

**Federal State Autonomous Educational Institution of Higher Education "Moscow
Institute of Physics and Technology
(National Research University)"**

APPROVED
**Head of the Phystech School of
Applied Mathematics and
Informatics**
A.M. Raygorodskiy

Work program of the course (training module)

course: Statistical Data Analysis/Математическая статистика и анализ данных
major: Applied Mathematics and Informatics
specialization: Modern State of Artificial Intelligence/Современные методы искусственного интеллекта
“Pusk” Online and Supplementary Education Centre
Chair of Machine Learning and Digital Humanities
term: 1
qualification: Master

Semester, form of interim assessment: 2 (spring) - Exam

Academic hours: 60 АЧ in total, including:

lectures: 30 АЧ.

seminars: 30 АЧ.

laboratory practical: 0 АЧ.

Independent work: 45 АЧ.

Exam preparation: 30 АЧ.

In total: 135 АЧ, credits in total: 3

Authors of the program:

R.G. Neychev, senior professor

A.M. Raygorodskiy, doctor of physics and mathematical sciences, associate professor, главный научный сотрудник

The program was discussed at the Chair of Machine Learning and Digital Humanities 05.03.2020

Annotation

In this course, mathematical statistics is considered as a scientific discipline that studies the theoretical foundations and methods for extracting and processing empirical data on mass phenomena, represented in the form of stochastic models. The problems of mathematical statistics also include the construction and study of decision-making procedures under conditions of "stochastic uncertainty".

The course contains basic information from mathematical statistics, used and supplemented in the future. The article considers classical probabilistic models of decision-making about the classes of observed objects according to the values of their attributes (models of classification or choice of hypotheses); it is assumed that the distributions of a feature for each class of objects are known exactly or with accuracy up to type. Goodness-of-fit criteria are discussed as a tool for testing the reliability of hypotheses and the problem of estimating distributions (in particular, the problem of the so-called parametric estimation).

Methods of nonparametric estimation of distributions are discussed separately, which impose much lower requirements on a priori information about their properties. The course also contains basic information about regression analysis, which serves to identify and evaluate the probabilistic relationships between the studied random variables.

1. Study objective

Purpose of the course

studying the mathematical and theoretical foundations of modern statistical analysis, as well as preparing students for further independent work in the field of analysis of statistical problems in applied mathematics, physics and economics.

Tasks of the course

studying the mathematical foundations of mathematical statistics;
acquisition of theoretical knowledge in the field of modern statistical analysis by students.

2. List of the planned results of the course (training module), correlated with the planned results of the mastering the educational program

Mastering the discipline is aimed at the formation of the following competencies:

Code and the name of the competence	Competency indicators
Gen.Pro.C-1 Address current challenges in fundamental and applied mathematics	Gen.Pro.C-1.2 Consolidate and critically assess professional experience and research findings
Pro.C-1 Become part of a professional community and conduct local research under scientific guidance using methods specific to a particular professional setting	Pro.C-1.2 Understand the verification process of software models used to solve related scientific problems
Pro.C-2 Understands and is able to apply modern mathematical apparatus and algorithms, the basic laws of natural science, modern programming languages and software; operating systems and networking technologies in research and applied activities	Pro.C-2.1 Demonstrate expert knowledge of research basics in the field of ICTs, philosophy and methodology of science, scientific research methods, and apply skills to use them

3. List of the planned results of the course (training module)

As a result of studying the course the student should:

know:

- basic concepts of mathematical statistics;
- basic approaches to comparing estimates of parameters of an unknown distribution;
- asymptotic and non-asymptotic properties of estimates of parameters of an unknown distribution;
- basic methods for constructing estimates with good asymptotic properties: method of moments, method of maximum likelihood, method of sample quantiles;
- the concept of effective estimates and inequality of information by Rao-Cramer;
- definition and main properties of the conditional mathematical expectation of a random variable relative to sigma-algebra or other random variable;
- definition of a general linear regression model and least squares method;
- multivariate normal distribution and its basic properties;
- basic concepts of the theory of testing statistical hypotheses;
- Neumann - Pearson lemma and monotonic likelihood ratio theorem;
- Pearson chi-square test for testing simple hypotheses in the Bernoulli scheme..

