Ministry of Science and Higher Education of the Russian Federation
NATIONAL RESEARCH
TOMSK STATE UNIVERSITY (NR TSU)

Institute of Applied Mathematics and Computer Science

APPROVE:
Director
_________________A. V. Zamyatin

Work program of the discipline

Mathematics & Statistics for Data Science - I

in the major of training

01.04.02 Applied mathematics and informatics

Orientation (profile) of training:

Big Data and Data Science

Form of study
full-time

Qualification
Master

Year of admission
2023

Code of discipline in the curriculum: B1.O.02

AGREED:
Head of EP
_____________ A.V. Zamyatin

Chairman of the EMC
_____________ _S.P. Sushchenko

Tomsk – 2023

1. Purpose and planned results of mastering the discipline

The purpose of mastering the discipline is the formation of the following competencies:

– GPC-1 – the ability to solve actual problems of fundamental and applied mathematics;

– GPC-2 – the ability to improve and implement new mathematical methods for solving applied problems.

The results of mastering the discipline are the following indicators of the achievement of competencies:

IOPC-1.3 Demonstrates the skills of using the basic concepts, facts, principles of mathematics, computer science and natural sciences to solve practical problems related to applied mathematics and computer science.

IOPC-2.1 Uses the results of applied mathematics to adapt new methods for solving problems in the field of his professional interests.

IOPC-2.2 Implements and improves new methods, solving applied problems in the field of professional activity.

IOPC-2.3. Carries out a qualitative and quantitative analysis of the resulting solution in order to build the best option.

2. Tasks of mastering the discipline

- To learn how to solve problems of statistical data analysis, starting from stating the initial problems of the corresponding subject area in terms of applied statistics, the choice of solution methods and quality criteria for the solutions developed, and ending with the interpretation of the findings in subject area terms.
- To study the main methods of statistical data analysis.
- To acquire skills required to use statistical data processing software.

3. The place of discipline in the structure of the educational program

Discipline belongs to the mandatory part of the educational program.

4. Semester of mastering and form of intermediate certification in the discipline

First semester, exam

5. Entrance requirements for mastering the discipline

Successful mastering of the discipline requires competencies formed during the development of educational programs of the previous level of education, knowledge of the basics of mathematical analysis, linear algebra, optimization methods, probability theory and mathematical statistics, as well as the basics of programming.

6. Implementation language

English.

7. Scope of discipline

The total labor intensity of the discipline is 6 credits, 216 hours, of which:
- lectures: 20 hours
- laboratory: 44 hours
  including practical training: 0 h.

The volume of independent work of the student is determined by the curriculum.

8. The content of the discipline, structured by topics

Topic 1. Mathematical foundations.

Elements of linear algebra. Fundamentals of mathematical analysis. Basic questions of optimization methods. Elements of the theory of probability.

Topic 2. Introduction to statistical analysis.

Data types. Graphical and tabular ways of presenting data. Data preprocessing. Estimates of parameters and numerical characteristics.

Topic 3. Tests for comparing groups.

Parametric tests. Student's t-test. Fisher's tet. Dispersion analysis. Nonparametric tests. Mann-Whitney, Wilcoxon, Kruskal-Wallis, Friedman tests.

Topic 4. Correlation analysis.

Pearson correlation coefficient. Fisher Z-transform. Rank correlation. Spearman coefficient, Kendall, Kendall concordance. Correlation analysis of categorical data.

Topic 5. Paired regression.

Definition of simple regression. The method of least squares (OLS) for estimating the parameters of a simple regression. Gauss-Markov assumptionы. Gauss-Markov theorem. Estimates of variances. Checking the quality of the regression model, the coefficient of determination, its interpretation, the quality of the model. Nonlinear models and linearization.

9. Ongoing evaluation

The ongoing evaluation is carried out by monitoring attendance, performing practical work and is recorded in the form of a checkpoint at least once a semester.

10. The procedure for conducting and criteria for evaluating the intermediate certification

The exam in the first semester is held in the form of final testing. The test consists of 15-20 questions. The duration of the exam is 1 academic hour (45 minutes).

An approximate list of theoretical questions and topics for preparing for the exam:

1. Solution of systems of linear equations.
2. Eigenvectors and eigenvalues of a matrix.
3. Functions of several variables. The concept of a gradient.
4. Method of gradient descent.
5. Total probability formula and Bayes formula.
6. Types of data and ways to represent them.
7. Parametric tests for comparing groups.
8. Nonparametric tests for comparing groups.
9. Correlation analysis of quantitative data.
10. Rank correlation.
11. Correlation analysis of categorical data.
12. Paired regression. Model. OLS-estimates of parameters.
13. Descriptive statistics of estimates of parameters of paired regression.
14. Gauss-Markov theorem for the case of paired regression.
15. Checking the quality of the paired regression equation.

