MINISTRY OF EDUCATION OF THE REPUBLIC OF BELARUS
Belarusian National Technical University

Department “Higher Mathematics”

MATHEMATICAL STATISTICS

Educational and methodical manual
for specialties:
1-37 01 02 “Automotive Industry (in areas)”
1-37 01 01 “Internal Combustion Engines”

Approved by the Educational and Methodological Association
in the field of transport end transport activities

Minsk
BNTU
2022

УДК 519.2(075.8)
ББК 22.172я7
М39

Authors:
T. I. Chepeleva, *N. I. Chepelev*,
A. N. Chepelev, *M. V. Shchukin*

Reviewers:
Department of General Mathematics and Informatics of the Faculty
of Mechanics and Mathematics of the Belarusian State University
(Head of the Department, Doctor of Economics, Professor *S. A. Samal*);
Doctor of Economics, Professor, Head of the Department of Business
Administration of the Institute of Business of BSU *G. A. Khatskevich*

M39 **Mathematical** statistics : Educational and methodical manual for specialties : 1-37 01 02 “Automotive Industry (in areas)” 1-37 01 01 “Internal Combustion Engines” / T. I. Chepeleva [and etc.]. – Minsk : BNTU, 2022. – 83 p.
ISBN 978-985-583-820-4.

This educational and methodological manual is intended for university students studying mathematical statistics; it can be used by graduate students, undergraduates, engineers to calculate statistical data of any structure.

The manual includes theoretical questions for preparing for the exam, tasks for classes and independent work, a typical calculation for mathematical statistics, as well as tables of function values, that are necessary for solving tasks.

The manual contains a brief description of access to the software tools of the STATISTICA system, which is designed for statistical data processing in the Windows environment.

УДК 519.2(075.8)
ББК 22.172я7

ISBN 978-985-583-820-4

INTRODUCTION

Mathematical statistics is a branch of mathematics devoted to mathematical methods of systematization, processing and use of statistical data for scientific and practical conclusions of justifications, where statistical data is information about the number of objects in any sample.

A statistical research method is a method based on the consideration of statistical data on samples of objects.

The subject of mathematical statistics is the formal mathematical side of statistical methods of research, which is related to the specific nature of the studied objects.

Statistics is a science that allows us to see patterns in the chaos of random data, highlight stable connections in them, determine the actions of a person in order to increase the proportion of correctly made decisions among all those made. “Statistics” – from the Latin word “status” – a state, a state of affairs.

Statistics was born from the needs of human practical activity. Confucius (5th century BC) talks about a census in China, conducted in 2238 BC. The book of Moses is one of the oldest in terms of the population capable of bearing weapons.

Statistical science is designed to study patterns in random data, which will be used in the future in the study of certain processes, phenomena. It makes possible to reveal cause-and-effect relationships. The development and improvement of civilizations inevitably leads to the expansion and strengthening of the needs of statistical methods of calculation.

In practice, the application of statistical methods is an extremely complex matter, requiring great intellectual effort and time.

THEORETICAL QUESTIONS FOR PREPARING FOR THE EXAM

1. General population. Variation series.
2. Polygon and histogram.
3. Empirical distribution function and its properties.
4. Sample mean and sample variance.
5. Estimates of distribution parameters. Point estimates and requirements for them.
6. Point estimates for the mathematical expectation and variance.
7. Interval estimates. Confidence interval.
8. Student's distribution.
9. Pearson's distribution.
10. Construction of a confidence interval for the mathematical expectation of a normally distributed RV (random variable) with a known standard deviation.
11. Construction of a confidence interval for the mathematical expectation of a normally distributed RV with an unknown standard deviation.
12. Construction of a confidence interval for the standard deviation of a normally distributed RV.
13. The concept of statistical hypotheses and goodness-of-fit test.
14. Pearson's goodness-of-fit test χ^2.
15. Kolmogorov's goodness-of-fit test.
16. Sample correlation coefficient and its properties.
17. Regression equation. Linear regression. Determination of linear regression coefficients by the least squares method.
18. Nonlinear regression. Determining the parameters of non-linear regression.

THE ANALYTICS SYSTEM TIBCO STATISTICA™

General information

With the introduction of information technologies in science, technology, production, medicine and in all areas of the national economy, the market for statistical software is rapidly developing, opening the way to new technologies for statistical data processing, minimizing the routine procedures associated with the processing of multidimensional, complex data and dependencies in the data.

Huge preference in this market is given to the Statistica software product from TIBCO Software Inc., USA (hereinafter referred to as STATISTICA), which is an integrated system for statistical analysis and data processing.

STATISTICA consists of the following main components:

– workbooks – are optimized ActiveX containers that can efficiently handle large numbers of documents. The documents can be organized into hierarchies of folders or document nodes (by default, one is created for each new analysis) using a tree view, in which individual documents, folders, or entire branches of the tree can be flexibly managed;

– spreadsheets (multimedia tables) – for entering and setting initial data, outputting numerical and multimedia results of analysis, etc.;

– graphs – a graphical system for visualizing data and the results of statistical analysis;

– a set of specialized statistical modules;

– special tools for preparing reports;

– built-in programming language STATISTICA VISUAL BASIC.

The specialized modules of the system are basic statistical functions, tables, non-parametric statistics, analysis of variance, multiple regression, non-linear estimation, cluster analysis, factorial analysis, discriminant functional analysis, canonical correlation, etc.

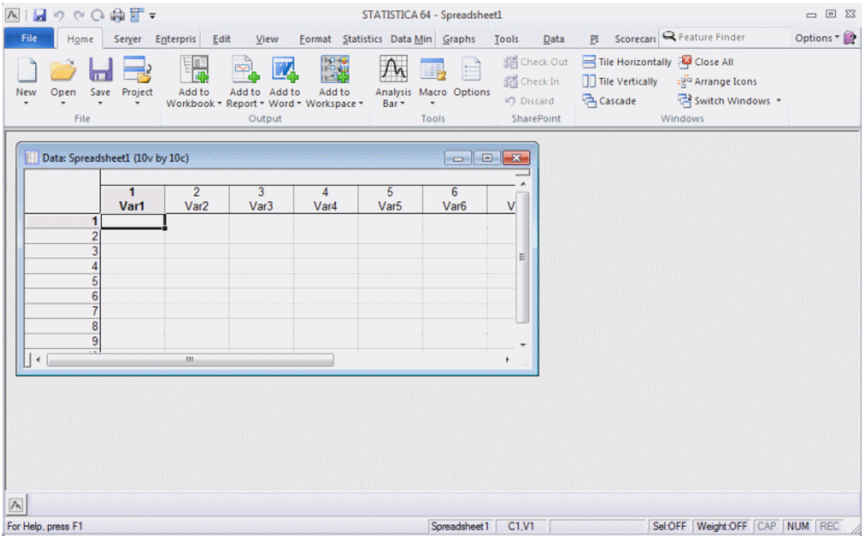


Fig. 1. Interface of STATISTICA^{tm}

Stages of statistical analysis:

– input of initial data into a spreadsheet and their preliminary transformation (selection, ranking, etc.);

– visualization of data using one of the types of graphs;

– application of a specific statistical processing procedure;

– output of analysis results in the form of spreadsheets with numerical and textual information and graphs;

– preparation and printing of the report.

STATISTICA has two interfaces:

1. Classic (consists of a menu of commands).

2. Ribbon.

The ribbon interface is enabled by default in STATISTICA 13 (fig. 1). Switching between modes is carried out by pressing the corresponding button in the title bar in ribbon mode or via the menu View-Ribbon in classic mode.

Using the system is interactive and is carried out by sequentially selecting the necessary commands from the menu (classic interface) or ribbon (ribbon interface).

STATISTICA Launch

Starting the Windows operating system, pressing the "Start" button, select the folder containing the STATISTICA system in the "Programs" menu. In this folder, select the STATISTICA shortcut and click on it.

After installing the system, the STATISTICA 13 folder will contain the following items:

– STATISTICA – a shortcut to launch the program;

– STATISTICA e-manual – help system.

After the first launch of STATISTICA 13, a window with a ribbon interface and an empty table of initial data appears on the screen.

For the purpose of compatibility with previous versions of the system, we will use the classic interface (fig. 2). Therefore, we recommend that you switch to it in advance through the corresponding item in the program title bar.

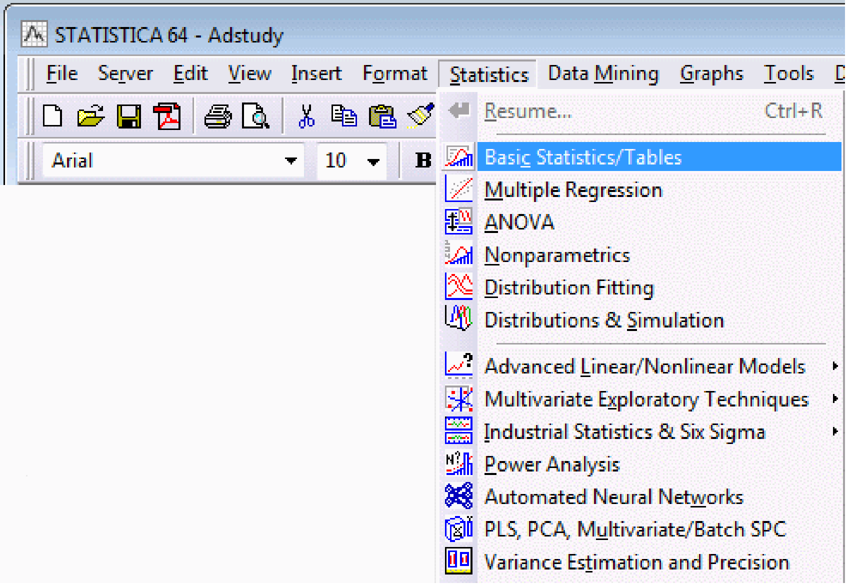

STATISTICA 64 - Adstudy
File
Server
Edit
View
Insert
Format
Statistics
Data Mining
Graphs
Tools
Arial
10
B
Resume... Ctrl+R
Basic Statistics/Tables
Multiple Regression
ANOVA
Nonparametrics
Distribution Fitting
Distributions & Simulation
Advanced Linear/Nonlinear Models
Multivariate Exploratory Techniques
Industrial Statistics & Six Sigma
Power Analysis
Automated Neural Networks
PLS, PCA, Multivariate/Batch SPC
Variance Estimation and Precision

Fig. 2. Classic Interface

STATISTICA Main Window

"Header line" contains the name of the active statistical module – STATISTICA Spreadsheet 1. In the right part of Header line – from left to right:

– Minimization of window sizes;
– Window restoration;
– Closing the window.

The second line, the "menu line", contains commands ordered by function.

STATISTICA works with different types of documents, and each document has its own menu and toolbar. For the "SpreadSheet" spreadsheet, the initial data entry menu contains the following commands:

– File;
– Server (if enabled);
– Edit;
– View;
– Insert;
– Format;
– Statistics;
– Data Mining;
– Graphs;
– Tools;
– Data;
– Window;
– Help.

"Toolbar" is located in the third row and in the left first column. Most part of the Main window of STATISTICA is occupied by the working area, which displays the documents with which you are working:

– Workspaces;
– Dashboards or Visualizations;
– Workbooks;
– Spreadsheets (multimedia tables);
– Reports;
– Graphs;
– Macros.

The document window consists of a title bar, a work area for entering and (or) displaying information, and tools for managing the window. In addition, STATISTICA has context menus that allows you to quickly access the most frequently used commands for working with a particular object in the active window.

The context menu of the Spreadsheet window provides access to groups of commands:

– graphics for the selected block of values;

– calculation of basic statistics;

– specification of variables, operations with text values;

– transformation of the spreadsheet structure (adding, deleting a row, etc.);

– operations with a selected block of values;

– operations with the "Windows Clipboard".

Spreadsheet – a Table with Raw Data

STATISTICA spreadsheets are based on multimedia table technology and are used to manage both input data and the numeric or text (and optionally any other type of) output (fig. 3). The basic form of the spreadsheet is a simple two-dimensional table that can handle a practically unlimited number of cases (rows) and variables (columns), and each cell can contain a virtually unlimited number of characters. Sound, video, graphs, animations, reports with embedded objects, or any ActiveX compatible documents can also be attached. Because Spreadsheets can also contain macros and any user-defined user interface, these multimedia tables can be used as a framework for custom applications (such as with a list box of options or a series of buttons placed in the upper left corner), self-running presentations, animations, simulations, etc.

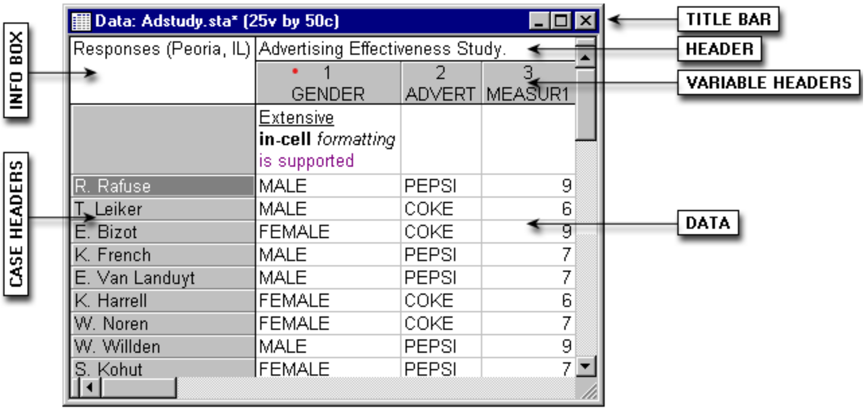
To work with source data tables, there are a large number of tools that are available through drop-down and context menus and from the toolbar, including:

– operations that change the structure of the spreadsheet (adding, deleting, copying, moving variables and cases);

– operations for setting specifications (names, formats, etc.) for variables and cases;

– operations with a selected block of values;

– operations implemented using Drag-and-Drop, including operations for copying, moving and auto-completing a block, etc.



Data: Adstudy.sta* (25v by 50c)

Advertising Effectiveness Study.

Responses (Peoria, IL)	1 GENDER	2 ADVERT	3 MEASUR1
	Extensive **in-cell** *formatting* is supported		
R. Rafuse	MALE	PEPSI	9
T. Leiker	MALE	COKE	6
E. Bizot	FEMALE	COKE	9
K. French	MALE	PEPSI	7
E. Van Landuyt	MALE	PEPSI	7
K. Harrell	FEMALE	COKE	6
W. Noren	FEMALE	COKE	7
W. Willden	MALE	PEPSI	9
S. Kohut	FEMALE	PEPSI	7

Fig. 3. Spreadsheet Window

Variables – have their own names, format and other attributes, which are called specifications and are set by the user. The variable is an observable value. The results of observations are recorded in the rows of the table – Cases.

Variable specifications can be set in Spreadsheet, which include:

– data display format (number of decimal places in number representation, date and time representation format, correspondence between numerical and text values, etc.);

– code assigned to missing data;

– long variable names and comments to them;

– formulas used to determine, recode or transform the values of variables;

– dynamic links between the STATISTICA data file and another Windows-compatible file, etc..

You can include any other embedded or linked objects in the Table (for example, multimedia objects, macros). The text in the cells can be practically unlimited in length (usually it is limited to 1,000 characters in the STATISTICA system settings to prevent accidental insertion of a large amount of unwanted information). The same is also true about the OLE/ActiveX and other objects that can be embedded into the spreadsheets; their size is not limited, and there can be virtually an unlimited number of them.

There are a lot of opportunities for formatting text in cells. You can change the format of variable names (such as bold, underline, italics, font color and font size) to increase their visibility in the spreadsheets. For a data set with several hundred variables, you could format variables of a particular type (such as dependent variables) in such a way as to make them easily recognizable in the spreadsheet.

