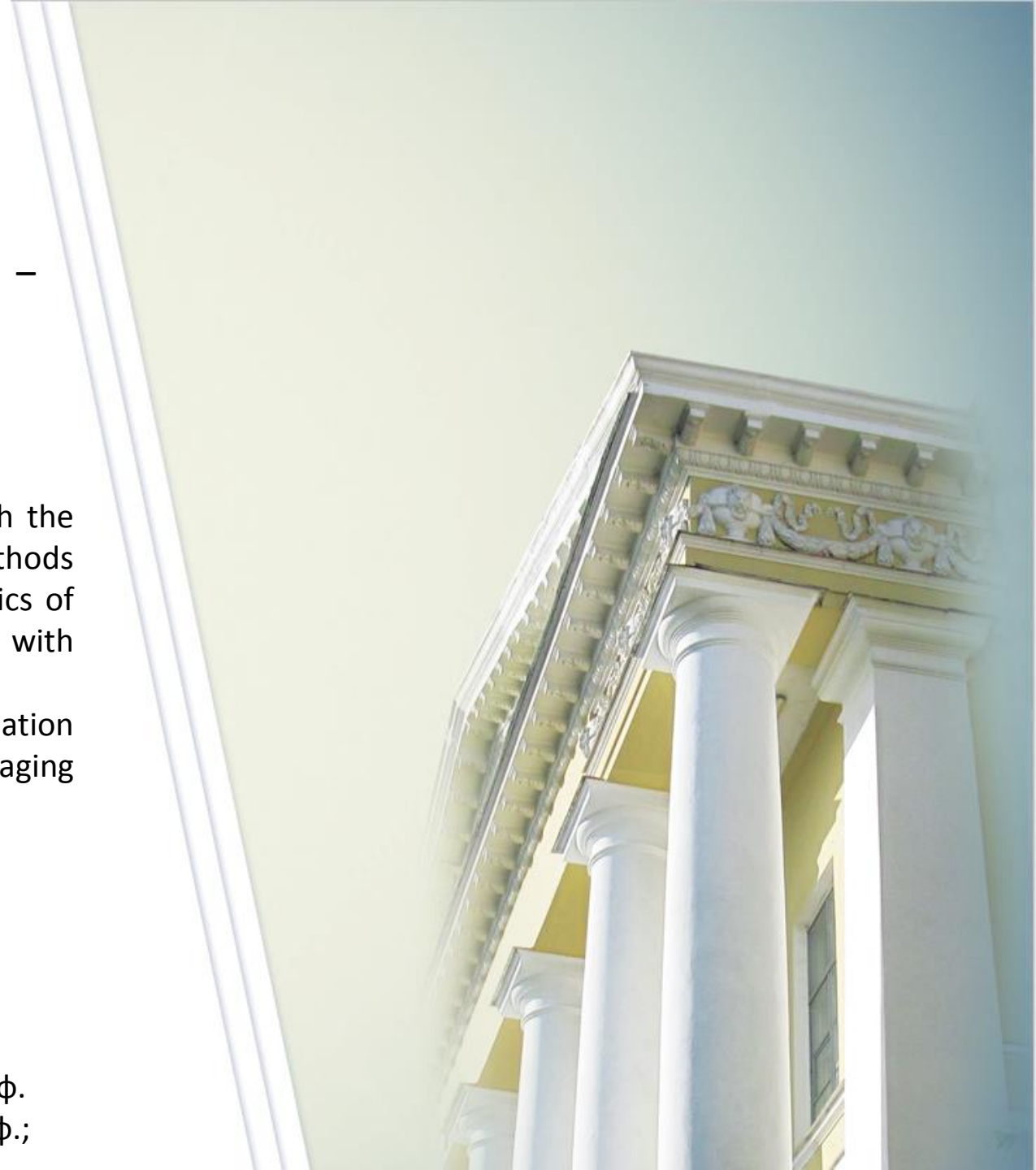
Національний технічний університет
ДНІПРОВСЬКА ПОЛІТЕХНІКА
1899

Пістунов І.М.

Economical statistic : навч. наоч. посіб. / І.М. Пістунов. – Дніпро : НТУ «ДП», 2022. – 35 с.



The purpose of the course "Statistics" is to acquaint students with the main categories and concepts of statistical science, with modern methods of processing and analyzing statistical information, with the specifics of the statistical study of socio-economic phenomena and processes, with the current system of macroeconomic indicators and models.
Mastering statistical methods for analyzing socio-economic information contributes to the development of specific solutions for managing economic and social processes in a market environment.
Designed for students of higher education specialties "Economics".
.

Рецензенти:
Кострицька С.І., завідувач кафедри іноземних мов НТУ “ДП”, проф.
Алексеєв М.О., декан факультету інформаційних технологій, проф.;

CONTENT

PREFACE

The statistics develops a special methodology for research and processing of materials: mass statistical observations, group method, averages, indices, balance method, graphic image method, and other methods for analyzing statistical data.