be able to:

- substantiate the asymptotic properties of estimates using the limit theorems of probability theory;
- construct estimates with good asymptotic properties for the parameters of an unknown distribution for a given sample from it;
- find Bayesian estimates for a given prior distribution;
- calculate conditional mathematical expectations using conditional distributions;
- find optimal estimates using complete sufficient statistics;
- build exact and asymptotic confidence intervals and areas for the parameters of the unknown distribution;
- find optimal estimates and confidence regions in a Gaussian linear model;
- build uniformly the most powerful criteria in the case of a parametric family with a monotonic likelihood ratio;
- Build an F-test to test linear hypotheses in a linear Gaussian model.

master:

- the main methods of mathematical statistics for constructing point and confidence estimates: the method of moments, sample quantiles, maximum likelihood, the method of least squares, the method of central statistics.
- skills of asymptotic analysis of statistical tests;
- skills of applying the theorems of mathematical statistics in applied problems of physics and economics.

4. Content of the course (training module), structured by topics (sections), indicating the number of allocated academic hours and types of training sessions

4.1. The sections of the course (training module) and the complexity of the types of training sessions

№	Topic (section) of the course	Types of training sessions, including independent work			
		Lectures	Seminars	Laboratory practical	Independent work
1	F-test for testing linear hypotheses in a Gaussian linear model.	10	10		15
2	Probabilistic-statistical model.	10	10		15
3	The main task of mathematical statistics.	10	10		15
AH in total		30	30		45
Exam preparation		30 AH.			
Total complexity		135 AH., credits in total 3			

4.2. Content of the course (training module), structured by topics (sections)

Semester: 2 (Spring)

1. F-test for testing linear hypotheses in a Gaussian linear model.

Binary search, Ternary search. Basic data structures: stack, queue, singly linked list, doubly linked list. Basic definitions of graph theory, DFS, BFS, topsort.

2. Probabilistic-statistical model.

Probabilistic-statistical model. Observation and sampling concepts. Parametric statistical model. Modeling a sample from an unknown distribution that belongs to a parametric family.

3. The main task of mathematical statistics.

The main task of mathematical statistics. Examples: sampling and linear model.

5. Description of the material and technical facilities that are necessary for the implementation of the educational process of the course (training module)

A standard classroom.

6. List of the main and additional literature, that is necessary for the course (training module) mastering

Main literature

1. Математическая статистика [Текст] : учеб. пособие для вузов / А. А. Натан, О. Г. Горбачев, С. А. Гуз ; Моск. физико-техн.ин-т (гос.ун-т) .— М : МЗ Пресс, 2004, 2005 .— 160 с.
2. Математическая статистика [Текст] : [учебник для вузов] / А. А. Боровков .— [3-е изд., испр.] .— М. : Физматлит, 2007 .— 704 с.
3. Введение в математическую статистику [Текст] : [учебник для вузов] / Г. И. Ивченко, Ю. И. Медведев .— М. : ЛКИ, 2010, 2014, 2015 .— 600 с.
4. Теория вероятностей и математическая статистика [Текст] : учеб. пособие для вузов / П. П. Бочаров, А. В. Печинкин .— М. : Физматлит, 2005 .— 295 с. : ил. + pdf-версия. - Библиогр.: с. 292. - ISBN 5-9221-0633-3. — Полный текст (Доступ из сети МФТИ / Удаленный доступ).

Additional literature

1. Наглядная математическая статистика [Текст] : учеб. пособие для вузов / М. Б. Лагутин .— 7-е изд. — М. : БИНОМ. Лаб. знаний, 2019 .— 472 с. : ил. - Библиогр.: с. 456-459. - Предм. указ.: с. 462-466. - ISBN 978-5-00101-105-7 (в пер.) .— Полный текст (Режим доступа : доступ из сети МФТИ).