To create a new spreadsheet, follow these instructions:

1. Press CTRL+N on your keyboard to display the Create New Document dialog box, or:

Ribbon bar. Select the Home tab. In the File group, click New to display the Create New Document dialog box.

Classic menus. On the File menu, select New, or click the New toolbar button to display the Create New Document dialog box.

2. In the Create New Document dialog box, select the Spreadsheet tab.

3. Use the microscrolls or enter the desired Number of variables and Number of cases.

4. Use the options in the Placement group box to create the spreadsheet in a new Workbook (to create a spreadsheet in a new, blank workbook) or As a stand-alone window (to create a blank spreadsheet in a new window).

5. Click the OK button.

To edit the contents of a cell without deleting the current contents, follow these steps:

1. Double-click in the cell that contains the data you want to edit. This will enter the editing mode and will position the cursor within the cell. Alternatively, you can follow the spreadsheet keyboard convention and press F2, which will also enter the edit mode for the currently highlighted cell.

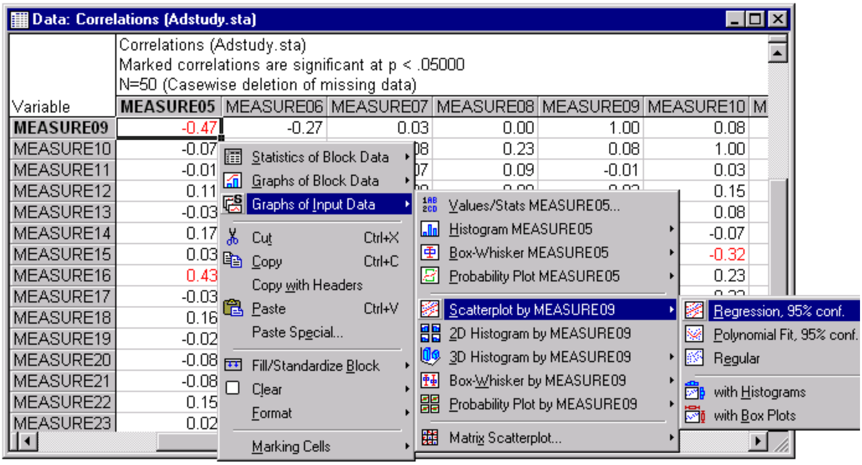
2. Make the desired changes to the cell contents.

3. Press ENTER.

To overwrite the contents of a cell, click on the cell and begin typing.

Building a Statistical Graph

Statistica includes a comprehensive selection of graphical methods for both data analysis and the presentation of results. All graphs in Statistica include a broad selection of built-in, interactive analytic techniques and extensive customization tools that enable you to interactively control virtually all aspects of the display (fig. 4). Also, flexible multi-graphics management facilities are available that are used to integrate various graphical displays and to build dynamic links between applications (e. g., using OLE-Object Linking and Embedding).



Data: Correlations (Adstudy.sta)

Correlations (Adstudy.sta)
Marked correlations are significant at p < .05000
N=50 (Casewise deletion of missing data)

Variable	MEASURE05	MEASURE06	MEASURE07	MEASURE08	MEASURE09	MEASURE10	M
MEASURE09	-0.47	-0.27	0.03	0.00	1.00	0.08	
MEASURE10	-0.07		08	0.23	0.08	1.00	
MEASURE11	-0.01		07	0.09	-0.01	0.03	
MEASURE12	0.11					0.15	
MEASURE13	-0.03					0.08	
MEASURE14	0.17					-0.07	
MEASURE15	0.03					-0.32	
MEASURE16	0.43					0.23	
MEASURE17	-0.03						
MEASURE18	0.16						
MEASURE19	-0.02						
MEASURE20	-0.08						
MEASURE21	-0.08						
MEASURE22	0.15						
MEASURE23	0.02						

Fig. 4. Graph Menu

Graph facilities in Statistica not only enable you to create a graph easily, you can also automatically update graphs in real-time whenever your input data change. If the spreadsheet is linked to an outside data source, graphs can be set to update automatically whenever the spreadsheet links are updated (e. g., from a database or data acquisition equipment connected to the serial port).

Once a graph is created, you can link the graph to different variables in the spreadsheet, or to a completely different spreadsheet. After the link update, the graph will retain its original customizations while linking to the new data.

The update can go the other way as well. When you change the case states of the data points of an existing graph (using the brushing tool), such as the label or color of the marker, or exclude or hide points from the graph, these changes can be updated to the spreadsheet.

Statistica graphical options can be accessed programmatically (using the built in Statistica Visual Basic or other compatible languages), which creates practically unlimited possibilities to produce highly customized graphical displays. These custom graphs can be permanently added to Statistica's user interface (e. g., assigned to buttons on toolbars or added to menus).

The Statistica system offers a variety of methods in which graphs can be requested or defined. These methods (constituting broad categories of graphs such as **input data, block data, and specialized**). They complement each other, providing a high level of integration between numbers (such as raw data, intermediate results, or final results) and graphical displays.

In addition to the specialized statistical graphs that are available from the output dialog boxes in all statistical procedures, there are **two general categories or classes of graphs both accessible from the Graphs tab or menu, Graphs toolbar (classic menus), shortcut menus, and the Statistica Start button menu**:

– Input Data Graphs;

– Graphs of Block Data.

The most important difference between these two general categories lies in the data that the graph types utilize for generating plots.

Input Data Graphs and their expanded version in the Graphs menu produce statistical summaries or other representations of the raw data in the current input data spreadsheet (typically for the entire variable(s) or its subsets if case selection conditions are used). If graphs of this general category are produced using a shortcut menu from within a spreadsheet of results that does not contain the actual data (e. g., a correlation matrix), Statistica will still reach to the respective input (raw) data in order to produce the graph (e. g., a scatterplot of the variables identified by the selected cell in the correlation matrix from which the shortcut menu was opened).

Graphs of Block Data, on the other hand, are entirely independent of the concept of input data or data file. They provide a general tool to visualize numeric values in the currently highlighted block of any spreadsheet (which can contain values from custom defined subsets of numerical output or arbitrarily selected subsets of raw data).

These two general categories of graphs offer the same customization options and the same selection of types of graphs. For example, you can create the same, highly specialized categorized ternary graph from the input (raw) data set, and from a custom defined block of values representing results of a particular test

Statistical Analysis

In the STATISTICA system, to obtain a set of numerical, textual and graphical information during the processing of initial data, the menu or the ribbon Statistics is used, depending on the interface.

Let's consider data analysis using two examples:

1. The option "Basic Statistics and Tables".

To work, you can open a training table from the Adstudy.sta file (Menu File – Open examples folder – Examples\Datasets – Adstudy.sta).

The data file contains 25 variables and 50 cases.

These (fictitious) data were collected in an advertising study where male and female respondents evaluated two advertisements.

Respondents' gender was coded in variable 1 (Gender: 1=MALE, 2=FEMALE).

Each respondent was randomly assigned to view one of the two ads (Advert: 1=COKE, 2=PEPSI). They were then asked to rate the appeal of the respective ad on 23 different scales (Measure01 to Measure23). On each of these scales, the respondents could give answers between 0 and 9.

Open the Adstudy.sta data file, and start the Basic Statistics and Tables module (fig. 5).

Ribbon bar – Select the Statistics tab. In the Base group, click Basic Statistics to display the Basic Statistics and Tables Startup Panel.

Classic menus – From the Statistics menu, select Basic Statistics/Tables to display the Basic Statistics and Tables Startup Panel.

The Startup Panel contains one tab: Quick.

OK – after you have selected the type of analysis to perform via the Quick tab, click the OK button to display the specified dialog box.

Cancel – click the Cancel button to close the Startup Panel without performing an analysis.

Options – see Options Menu for descriptions of the commands on this menu.

Open Data – click the Open Data button to display the Select Data Source dialog box, which is used to choose the spreadsheet on which to perform the analysis. The Select Data Source dialog box contains a list of the spreadsheets that are currently active.

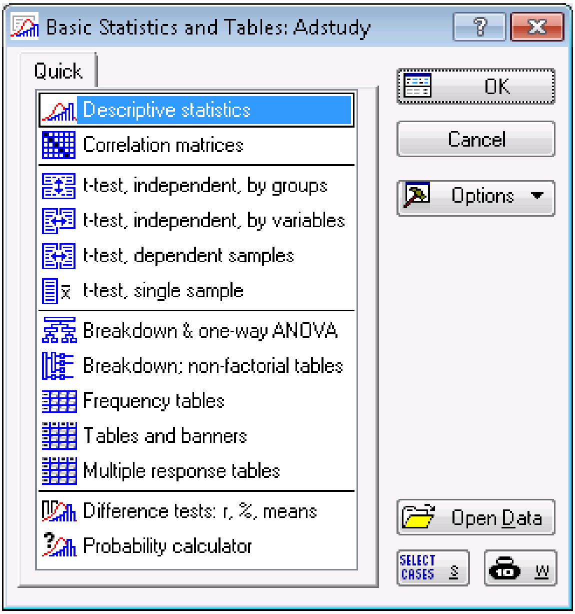

Basic Statistics and Tables: Adstudy
Quick
Descriptive statistics
Correlation matrices
t-test, independent, by groups
t-test, independent, by variables
t-test, dependent samples
t-test, single sample
Breakdown & one-way ANOVA
Breakdown; non-factorial tables
Frequency tables
Tables and banners
Multiple response tables
Difference tests: r, %, means
Probability calculator
OK
Cancel
Options
Open Data
SELECT CASES
W

Fig. 5. Basic Statistics and Tables

Select Cases – click the Select Cases button to display the Analysis/Graph Case Selection Conditions dialog box, which is used to create conditions for which cases will be included (or excluded) in the current analysis. More information is available in the case selection conditions overview, syntax summary, and dialog box description.

W – click the W (Weight) button to display the Analysis/Graph Case Weights dialog box, which is used to adjust the contribution of individual cases to the outcome of the current analysis by weighting those cases in proportion to the values of a selected variable.

We will check whether ratings on individual scales are correlated (i. e., whether some scales measured the same thing). In the Basic Statistics and Tables dialog box, select Correlation matrices and then click the OK button (or double-click Correlation matrices). The Product-Moment and Partial Correlations dialog box is displayed.

You can select variables either in one list (i. e., a square matrix) or in two lists (rectangular matrix). For this example, click the One variable list button; a variable selection dialog box is displayed. Ensure that the "Show appropriate variables only" check box is cleared, select all variables (click the Select all button), and click the OK button.

Since the analysis expects continuous variables, but text variables were selected, the Variables contain text values/text labels dialog box is displayed. For this example we want to retain the text variables. Click the Continue with current selection button. In the Product-Moment and Partial Correlations dialog box, click the Summary button to display results.

2. The option "Multiple Regression".

The general purpose of multiple regression (the term was first used by Pearson, 1908) is to learn more about the relationship between several independent (or predictor) variables and a dependent (or criterion) variable.

In the social and natural sciences, multiple regression procedures are very widely used in research. In general, multiple regression allows the researcher to ask (and hopefully answer) the general question "what is the best predictor of ...".

For example, educational researchers might want to learn what are the best predictors of success in high-school.

Psychologists may want to determine which personality variable best predicts social adjustment. Sociologists may want to find out which of the multiple social indicators best predict whether a new immigrant group will adapt and be absorbed into society.

As example we will use the data file Poverty.sta (Menu File – Open examples folder – Examples\Datasets – Poverty.sta).

The data are based on a comparison of 1960 and 1970 Census figures for a random selection of 30 counties. The names of the counties were entered as case names.

We can additionally view the information for each variable in the Variable Specifications Editor (accessible by selecting All Variable Specs from the Data menu).

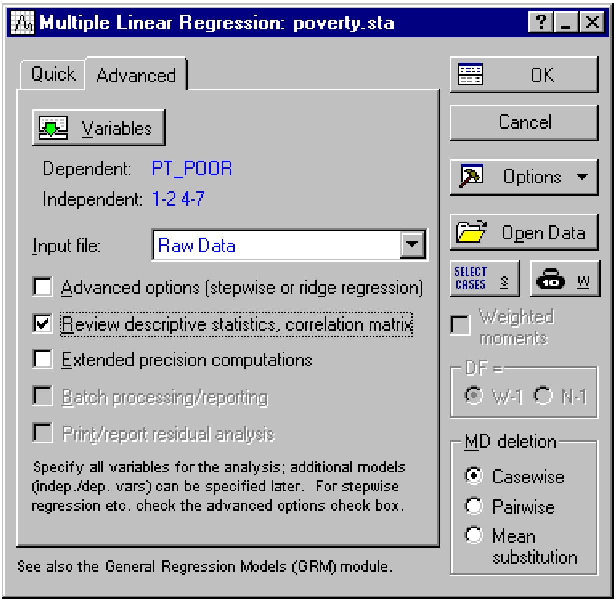
Our task is to analyze the correlates of poverty, that is, the variables that best predict the percent of families below the poverty line in a county. Thus, we will treat variable 3 (Pt_Poor) as the dependent or criterion variable, and all other variables as the independent or predictor variables.

To start the analysis, select Multiple Regression from the Statistics menu.

Specify the regression equation by clicking the Variables button on the Multiple Linear Regression dialog box – Quick tab to display the variable selection dialog.

Select PT_POOR as the Dependent variable and all of the other variables in the data file from the Independent variable list (select them while pressing Ctrl button on keyboard), and then click the OK button.

Also, on the Multiple Linear Regression dialog box – Advanced tab, select the Review descriptive statistics, correlation matrix check box (Fig. 6).



Dependent: PT_POOR
Independent: 1-2 4-7
Input file: Raw Data

□ Advanced options (stepwise or ridge regression)
☑ Review descriptive statistics, correlation matrix
□ Extended precision computations
□ Batch processing/reporting
□ Print/report residual analysis

Specify all variables for the analysis; additional models (indep./dep. vars) can be specified later. For stepwise regression etc. check the advanced options check box.

See also the General Regression Models (GRM) module.

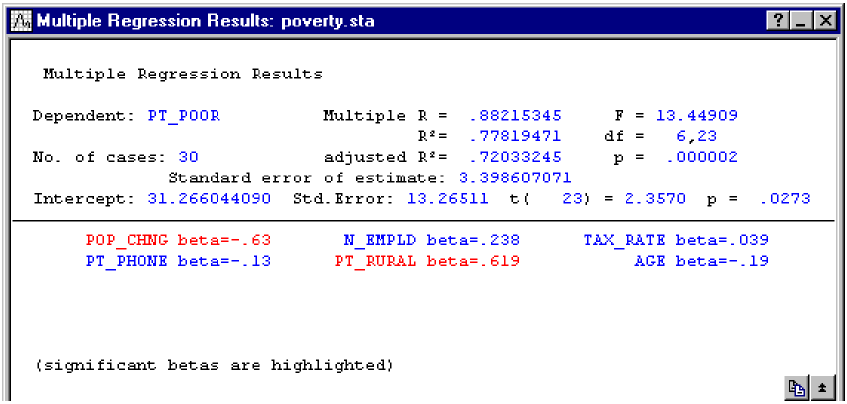
□ Weighted moments

DF =
● W-1 ○ N-1

MD deletion
● Casewise
○ Pairwise
○ Mean substitution

Fig. 6. Multiple Linear Regression

When everything is ready – press OK button. If everything is correct – Multiple Regression Results window will appear (fig. 7). At the bottom of that window you will have all necessary functions for further statistical analysis.



```
 Multiple Regression Results

Dependent: PT_POOR          Multiple R =  .88215345     F = 13.44909
                                   R²=  .77819471    df =   6,23
No. of cases: 30            adjusted R²=  .72033245     p =  .000002
             Standard error of estimate: 3.398607071
Intercept: 31.266044090  Std.Error: 13.26511  t(   23) = 2.3570  p =  .0273

     POP_CHNG beta=-.63        N_EMPLD beta=.238      TAX_RATE beta=.039
     PT_PHONE beta=-.13       PT_RURAL beta=.619           AGE beta=-.19



(significant betas are highlighted)
```

Fig. 7. Multiple Regression Results

Output from Analysis

When you perform an analysis, Statistica generates output in the form of multimedia tables (spreadsheets) and graphs. There are five basic channels to which you can direct all output:

1. Workbooks.
2. Stand-Alone Windows.
3. Reports.
4. Microsoft Word.
5. The Web Resources.

The first four output channels listed above are controlled by the options on the Output Manager tab of the Options dialog box.