Statistics are divided according to their content into demographic, economic, financial, social, sanitary, judicial, biological, technical, etc.; mathematical statistics studies mathematical methods of systematization, processing and use of statistical data for scientific and practical conclusions.

Statistics consists of three sections:

- collection of statistical information, that is, information characterizing separate units of some mass aggregates;
- statistical study of the received data, which consists in elucidation of those regularities that can be established on the basis of mass observation data;
- development of methods of statistical observation and analysis of statistical data. The last section, in fact, is the content of mathematical statistics.

The term "statistics" is used in two different meanings. First, in everyday life under the "statistics" is often understood as a set of quantitative data about any phenomenon or process. Secondly, the statistic is called a function from the results of observations used to evaluate the characteristics and parameters of distributions and check hypotheses.

Basic concepts (categories) of statistics

Statistical aggregate is a mass of elements that are homogeneous in a certain respect, having a single qualitative basis, but differ from each other by certain attributes and subject to a certain distribution law. Statistical aggregate is a certain set of elements combined with the conditions of existence and development.

A homogeneous set - if one or more of the features being studied are common to all units.

A diverse set combines phenomena of different types.

The aggregate unit is the primary element of a statistical aggregate that is the bearer of the characteristics that are subject to registration and is the basis of accounting.

The attribute is the property of a separate unit of the population.

Qualitative attributes (attributes) are expressed in the form of concepts, definitions that characterize their essence, state or quality. For example, a variety of products, occupation, family status.

Quantitative attributes express the individual values of qualitative signs in a numerical expression.

Discrete - signs, expressed by separate integers, without intermediate values.

Continuous - signs that can acquire any values in certain numbers.

Direct - characterize the object of research directly (age of persons, number of present in the audience).

Indirect - signs that do not belong directly to the investigated object (or aggregate), but belong to another group included in this.

Multivariate are primarily characterized by ranks (scale of ranks) from greater to smaller (eg very low, low, medium, high, very high).

Alternative - mutually exclusive values: yes-no, positive-negative.

Intervals are signs that characterize the outcome of processes.

Momentum - characterize the object at a certain point in time.

Separate values of quantitative attributes are called variants.

The primary variants characterize the unit of the whole as a whole: absolute values, measured, calculated.

Secondary variants (derivative, estimated) are data that can not be verified because they are derived from certain sources.

Statistical indicators are numbers in combination with a set of attributes that characterize the circumstances to which they relate, what, where, when, and how they are measurable. The statistical indicator is a quantitative characteristic of socio-economic phenomena and processes in conditions of high-quality certainty.

Statistical data is a set of indicators obtained as a result of statistical observation or data processing.

Statistical regularity is a pattern in which the necessity is connected in each individual phenomenon with chance, and only in the aggregate of phenomena it manifests itself as a law.

The system of statistical indicators - a set of statistical indicators, which reflect the relationships that objectively exist between phenomena.

The overall purpose of the statistical research of the project is to study causation, and in particular, to draw a conclusion on the effect of changing the values of predictors or independent variables into dependent variables or responses. There are two main types of causative statistical research: experimental studies and observational studies. The difference between these two types of research is how the research was actually conducted. Each of these studies can be very effective. An experimental study involves measuring the measurement of this system by manipulating the system, and then taking additional measurements using the same procedure to determine the manipulation of the modified measurement values. In contrast, supervisory research does not involve experimental manipulation. Instead, data is collected and correlations between predictors and responses are researched.

Predictor - the predictive parameter; means of forecasting

1. ABSOLUTE AND RELATIVE INDICATORS

Depending on the nature and method of calculation, there are seven types of relative variables: planned task, plan execution, dynamics, structure, coordination, comparison and intensity.

$$ВВ_{пз} = \frac{П_1}{Ф_0}$$

$$ВВ_{пз} = \frac{П_1}{Ф_0} x100$$

The relative value of the dynamics characterizes the change in the time index and is defined as the ratio of the value in the next period to the value in the previous period:

Such indicators are called the rate of growth

$$BB_{\partial} = \frac{\Phi_1}{\Phi_0}$$

And such - the growth factor

$$BB_{\partial} = \frac{\Phi_1}{\Phi_0} x100$$

And such - the overgrowth factor

$$BB_{\partial} = \left(\frac{\Phi_1}{\Phi_0} - 1 \right) x100\%$$

In statistical practice, several types of averages are used:
- the arithmetic mean;
- the average harmonic;
- mean square;
- geometric mean.

The arithmetic mean - one of the most common types of the average, is used in cases where the volume of the variating feature for the whole population is the sum of the individual values of its individual elements. Arithmetic mean used to average the direct values of the sign by summing them.