7. List of web resources that are necessary for the course (training module) mastering

<http://dm.fizteh.ru/>

8. List of information technologies used for implementation of the educational process, including a list of software and information reference systems (if necessary)

Multimedia technologies can be employed during lectures and practical lessons, including presentations.

9. Guidelines for students to master the course

1. It is recommended to successfully pass test papers, as this simplifies the final certification in the subject.
2. To prepare for the final certification in the subject, it is best to use the lecture materials.

Assessment funds for course (training module)

major: Applied Mathematics and Informatics
specialization: Modern State of Artificial Intelligence/Современные методы искусственного интеллекта
“Pusk” Online and Supplementary Education Centre
Chair of Machine Learning and Digital Humanities
term: 1
qualification: Master

Semester, form of interim assessment: 2 (spring) - Exam

Authors:

R.G. Neychev, senior professor

A.M. Raygorodskiy, doctor of physics and mathematical sciences, associate professor, главный научный сотрудник

1. Competencies formed during the process of studying the course

Code and the name of the competence	Competency indicators
Gen.Pro.C-1 Address current challenges in fundamental and applied mathematics	Gen.Pro.C-1.2 Consolidate and critically assess professional experience and research findings
Pro.C-1 Become part of a professional community and conduct local research under scientific guidance using methods specific to a particular professional setting	Pro.C-1.2 Understand the verification process of software models used to solve related scientific problems
Pro.C-2 Understands and is able to apply modern mathematical apparatus and algorithms, the basic laws of natural science, modern programming languages and software; operating systems and networking technologies in research and applied activities	Pro.C-2.1 Demonstrate expert knowledge of research basics in the field of ICTs, philosophy and methodology of science, scientific research methods, and apply skills to use them

2. Competency assessment indicators

As a result of studying the course the student should:

know:

- basic concepts of mathematical statistics;
- basic approaches to comparing estimates of parameters of an unknown distribution;
- asymptotic and non-asymptotic properties of estimates of parameters of an unknown distribution;
- basic methods for constructing estimates with good asymptotic properties: method of moments, method of maximum likelihood, method of sample quantiles;
- the concept of effective estimates and inequality of information by Rao-Cramer;
- definition and main properties of the conditional mathematical expectation of a random variable relative to sigma-algebra or other random variable;
- definition of a general linear regression model and least squares method;
- multivariate normal distribution and its basic properties;
- basic concepts of the theory of testing statistical hypotheses;
- Neumann - Pearson lemma and monotonic likelihood ratio theorem;
- Pearson chi-square test for testing simple hypotheses in the Bernoulli scheme..

be able to:

- substantiate the asymptotic properties of estimates using the limit theorems of probability theory;
- construct estimates with good asymptotic properties for the parameters of an unknown distribution for a given sample from it;
- find Bayesian estimates for a given prior distribution;
- calculate conditional mathematical expectations using conditional distributions;
- find optimal estimates using complete sufficient statistics;
- build exact and asymptotic confidence intervals and areas for the parameters of the unknown distribution;
- find optimal estimates and confidence regions in a Gaussian linear model;
- build uniformly the most powerful criteria in the case of a parametric family with a monotonic likelihood ratio;
- Build an F-test to test linear hypotheses in a linear Gaussian model.

master:

- the main methods of mathematical statistics for constructing point and confidence estimates: the method of moments, sample quantiles, maximum likelihood, the method of least squares, the method of central statistics.
- skills of asymptotic analysis of statistical tests;
- skills of applying the theorems of mathematical statistics in applied problems of physics and economics.