There are a number of ways to output to the Web, depending on the version of STATISTICA you have.

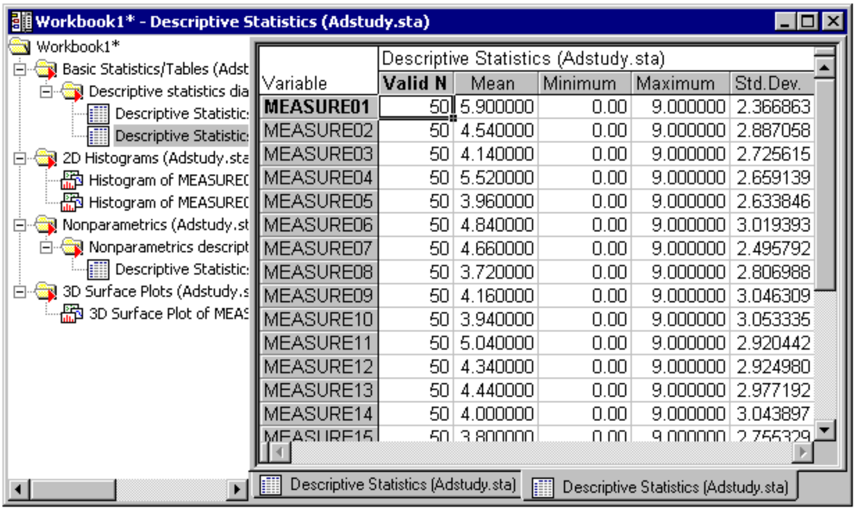
These means for output channels can be used in many combinations (a workbook and report simultaneously) and can be customized in a variety of ways.

All output objects (spreadsheets and graphs) placed in each of the output channels can contain other embedded and linked objects and documents, so Statistica output can be hierarchically organized in a variety of ways.

Each of the Statistica output channels has its unique advantages.

Workbooks are the default way of managing output.

Each output document (a Statistica Spreadsheet or Graph, as well as a Microsoft Word or Excel document) is stored as a tab in the workbook (fig. 8).



Workbook1* - Descriptive Statistics (Adstudy.sta)

- Workbook1*
  - Basic Statistics/Tables (Adst
    - Descriptive statistics dia
      - Descriptive Statistic:
      - Descriptive Statistic:
  - 2D Histograms (Adstudy.sta
    - Histogram of MEASUREC
    - Histogram of MEASUREC
  - Nonparametrics (Adstudy.st
    - Nonparametrics descript
      - Descriptive Statistic:
  - 3D Surface Plots (Adstudy.s
    - 3D Surface Plot of MEAS

	Descriptive Statistics (Adstudy.sta)				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
MEASURE01	50	5.900000	0.00	9.000000	2.366863
MEASURE02	50	4.540000	0.00	9.000000	2.887058
MEASURE03	50	4.140000	0.00	9.000000	2.725615
MEASURE04	50	5.520000	0.00	9.000000	2.659139
MEASURE05	50	3.960000	0.00	9.000000	2.633846
MEASURE06	50	4.840000	0.00	9.000000	3.019393
MEASURE07	50	4.660000	0.00	9.000000	2.495792
MEASURE08	50	3.720000	0.00	9.000000	2.806988
MEASURE09	50	4.160000	0.00	9.000000	3.046309
MEASURE10	50	3.940000	0.00	9.000000	3.053335
MEASURE11	50	5.040000	0.00	9.000000	2.920442
MEASURE12	50	4.340000	0.00	9.000000	2.924980
MEASURE13	50	4.440000	0.00	9.000000	2.977192
MEASURE14	50	4.000000	0.00	9.000000	3.043897
MEASURE15	50	3.800000	0.00	9.000000	2.755329

Descriptive Statistics (Adstudy.sta)
Descriptive Statistics (Adstudy.sta)

Fig. 8. Workbook Window

Documents can be organized into hierarchies of folders or document nodes (by default, one is created for each new analysis) using a tree view, in which individual documents, folders, or entire branches of the tree can be flexibly managed.

Lesson 1.
Statistical Distribution. Empirical Distribution Function and Its Properties. Polygon and Histogram. Numerical Characteristics of the Sample

1.1. Brief Theoretical Information

The general population is a set of elements united by some attribute, from which a sample is made.

A sample set or sample is a set of objects randomly selected for research.

The sample size is the number of objects included in the sample.

Let a sample of size n be taken from the population.

A sample set arranged in ascending or descending order of the attribute value is called a *variational series*, and its objects are *variants*.

If the values of the variant coincide or differ slightly, then they can be grouped by giving a frequency to each variant. As a result, we obtain a *grouped variational series*.

The frequency or relative frequency of options is the ratio of the frequency of options to the sample size.

$$\omega_i = \frac{m_i}{n}. \tag{1.1}$$

A statistical distribution is a correspondence, according to which each possible value of the variants is assigned the frequency (relative frequency) of its occurrence. The statistical distribution is written in the form of a table, in which the first line lists all the values of the option, and the second line lists the frequencies that correspond to the options

x_i	x_1	x_2	x_3	…	x_k
m_i	m_1	m_2	m_3	…	m_k

$$\sum_{i=1}^{k} m_i = n$$

To build an interval statistical series, the set of options is divided into half-intervals $[a_i; a_{i+1})$, i. e. produce grouping. It is recommended to determine the number of intervals k by the formula

$$k = 1 + 1.4 \cdot \ln n. \qquad (1.2)$$

The length of the interval is

$$\Delta = \frac{x_{\max} - x_{\min}}{k}. \qquad (1.3)$$

For clarity, graphic representations of variational series are used in the form of a polygon and a histogram.

A polygon of frequencies is a broken line connecting points with coordinates $(x_i; m_i)$ or $(x_i; \omega_i)$.

A histogram of frequencies is called a stepped figure, composed of rectangles with a base Δ and height $\frac{m_i}{\Delta}$ or $\frac{\omega_i}{\Delta}$.

The empirical distribution function is called the function $F^*(x)$, defining for each value x the relative frequency of the event $X < x$:

$$F^*(x) = \omega(X < x) = \frac{m_x}{n}, \qquad (1.4)$$

where m_x – is the number of options (taking into account their multiplicities) less than x;

n – is the sample size.

The empirical distribution function has the following properties:

1. The values of the empirical function belong to the segment $[0; 1]$.

2. The empirical function is a non-decreasing function.

3. If x_1 – is the smallest value of the options, and x_κ – is the largest value of the options, then $F^*(x) = 0$ at $x \le x_1$ и $F^*(x) = 1$ at $x > x_k$.

To describe the sample, such numerical characteristics as the sample mean, sam-ple variance, and sample standard deviation are used.

The sample mean is the mean value of the options calculated from the sample data

$$\overline{x}_s = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ or } \overline{x}_s = \frac{1}{n}\sum_{i=1}^{k} m_i x_i,$$

where m_i – frequency of the option x_i.

The sample variance is the variance calculated from the sample data

$$D_s = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}_s\right)^2 \text{ or } D_s = \frac{1}{n}\sum_{i=1}^{k} m_i\left(x_i - \overline{x}_s\right)^2 .$$

The sample variance is equal to the difference between the mean of the square of the option and the square of the sample mean

$$D_s = \overline{X^2} - \left(\overline{x}_s\right)^2, \text{ where } \overline{X^2} = \frac{1}{n}\sum_{i=1}^{k} x_i{}^2 m_i .$$

The sample standard deviation is the square root of the sample variance:

$$\sigma_s = \sqrt{D_s} .$$

Task 1.1.1. Based on the given sample distribution, find the empirical distribution function and construct a frequency polygon.

x_i	1	3	5	7	9
m_i	6	11	23	7	3

Solution

Determine the sample size $m = \sum_{i=1}^{k} m_i = 6 + 11 + 23 + 7 + 3 = 50.$

Define the relative frequencies of the variant $\omega_i = \frac{m_i}{n}.$

x_i	1	3	5	7	9
ω_i	0.12	0.22	0.46	0.14	0.06

Write the empirical distribution function:

$$F^*(x) = \omega(X < x) = \begin{cases} 0, & x \le 1; \\ 0,12, & 1 < x \le 3; \\ 0,34, & 3 < x \le 5; \\ 0,80, & 5 < x \le 7; \\ 0,94, & 7 < x \le 9; \\ 1, & x > 9. \end{cases}$$

Build a polygon of frequencies (fig. 1.1).

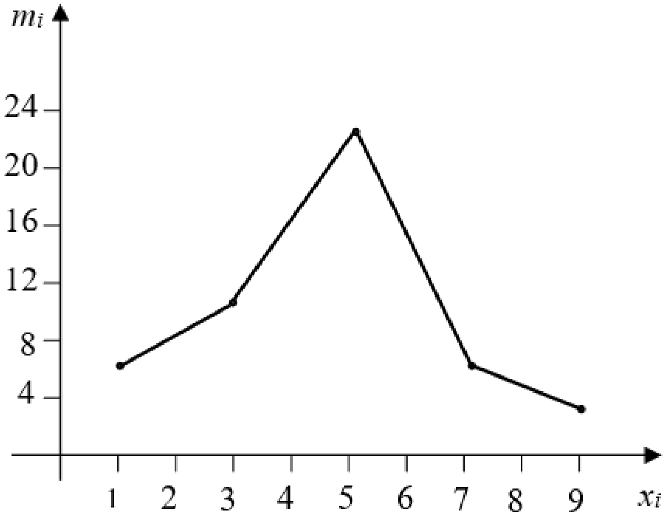


Fig. 1.1

Task 1.1.2. Construct a frequency histogram based on the sample size 100 and calculate the numerical characteristics of the sample.

$x_i - x_{i+1}$	1–5	5–9	9–13	13–17	17–21
m_i	10	20	50	12	8

Solution

Calculate the relative frequencies using the formula $\omega_i = \frac{m_i}{n}$ and find the heights of the rectangles using the formula $h_i = \frac{\omega_i}{h}$, where $h = 4$. Summarize the calculations in a table.

$x_i - x_{i+1}$	1–5	5–9	9–13	13–17	17–21
ω_i	0.1	0.2	0.5	0.12	0.08
h_i	0.025	0.05	0.125	0.03	0.02

Build a histogram of frequencies (fig. 1.2).

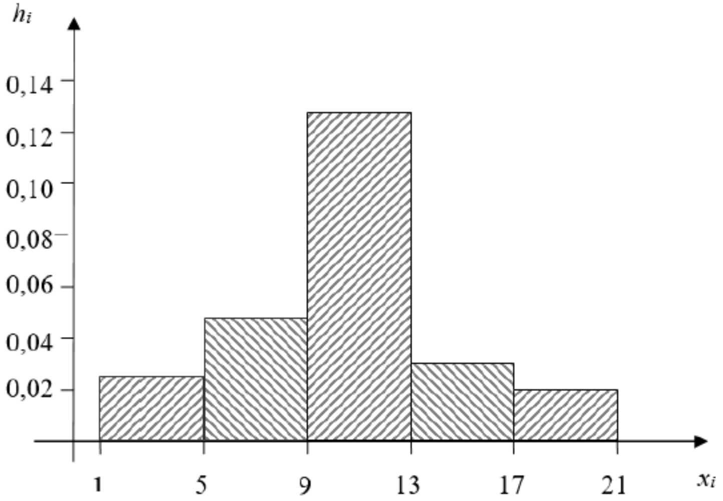


Fig. 1.2

Calculate the numerical characteristics of the sample:

$$\overline{x}_s = \frac{1}{n}\sum_{i=1}^{k} x_i^* m_i = \frac{1}{100}\left(3\cdot 10 + 7\cdot 20 + 11\cdot 50 + 15\cdot 12 + 19\cdot 8\right) =$$
$$= \frac{1}{100}\cdot\left(30 + 210 + 550 + 180 + 152\right) = \frac{1122}{100} = 11.22.$$

$$D_s = \overline{X^2} - \left(\overline{x}_в\right)^2.$$

Calculate D_s and σ_s

$$\overline{X^2} = \frac{1}{n}\sum_{i=1}^{k}\left(x_i^*\right)^2 m_i = \frac{1}{100}\left(9\cdot 10 + 49\cdot 20 + 121\cdot 50 + 225\cdot 12 + 361\cdot 8\right) =$$
$$= \frac{1}{100}\left(90 + 980 + 6050 + 2700 + 2888\right) = \frac{12\,708}{100} = 127.08.$$

$$D_s = 127.08 - \left(11.22\right)^2 = 1.1916.$$

$$\sigma_s = 1.092.$$

1.2. Tasks for Classroom Work

1.2.1. Measurements of deviations from the nominal value of 50 bearings in microns are given:

–1.752; –0.291; –0.933; –0.450; 0.512;
–1.256; 1.701; 0.634; 0.720; 0.490;
1.531; –0.433; 1.409; 1.730; –0.266;
–0.058; 0.248; –0.095; –1.488; –0.361;
0.415; –1.382; 0.129; –0.361; –0.087;
–0.329; 0.086; 0.130; –0.244; –0.882;
0.318; –1.087; 0.899; 1.028; –1.304;
0.349; –0.293; –0.883; –0.056; 0.757;
–0.059; –0.539; –0.078; 0.229; 0.194;
–1.084; 0.318; 0.367; –0.992; 0.529.

Build an interval statistical series for this sample.

$x_i - x_{i+1}$	–1.75– (–1.25)	–1.25– (–0.75)	–0.75 – (–0.25)	–0.25– 0.25	0.25– 0.75	0.75– 1.25	1.25– 1.75
m_i	5	8	9	12	9	3	4

1.2.2. The height is measured (with an accuracy of up to cm) of 30 randomly selected students:

178; 160; 154; 183; 155; 153; 167; 186; 163; 155;
157; 175; 170; 166; 159; 173; 182; 167; 171; 169;
179; 165; 156; 179; 158; 171; 175; 173; 164; 172.

Build an interval statistical series.