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

X - average value of the studied feature
x_i - separate values of the averaging attribute
n - the number of units of the studied population

An average harmonic is used in cases where we know the values of the sampling elements and their inverse numbers.

$$\overline{x} = \frac{n}{\sum \frac{1}{x}}$$

Medium quadratic

$$\overline{X} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$$

Average geometric

$$\overline{X} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

The median is called the option dividing the variation line into two parts with an equal number of options. If the number of options is odd, $n = 2k + 1$ then $\tilde{x} = x_{k+1}$ in the case of a pair number option $n = 2k$ median equals

$$\tilde{x} = \frac{(x_k + x_{k+1})}{2}$$

2. CONCEPT of VARIATIONS and ITS MAIN INDICATORS

Swing variation: $R = x_{\max} - x_{\min}$

Simple linear deviation:

$$d = \frac{\sum_{i=1}^{n} \left| x_i - \bar{x} \right|}{n}$$

Weighted linear deviation

$$d = \frac{\sum_{i=1}^{n} \left| x_i - \bar{x} \right| \cdot f_i}{\sum_{i=1}^{n} f_i}$$

The formula of the weighted average dispersion

$$\sigma_x^2 = \frac{\sum_{i=1}^{n} (x - \bar{x})^2 f_i}{\sum_{i=1}^{n} f_i}$$

Weighted mean square deviation

$$\sigma_x = \sqrt{\frac{\sum (x - \bar{x})^2 f_i}{\sum f_i}}$$

Relative variations

Oscillation coefficient, which shows relative the oscillation of the extreme values of the sign around the average :

$$V_R = \frac{R}{\bar{x}} 100\%$$

Linear coefficient of variation:

$$V_{\bar{d}} = \frac{\bar{d}}{\bar{x}} 100\%$$

Coefficient of variation:

$$V_\sigma = \frac{\sigma_x}{\bar{x}} 100\%$$

The coefficient of variation is to some extent a criterion for the average type. If the coefficient is very large - this means that the average characterizes the population on a basis that varies significantly in individual units. The typicality of such an average is doubtful, that is, small.

Selective observation is a kind of observation, in which the survey is subjected to a number of units of the studied population, which allows it to get data for the characteristics of the entire population.

All studied sets of phenomena called general (denoted by N). That part of the units selected from the general population is called a sample population (denoted by n).

There are average and marginal sample errors. These two types of errors are related to the following relationship:

$$\Delta = t \cdot \mu$$

Average sample error

$$\mu = \sqrt{\frac{\sigma^2}{n}}$$

The formula of the average error in a random non-repeat sample

$$\mu = \sqrt{\left[\frac{\sigma^2}{n} \cdot \left(1 - \frac{n}{N}\right)\right]}$$

Range of existence of the average

$$\tilde{X} - \Delta_{\overline{x}} \leq \overline{X} \leq \tilde{X} + \Delta_{\overline{x}}$$

Indices

Economic index - a relative value that characterizes the change of the phenomenon under study in time, space or in comparison with a certain standard. The amount of this type of product is determined by the letter - *q* (physical volume); unit price of the product - *p*; cost per unit of product - *z*; complexity of unit of product– *t*, P_0 - value of the indicator in the base period; P_1 - value of the indicator in the current period.

Individual index :

$$i_p = \frac{P_1}{P_0}$$

Index of physical volume of sales :

$$J_q = \frac{\sum q_1 \cdot p_0}{\sum q_0 \cdot p_0}$$

General price index

$$J_p = \frac{\sum p_1 \cdot q_1}{\sum p_0 \cdot q_1}$$

General index of goods turnover (sales of products)

$$J_{qp} = \frac{\sum p_1 \cdot q_1}{\sum p_0 \cdot q_0}$$

Individual cost index

$$i_z = \frac{z_1}{z_0}$$

The general index of physical output, weighted at cost

$$J_q = \frac{\sum q_1 \cdot z_0}{\sum q_0 \cdot z_0}$$

Average cost levels

$$\bar{z}_0 = \frac{\sum z_0 \cdot q_0}{\sum q_0},$$

$$\bar{z}_1 = \frac{\sum z_1 \cdot q_1}{\sum q_1}.$$

Total Cost Index

$$J_z = \frac{\sum z_1 \cdot q_1}{\sum z_0 \cdot q_1}$$

Total cost of production

$$J_{zq} = \frac{\sum z_1 \cdot q_1}{\sum z_0 \cdot q_0}$$

Index of variable composition

$$J_{п.с.} = \frac{\bar{z}_1}{\bar{z}_0}$$

Fixed-stock index (or permanent warehouse)

$$J_{f-s} = \frac{\sum z_1 \cdot q_1}{\sum q_1} : \qquad \frac{\sum z_0 \cdot q_1}{\sum q_1} = \frac{\sum z_1 \cdot q_1}{\sum z_0 \cdot q_1}$$

Index of structural shifts

$$J_{ss} = \frac{\sum z_0 \cdot q_1}{\sum q_1} : \frac{\sum z_0 \cdot q_0}{\sum q_0}$$

Statistical study of interconnection

The calculation of the empirical correlation relationship is based on the use of the well-known theorem of adding dispersions. Overall variance of the resultant trait (σ_0^2) can be divided into two components. The first component is intergroup variance (δ^2), xdescribes the part of the fluctuation of the resultant characteristic, which is formed by the change in the sign-factor, which is the basis of the group:

$$\delta^2 = \frac{\sum_{j=1}^{k} (\bar{y}_j - \bar{y}_0)^2 \cdot n_j}{\sum_{j=1}^{k} n_j}$$

where $\bar{y}_j$ - the average value of the effect in the respective groups;
$\bar{y}_0$ - the total average value of the resultant attribute for the whole population;
n_j - number of observations in the corresponding group;
k - number of selected groups.