3. List of typical control tasks used to evaluate knowledge and skills

Examples of homework assignments:

1. Find the optimal estimate of the parameter $\theta > 0$ based on a sample from the distribution: a) $N(\theta, 1)$, b) $R(0, \theta)$, c) $Pois(\theta)$, d) $Bin(1, \theta)$ (here $(0, 1)$) ... 6. Let X_1, \dots, X_n be a sample from a uniform distribution on the interval $[0, \theta]$, $\theta > 0$. Construct the confidence interval for the confidence level using the statistics a) X , b) $X(1)$, c) $X(n)$...
2. Let X_1, \dots, X_n be a sample from a normal distribution with parameters $(\theta, 1)$. Find the Bayesian estimate of the parameter if the prior distribution is $Bin(1, p)$. Will the resulting estimate be a consistent estimate of the parameter?
3. There are 2 objects with weights a and b . We weighed the first, second, and both objects together with errors, and the variance of the error in the latter case was 4 times greater. Reduce the problem to a linear regression model and find the least squares estimates for a and b .
4. X_1, \dots, X_n sample from exponential distribution with parameter θ . Construct evenly the most powerful criterion for the significance level of testing the hypothesis $H_0: \theta = 0$ against the alternative a) $H_1: \theta > 0$, b) $H_1: \theta < 0$.

4. Evaluation criteria

1. Types of convergence of random vectors: with probability 1, in probability, in distribution. Relationships between different types of convergence. Strong law of large numbers for random vectors. Multidimensional Central Limit Theorem.
2. The theorem on the inheritance of convergence and Slutsky's lemma. An example of their application.
3. Gaussian random vectors (multivariate normal distribution). A theorem on three equivalent definitions. The meaning of the parameters of the Gaussian vector.
4. Basic properties of Gaussian vectors: linear transformations and a criterion for the independence of components. The theorem on the orthogonal decomposition of a Gaussian vector.
5. Probabilistic-statistical model, parametric model. Sample, empirical distribution. Glivenko-Kantelli theorem.
6. Statistics and estimates. Unbiasedness, consistency, strong consistency and asymptotic normality. Lemma on the inheritance of asymptotic normality.
7. Estimation of the parameter by the substitution method. Parameter estimation by the method of moments. Consistency theorem for the method of moments.
8. Quantiles of distribution, sample quantiles, sample median. The theorem on the asymptotic normality of the sample p -quantile.
9. Loss function and risk function. A uniform approach to comparing grades. Bayesian, minimax and asymptotic approaches to the comparison of estimates.
10. Fisher's information and observation contribution. Rao-Cramer inequality. Effective assessments and performance criteria.
11. The concept of density for a discrete random variable. Dominated family of distributions. Likelihood function and maximum likelihood estimate. Theorem on the extreme likelihood property.
12. A theorem on the existence of a consistent solution to the likelihood equation. Consistency of the maximum likelihood estimate. A theorem on the asymptotic normality of the solution to the likelihood equation.
13. Bahadur's theorem. Asymptotically effective estimates. Asymptotic efficiency and efficiency of maximum likelihood estimation.
14. Conditional expectation of a random variable with respect to sigma algebra. Charge on probability space. Radon-Nikodym theorem.
15. Properties of conditional mathematical expectation (9 items).
16. Conditional expectation. Conditional distribution and conditional density of one random variable relative to another. A theorem on calculating the conditional mathematical expectation using conditional density. A theorem on a sufficient condition for the existence of a conditional density.

17. Sufficient statistics. Neumann-Fischer factorization criterion. Kolmogorov-Blackwell-Rao theorem and a corollary from it.
18. Complete statistics. Optimal estimate theorem. Exponential family of distributions. A theorem on complete sufficient statistics in an exponential family.
19. Confidence intervals and confidence regions. The concept of central statistics and the method for constructing confidence regions with its help. Asymptotic confidence intervals.
20. Linear regression model. Least squares estimate, the formula for its calculation. Unbiased estimates for parameters of a linear regression model.
21. Testing statistical hypotheses: hypothesis and alternative, hypothesis test criterion, first errors

Examination ticket

1. Fisher's information and observation contribution. Rao-Cramer inequality. Effective assessments and performance criteria.
2. Types of convergence of random vectors: with probability 1, in probability, in distribution. Relationships between different types of convergence. Strong law of large numbers for random vectors. Multidimensional Central Limit Theorem.

