

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

Institute of Applied Mathematics and Computer Science



A. V. Zamyatin

Evaluation materials of the current control and intermediate certification in the discipline

(Evaluation tools by discipline)

Mathematics and Statistics for Data Science – II

in the major of training

01.04.02 Applied mathematics and informatics

Orientation (profile) of training:

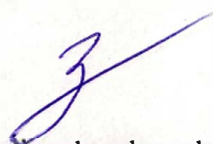
Big Data and Data Science

ET was implemented:
PhD,
Associate Professor of the Department
of Probability Theory and Mathematical Statistics



T.V. Kabanova

Reviewer:
PhD,
Associate Professor of the Department
of System Analysis and Mathematical Modeling

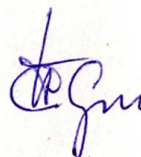


Zh.N. Zenkova

Evaluation tools were approved at a meeting of the educational and methodological commission of the Institute of Applied Mathematics and Computer Science (EMC IAMCS).

Protocol dated 20.05.2024 № 2

Chairman of the EMC IAMCS,
Dr. tech. Sciences, Professor



S.P. Sushchenko

Evaluation tools (ET) are an element of the system for assessing the formation of competencies among students in general or at a certain stage of its formation.

The ET is developed in accordance with the work program (WP) of the discipline.

1. Competencies and training outcomes, obtained upon the discipline mastery

Competencies	Competence indicator	Code and name of planned training outcomes that characterize the stages of competency formation	Criteria for evaluating training outcomes			
			Excellent	Good	Satisfactory	Unsatisfactory
GPC-1 – the ability to solve actual problems of fundamental and applied mathematics	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.	TO-1.1.1. the student will be able to: - choose an adequate method for solving the problem; - implement the selected method in the data analysis program; - draw conclusions and interpret the results.	Demonstration of a high level of knowledge of the mathematical foundations and basic concepts that are necessary to understand the statistical methods of data analysis.	In general, successful, but containing some gaps, knowledge of the mathematical foundations and basic concepts that are necessary to understand the statistical methods of data analysis.	Fragmentary, incomplete knowledge without gross errors of the mathematical foundations and basic concepts that are necessary to understand the statistical methods of data analysis.	Does not know the mathematical foundations and basic concepts that are necessary to understand the statistical methods of data analysis.
GPC-2 – the ability to improve and implement new mathematical methods for solving	IOPC-2.1 Uses the results of applied mathematics to adapt new methods for solving problems in the field of his professional interests.	TO-2.1.1. the student will be able to: adapt models to describe the processes of a real subject area	Demonstration of a high level of knowledge of the mathematical foundations and	In general, successful, but containing some gaps, knowledge of	Fragmentary, incomplete knowledge without gross errors of the	Does not know the mathematical foundations and basic concepts that are necessary to

applied problems.	IOPC-2.1 Uses the results of applied mathematics to adapt new methods for solving problems in the field of his professional interests.	<p>TO-2.2.1. the student will be able to: implement and interpret the constructed models to describe the processes of a real subject area.</p> <p>TO-2.3.1. The student will be able to implement a qualitative and quantitative analysis of the constructed models and the forecasts obtained on their basis and choose the most optimal one in accordance with the selected metric</p>	basic concepts that are necessary to understand the statistical methods of data analysis.	the mathematical foundations and basic concepts that are necessary to understand the statistical methods of data analysis.	mathematical foundations and basic concepts that are necessary to understand the statistical methods of data analysis.	understand the statistical methods of data analysis.
	IOPC-2.2 Implements and improves new methods, solving applied problems in the field of professional activity.					

2. Stages of competency formation and types of evaluation tools

№	Stages of competency formation (discipline sections)	Code and name of training outcomes	Type of evaluation tool (tests, assignments, cases, questions, etc.)
1.	Multiple regression	TO-1.1.1, TO-2.1.1, TO-2.2.1, TO-2.3.1.	Practical works; answers to questions in an exam or test
2.	Additional issues of regression analysis	TO-1.1.1, TO-2.1.1, TO-2.2.1, TO-2.3.1.	Practical works; answers to questions in an exam or test
3.	Tasks of classification	TO-1.1.1, TO-2.1.1, TO-2.2.1, TO-2.3.1.	Practical works; answers to questions in an exam or test

3. Typical control tasks or other materials necessary for the assessment of educational training outcomes

3.1. Typical tasks for conducting ongoing monitoring of progress in the discipline: tests, questions for colloquia, assignments for laboratory work.

Examples of tasks for practical work

Practical work. Data preprocessing

Exercise.

Import the given data set.

1. Build graphs to visualize data and their relations.
2. Check the relations of features with each other and their influence on the dependent target variable.
3. Build and analyze a multiple regression model of the target variable from all the presented quantitative and ordinal factors.
4. Carry out processing and coding of categorical factors.
5. Build and analyze a multiple regression model on all the proposed features.
6. Remove insignificant factors. Build the final model.
7. Check the residuals of the model for normality.
8. Set a new observation with your own feature values and build a target variable forecast for it.

Practical work. Logistic Regression.

Exercise

Generate observations related by one-way logistic regression.

1. Set the sample size $n = 20:50$.

2. Form the values of the factor x as an integer uniformly distributed random variable in the interval $[a, b]$.
3. Specify normally distributed noise $\varepsilon \sim N(0, \sigma)$.
4. Define the regression model

$$\Pi(x) = \frac{e^{\theta_0 + \theta_1 x + \varepsilon}}{1 + e^{\theta_0 + \theta_1 x + \varepsilon}}.$$

5. The value of the binary dependent variable is determined as

$$y_i = \begin{cases} 0, & \Pi(x_i) < \frac{1}{2}; \\ 1, & \Pi(x_i) \geq \frac{1}{2}. \end{cases}$$

Set all parameters yourself, depending on the scatterplot.

6. Estimate the model parameters.
7. Check the quality of the model.

3.2. Typical tasks for conducting intermediate certification in the discipline.

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

1. Nonlinear models and linearization.
2. The case of shifted noise.
3. The case of correlated homoscedastic observations.
4. The case of uncorrelated heteroscedastic observations.
5. Multicollinearity.
6. Dummy variables.
7. Classification Problem Statement.
8. Logistic Regression.
9. Quality metrics of a binary classifier.
10. ROC analysis.

4. Methodological materials that determine the procedures for evaluating training outcomes

4.1. Methodological materials for assessing the progress in the discipline.

For the ongoing certification, it is necessary to have attendance of at least 75% of all classes held at the time of certification and pass all practical work given at the time of certification. The practical work is graded for pass/fail.

4.2. Methodological materials for conducting intermediate certification in the discipline.

Final assessment in the second semester is carried out as a test. The test consists of 10-15 questions. The duration of the exam is 30 minutes.

The test is graded “passed” or “not passed”.

For a 10 questions test. For each question, depending on its complexity, you can get from 1 to 3 points. Max 20.

passed	from 11 to 20
not passed	from 0 to 10

To complete the course successfully it is necessary to score more than 10 points in a test and complete all the lab works throughout the semester.