The second component - the average of intra-group variances - ($\overline{\sigma}^2$) evaluates that part of the variation of the resultant characteristic, which is due to the effect of other "random" causes

$$\overline{\sigma}^2 = \frac{\sum \sigma_j^2 \cdot n_j}{\sum n_j}$$

where σ_j^2 - dispersion of the resultant characteristic in the corresponding group:

$$\sigma_j^2 = \frac{\sum (y_j - \bar{y})^2}{n}$$

Total dispersion $\sigma_0^2 = \delta^2 + \overline{\sigma}^2$.

The proportion of intergroup dispersion in the overall dispersion

$$\eta^2 = \frac{\delta^2}{\delta^2 + \overline{\sigma^2}}$$

Selective correlation relation

$$\eta = \sqrt{\eta^2}$$

Measures of connection of samples

Correlation coefficient $$r = \frac{n \cdot \sum x_i \cdot y_i - \sum x_i \cdot \sum y_i}{\sqrt{\left(n \cdot \sum x_i^2 - \left(\sum x_i\right)^2\right) \cdot \left(n \cdot \sum y_i^2 - \left(\sum y_i\right)^2\right)}}$$

The determination coefficient is called the squared coefficient of correlation (r^2).

In practice, the following empirical rules can be used to assess the degree of interconnection:

1) $|r| > 0{,}95$ - there is practically linear dependence;
2) $0{,}8 < |r| < 0{,}95$ - strong degree of linear dependence;
3) $0{,}6 < |r| < 0{,}8$ - affiliation of linear communication;
4) $|r| < 0{,}4$ - the linear connection could not be detected.

The calculation of the empirical correlation relationship is based on the use of the well-known theorem of adding dispersions. Overall variance of the resultant trait (σ_0^2) can be divided into two components. The first component is intergroup variance (δ^2), characterizes that part of the fluctuation of the resultant characteristic, which is formed under the influence of the change of the sign-factor, which is the basis of the grouping:

Trust interval

- for average

$$\varepsilon_m = \ddot{\sigma}_x \Phi^{-1}(\beta)$$

- for dispersion

$$\varepsilon_D = \ddot{D}_x \Phi^{-1}(\beta) \sqrt{\frac{0,8N + 1,2}{N(N-1)}}$$

- for the relative frequency in the range of the histogram

$$\varepsilon_{p_i} = \Phi^{-1}(\beta) \sqrt{\frac{k_i(1-k_i)}{N}}$$

where $\ddot{\sigma}_m = \sqrt{\frac{D_x}{N}}$

$\Phi^{-1}(\beta)$ - the inverse of the Laplace function, that is, the value of an argument (quantum z), in which the Laplace function is equal to β.

- Statistical decisions are probabilistic, that is, there is always a possibility that the decisions made will be erroneous.
- The main value of making statistical decisions is that within the probabilistic categories one can objectively measure the degree of risk that corresponds to one or another decision.
- Any statistical conclusions obtained on the basis of sample processing are called statistical hypotheses.
- Statistical hypotheses about the value of the parameters of the signs of the general population are called parametric.
- For example, the statistical hypothesis is proposed about the numerical values of the general average Hg, the general dispersion DG, the general mean square deviation of 6G, and others.
- Statistical hypotheses, advanced on the basis of sampling processing on the law of distribution of the feature of the general population, are called nonparametric.
  For example, based on the processing of a sample, a hypothesis may be put forward that the sign of the general population has a normal distribution law, an exponential law, etc.

- Zero and alternative hypotheses
- The hypothesis to be checked is called the main one. Since this hypothesis implies the absence of systematic differences (zero divergences) between an unknown parameter of the general population and the value obtained as a result of processing the sample, it is called the null hypothesis and denoted H_0.
- The content of the null hypothesis is written as follows:

$$H_0: \; X_h = a;$$
$$H_0: \; 6_h = 2;$$
$$H_0: \; r_{xy} = 0{,}95$$

3. STATISTICAL CRITERION

The empirical value of the criterion

To verify the validity of the advanced statistical hypothesis, choose the so-called statistical criterion, guided by which they reject or reject the null hypothesis. The statistical criterion, conventionally denoted by *K*, is a random variable whose law of distribution of probabilities is known to us in advance.

The observed criterion value, denoted by *K**, is calculated by the result of the sampling.