$x_i - x_{i+1}$	150–156	156–162	162–168	168–174	174–180	180–186
m_i	4	5	6	7	5	3

1.2.3. Based on a sample of 100, find an empirical function and build a frequency polygon.

$x_i - x_{i+1}$	9–12	12–15	15–18	18–21	21–24	24–27
m_i	6	12	33	22	19	8

$$F^*(x)=\begin{cases}0. & x\le 9;\\ 0.06, & 9<x\le 12;\\ 0.18, & 12<x\le 15;\\ 0.51, & 15<x\le 18;\\ 0.73, & 18<x\le 21;\\ 0.95, & 21<x\le 24;\\ 1, & x>24.\end{cases}$$

1.2.4. Find the empirical distribution function and build a polygon of frequencies according to the following data.

x_i	1	2	3	4	5
m_i	4	6	16	26	48

$$F^*(x)=\begin{cases}0, & x\le 1;\\ 0.04, & 1<x\le 2;\\ 0.1, & 2<x\le 3;\\ 0.26, & 3<x\le 4;\\ 0.52, & 4<x\le 5;\\ 1, & x>5.\end{cases}$$

1.2.5. Build frequency histogram from sample data.

$x_i - x_{i+1}$	2–7	7–12	12–17	17–22	22–27
m_i	5	10	25	6	4

1.2.6. Construct a histogram of frequencies and find the empirical distribution function for a sample of 100 data.

$x_i - x_{i+1}$	154–158	158–162	162–166	166–170	170–174	174–178	178–182
m_i	10	14	26	28	14	6	2

$$F^{*}(x)=\begin{cases}0, & x\le 154;\\ 0.1, & 154<x\le 158;\\ 0.24, & 158<x\le 162;\\ 0.5, & 162<x\le 166;\\ 0.78, & 166<x\le 170;\\ 0.92, & 170<x\le 174;\\ 0.98, & 174<x\le 178;\\ 1, & x>178.\end{cases}$$

1.2.7. Find numerical characteristics from sample data.

a)

x_i	1	3	6	26
m_i	8	40	10	2

$$\left(\overline{x}_s=4;\quad D_s=18.67;\quad \sigma_s=4.32\right)$$

b)

x_i-x_{i+1}	40.1–40.2	40.2–40.3	40.3–40.4	40.4–40.5	40.5–40.6
m_i	7	24	34	26	9

$$\left(\overline{x}_s=40.356;\quad D_s=0.011;\quad \sigma_s=0.0105\right)$$

1.2.8. To test the ore crushing equipment, 50 samples of the processed mineral were randomly selected and measured. Find the sample mean, sample variance and sample standard deviation:

0.030; 0.559; 0.407; 2.784; 0.518; 1.185; 1.297;
0.614; 0.171; 0.155; 0.081; 30.02; 3.554; 1.155;
2.664; 1.889; 0.114; 6.038; 7.815; 0.074; 21.370;
0.412; 16.740; 31.820; 0.587; 2.010; 0.558; 0.171;
0.894; 4.545; 0.147; 1.642; 0.827; 0.051; 0.486;
0.889; 0.340; 0.856; 1.581; 1.474; 2.293; 0.063;
1.294; 0.009; 0.114; 1.889; 2.083; 0.138; 2.881;
0.114.

$$\left(\overline{x}_s=4.67;\quad D_s=39.25;\quad \sigma_s=6.26\right)$$

1.3. Tasks for Independent Work

1.3.1. Compile an empirical distribution function and build a frequency polygon from the sample data.

x_i	15	16	17	18	19
m_i	1	4	5	6	4

$$F^*(x)=\begin{cases}0, & x\le 15;\\ 0.05, & 15<x\le 16;\\ 0.25, & 16<x\le 17;\\ 0.5, & 17<x\le 18;\\ 0.8, & 18<x\le 19;\\ 1, & x>19.\end{cases}$$

1.3.2. Compile an empirical distribution function and build a histogram of frequencies according to the sample data.

$x_i - x_{i+1}$	10–20	20–30	30–40	40–50	50–60	60–70	70–80
m_i	1	2	7	18	12	8	2

$$F^*(x)=\begin{cases}0, & x\le 10;\\ 0.02, & 10<x\le 20;\\ 0.06, & 20<x\le 30;\\ 0.20, & 30<x\le 40;\\ 0.56, & 40<x\le 50;\\ 0.80, & 50<x\le 60;\\ 0.96, & 60<x\le 70;\\ 1, & x>70.\end{cases}$$

1.3.3. The interval of trains in the subway is 2 minutes. The values of a random variable X are given – the time the passenger waits for the train. Compose an interval variation series and find the average waiting time:

0.000; 0.002; 0.007; 0.025; 0.089; 0.312; 1.068; 1.604; 0.014;
0.045; 1.747; 1.677; 0.341; 0.952; 0.645; 1.297; 1.981; 0.214;
1.452; 0.787; 1.654; 0.838; 0.143; 1.317; 0.618; 1.853; 1.555;
0.653; 1.922; 1.653; 0.617; 0.828; 1.413; 1.030; 1.459; 1.483;
1.769; 1.265; 1.669; 0.635; 0.787; 1.004; 0.941; 0.612; 1.200;
1.692; 1.356; 0.908; 1.245; 1.295.

$$\left(\overline{x}_s = 1.022\right)$$

1.3.4. Calculate sample variance from sample data.

x_i	340	360	375	380
m_i	20	50	18	12

$$\left(D_s = 167.29\right)$$

1.3.5. Calculate numerical characteristics of the sample.

$x_i - x_{i+1}$	10–20	20–30	30–40	40–50	50–60
m_i	1	8	10	3	3

$$\left(\overline{x}_s = 34.6; \quad D_s = 107.84; \quad \sigma_s = 10.35\right)$$

Lesson 2.
Point Estimates of Unknown Distribution Parameters

2.1. Brief Theoretical Information

Let random variable X be studied with a distribution law depending on one or more parameters. It is required to evaluate the unknown parameter from the sample obtained as a result of the tests θ.

A point estimate of an unknown parameter θ of a theoretical distribution is its approximate value, which depends on the sample data:

$$\overline{\theta} = \overline{\theta}\left(x_1, x_2, x_3, ..., x_n\right).$$

A point estimate must meet the following requirements:

– the estimate must be unbiased, i. e. $M\left(\overline{\theta}\right) = \theta$;

– the assessment must be consistent, i. e. it must converge in probability to the estimated parameter: for $\forall \varepsilon > 0 \quad \lim\limits_{n\to\infty} P\left(\left|\overline{\theta} - \theta\right| < \varepsilon\right) = 1$;

– the estimate must be effective: if an unknown parameter has several estimates, then the estimate with the smallest variance should be taken as an estimate.

The sample mean $\overline{x}_s$ is an unbiased and consistent estimate of the population mean.

The unbiased and consistent estimate for the population variance is the corrected sample variance

$$D_c = s^2 = \frac{n}{n-1} D_s.$$

The corrected standard deviation is the square root of the corrected variance.

$$s = \sqrt{D_c}.$$

Many methods have been developed for calculating $\overline{x}_s$ and D_s. One of the most common methods is **the product method**. When calculating the sample mean and sample variance, proceed as follows:

– choose “false zero” c. As a “false zero” is taken the variant standing in the middle of the variational series or variant, having the maximum frequency;

– pass to conditional variants U_i according to the formulas $U_i = \frac{x_i - c}{h}$, where h – is the partitioning step;

– calculate the conditional moments of the 1st and 2nd orders:

$$M_1^* = \frac{1}{h}\sum_{i=1}^{k} U_i m_i; \qquad M_2^* = \frac{1}{h}\sum_{i=1}^{k} U_i^2 m_i;$$

– calculate the sample mean $\overline{x}_s$ and sample variance D_s:

$$\overline{x}_s = M_1^* h + c; \qquad D_s = \left(M_2^* - \left(M_1^*\right)^2 \right) h^2.$$

Task 2.1.1. Use the product method to calculate the sample mean and sample variance from the sample data.

x_i	65	70	75	80	85
m_i	2	5	25	15	3

Solution

As a “false zero” take option 75, $c = 75$. Move on to conditional options according to the formula $U_1 = \frac{x_i - c}{h}$. The results of the calculations are summarized in a table.

x_i	m_i	U_i	$U_i n_i$	$U_i^2 n_i$	$(U_i + 1)^2 m_i$
65	2	–2	–4	8	2
70	5	–1	–5	5	0
$\boxed{75}$	25	0	0	0	25
80	15	1	15	15	60
85	3	2	6	12	27
Σ	50	0	12	40	114

The calculation results can be checked by the equality:

$$\sum_{i=1}^{k}(U_i+1)m_i=\sum_{i=1}^{k}m_i+2\sum_{i=1}^{k}m_iU_i+\sum_{i=1}^{k}U_i{}^2m_i.$$

$$114=50+2\cdot 12+40.$$

The equality is satisfied, therefore, the table is filled correctly.
Let's calculate the conditional moments:

$$M_1^*=\frac{1}{50}\cdot 12=0.24;\qquad M_2^*=\frac{1}{50}\cdot 40=0.8.$$

Calculate the sample mean and sample variance:

$$\overline{x}_s=M_1^*h+C=0{,}24\cdot 5+75=76{,}2;$$
$$D_s=\left(M_2^*-\left(M_1^*\right)^2\right)h^2=(0.8-0.0576)\cdot 25=18.56.$$

2.2. Tasks for Classroom Work

2.2.1. Find unbiased estimates for the mean and variance from the sample data.

$x_i - x_{i+1}$	7,8–8,0	8,0–8,2	8,2–8,4	8,4–8,6	8,6–8,8	8,8–9,0
m_i	5	20	80	95	40	10

$$\left(\overline{x}_s=8.44;\quad D_c=0.042\right)$$

2.2.2. Find unbiased estimate for variance from sample data.

x_i	102	104	108
m_i	2	3	5

$$\left(D_c=6.93\right)$$

2.2.3. Based on the sample data, find unbiased estimates for the mathematical expectation and variance of the general population:

a) positive deviations from the nominal size in the batch of parts (in microns):

17; 21; 8; 20; 23; 18; 22; 20; 17; 12;
20; 11; 9; 19; 20; 9; 19; 17; 21; 13;
17; 22; 22; 10; 20; 20; 15; 19; 20; 20;
13; 21; 21; 9; 14; 11; 19; 18; 23; 19;

$$\left(\overline{x}_s = 17.2;\quad D_c = 19.7\right)$$

b) response time (in seconds):

8.5; 7.1; 6.7; 6.2; 2.9; 4.4; 6.0; 5.8; 5.4;
8.2; 6.9; 6.5; 6.1; 3.8; 6.0; 6.0; 5.6; 5.3;
7.7; 6.8; 6.5; 6.1; 4.2; 4.7; 5.6; 5.4; 5.3;
7.4; 6.7; 6.4; 6.1; 4.5; 6.0; 5.8; 5.6; 5.1.

$$\left(\overline{x}_s = 5.9;\quad D_c = 1.3\right)$$

2.2.4. The results of observations of the service life of 100 similar machines before going beyond the accuracy standards are given.

$x_i - x_{i+1}$	20–25	25–30	30–35	35–40	40–45
m_i	9	24	35	22	10

Find unbiased estimate for lifetime variance.

$$\left(D_c = 30{,}14\right)$$

2.2.5. The results of measuring the diameter of bushings processed by the machine are given.

$x_i - x_{i+1}$	20.00–20.04	20.04–20.08	20.08–20.12	20.12–20.16	20.16–20.20
m_i	8	18	45	20	9

Find estimates for the mathematical expectation and variance.

$$\left(\overline{x}_s = 20.1016;\quad D_c = 0.002\right)$$

2.3. Tasks for Independent Work

2.3.1. Voltage deviations from nominal (mV) are given.

$x_i - x_{i+1}$	0.00–0.02	0.02–0.04	0.04–0.06	0.06–0.08	0.08–0.10	0.10–0.12	0.12–0.14	0.14–0.16
m_i	9	15	29	35	32	19	8	3

Find estimates for the mathematical expectation and variance.

$$\left(\overline{x}_s = 0.05; \quad D_c = 0.02\right)$$

2.3.2. Given the yield of rye in different parts of the field.

Yield, q/ha	9–12	12–15	15–18	18–21	21–24	24–27
Quantity of parts	6	12	33	22	19	8

Find an estimate for the average yield of the entire field.

$$\left(\overline{x}_s = 18.3\right)$$

2.3.3. For this sample, find estimates for the mathematical expectation and variance of the general population.

x_i	12	14	16	18	20	22
m_i	5	15	50	16	10	4

$$\left(\overline{x}_s = 16.46; \quad D_c = 4.92\right)$$

Lesson 3.
Interval Estimates

3.1. Brief Theoretical Information

Let $\theta^* = \theta^*\left(x_1,...,x_n\right)$ – be a sampling function. This is a random variable called **a statistic**.

An interval estimate – is estimate, which is determined by a random interval $\left(\theta_1^*, \theta_2^*\right)$, $\theta_1^* < \theta_2^*$. Confidence intervals are used as an interval estimate.

A confidence interval for an unknown parameter θ, is a random interval $\left(\theta_1^*, \theta_2^*\right)$, which with a given probability γ (reliability) covers the unknown parameter, θ.

If the studied RV is distributed according to the normal law with a known standard deviation σ, then the confidence interval for the mathematical expectation is determined by the inequality

$$\overline{x}_s - t_\gamma \frac{\sigma}{\sqrt{n}} < a < \overline{x}_s + t_\gamma \frac{\sigma}{\sqrt{n}}, \tag{3.1}$$

where $\delta = t_\gamma \frac{\sigma}{n}$ – is the estimation accuracy;

n – is the sample size;

t_γ – is the value of the argument of the Laplace function, at which $\Phi\left(t_\gamma\right) = \frac{\gamma}{2}$.

If the standard deviation is unknown, then the confidence interval for the mathematical expectation of the investigated RV is determined by the inequality:

$$\overline{x}_s - t_{\gamma,n} \frac{s}{\sqrt{n}} < a < \overline{x}_s + t_{\gamma,n} \frac{s}{\sqrt{n}}, \quad \text{where } s = \sqrt{D_c}. \tag{3.2}$$

The values $t_{\gamma,n}$ are found according to the table of Appendix 5 for the given n and γ. The number $\delta = t_{\gamma,n}\frac{s}{\sqrt{n}}$ is called the accuracy of the estimate of the mathematical expectation.

The confidence interval for the standard deviation of the studied RV is determined by the inequality

$$s\,q_1 < \sigma < s\,q_2, \qquad (3.3)$$

The values q_1 and q_2 are found according to the table in Appendix 6 for the given γ and n.

Task 3.1.1. Find the confidence interval for estimating with reliability $\gamma = 0.99$ of the unknown mathematical expectation of a normally distributed feature X, if is known $\sigma = 4$, and according to a sample size of 100 calculated $\overline{x}_s = 12{,}4$.

Solution

Since the standard deviation of the RV is known, we use inequality (3.1) to determine the confidence interval for the mathematical expectation. Let's define the value t_γ: $\Phi\left(t_\gamma\right) = \frac{\gamma}{2} = \frac{0.99}{2} = 0.495 \Rightarrow t_\gamma = 2.58$. Substitute into inequality (3.1):

$$12.4 - 2.58\frac{4}{10} < a < 12.4 + 2.58\frac{4}{10}; \qquad 11.08 < a < 13.432.$$

Task 3.1.2. To study a normally distributed RV, a sample of 25 is chosen:

$x_i - x_{i+1}$	10–20	20–30	30–40	40–50	50–60
m_i	1	8	10	3	3

Find with reliability $\gamma = 0{,}95$ the confidence intervals for the mathematical expectation and the standard deviation of the studied RV.

Solution

According to the sample data, by the method of products, we determine $\overline{x}_s$ and s:

x_i^*	m_i	U_i	$U_i m_i$	$U_i^2 m_i$	$(U_i+1)^2 m_i$
15	1	–2	–2	4	1
25	8	–1	–8	8	0
35	10	0	0	0	10
45	3	1	3	3	12
55	3	2	6	12	27
$\sum$	25	0	–1	27	50

Verification:

$$50 = 25 + 2(-1) + 27 = 50.$$

$$M_1^* = \frac{-1}{25} = -0.04; \quad M_2^* = \frac{27}{25} = 1.08.$$

$$\overline{x}_s = 35 + (-0.04)\cdot 10 = 34.6; \quad D_s = \left(1.08 - (0.04)^2\right)\cdot 100 = 107.84;$$

$$D_c = \frac{n}{n-1} D_s = \frac{25}{24}\cdot 107.84 = 112.33; \quad s = \sqrt{D_c} = 10.6.$$

To determine the confidence interval for the mathematical expectation, we use inequality (3.2):

$$t_{\gamma,n} = t(0.95; 24) = 2.064;$$

$$34.6 - 2.064\frac{10.6}{5} < a < 34.6 + 2.064\frac{10.6}{5};$$

$$34.6 - 4.38 < a < 34.6 + 4.38;$$

$$30.22 < a < 38.98.$$

To determine the confidence interval for the standard deviation, we use inequality (3.3):

$q_1 = 0.781; \quad q_2 = 1.391;$

$10.6 \cdot 0.781 < \sigma < 10.6 \cdot 1.391;$

$8.28 < \sigma < 14.74.$

3.2. Tasks for Classroom Work

3.2.1. Selective studies were carried out in one of the fish farms to determine the weight gain of fish per year. The carps bred in the pond were weighed and released back.