Assessment “excellent (10)” is given to a student who has displayed comprehensive, systematic and deep knowledge of the educational program material, has independently performed all the tasks stipulated by the program, has deeply studied the basic and additional literature recommended by the program, has been actively working in the classroom, and understands the basic scientific concepts on studied discipline, who showed creativity and scientific approach in understanding and presenting educational program material, whose answer is characterized by using rich and adequate terms, and by the consistent and logical presentation of the material;

Assessment “excellent (9)” is given to a student who has displayed comprehensive, systematic knowledge of the educational program material, has independently performed all the tasks provided by the program, has deeply mastered the basic literature and is familiar with the additional literature recommended by the program, has been actively working in the classroom, has shown the systematic nature of knowledge on discipline sufficient for further study, as well as the ability to amplify it on one's own, whose answer is distinguished by the accuracy of the terms used, and the presentation of the material in it is consistent and logical;

Assessment “excellent (8)” is given to a student who has displayed complete knowledge of the educational program material, does not allow significant inaccuracies in his answer, has independently performed all the tasks stipulated by the program, studied the basic literature recommended by the program, worked actively in the classroom, showed systematic character of his knowledge of the discipline, which is sufficient for further study, as well as the ability to amplify it on his own;

Assessment “good (7)” is given to a student who has displayed a sufficiently complete knowledge of the educational program material, does not allow significant inaccuracies in the answer, has independently performed all the tasks provided by the program, studied the basic literature recommended by the program, worked actively in the classroom, showed systematic character of his knowledge of the discipline, which is sufficient for further study, as well as the ability to amplify it on his own;

Assessment “good (6)” is given to a student who has displayed a sufficiently complete knowledge of the educational program material, does not allow significant inaccuracies in his answer, has independently carried out the main tasks stipulated by the program, studied the basic literature recommended by the program, showed systematic character of his knowledge of the discipline, which is sufficient for further study;

Assessment “good (5)” is given to a student who has displayed knowledge of the basic educational program material in the amount necessary for further study and future work in the profession, who while not being sufficiently active in the classroom, has nevertheless independently carried out the main tasks stipulated by the program, mastered the basic literature recommended by the program, made some errors in their implementation and in his answer during the test, but has the necessary knowledge for correcting these errors by himself;

Assessment “satisfactory (4)” is given to a student who has discovered knowledge of the basic educational program material in the amount necessary for further study and future work in the profession, who while not being sufficiently active in the classroom, has nevertheless independently carried out the main tasks stipulated by the program, learned the main literature but allowed some errors in their implementation and in his answer during the test, but has the necessary knowledge for correcting these errors under the guidance of a teacher;

Assessment “satisfactory (3)” is given to a student who has displayed knowledge of the basic educational program material in the amount necessary for further study and future work in the profession, not showed activity in the classroom, independently fulfilled the main tasks envisaged by the program, but allowed errors in their implementation and in the answer during the test, but possessing necessary knowledge for elimination under the guidance of the teacher of the most essential errors;

Assessment “unsatisfactory (2)” is given to a student who showed gaps in knowledge or lack of knowledge on a significant part of the basic educational program material, who has not performed independently the main tasks demanded by the program, made fundamental errors in the fulfillment of the tasks stipulated by the program, who is not able to continue his studies or start professional activities without additional training in the discipline in question;

Assessment “unsatisfactory (1)” is given to a student when there is no answer (refusal to answer), or when the submitted answer does not correspond at all to the essence of the questions contained in the task.

5. Methodological materials defining the procedures for the assessment of knowledge, skills, abilities and/or experience

During examination the student are allowed to use the program of the discipline.