The set of values of the statistical criterion *K*, in which the null hypothesis is not rejected, is called the domain of acceptance of the null hypothesis.

The set of values of the statistical criterion *K*, in which the null hypothesis is not accepted, is called a critical area.

Verification of statistical hypotheses with respect to averages

In order to test the hypothesis of the equality of averages in the general population, it is necessary to formulate a null hypothesis. In this case, as a rule, it follows from the fact that both samples are taken from a normally distributed general population with mathematical expectations equal to X and with a dispersion equal to c0.

Then the problem of checking the hypothesis is reduced to checking the

$$\left|\tilde{x}_1 - \bar{x}\right| - \left|\tilde{x}_2 - \bar{x}\right| = \left|\tilde{x}_1 - \tilde{x}_2\right| = \varepsilon_{\phi\text{экспер}}.$$

Each sample average has its own error:

$$\mu_1^2 = \frac{S_1^2}{n_1}; \qquad \mu_2^2 = \frac{S_2^2}{n_2}.$$

where the corrected dispersions

$$S_1^2 = \frac{\Sigma(x_{1i} - \tilde{x}_1)^2}{n_1 - 1}; \qquad S_2^2 = \frac{\Sigma(x_{2i} - \tilde{x}_2)^2}{n_2 - 1},$$

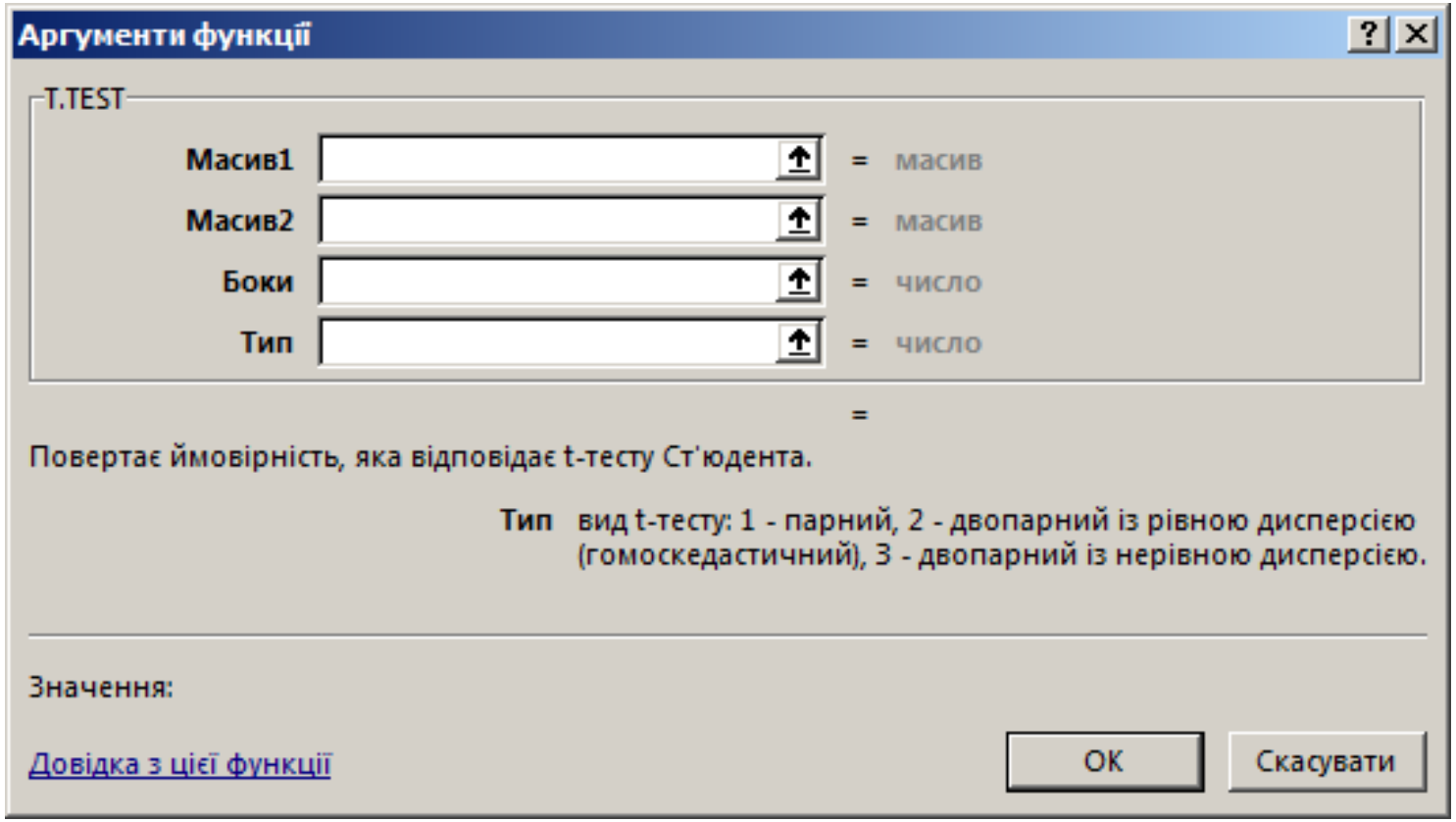
or other formulas

$$S_1^2 = \frac{\Sigma x_{1i}^2 - n_1 \tilde{x}_1^2}{n_1 - 1}; \qquad S_2^2 = \frac{\Sigma x_{2i}^2 - n_2 \tilde{x}_2^2}{n_2 - 1}.$$