The results of 100 such measurements showed that the annual weight gain of fish was on average 200 g, and the variance was 320 g. Find, with a reliability of 0.95, a confidence interval for the annual weight gain of fish ΔP.

$(196.49 < \Delta P < 203.51)$

3.2.2. The same device with the standard deviation of random measurement errors $\sigma = 40$ m carried out five different distance measurements. Find with a reliability of 0.95 a confidence interval for estimating the true distance, if the average of all measurements taken is $\overline{x}_s = 2000$ m.

$(1964.94 < a < 2035.06)$

3.2.3. A sample from a large batch of electric lamps contains 100 lamps. The average burning time of the lamps in the sample is 1000 hours. Find, with a reliability of 0.95, a confidence interval for the burning time of the lamps of the entire batch, if it is known that the standard deviation of the lamp burning time is $\sigma = 40$ hours.

$(992.16 < a < 1007.84)$

3.2.4. Find the minimum sample size at which, with a reliability of 0.925, the accuracy of estimating the mathematical expectation of a normally distributed RV is 0.2, if the standard deviation is $\sigma = 1.5$.

$(n = 179)$

3.2.5. Based on the sample data, find a confidence interval covering the standard deviation with a reliability of 0.99.

x_i	0,1	0,2	0,3	0,4	0,5
m_i	2	4	7	6	1

$(0.077 < \sigma < 0.583)$

3.2.6. From the general population of RV, distributed according to the normal law, 100 RV values were selected. Find confidence intervals for mathematical expectation and standard deviation with a reliability of 0.95.

$x_i - x_{i+1}$	100–120	120–140	140–160	160–180	180–200
m_i	17	40	32	8	3

$$\begin{pmatrix} 125.78 < a < 127.42 \\ 3.64 < \sigma < 4.82 \end{pmatrix}$$

3.2.7. In order to determine the average amount Q of deposits in a bank, a sample was made:

Sum, thousands of USD	10–30	30–50	50–70	70–90	90–110	110–130
m_i	1	3	10	30	60	7

Find the limits of the average contribution with a reliability of 0.95.

$$\left(86.45 < Q < 93.37\right)$$

3.3. Tasks for Independent Work

3.3.1. He automatic machine stamps rollers. Based on a sample size of 100, the sample mean $\overline{x}_s = 12{,}5$ and $s = 2{,}1$ are calculated. Find with a reliability of 0.95 confidence intervals for the mathematical expectation and standard deviation.

$$\begin{pmatrix} 12.08 < a < 12.92 \\ 1.84 < \sigma < 2.44 \end{pmatrix}$$

3.3.2. Find the minimum sample size at which, with a reliability of 0.975, the accuracy of estimating the mathematical expectation a by the sample mean is equal $\delta = 0{,}3,$ if $\sigma = 1{,}2.$

$$\left(n = 81\right)$$

3.3.3. Find with a reliability of 0.99 confidence intervals for the mathematical expectation and the standard deviation from the sample data:

a)

x_i	12	14	16	18	20	22
m_i	5	15	50	16	10	4

$$\left(\begin{array}{l} 15.88 < a < 17.04 \\ 1.88 < \sigma < 2.71 \end{array}\right)$$

b)

$x_i - x_{i+1}$	46–50	50–54	54–58	58–62	62–66	67–70	70–74	74–78	78–82	82–66
m_i	2	4	6	8	12	30	18	8	7	5

$$\left(\begin{array}{l} 65.99 < a < 70.25 \\ 6.84 < \sigma < 9.86 \end{array}\right)$$

Lesson 4.
Statistical testing of hypotheses.
Pearson's and Kolmogorov's goodness-of-fit criteria

4.1. Brief Theoretical Information

Statistical is the hypothesis about the supposed form of the unknown distribution RV or about the values of the parameters of the known type of distribution. **The null hypothesis** H_0 is the proposed hypothesis. **A competing (alternative) hypothesis** is one that contradicts the null hypothesis.

When testing a statistical hypothesis, errors of two kinds can be made. **Type I error** – the correct hypothesis will be rejected. **Type II error** – the wrong hypothesis will be accepted.

The probability of making a Type I error is called the level of significance. To test a statistical hypothesis, a special statistic is used, which is called a criterion (or test). According to the calculated value of the criterion, it is determined to accept or reject the null hypothesis. **The goodness-of-fit test** – is a test of the hypothesis about the type of distribution of RV.

The main goodness-of-fit criteria are the Pearson's χ^2 and Kolmogorov's criteria.

When testing a hypothesis using the Pearson's test, proceed as follows:

– a sample of volume n is taken from the general population;

– according to sample calculate $\overline{x}_s$ and σ_s;

– go to the normalized RV according to the formula $U_i = \dfrac{x_i - \overline{x}_s}{\sigma_s}$;

– find the probability of getting into the interval $(U_i,\ U_{i+1})$, $P_i = \Phi(U_{i+1}) - \Phi(U_i)$;

– calculate theoretical treguensies $m_i' = nP_i$;

– calculate the Pearson's statistics $\chi^2_{obs} = \sum\limits_{i=1}^{k} \dfrac{(m_i - m_i')^2}{m_i'}$;

– from the table of critical points of the Pearson's distribution (Appendix 3) by the level of significance α and the number of degrees of freedom $\nu = k - 1 - r$ determine χ^2_{cr}, where k – is the number of intervals in the variational series; r – is the number of parameters of the distribution law that are estimated from the sample (for the normal law $r = 2$);

– if $\chi^2_{obs} \le \chi^2_{cr}$, then there is no need to reject the null hypothesis, i. e. empirical and theoretical frequencies are consistent;

– if $\chi^2_{obs} > \chi^2_{cr}$, then the hypothesis is rejected, i. e. the discrepancy between theoretical and empirical frequencies is significant;

– if a discrete RV distributed according to the normal law is being investigated, then the theoretical probabilities are determined by the formula $p_i = \dfrac{h}{\sigma_в}\varphi(U_i)$, where h – is the step, $U_i = \dfrac{x_i - \overline{x}_s}{\sigma_s}$, $\varphi(x) = \dfrac{1}{\sqrt{2\pi}}e^{-x^2/2}$.

Task 4.1.1. Using the Pearson's test, at the significance level $\alpha = 0.05$ check whether the hypothesis of the normal distribution of the general population is consistent with the sample data

$x_i – x_{i+1}$	3–8	8–13	13–18	18–23	23–28	28–33	33–38
m_i	6	8	15	40	16	8	7

Solution

Using the sample data, calculate by the method of products $\overline{x}_s$ and D_s.

x_i^*	m_i	U_i	$U_i m_i$	$U_i^2 m_i$	$(U_i + 1)^2 m_i$
5.5	6	–3	–18	54	24
10.5	8	–2	–16	32	8
15.5	15	–1	–15	15	0
20.5	40	0	0	0	40
25.5	16	1	16	16	64
30.5	8	2	16	32	72
35.5	7	3	21	63	112
Σ	100		4	212	320

Verification:

$$320 = 100 + 8 + 212 = 320.$$

$$M_1^* = \frac{4}{100} = 0.04; \quad M_2^* = \frac{212}{100} = 2.12;$$

$$\overline{x}_s = M_1^* h + c = 20.5 + 0.04 \cdot 5 = 20.7;$$

$$D_s = \left(M_2^* - \left(M_1^* \right)^2 \right) h^2 = \left(2.12 - 0.0016 \right) \cdot 25 = 52.96;$$

$$\sigma_s = \sqrt{D_s} = 7.28.$$

Calculate the probabilities of getting into the intervals

x_i	$x_i - \overline{x}_s$	$\dfrac{x_i - \overline{x}_s}{\sigma_s}$	$\Phi\left(\dfrac{x_i - \overline{x}_s}{\sigma_s}\right)$
$-\infty$	$-\infty$	$-\infty$	–0.5
8	–12.7	–1.74	–0.4591
13	–7.7	–1.06	–0.3554
18	–2.7	–0.37	–0.1443
23	2.3	0.32	0.1255
28	7.3	1.00	0.3413
33	12.3	1.69	0.4545
$+\infty$	$+\infty$	$+\infty$	0.5

$$P_1 = -0.4591 + 0.5 = 0.0409;$$

$$P_2 = -0.3554 + 0.4591 = 0.1037;$$

$$P_3 = -0.1443 + 0.3554 = 0.2111;$$

$$P_4 = 0.1255 + 0.1443 = 0.2698;$$

$$P_5 = 0.3413 - 0.1255 = 0.2158;$$

$$P_6 = 0.4545 - 0.3413 = 0.1132;$$

$$P_7 = 0.5 - 0.4545 = 0.0455.$$

Calculate χ^2_{obs}.

m_i	P_i	$m'_i = n\text{P}_i$	$m_i - m'_i$	$(m_i - m'_i)^2$	$(m_i - m'_i)^2 / m'_i$
6	0,0409	4,09	1,91	3,648	0,892
8	0,1037	10,37	–2,37	5,617	0,542
15	0,2111	21,11	–6,11	37,332	1,768
40	0,2698	26,98	13,02	169,520	6,283
16	0,2158	21,58	–5,58	31,136	1,443
8	0,1132	11,32	–3,32	11,022	0,974
7	0,0455	4,55	2,45	6,003	1,319
					$\chi^2_{obs} = 13{,}221$

Determine the number of degrees of freedom:

$\nu = k - 1 - r = 7 - 1 - 2 = 4.$

By the level of significance $\alpha = 0{,}05$ and the number of degrees of freedom $\nu = 4$ find the critical point of the right-handed critical region of the Pearson's distribution (Appendix 3):

$\chi^2_{cr} = \chi^2(0.05;\, 4) = 9.5.$

Since $\chi^2_{obs} > \chi^2_{cr}$, then the hypothesis of a normal population distribution is rejected.

Kolmogorov's goodness-of-fit test is used to test the hypothesis about the law of distribution of continuous RV. To statistically test the hypothesis using the Kolmogorov's goodness-of-fit test, proceed as follows:

– select a sample from the general population;

– according to the sample, make up the empirical distribution function $F^*(x)$;

– write down the theoretical distribution function $F(x)$;

– calculate the value $D = \max\left|F^*(x) - F(x)\right|$;

– calculate the Kolmogorov statistics $\lambda = D\sqrt{n}$, where n – is the sample size. RV λ has a distribution function $K(x) = \sum\limits_{i=-\infty}^{\infty} (-1)^i e^{-2i^2x^2}$, $x > 0$ – called the Kolmogorov's function;

– find λ_α by significance level α (Appendix 7);

– if $\lambda \geq \lambda_\alpha$, then the hypothesis about the RV distribution law is rejected;

– if $\lambda < \lambda_\alpha$, then there is no reason to reject the null hypothesis.

Consider the application of the Kolmogorov's goodness-of-fit test by an example.

Task 4.1.2. Test the hypothesis of the normal distribution of RV according to the sample data using the Kolmogorov's goodness-of-fit test at a significance level $\alpha = 0.05$.

x_{i-1}–x_i	0–0,5	0,5–1	1–1,5	1,5–2	2–2,5	2,5–3	3–3,5	3,5–4
m_i	17	11	9	8	2	1	1	1

Solution

Calculate the sample mean $\overline{x}_s$ and the corrected standard deviation s.

$$\overline{x}_s = \frac{1}{n}\sum_{i=1}^{n} x_i m_i = \frac{1}{50}(0.25 \cdot 17 + 0.75 \cdot 11 + 1.25 \cdot 9 + 1.75 \cdot 8 + 2.25 \cdot 2 +$$

$$+ 2.75 + 3.25 + 3.75) = 1.04;$$

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 m_i = 1.7625;$$

$$D_s = \overline{x}^2 - \left(\overline{x}_s\right)^2 = 1.7625 - 1.04^2 = 0.6809;$$

$$D_c = \frac{n}{n-1} D_s = \frac{50}{49} \cdot 0.6809 = 0.6948;$$

$$s = \sqrt{D_c} = \sqrt{0.6948} = 0,834.$$

Then the theoretical distribution function, assuming that RV is distributed according to the normal law, has the form:

$$F(x) = \frac{1}{2} + \Phi\left(\frac{x - \overline{x}_s}{s}\right) = \frac{1}{2} + \Phi\left(\frac{x - 1,04}{0,834}\right),$$

where $\Phi(x)$ – is the Laplace function.

The empirical distribution function is determined by the formula

$$F^*(x) = \frac{m_x}{n},$$

where m_x – is the sum of frequencies of variants that smaller, then x.

$$F^*(x) = \begin{cases} 0, & x \le 0; \\ 0.34, & 0 < x \le 0.5; \\ 0.56, & 0.5 < x \le 1; \\ 0.74, & 1 < x \le 1.5; \\ 0.90, & 1.5 < x \le 2; \\ 0.94, & 2 < x \le 2.5; \\ 0.96, & 2.5 < x \le 3; \\ 0.98, & 3.0 < x \le 3.5; \\ 1, & x > 3.5. \end{cases}$$

Calculate the value $D = \max\left|F^*(x) - F(x)\right|$.

x_i	$x_i - \overline{x}_s$	$\dfrac{x_i - \overline{x}_s}{s}$	$\Phi\left(\dfrac{x_i - \overline{x}_s}{s}\right)$	$F(x_i) = \dfrac{1}{2} + \Phi\left(\dfrac{x_i - \overline{x}_s}{s}\right)$	$F^*(x_i)$	$\left\|F(x_i) - F^*(x)\right\|$
0	–1.04	–1.25	–0.3944	0.1056	0	0.1056
0.5	–0.55	–0.65	–0.2422	0.2578	0.34	0.0822
1.0	–0.04	–0.05	–0.0199	0.4801	0.56	0.0799
1.5	0.46	0.55	0.2088	0.7088	0.74	0.0312
2.0	0.96	1.15	0.3749	0.8749	0.90	0.0251
2.5	1.46	1.75	0.4599	0.9599	0.94	0.0199
3.0	1.96	2.35	0.4908	0.9906	0.96	0.0306
3.5	2.46	2.95	0.4985	0.9985	0.98	0.0185
4.0	2.96	3.55	0.4999	0.9999	1	0.0001

$D = 0.1056.$

Calculate the Kolmogorov's statistics:

$\lambda = D\sqrt{n} = \sqrt{50} \cdot 0{,}1056 = 0{,}747.$

By the level of significance $\alpha = 0{,}05$ find according to the table (Appendix 7) $\lambda_\alpha = 1{,}358$.

Because $\lambda < \lambda_\alpha$, then there is no reason to reject the hypothesis of a normal distribution.

4.2. Tasks for Classroom Work

4.2.1. Using Pearson's test, at the significance level α check whether it is random or significant discrepancy between the theoretical and empirical frequencies, which are calculated based on the hypothesis of a normal distribution of RV.

a) $\alpha = 0{,}01$

m_i	8	16	40	72	36	18	10
m_i'	6	18	36	76	39	18	7

(Random)

b) $\alpha = 0{,}05$

m_i	5	10	20	8	7
m_i'	6	14	18	7	5

(Random)

4.2.2. Using the Pearson's test, at the significance level $\alpha = 0{,}05$ check whether the hypothesis of a normal population distribution is consistent with the sample data.

a)

$x_i - x_{i-1}$	–20–(–10)	–10–0	0–10	10–20	20–30	30–40	40–50
m_i	20	47	80	89	40	16	8

(Consistent)

b)

x_i	0,3	0,5	0,7	0,9	1,1	1,3	1,5	1,7	1,9	2,1	2,3
m_i	6	9	26	25	30	26	21	24	20	8	5

(Consistent)

4.2.3. Observations of the overhaul intervals T (in months) of the work of the grain harvesting complex gave the following results:

0.000; 0.001; 0.003; 0.012; 0.044; 0.156; 0.534;
0.802; 0.007 0.822; 0.873; 0.838; 0.170; 0.476;
0.322; 0.648; 0.991; 0.107; 0.726; 0.393; 0.827
0.419; 0.071; 0.659; 0.309; 0.927; 0.778; 0.327;
0.961; 0.826; 0.308; 0.414; 0.707; 0.515; 0.729;
0.742; 0.884; 0.632; 0.835; 0.318; 0.394; 0.502;
0.471; 0.306; 0.600; 0.846; 0.678; 0.454; 0.623;
0.648.