Examples of practical tasks on mathematical foundations

Exercise. Solving systems of linear equations.
For a given system of equations, find a solution using the inverse matrix method.

Exercise. Eigenvalues and eigenvectors.
For a given matrix, find eigenvalues and eigenvectors.

Exercise. Functions of many variables.
For a given function of many variables, calculate the gradient, find the extrema of the function.

Examples of tasks for laboratory work on statistical analysis

Practical work. Data preprocessing

Exercise.

1. Import the given data set.
2. Check for gaps and outliers.
3. For quantitative indicators, build histograms.
4. Find estimates of numerical characteristics.
5. Test the hypothesis of normality.
6. Construct range diagrams by groups based on the breakdown of quantitative indicators by levels of categorical features.

Practical work. Exploratory Analysis

Runs in R.

Exercise.

Import a table with data into R.

1. Build graphs to visualize data and its relationships.
2. Check the relationships of factors with each other and their influence on the dependent target variable, choosing the appropriate criterion, depending on the types of data.
3. Test hypotheses about the significance of the relationship.

Practical work. Paired regression. Generation.

Runs in R.

Exercise.

1. Set the sample size n (from 50 to 150).
2. Generate a vector of predictor variable values.
3. Set the noise vector corresponding to the Gauss-Markov assumptions.

4. Set regression parameters.
5. Form a vector of values of the dependent variable according to the linear regression model.
6. Build a scatterplot and, if necessary, change the parameters.
7. Build OLS estimates of parameters, check their significance, compare with initial values
8. Find the SD of the residuals.
9. Check the quality of the model.

Practical work. Paired regression for real data. Linear and non-linear models.

Runs in R.

Exercise.

Import a table with data into R.

1. Build graphs to visualize data and their relationships.
2. Check the relationship of the factor with the dependent target variable.
3. Build and analyze a paired regression model of the target variable from the factor.
4. Build linear, exponential, exponential, logarithmic and inverse relationships.
5. Assess the quality of each model.
6. Choose the most adequate model.

The results of the examination of ratings are "excellent", "good", "satisfactory", "unsatisfactory".

For a 15 question test. For each question, depending on its complexity, you can get from 1 to 3 points. Max 30.

excellent	from 26 to 30
good	from 21 to 25
satisfactory	from 16 to 20
unsatisfactory	from 0 to 15

To complete the course successfully it is necessary to score more than 15 points in a test and complete all the lab works throughout the semester.

11. Educational and methodological support

a) Electronic training course on the discipline at the electronic university "Moodle" - https://moodle.tsu.ru/course/view.php?id=00000

b) Assessment materials of the ongoing evaluation and intermediate certification in the discipline.

c) Guidelines for practical work.

12. List of educational literature and Internet resources

a) basic literature

1. An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics) 1st ed. 2013, Corr. 7th printing 2017 Edition.
2. https://book.stat420.org/applied_statistics.pdf
3. http://thuvienso.bvu.edu.vn/bitstream/TVDHBRVT/15780/1/Applied-Statistics.pdf
4. http://wpage.unina.it/cafiero/books/stat.pdf
5. https://www.researchgate.net/publication/242692234_Statistical_foundations_of_machine_learning_the_handbook

б) additional literature:

6. https://bookdown.org/ndphillips/YaRrr/
7. https://mml-book.github.io/book/mml-book.pdf

13. List of information technologies

a) licensed and freely distributed software:

- Microsoft Office Standard 2013 Russian: software package. Includes applications: MS Office Word, MS Office Excel, MS Office PowerPoint, MS Office On-eNote, MS Office Publisher, MS Outlook, MS Office Web Apps (Word Excel MS PowerPoint Outlook);
- publicly available cloud technologies (Google Docs, Yandex disk, etc.)
- R The R Foundation, USA freeware.
- RStudio RStudio, PBC, USA freeware.
- JASP University of Amsterdam, Netherlands freeware.

14. Logistics

Halls for lectures.

Classrooms for seminars, individual and group work, ongoing evaluation and intermediate certification.

Classrooms for independent work, equipped with computer technology and access to the Internet, to the electronic information and educational environment and to information reference systems.

Halls for lectures and seminars, individual and group consultations, ongoing evaluation and intermediate certification in a mixed format (“Aktru”).

15. Authors information

Tatiana Valerievna Kabanova, PhD, Associate Professor, Department of Probability Theory and Mathematical Statistics, Institute of Applied Mathematics and Computer Science TSU.

.