Generalized mean error of two selective averages

$$\overline{\mu}_{1,2} = \sqrt{\mu_1^2 + \mu_2^2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.$$

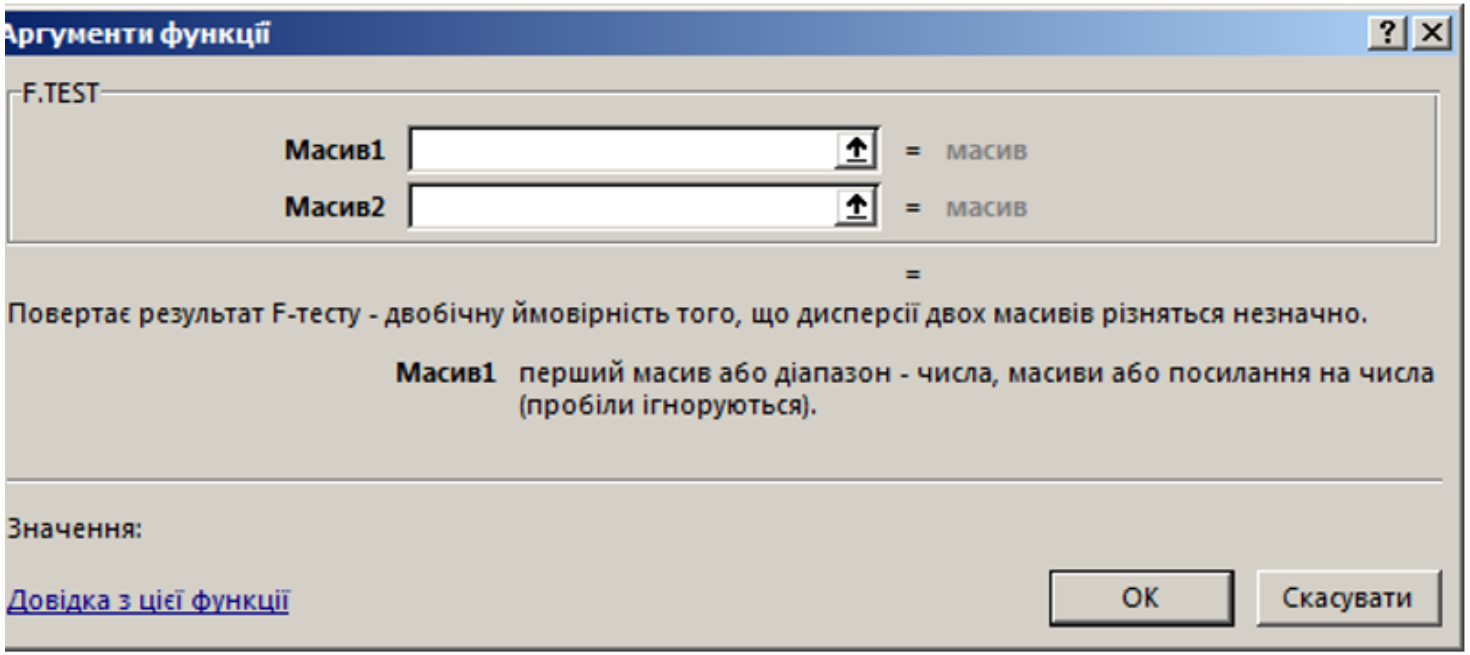
By determining the variance and the mean error of the sample mean, one can calculate the actual value of the I-criterion and compare it with the critical (tabular) values at the appropriate level of significance and the number of degrees of freedom of variation (for samples with the number of n> 30, the criterion of the normal distribution is used, and for samples with the number n <30 - and-criterion of Student).



Dispersion hypotheses arise quite often, because the variance characterizes such extremely important indicators as the accuracy of machines, technological processes, devices, the degree of homogeneity of aggregates, the risk associated with the deviation of the return on assets from the expected level, etc.