Check at the significance level $\alpha = 0,01,$, using the Kolmogorov's test, the hypothesis of the exponential distribution of the population.

(Consistent)

4.2.4. Given the measurement results of 1000 parts.

$x_i - x_{i+1}$	97.25–98.25	98.25–98.75	98.75–99.25	99.25–99.75	99.75-100.25	100.25–100.75	100.75–101.25	101.25–101.75	101.75–102.25	102.25–102.75
m_i	21	47	87	158	181	201	142	97	41	25

At the significance level $\alpha = 0,05$ check whether the sample data are consistent with the hypothesis of a normal distribution using the Kolmogorov test.

(Not consistent)

4.3. Tasks for Independent Work

4.3.1. Using the Pearson's test, at the level of significance $\alpha = 0.05$ check whether the discrepancy between the empirical (m_i) and theoretical $\left(m_i'\right)$ frequencies is random or significant. Frequancies are calculated under the assumption that the general population is distributed according to the normal law.

m_i	14	18	32	70	20	36	10
m_i'	10	24	34	80	18	22	12

(Significant)

4.3.2. Using the Pearson's test, at the significance level $\alpha = 0.05$ check whether the hypothesis of a normal distribution of the population is consistent with the sample data.

$x_i - x_{i+1}$	6–16	16–26	26–36	36–46	46–56	56–66	66–76	76–86
m_i	8	7	16	35	15	8	6	5

(Not consistent)

4.3.3. At the significance level $\alpha = 0{,}05$, use the Kolmogorov's test to check whether the hypothesis of the normal distribution of RV is consistent with the sample data.

$x_i - x_{i-1}$	–20–(–10)	–10–0	0–10	10–20	20–30	30–40	40–50
m_i	20	47	80	89	40	16	8

(Consistent)

Lesson 5.
Sample Correlation Coefficient and Its Properties. Testing the Hypothesis that the Correlation Coefficient Is Equal to Zero

5.1. Brief Theoretical Information

To calculate the sample correlation coefficient, the data are presented in the form of a correlation table. The correlation table is a table of the following form: the first row contains the observed values of RV *X*, the first column contains the observed values of RV *Y*, at the intersection of the *i*-th row and the *j*-th column, the frequency m_{ij} of the occurrence of the pair (y_i, x_j) is recorded. The last column contains the frequency of occurrence of variants y_i, the last line – the frequency of occurrence of variants x_j, at the intersection of the last row and the last column the total number of observations is recorded. The correlation table looks like.

$Y \backslash X$	x_1	x_2	x_3	…	x_k	n_y
y_1	m_{11}	m_{12}	m_{13}	…	m_{1k}	n_{y1}
y_2	m_{21}	m_{22}	m_{23}	…	m_{2k}	n_{y2}
…	…	…	…	…	…	…
y_ℓ	$m_{\ell 1}$	$m_{\ell 2}$	$m_{\ell 3}$	…	$m_{\ell k}$	$n_{y\ell}$
n_x	n_{x1}	n_{x2}	n_{x3}	…	n_{xk}	n

The main estimate of the closeness of the connection between the random variables *X* and *Y* is the sample correlation coefficient r_s, which is defined as follows

$$r_s = \frac{\overline{XY} - \overline{x_s} \cdot \overline{y_s}}{\sigma_x \cdot \sigma_y},$$

where $\overline{XY}$ – is the arithmetic mean of the products of RV *X*, *Y* values.

The properties of the sample correlation coefficient are similar to the properties of the correlation coefficient between RVs:

1. $-1 \le r_s \le 1$.

2. If the variables X and Y are multiplied by the same number, then the correlation coefficient will not change.

3. If $r_s = \pm 1$, then the correlation between the values of X and Y is a linear functional dependence.

To calculate the sample correlation coefficient, the formula is used

$$r_s = \frac{\sum_i \sum_j x_i y_j m_{ij} - n \overline{x_s}\, \overline{y_s}}{n \sigma_x \sigma_y}. \tag{5.1}$$

If $r_s = 0$, then there is no correlation dependence between the observed values X and Y; the closer the modulus of the correlation coefficient approaches *1*, the closer the relationship between the variables X and Y. The sample correlation coefficient is calculated from the sample data, then, unlike the correlation coefficient of the general population r_s, it is a random variable. If $r_s \ne 0$, then the question arises whether this is due to a really existing relationship between RV X and Y or caused by random factors. To clarify this issue, the hypothesis H_0 is tested about the equality of the correlation coefficient r of the general population to zero.

In order to test the null hypothesis about the equality of the correlation coefficient of the general two-dimensional normal population at the significance level α, the statistics are calculated

$$T_{obs} = \frac{r_s \sqrt{n-2}}{\sqrt{1 - r_s^2}}$$

and according to the table of critical points of the Student's distribution (Appendix 4), the critical point of the two-sided critical region $t_{cr} = t\left(\frac{\alpha}{2}, \nu\right)$ is found by the significance level α and the number of degrees of freedom $\nu = n-2$. If $|T_{obs}| < t_{cr}$ – there is no reason to reject the null hypothesis, i. e. $r = 0$; if $|T_{obs}| > t_{cr}$ – the null hypothesis is rejected, i. e. $r \neq 0$. Consider the calculation of the sample correlation coefficient and testing the hypothesis that the correlation coefficient of the general population is equal to zero using an example.

Task 5.1.1. Using this correlation table, calculate the sample correlation coefficient and, at the significance level $\alpha = 0.05$, test the hypothesis that the correlation coefficient of the general population is equal to zero.

x_i y_j	12.5	17.5	22.5	27.5	n_y
20.5	1				1
21.5		2			2
22.5		1	2		3
23.5			3	3	6
24.5				8	8
n_x	1	3	5	11	20

Solution

Compute the components included in formula (5.1) to calculate r_s.

$$\sum x_i n_i = 12.5 \cdot 1 + 17.5 \cdot 3 + 22.5 \cdot 5 + 27.5 \cdot 11 = 480;$$

$$\sum y_j n_j = 468; \quad \sum x_i^2 n_i = 11\,925; \quad \sum y_j^2 n_j = 10\,979;$$

$$\sum \sum x_i y_j n_{ij} = 20.5 \cdot 12.5 + 21.5 \cdot 17.5 \cdot 2 + 22.5 \cdot 17.5 +$$

$$+ 22.5 \cdot 22.5 \cdot 2 + 23.5 \cdot 22.5 \cdot 3 + 23.5 \cdot 27.5 \cdot 3 +$$

$$+ 24.5 \cdot 27.5 \cdot 8 = 11\,330;$$

$$\overline{x}_s = \frac{1}{n} \sum x_i n_i = \frac{480}{20} = 24; \quad \overline{y}_s = \frac{1}{n} \sum y_j n_j = \frac{468}{20} = 23.4;$$

$$\sigma_x = \sqrt{\frac{1}{n} \sum x_i^2 n_i - \left(\overline{x_s}\right)^2} = \sqrt{596.25 - 576} = \sqrt{20.25} = 4.5;$$

$$\sigma_y = \sqrt{\frac{1}{n} \sum y_j^2 n_j - \left(\overline{y_s}\right)^2} = \sqrt{548.95 - 547.56} = \sqrt{1.39} = 1.18.$$

Calculate the sample correlation coefficient:

$$r_s = \frac{\sum\sum x_i y_j m_{ij} - n\overline{x_s}\,\overline{y_s}}{n\sigma_x\sigma_y} = \frac{11\,330 - 20\cdot 24\cdot 23.4}{20\cdot 4.5\cdot 1.18} = 0.923.$$

Test the hypothesis that the correlation coefficient of the general population is equal to zero. Compute:

$$T_{obs} = \frac{r_в\sqrt{n-2}}{\sqrt{1-r_в^{\,2}}} = \frac{0.923\cdot\sqrt{18}}{\sqrt{1-0.923^2}} = 10.25.$$

According to the table of critical points of the Student's distribution (Appendix 4), by the significance level α = 0,05 and the number of degrees of freedom $\nu = n-2 = 18$ we find $t_{cr} = t(0.025;18) = 2.101.$

So $|T_{obs}| > t_{cr}$, then the hypothesis that the correlation coefficient of the general population is equal to zero is rejected, i. e. the chosen correlation coefficient is significant.

5.2. Tasks for Classroom Work

5.2.1. Determine the closeness of the relationship and the significance of the total weight *X* (g) of the plant and the weight *Y* (g) of its seeds based on the data.

x_i	40	50	60	70	80	90	100
y_i	20	25	28	30	35	40	45

$(r_s = 0.992; T_{obs} = 17.6;\text{significant})$

5.5.2. To study the impact of investment volume *X* (billion of USD) on the annual profit *Y* (billion of USD), statistics were collected for *20* large enterprises, which were summarized in a correlation table.

Y \ X	0–10	10–20	20–30	30–40	40–50	n_y
1.5–2.5	1					1
2.5–3.5	2	5	2			9
3.5–4.5		3	3	2		8
4.5–5.5					2	2
n_x	3	8	5	2	2	20

Calculate the sample correlation coefficient.

($r_s = 0.782$)

5.2.3. Using a sample of size n = 100, extracted from a two-dimensional normal population (X, Y), calculate the sample correlation coefficient and, at a significance level $\alpha = 0.05$, test the hypothesis that the correlation coefficient of the general population is equal to zero.

Y \ X	10	15	20	25	30	35	n_y
35	5	1					6
45		6	2				8
55			5	40	5		50
65			2	8	7		17
75				4	7	8	19
n_x	5	7	9	52	19	8	100

$(r_s = 0.817; T_{obs} = 14.03; r \neq 0)$

5.2.4. Determine the closeness of the relationship between the cost of production Y (thousand USD) and the number of products X (thousand pieces) according to 7 enterprises.

X	2	3	4	5	6	7	8
Y	2	1.9	2.2	2.4	2.3	2.5	2.5

Find out the significance of the sample correlation coefficient at $\alpha = 0{,}05$.

($r_s = 0.925$; $T_{obs} = 5.44$; significant)

5.2.5. To determine the dependence of crop yields on soil moisture, 20 identical plots of land in the floodplain of the river were investigated (X is the distance of the plot from the river; Y is the yield).

Y \ X	0.4	0.8	1.0	1.2	1.8	2.0	n_y
3.0	2	3					5
3.5		4	2	1			7
4.5			1	2	2		5
5.0					2	1	3
n_x	2	7	3	3	4	1	20

Calculate the sample correlation coefficient.

($r_s = 0.871$)

5.3. Tasks for Independent Work

5.3.1. When debugging a lathe, machining errors X (μm) were measured for different diameters of workpieces Y (cm).

X	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
Y	3	3	4	4	4	5	5	5	6	8

Find the sample correlation coefficient between X and Y and determine its significance at $\alpha = 0.05$.

($r_s = 0.921$; $T_{obs} = 6.68$; significant)

5.3.2. To study the reliability of machines, statistical material was collected depending on the time of continuous operation Y (in months) and the number of previous repairs X.

X Y	0	1	2	3	n_y
2–6			1	2	3
6–10		1	3	1	5
10–14	1	2	1		4
14–18	2	1	1		4
18–22	1	3			4
n_x	4	7	6	3	20

Calculate r_v and establish the closeness of the connection at $\alpha = 0{,}05$.

$(r_s = -0.812; T_{obs} = 5.9;$ significant)

5.3.3. Calculate r_v according to the correlation table.

X Y	5	10	15	20	n_y
10	2				2
20	5	4	4		13
30	3	8	3	3	17
40		3	6	6	15
50			2	1	3
n_x	10	15	15	10	50

Lesson 6.
Linear Regression.
Determining Linear Regression Parameters

6.1. Brief Theoretical Information

If both regression lines *Y* on *X* and *X* on *Y* are straight lines, then the correlation is said to be linear.

The sample equation of the straight line regression *Y* on *X* has the form:

$$\overline{y}_x - \overline{y}_s = r_s \frac{\sigma_y}{\sigma_x}\left(x - \overline{x}_s\right). \tag{6.1}$$

The direct regression equation *X* on *Y* has the form:

$$\overline{x}_y - \overline{x}_s = r_s \frac{\sigma_x}{\sigma_y}\left(y - \overline{y}_s\right). \tag{6.2}$$

Here *x*, *y* – are the RV *X*, *Y* values; $\overline{y}_x, \overline{x}_y$ – are their sample means.

The coefficient of equations (6.1)–(6.2) can also be determined from the formulas obtained by the least squares method. For example, if equation (6.1) is taken as $\overline{y}_x = ax + b,$ then the parameters *a* and *b* of the linear regression are:

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2};$$

$$b = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}. \tag{6.3}$$

Task 6.1.1. The distribution of 40 factories in the industry by the number of fitters X and the number of machine shifts Y is given by the correlation table.

Y \ X	10–15	15–20	20–25	25–30	30–35	35–40	n_y
0–0.2	4	–	–	–	–	–	4
0.2–0.4	2	2	–	–	–	–	4
0.4–0.6	–	–	2	–	–	–	2
0.6–0.8	–	6	–	4	4	–	14
0.8–1.0	–	–	–	–	6	6	12
1.0–1.2	–	–	–	–	–	4	4
n_x	6	8	2	4	10	10	40

Write a direct regression equation for Y on X.

Solution

According to the correlation table, calculate.

$$\overline{x_S} = \frac{1}{n}\sum_{i=1}^{n} x_i n_{xi} = \frac{1}{40}(12.5\cdot 6 + 17.5\cdot 8 + 22.5\cdot 2 + 27.5\cdot 4 +$$

$$+ 32.5\cdot 10 + 37.5\cdot 10) = 26.75;$$

$$\overline{y_S} = \frac{1}{n}\sum_{i=1}^{n} y_i n_{yi} = \frac{1}{40}(0.4\cdot 4 + 0.3\cdot 4 + 0.5\cdot 2 + 0.7\cdot 14 +$$

$$+ 0.9\cdot 12 + 1.1\cdot 4) = 0.69;$$

$$\sigma_x = 0.29; \quad \sigma_y = 9.25; \quad r_s = 0.85.$$

Substitute the calculated values into equation (6.1):

$$\overline{y_x} - 0.69 = 0.85\cdot\frac{9.25}{0.29}(x - 26.75);$$

$$\overline{y_x} - 0.69 = 22.8(x - 26.75);$$

$$\overline{y_x} = 22.8x - 609.2.$$

Task 6.1.2. When standardizing a copper thermometer, the dependence of the electrical resistance *Y* on temperature *X* was studied. The following results were obtained.

i	1	2	3	4	5	6
x_i	0	10	20	30	40	50
y_i	0.533	0.553	0.574	0.596	0.619	0.645

Estimate the parameters of the regression equation using the least squares method and write the regression equation *Y* to *X*.

Solution

Let's summarize the calculation results in a table.