F-distribution of Fischer-Snedekor

$$F = \frac{\frac{1}{n_1 - 1}\left[(n_1 - 1)\frac{\hat{s}_1^2}{\sigma^2}\right]}{\frac{1}{n_2 - 1}\left[(n_2 - 1)\frac{\hat{s}_2^2}{\sigma^2}\right]} = \frac{\hat{s}_1^2}{\hat{s}_2^2}$$



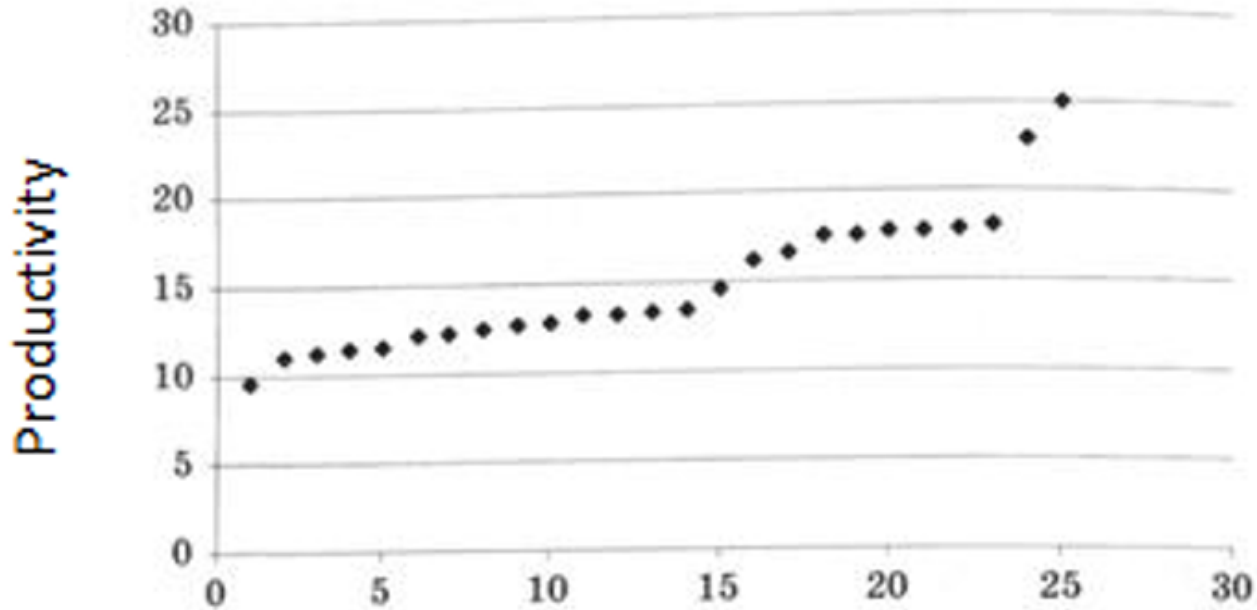
Regression analysis is a method of determining the separated and co-influence of factors on a resultant trait and quantifying this effect by using appropriate criteria.

Regression analysis is based on the constructed regression equation and determines the contribution of each independent variable in the variation of the investigated (predicted) dependent variable.

The main task of regression analysis is to determine the influence of factors on the performance indicator (in absolute terms). First of all, it is necessary to select and justify the equation of communication, which corresponds to the nature of the analytical stochastic relationship between the investigated features. The regression equation shows how the resultant sign (Y) changes on average, under the influence of the change of factor characteristics (x_i).

$$Y_x = f(x_1, x_2, \ldots, x_n)$$

The parallel comparison of the series of values of the levels of labor force by the basic means and its productivity, as well as the dotted graph of the "correlation field", indicate the presence and direction of communication (direct) between the given indicators. Moreover, the change in labor armament (factor factor *x*) leads to a relatively uniform change in labor productivity (the effect of *y*), as can be seen from the graph.



Synthesis of statistical linear and quasilinear models

Possible transformations

$$y = a_0 + \sum_{i=1}^{K} a_i x_i \qquad y = a_0 \ell^{a_i x_i} \qquad y = a_0 \log_n x$$

$$y = a_0 + \sum_{i=-n}^{-1} a_i x^i$$

$x_1\ x_2,\ x_1/\ x_2,\ x_1\text{-}x_2,\ \log\ x_1\ x_2,$

An example of normalization-denormalization

$y = a_0 + a_1 x + a_2 x^2$

$$y = \left[M_y + \sigma_y \left(a_0 - \frac{a_1 M_x}{\sigma_x} - \frac{a_2 M_{x2}}{\sigma_{x2}} \right) \right] + \frac{a_1 \sigma_y}{\sigma_x} x + \frac{a_2 \sigma_y}{\sigma_{x2}} x^2$$

$\mathrm{Ln}y = a_0 + a_1 * \mathrm{Ln}x_1 + a_2 * \mathrm{Ln}x_2$

$$\mathrm{Ln}y = \left[M_y + \sigma_y \left(a_0 - \frac{a_1 M_x}{\sigma_x} - \frac{a_2 M_{x2}}{\sigma_{x2}} \right) \right] + \frac{a_1 \sigma_y}{\sigma_x} \mathrm{Ln}x_1 + \frac{a_2 \sigma_y}{\sigma_{x2}} \mathrm{Ln}x_2$$

The F-test is used to evaluate the significance of the regression equation. To do this, the actual Fact and the critical (tabular) Fabel F-criterion are compared.