Experiment number i	Initial data		Calculated data		
	x_i	y_i	x_iy_i	x_i^2	y_i^2
1	0	0.533	0	0	0.2841
2	10	0.553	5.53	100	0.3047
3	20	0.574	11.48	400	0.3295
4	30	0.596	17.88	900	0.3552
5	40	0.619	24.75	1600	0.3832
6	50	0.645	32.25	2500	0.4160
$n = 6$	$\sum x_i$	$\sum y_i$	$\sum x_iy_i$	$\sum x_i^2$	$\sum y_i^2$
	150	3.519	91.89	5500	2.0727

The linear regression parameters are determined by the formulas (6.3):

$$a = \frac{6 \cdot 91.89 - 150 \cdot 3.519}{6 \cdot 5500 - 150^2} = 0,002237;$$

$$b = \frac{3.519 \cdot 5500 - 150 \cdot 91.89}{6 \cdot 5500 - 150^2} = 0.53067.$$

The empirical regression equation Y on X will take the form

$$\overline{y_x} = 0.53067 + 0.002237x.$$

6.2. Tasks for Classroom Work

6.2.1. Find a sample regression equation Y on X according to the data given in the correlation table.

X Y	5	10	15	20	25	30	35	40	n_y
100	2	1							3
120	3	4	3						10
140			5	10	8				23
160				1		6	1	1	9
180							4	1	5
n_x	5	5	8	11	8	6	5	2	50

$$\left(\overline{y}_x = 0.92x + 112.72\right)$$

6.2.2. To study the dependence of the annual volume of production Y on fixed assets X, statistical data were obtained for 20 enterprises.

X Y	12.5	17.5	22.5	27.5	n_y
20–21	1				1
21–22		2			2
22–23		1	2		3
23–24			3	3	6
24–25				8	8
n_x	1	3	5	11	20

Write a sample regression equation for Y on X.

$$\left(\overline{y}_x = 17.524 + 0.2447x\right)$$

6.2.3. According to measurements of two variables find a linear regression equation for Y on X.

x_i	66	70	75	80	82	85	90	92	95	98
y_i	60	78	65	87	74	70	78	95	88	90

$$\left(\bar{y}_x = 12.25 + 0.8x\right)$$

6.2.4. The table contains data on the decay of 10 g of radioactive material, where t is the time (in months), X is the amount (g) of the remaining substance at time t.

t	1	2	3	4	5	6	7	8	9	10	11	12
X	8.45	7.67	5.08	3.63	3.46	2.43	1.91	1.17	0.98	0.81	0.76	0.72

Write a linear regression equation for X on t.

6.3. Tasks for Independent Work

6.3.1. The table shows data on the relationship between the price of oil X (currency units) and the index of oil companies Y (conventional units).

X	11.0	11.5	12.0	12.5	13.0	13.5
Y	1.5	1.5	1.6	1.7	1.9	1.9

Write a direct regression equation for Y on X.

$$\left(\bar{y}_x = 0.189x - 0.677\right)$$

6.3.2. Find sample equations of regression lines Y on X and X on Y according to the data given in the correlation table.

Y \ X	5	10	15	20	25	30	35	40	n_y
100	2	1							3
120	3	4	3						10
140			5	10	8				23
160				1		6	1	1	9
180							4	1	5
n_x	5	5	8	11	8	6	5	2	50

$$\left(\overline{y_x} = 1.92x + 100.9; \quad \overline{x_y} = 0.42y - 38.3\right)$$

TYPICAL CALCULATION FOR MATHEMATICAL STATISTICS

In tasks 1–20, an interval statistical series of the frequency distribution of the experimental values of the random variable X is given.

Required:

1) build a polygon and a histogram of frequencies (relative frequencies) of the RV X;

2) according to the type of the polygon and the histogram and, based on the mechanism of formation of RV, make a preliminary choice of the distribution law;

3) calculate the sample mean $\overline{x_s}$ and corrected standard deviation s;

4) write down the hypothetical distribution function and distribution density;

5) find the confidence intervals for the mathematical expectation and the standard deviation at the confidence level $\gamma = 0.95$;

6) find the theoretical frequencies of the normal distribution law and test the hypothesis of the normal distribution of RV using Pearson's test at a significance level of $\alpha = 0.05$.

1. The results of testing the resistance of 200 elongated drills with a diameter of 4 mm, hours, are given:

Drill life x_i	3–3.2	3.2–3.4	3.4–3.6	3.6–3.8	3.8–4
Frequency m_i	16	50	70	44	20

2. The results of a study of 100 sprayed samples on the strength of the sprayed layer, kg/mm^2, are given:

Strength x_i	2.0–2.2	2.2–2.4	2.4–2.6	2.6–2.8	2.8–3.0
Frequency m_i	7	22	38	23	10

3. The results of a study on the rupture of 100 samples of duralumin, kg/mm^2, are given:

Tensile strength x_i, kg/mm^2	42–43	43–44	44–45	45–46	46–47
Frequency m_i	8	25	36	22	9

4. The results of the phosphorus content (6 %) in 100 cast iron samples are given:

Phosphorus content x_i	0.1–0.2	0.2–0.3	0.3–0.4	0.4–0.5	0.5–0.6
Frequency m_i	7	22	38	24	9

5. The results of testing the resistance of 100 drills, h, are given:

Resistance x_i, h	17.5–22.5	22.5–27.5	27.5–32.5	32.5–37.5	37.5–42.5
Frequency m_i	7	20	44	20	9

6. Data are given on the average daily mileage of 100 motor vehicles, hundreds of km:

x_i, hundreds of km	1.2–1.6	1.6–2.0	2.0–2.4	2.4–2.8	2.8–3.2
Frequency m_i	8	19	47	20	6

7. A sample of products with a volume of 100 was taken from the machine for processing bushings with a diameter of d = 40 mm. The results of measuring the diameters of the bushings are given in the table:

Diameter x_i, mm	40.00–40.04	40.04–40.08	40.08–40.12	40.12–40.16	40.16–40.20
Frequency m_i	8	19	44	20	9

8. The table shows the statistical data on the labor intensity, min, of the operation “Monitoring the mechanical condition of the car after returning to the garage”:

Labor intensity x_i, min	3–4	4–5	5–6	6–7	7–8	8–9
Frequency m_i	6	8	33	35	11	7

9. The table shows the statistical data on the labor intensity, min, of the operation “repair of the car's water pump roller”:

Labor intensity x_i, min	0–10	10–20	20–30	30–40	40–50
Frequency m_i	17	47	70	46	20

10. The results of testing the durability of 100 cutters, h, are given:

Resistance x_i, h	21–26	26–31	31–36	36–41	41–46
Frequency m_i	8	21	43	21	7

11. Information is given on the consumption of water used by the workshop for technical needs within 100 days, m^3:

Flow rate x_i, m^3	8–12	12–16	16–20	20–24	24–28
Frequency m_i	7	25	36	22	10

12. Given quarterly data on the average daily mileage of 100 cars, km:

Daily mileage x_i	120–140	140–160	160–180	180–200	200–220
Frequency m_i	9	21	40	18	12

13. The values of the oil temperature in the BelAZ car engine at medium speeds are given:

Temperaturex_i, degr.	40–45	45–50	50–55	55–60	60–65
Frequency m_i	8	17	46	18	11

14. Given the dimensions of the inner diameter of the screw, mm:

Diameter x_i, mm	10.00–10.02	10.02–10.04	10.04–10.06	10.06–10.08	10.08–10.10
Frequency m_i	9	16	47	21	7

15. Given the dimensions of the diameters of 100 holes drilled with the same drill:

Diameter x_i, mm	8.02–8.07	8.07–8.12	8.12–8.17	8.17–8.22	8.22–8.27
Frequency m_i	10	19	38	21	12

16. The results of measuring the diameter of the roller processed by a single-spindle machine are given:

Diameter x_i, mm	19.80–19.85	19.85–19.90	19.90–19.95	19.95–20.00	20.05–20.10	20.10–20.15
Frequency m_i	6	15	27	32	14	6

17. The results of a study of the granulation of a batch of powder (in microns) are given:

Granulation x_i, μm	0–40	40–80	80–120	120–160	160–200
Frequency m_i	7	23	35	26	9

18. The results of observations of the service life of 150 machines of the same type before going beyond the limits (in months of two-shift operation) are given:

Term x_i, months	18–20	20–22	22–24	24–26	26–28
Frequency m_i	15	27	61	29	18

19. The results of measuring the thickness, cm, of 100 mica gaskets are given:

Thickness x_i, cm	0.20–0.26	0.26–0.32	0.32–0.38	0.38–0.44	0.44–0.50
Frequency m_i	13	19	48	12	8

20. Given the diameters of 100 rollers after grinding, mm:

Diameter x_i, mm	20.0–20.1	20.1–20.2	20.2–20.3	20.3–20.4	20.4–20.5
Frequency m_i	11	23	49	10	7

In tasks 21–40, the results of observations on RV X and Y are presented. Using these experimental data, it is necessary:

1. Build a correlation field. Select a mathematical model of the regression dependence of Y on X (it is recommended to use a linear regression model).
2. Estimate the parameters a and b of the model regression equation (6.1) by the least squares method.
3. Write the empirical regression equation Y on X.

21. RV X and RV Y – fluid levels in different cylinders of the same hydraulic system after control tests.

x_i, cm	12.1	11.2	9.8	10.4	9.2	8.5	8.8	7.4
y_i, cm	10.5	9.3	8.3	9.6	8.6	7.1	6.9	5.8

22. RV X – the magnitude of the stress of the steel bar; RV Y – the value of the load during compression of the steel bar.

x_i, kN	5	10	20	40	60
y_i, MPa	51.33	78.00	144.3	263.6	375.2

23. RV X – deepening of the cutter; RV Y – specific energy.

x_i	4	8	10	14	16	20	19	23
y_i	41	50	81	104	120	139	154	180

24. RV X, RV Y – fluid levels in different cylinders of the same hydraulic system after control tests.

x_i, cm	7.4	6.6	7.0	6.4	6.0	6.5	5.8	5.4
y_i, cm	5.8	5.2	5.0	5.1	4.6	5.0	4.4	3.9

25. RV X – electrical resistance of molybdenum; RV Y – is the temperature.

x_i, cm	61.97	57.32	52.70	47.92	37.72	32.09	28.09
y_i, K	2289	2132	1988	1830	1489	1286	1178

26. RV X – labor power per worker; RV Y – output per worker.

x_i, kWh	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5
y_i, thousand of USD	4.3	4.8	5.0	5.7	6.5	7.0	7.5	8.1

27. RV X – is the temperature; RV Y – is the resistance of the copper thermometer.

x_i, °C	0	10	20	30	40	50	60	70
y_i, Ohm	0.533	0.552	0.574	0.596	0.619	0.645	0.687	0.690

28. RV X – mass of the part; RV Y – the time spent on fixing the part on the lathe.

x_i, kg	0.7	0.8	1.0	1.2	1.3	1.4	1.5	1.7
y_i, sec	2.2	2.3	2.4	2.5	2.6	2.7	2.8	3.0

29. RV X – density of briquettes from carbonyl iron powder; RV Y – tensile strength at the mill of two such briquettes.

x_i, %	75	76	77	80	82	85	88	90
y_i, GPa	2.1	2.0	2.5	2.4	3.6	4.0	4.1	5.0

30. RV X – the speed of the car ZIL-130; RV Y – is the length of its braking distance.

x_i, km/h	10	15	20	25	30	35	40	45
y_i, m	1.8	2.7	2.5	4.5	4.4	6.3	6.5	6.5

31. RV X – the speed of the car VAZ-2301; RV Y – is the length of its braking distance.

x_i, km/h	31	30	42	40	55	48	64	59
y_i, m	2.0	2.6	3.9	5.2	7.0	6.2	7.5	8.6

32. RV X – helium pressure; RV Y – is the volume of one mole of helium.

$x_i \cdot 10^8$, Pa	3.0	3.6	4.0	4.5	5.2	5.6	6.0	6.4
$y_i \cdot 10^{-2}$, m^3	1.98	1.92	1.93	1.81	1.83	1.70	1.73	1.68

33. RV X – the mass of the load suspended on an elastic cord; RV Y – is the extension of this cord.

x_i, kg	0.05	0.07	0.100	0.125	0.150	0.175	0.200	0.250
y_i, cm	0.005	0.052	0.012	0.016	0.017	0.025	0.027	0.034

34. RV X – temperature when pressing fiberglass bolts; RV Y – their ultimate strength.

x_i, °C	1.30	1.35	1.40	1.45	1.50	1.55	1.60	1.65
$y_i \cdot 10^8$, Pa	10.8	10.2	9.2	8.9	8.3	8.3	8.0	7.3

35. RV X – impact strength of tool high-speed steels; RV Y – coefficient of their machinability.

$x_i \cdot 10^{-3}$, J/m^2	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4
y_i,	0.6	0.62	0.64	0.67	0.69	0.73	0.75	0.8

36. RV X – work experience; RV Y – average annual overfulfillment of the norm.

x_i, years	2	3	4	5	6	7
y_i, %	6	6	7	8	9	10

37. RV X – deviation of the dimensions of the rollers from the nominal value during roughing; RV Y – when finishing.

x_i, µm	–30	–25	–20	–15	–10	–5	0
y_i, µm	–8	–4	0	2	4	8	12

38. RV X – the speed of the BelAZ car; RV Y – is the temperature of the lubricating oil in the engine of this vehicle

x_i, km/h	20	25	30	35	40	45	50	55
y_i, °C	43.5	43.9	44.2	45.0	46.0	46.9	47.5	49.0

39. RV X – cutting speed; RV Y – is the cross-sectional area of the chip during machining.

x_i, m/min	25.0	22.7	22.1	19.8	17.0	12.3	10.7	10.0	8.2
y_i, mm^2	1.1	1.4	1.7	2.1	2.6	4.7	6.1	7.0	10.0

40. RV X – temperature; RV Y – is the coefficient of friction in the bearing.

x_i, °C	60	70	80	90	100	110	120
y_i,	0.0148	0.0124	0.0102	0.0085	0.0071	0.0059	0.0051

REFERENCE LIST

1. Borovikov, V. P. STATISTICA: Statistical analysis and data processing in the Windows environment / V. P. Borovikov, I. P. Borovikov. – M. : Information and publishing house “Filin”, 1998. – 608 p.

2. Gmurman, V. E. Probability Theory and Mathematical Statistics / V. E. Gmurman. – M. : Higher School, 2003. – 479 p.

3. Gmurman V.E. Guide to solving tasks in probability theory and mathematical statistics / V. E. Gmurman. – M. : Higher School, 1997. – 400 p.

4. Mikulik, N. A. Probability theory and mathematical statistics / N. A. Mikulik, A. V. Metelsky. – Minsk : Pion, 2002. – 192 p.

5. Mikulik, N. A. Mathematics for engineers / Under the scientific editorship of N. A. Mikulik. – Minsk : Elaida, 2006. – V. 2. – 496 p.

6. Mikulik, N. A. Solving technical tasks in probability theory and mathematical statistics / N. A. Mikulik, G. N. Reizina. – Minsk : Higher School, 1991. – 163 p.

7. Belko, I. V. Theory of Probability and Mathematical Statistics. Examples and tasks / I. V. Belko, G. L. Svirid. – Minsk : New Knowledge, 2002. – 250 p.

8. Matalytsky, M. A. Probability theory and mathematical statistics: textbook / M. A. Matalytsky, G. A. Khatskevich. – Minsk : Higher School, 2017. – 591 p. : ill.

9. Collection of tasks on the theory of probability, random processes and mathematical statistics: [textbook] / Yu. S. Kharin, G. A. Khatskevich, V. I. Lobach. – Minsk : BSU, 1995. – 99 p.

Appendix 1

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

x	0	1	2	3	4	5	6	7	8	9
0.0	0.3989	3989	3989	3988	3986	3084	3982	3980	3977	3973
0.1	3970	3965	3961	3956	3951	3945	3939	3932	3025	3918
0.2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0.3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0.4	3683	3668	3652	3637	3621	3605	3589	3572	3555	3538
0.5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0.6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0.7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0.8	2897	2874	2850	2827	2804	2780	2756	2732	2709	2685
0.9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1.0	0.2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1.1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1.2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1.3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1.4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1.5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1.6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1.7	0940	0925	0909	0893	0878	0863	0846	0833	0818	0804
1.8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1.9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2.0	0.0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2.1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2.2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2.3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2.4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180

x	0	1	2	3	4	5	6	7	8	9
2.5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2.6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2.7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2.8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2.9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3.0	0.0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
3.1	0033	0032	0032	0030	0029	0028	0027	0026	0025	0025
3.2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3.3	0017	0017	0012	0016	0015	0015	0014	0014	0013	0013
3.4	0012	0012	0010	0011	0011	0010	0010	0010	0009	0009
3.5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3.6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3.7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3.8	0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3.9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001

Appendix 2

Values of the Laplace function $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{0}^{x} e^{-\frac{t^2}{2}}\,\mathrm{d}t$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0.00	0.0000	0.32	0.1255	0.64	0.2389	0.96	0.3315
0.01	0.0040	0.33	0.1293	0.65	0.2422	0.97	0.3340
0.02	0.0080	0734	0.1331	0.66	0.2454	0.98	0.3365
0.03	0.0120	0.35	0.1368	0.67	0.2486	0.99	0.3389
0.04	0.0160	0.36	0.1406	0.68	0.2517	1.00	0.3413
0.05	0.0199	0.37	0.1443	0.69	0.2549	2.01	0.3438
0.06	0.0239	0.38	0.1480	0.70	0.2580	1.02	0.3461
0.07	0.0279	0.39	0.1517	0.71	0.2611	1.03	0.3485
0.08	0.0319	0.40	0.1554	0.72	0.2642	1.04	0.3508
0.09	0.0359	0.41	0.1591	0.73	0.2673	1.05	0.3531
0.10	0.0398	0.42	0.1628	0.74	0.2703	1.06	0.3554
0.11	0.0438	0.43	0.1664	0.75	0.2734	1.07	0.3577
0.12	0.0478	0.44	0.1700	0.76	0.2764	1.08	0.3599
0.13	0.0517	0.45	0.1736	0.77	0.2794	1.09	0.3621
0.14	0.0557	0.46	0.1772	0.78	0.2823	1.10	0.3643
0.15	0.0596	0.47	0.1808	0.79	0.2852	1.11	0.3665
0.16	0.0636	0.48	0.1844	0.80	0.2881	1.12	0.3686
0.17	0.0675	0.49	0.1879	0.81	0.2910	1.13	0.3708
0.18	0.0714	0.50	0.1915	0.82	0.2939	1.14	0.3729
0.19	0.0753	0.51	0.1950	0.83	0.2967	1.15	0.3749
0.20	0.0793	0.52	0.1985	0.84	0.2995	1.16	0.3770
0.21	0.0832	0.53	0.2019	0.85	0.3023	1.17	0.3790
0.22	0.0871	0.54	0.2054	0.86	0.3051	1.18	0.3810
0.23	0.0910	0.55	0.2088	0.87	0.3078	1.19	0.3830
0.24	0.0948	0.56	0.2123	0.88	0.3106	1.20	0.3849
0.25	0.0987	0.57	0.2157	0.89	0.3133	1.21	0.3869
0.26	0.1026	0.58	0.2190	0.90	0.3159	1.22	0.3883
0.27	0.1064	0.59	0.2224	0.91	0.3186	1.23	0.3907
0.28	0.1103	0.6.	0.2257	0.92	0.3212	1.24	0.3925
0.29	0.1141	0.61	0.2291	0.93	0.3238	1.25	0.3944
0.30	0.1179	0.62	0.2324	0.94	0.3264		
0.31	0.1217	0.63	0.2357	0.95	0.3289		

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
1.26	0.3962	1.59	0.4441	1.92	0.4726	2.50	0.4938
1.27	0.3980	1.60	0.4452	1.93	0.4732	2.52	0.4941
1.28	0.3997	1.61	0.4463	1.94	0.4738	1.54	0.4945
1.29	0.4015	1.62	0.4474	1.95	0.4744	2.56	0.4948
1.30	0.4032	1.63	0.4484	1.96	0.4750	2.58	0.4951
1.31	0.4049	1.64	0.4495	1.97	0.4756	2.60	0.4953
1.32	0.4066	1.65	0.4505	1.98	0.4761	2.62	0.4956
1.33	0.4082	1.66	0.4515	1.99	0.4767	2.64	0.4959
1.34	0.4099	1.67	0.4525	2.00	0.4772	2.66	0.4961
1.35	0.4115	1.68	0.4535	2.02	0.4783	2.68	0.4963
1.36	0.4131	1.69	0.4545	2.04	0.4793	2.70	0.4965
1.37	0.4147	1.70	0.4554	2.06	0.4803	2.72	0.4967
1.38	0.4162	1.71	0.4564	2.08	0.4812	2.74	0.4969
1.39	0.4177	1.72	0.4573	2.10	0.4821	2.76	0.4971
1.40	0.4192	1.73	0.4582	2.12	0.4830	2.78	0.4973
1.41	0.4207	1.74	0.4591	2.14	0.4838	2.80	0.4974
1.42	0.4222	1.75	0.4599	2.16	0.4846	2.82	0.4976
1.43	0.4236	1.76	0.4608	2.18	0.4854	2.84	0.4977
1.44	0.4251	1.77	0.4616	2.20	0.4861	2.86	0.4979
1.45	0.4265	1.78	0.4625	2.22	0.4868	2.88	0.4980
1.46	0.4279	1.79	0.4633	2.24	0.4875	2.90	0.4981
1.47	0.4292	1.80	0.4641	2.26	0.4881	2.92	0.4982
1.48	0.4306	1.81	0.4649	2.28	0.4887	2.94	0.4984
1.49	0.4319	1.82	0.4656	2.30	0.4893	2.96	0.4985
1.50	0.4332	1.83	0.4664	2.32	0.4898	2.98	0.4986
1.51	0.4345	1.84	0.4671	2.34	0.4904	3.00	0.49865
1.52	0.4357	1.85	0.4678	2.36	0.4909	3.20	0.49931
1.53	0.4370	1.86	0.4686	2.38	0.4913	3.40	0.49966
1.54	0.4382	1.87	0.4693	2.40	0.4918	3.60	0.499841
1.55	0.4394	1.88	0.4699	2.42	0.4922	3.80	0.499928
1.56	0.4406	1.89	0.4706	2.44	0.4927	4.00	0.499968
1.57	0.4418	1.90	0.4713	2.46	0.4931	4.50	0.499997
1.58	0.4429	1.91	0.4719	2.48	0.4934	5.00	0.499997

Appendix 3

$\chi^2_{\alpha;\nu}; \quad P(\chi^2 \geq \chi^2_{\alpha;\nu}) = \alpha$ function values

$\nu \setminus \alpha$	0.20	0.10	0.05	0.02	0.01	0.001
1	1.642	2.706	3.841	5.412	6.635	10.827
2	3.219	4.605	5.991	7.824	9.210	13.815
3	4.642	6.251	7.815	9.837	11.345	16.266
4	5.989	7.779	9.488	11.668	13.277	18.467
5	7.289	9.236	11.070	13.388	15.086	20.515
6	8.558	10.645	12.592	15.033	16.812	22.457
7	9.803	12.017	14.067	16.622	18.475	24.322
8	11.030	13.362	15.507	18.168	20.090	26.125
9	12.242	14.684	16.919	19.679	21.666	27.877
10	13.442	15.987	18.307	21.161	23.209	29.588
11	14.631	17.275	19.675	22.618	24.725	31.264
12	15.812	18.549	21.026	24.054	26.217	32.909
13	16.985	19.812	22.362	25.472	27.688	34.528
14	18.151	21.064	23.685	26.683	29.141	36.123
15	19.311	22.307	24.996	28.259	30.578	37.697
16	20.465	23.542	26.296	29.633	32.000	39.252
17	21.615	24.769	27.587	30.995	33.409	40.790
18	22.760	25.989	28.869	32.346	34.805	42.312
19	23.900	27.204	30.144	33.687	36.191	43.820
20	25.038	28.412	31.410	35.020	37.566	45.315
21	26.171	29.615	32.671	36.343	38.932	46.797
22	27.301	30.813	33.924	37.659	40.289	48.268
23	28.429	32.007	35.172	38.968	41.638	49.728
24	29.553	33.196	36.415	40.270	42.980	51.179
25	30.675	34.382	37.652	41.566	44.312	52.620
26	31.795	35.563	38.885	42.856	45.642	54.052
27	32.912	36.741	40.113	44.140	46.963	55.476
28	34.027	37.916	41.337	45.419	48.278	56.893
29	35.139	39.087	42.557	46.693	49.588	58.302
30	36.250	40.256	43.773	47.962	50.892	59.703

Appendix 4

Student's distribution

Values $t_{\alpha;\nu}$ satisfy the condition $P(t \geq t_{\alpha;\nu}) = \int\limits_{t_{\alpha;\nu}}^{\infty} S(t,\nu)\mathrm{d}t = \alpha$

$\nu \setminus \alpha$	0.40	0.30	0.20	0.10	0.05	0.025	0.010	0.005	0.001	0.0005
1	0.325	0.727	1.376	3.078	6.314	12.71	31.82	63.66	318.3	636.6
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925	22.33	32.60
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841	10.22	12.94
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	5.032	5.893	6.859
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499	4.785	5.405
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878	3.611	3.922
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797	3.467	3.745

ν \ α	0.40	0.30	0.20	0.10	0.05	0.025	0.010	0.005	0.001	0.0005
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.256	0.531	0.856	1.315	1.716	2.056	2.479	2.779	3.435	3.707
27	0.256	0.531	0.855	1.314	1.713	2.052	2.473	2.771	3.421	3.690
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	7.756	3.396	3.659
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	0.255	0.528	0.849	1.298	1.676	2.009	2.403	2.678	3.262	3.495
60	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.254	0.527	0.846	1.292	1.664	1.990	2.374	2.639	3.195	3.415
100	0.254	0.526	0.845	1.290	1.660	1.984	2.365	2.626	3.174	3.389
200	0.254	0.525	0.843	1.286	1.653	1.972	2.345	2.601	3.131	3.339
500	0.253	0.525	0.842	1.283	1.648	1.965	2.334	2.586	3.106	3.310
∞	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Appendix 5

$t_{\gamma;n}$ function values: $\overline{x}_s - t_{\gamma;n}\frac{\overline{s}}{\sqrt{n}} < a < \overline{x}_s + t_{\gamma;n}\frac{s}{\sqrt{n}}$

$n \backslash \gamma$	0.95	0.99	0.999	$n \backslash \gamma$	0.95	0.99	0.999
5	2.78	4.60	8.61	20	2.093	2.861	3.883
6	2.57	4.03	6.86	25	2.064	2.797	3.745
7	2.45	3.71	5.96	30	2.045	2.756	3.659
8	2.37	3.50	5.41	35	2.032	2.720	3.600
9	2.31	3.36	5.04	40	2.023	2.708	3.558
10	2.26	3.25	4.78	45	2.016	2.692	3.527
11	2.23	3.17	4.59	50	2.009	2.679	3.502
12	2.20	3.11	4.44	60	2.001	2.662	3.464
13	2.18	3.06	4.32	70	1.996	2.649	3.439
14	2.16	3.01	4.22	80	1.991	2.640	3.418
15	2.15	2.98	4.14	90	1.987	2.633	3.403
16	2.13	2.95	4.07	100	1.984	2.627	3.392
17	2.12	2.92	4.02	120	1.980	2.617	3.374
18	2.11	2.90	3.97	∞	1.960	2.576	3.291
19	2.10	2.88	3.92				

Appendix 6

The values of the coefficients q_1 and q_2; $q_1 S < \sigma < q_2 S$

	0.99		0.98		0.95		0.00	
	q_1	q_2	q_1	q_2	q_1	q_2	q_1	q_2
1	0.356	15.0	0.388	79.8	0.446	31.9	0.510	15.9
2	0.434	14.1	0.466	9.97	0.521	6.28	0.578	4.40
3	0.483	6.47	0.514	5.11	0.566	3.73	0.620	2.92
4	0.519	4.39	0.549	3.67	0.599	2.87	0.649	2.37
5	0.546	3.48	0.576	3.00	0.624	2.45	0.672	2.090
6	0.569	2.98	0.597	2.62	0.644	2.202	0.690	1.916
7	0.588	2.66	0.616	2.377	0.661	2.035	0.705	1.797
8	0.604	2.440	0.631	2.205	0.675	1.916	0.718	1.711
9	0.618	2.277	0.644	2.076	0.688	1.826	0.729	1.645
10	0.630	2.154	0.656	1.977	0.699	1.755	0.739	1.593
11	0.641	2.056	0.667	1.898	0.708	1.698	0.748	1.550
12	0.651	1.976	0.676	1.833	0.717	1.651	0.755	1.515
13	0.660	1.910	0.685	1.779	0.725	1.611	0.762	1.485
14	0.669	1.854	0.693	1.733	0.732	1.577	0.769	1.460
15	0.676	1.806	0.700	1.694	0.739	1.548	0.775	1.437
16	0.683	1.764	0.707	1.659	0.745	1.522	0.780	1.418
17	0.690	1.727	0.713	1.629	0.750	1.499	0.785	1.400
18	0.696	1.695	0.719	1.602	0.756	1.479	0.790	1.385
19	0.702	1.668	0.725	1.578	0.760	1.460	0.794	1.370
20	0.707	1.640	0.730	1.556	0.765	1.414	0.798	1.358
21	0.712	1.617	0.734	1.536	0.769	1.429	0.802	1.346
23	0.722	1.576	0.743	1.502	0.777	1.402	0.809	1.326
24	0.726	1.558	0.747	1.487	0.781	1.391	0.812	1.316
25	0.730	1.541	0.751	1.473	0.784	1.380	0.815	1.308
26	0.734	1.526	0.755	1.460	0.788	1.371	0.818	1.300
27	0.737	1.512	0.758	1.448	0.791	1.361	0.820	1.293
29	0.744	1.487	0.765	1.426	0.796	1.344	0.825	1.279
30	0.748	1.475	0.768	1.417	0.799	1.337	0.828	1.274
40	0.774	1.390	0.792	1.344	0.821	1.279	0.847	1.228
50	0.793	1.336	0.810	1.297	0.837	1.243	0.861	1.199
60	0.808	1.299	0.824	1.265	0.849	1.217	0.871	1.179
70	0.820	1.272	0.835	1.241	0.858	1.198	0.879	1.163
80	0.829	1.250	0.844	1.222	0.866	1.183	0.886	1.151
90	0.838	1.233	0.852	1.207	0.873	1.171	0.892	1.141
100	0.845	1.219	0.858	1.195	0.878	1.161	0.897	1.133
200	0.887	1.15	0.897	1.13	0.912	1.11	0.925	1.09

Appendix 7

Critical values of Kolmogorov's distribution

$$P(\lambda > \lambda_\alpha) = \alpha$$

α	0.2	0.1	0.05	0.02	0.01	0.001
λ_α	1.073	1.224	1.358	1.520	1.627	1.950

TABLE OF CONTENTS

Учебное издание

ЧЕПЕЛЕВА Тереса Иосифовна
ЧЕПЕЛЕВ Николай Иосифович
ЧЕПЕЛЕВ Андрей Николаевич и др.

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебно-методическое пособие
для обучающихся по специальностям
1-37 01 02 «Автомобилестроение (по направлениям)»,
1-37 01 01 «Двигатели внутреннего сгорания»

MATHEMATICAL STATISTICS

Educational and methodical manual
for specialties:
1-37 01 02 “Automotive Industry (in areas)”
1-37 01 01 “Internal Combustion Engines”

Редактор *Н. А. Костешева*
Компьютерная верстка *Н. А. Школьниковой*

Подписано в печать 05.12.2022. Формат 60×84 $^{1}/_{16}$. Бумага офсетная. Ризография.
Усл. печ. л. 4,88. Уч.-изд. л. 3,10. Тираж 100. Заказ 604.

Издатель и полиграфическое исполнение: Белорусский национальный технический университет.
Свидетельство о государственной регистрации издателя, изготовителя, распространителя
печатных изданий № 1/173 от 12.02.2014. Пр. Независимости, 65. 220013, г. Минск.