$$F_{факт} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}(n-2) = \frac{r^2}{1-r^2}(n-2)$$

The t-test is used to assess the significance of the regression coefficients b_0, b_1. For this, the actual TFACT and the critical (tabular) TTABL value of the Student t-criterion are compared. TFACT for the coefficients b_0, b_1 is determined by the following formulas:

$$t_{b_0} = \frac{|b_0|}{S}\sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{\overline{x^2}}}$$

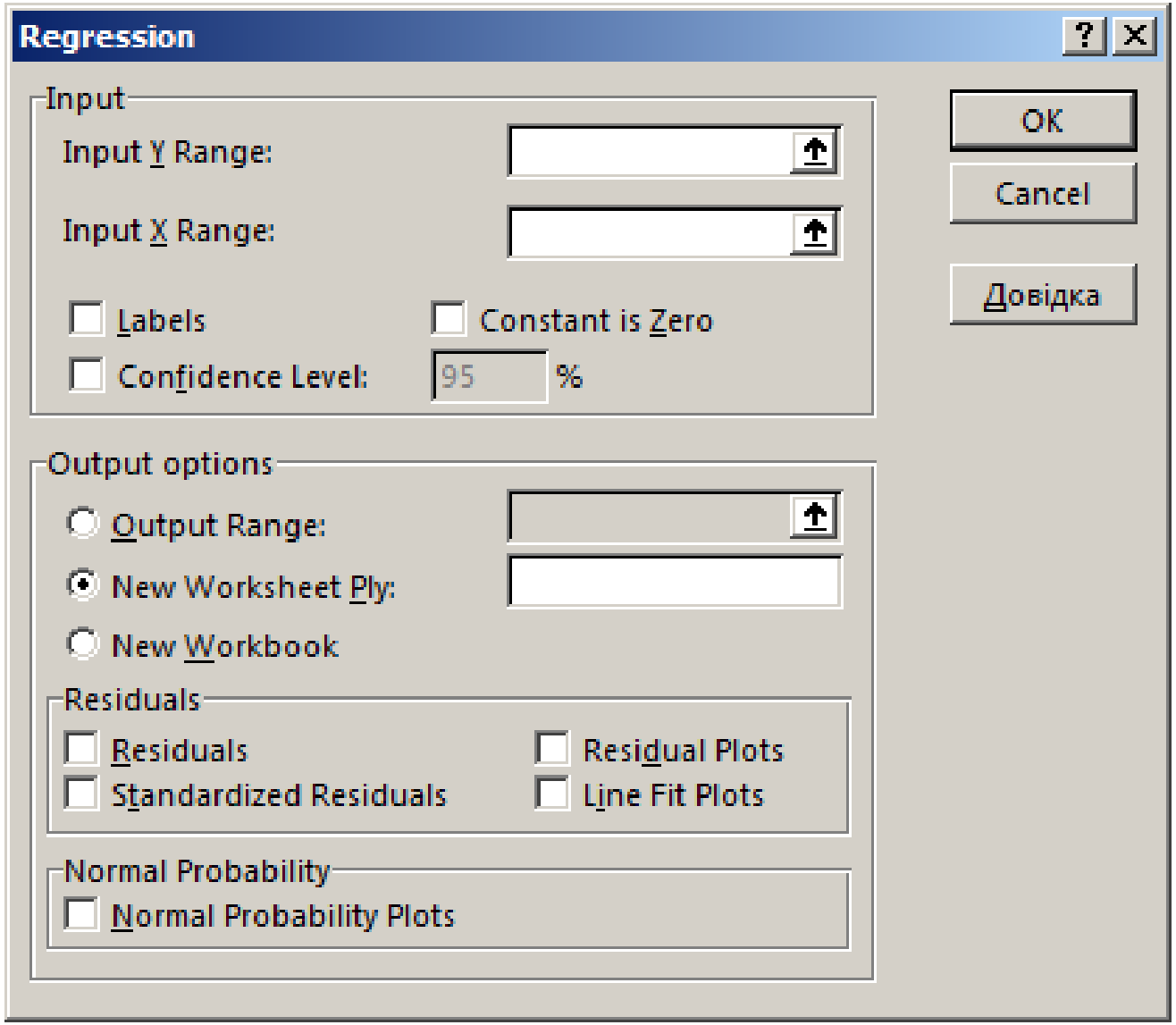
The assessment of the adequacy of the regression model is made on the basis of the determination coefficient:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

Significance of the correlation coefficient

according to Student statistics. :

$$|t| = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} > t_{1-\alpha;n-2}$$


Regression
Input
Input Y Range:
Input X Range:
Labels
Constant is Zero
Confidence Level:
95
%
OK
Cancel
Довідка
Output options
Output Range:
New Worksheet Ply:
New Workbook
Residuals
Residuals
Residual Plots
Standardized Residuals
Line Fit Plots
Normal Probability
Normal Probability Plots