МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ СЕМЕНА КУЗНЕЦЯ

ЗАТВЕРДЖЕНО
на засіданні кафедри
статистики і економічного прогнозування
Протокол № 2 від 2.09.2024 р.

ПОГОДЖЕНО
Проректор з навчально-методичної роботи
Карина НЕМАШКАЛО

СТАТИСТИЧНЕ МИСЛЕННЯ ДЛЯ НАУКИ ПРО ДАНІ
робоча програма навчальної дисципліни (РПНД)

Галузь знань	**12 Інформаційні технології**
Спеціальність	**122 Комп'ютерні науки**
Освітній рівень	**другий (магістерський)**
Освітня програма	**Комп'ютерні науки**

Статус дисципліни	**вибіркова**
Мова викладання, навчання та оцінювання	**англійська**

Розробники д.е.н., професор д.е.н., професор к.е.н., доцент викладач	підписано КЕП	Олена РАЄВНЄВА Костянтин СТРИЖИЧЕНКО Ольга БРОВКО Мар'яна СЕМКІВ
Завідувач кафедри статистики і економічного прогнозування	підписано КЕП	Олена РАЄВНЄВА
Гарант програми	підписано КЕП	Сергій МІНУХІН

Харків
2024

MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
SIMON KUZNETS KHARKIV NATIONAL UNIVERSITY OF ECONOMICS

APPROVED
at the meeting of the department
statistics and economic forecasting
Protocol № 2 of 2.09.2024.

AGREED
Vice-rector for Educational and Methodical work

Karina NEMASHKALO

STATISTICAL THINKING FOR DATA SCIENCE

Program of the course

Field of knowledge	**12 "Information technologies"**
Specialty	**122 Computer sciences**
Study cycle	**Second (master's)**
Study programme	**"Computer sciences"**

Course status	**Elective**
Language	**English**

Developer:

Doctor of Economics, professor	digitally signed	Olena RAYEVNYEVA
Doctor of Economics, professor		Konstantin STRYZHYCHENKO
PhD, associate professor		Olha BROVKO
teacher		Mariana SEMKIV

Head of the Department of Statistics and economic forecasting	digitally signed	Olena RAYEVNYEVA
Head of Study Programme	digitally signed	Serhii MINUKHIN

Kharkiv
2024

INTRODUCTION

The rapid development and widespread use of the latest packages of application programs and computing tools necessitate the formation of new competencies in a specialist in business analytics and information systems, aimed at acquiring knowledge and the ability to use economic and mathematical modeling for the analysis of complex, mass socio-economic phenomena and processes in various fields of activity.

Moreover, the explosive growth of information in every area of business and social life, from the environment to a variety of market research, is the main driver for creating an unprecedented global demand for information technology professionals with in-depth knowledge of business intelligence and working with Big Data. Today, there is a demand for specialists who combine competencies in the application of intelligent information processing systems and computer systems, the use of modern software products, IT technologies and technological tools in professional, in particular, entrepreneurial activities with the competencies of a business analyst for substantiation and decision-making in the fields of management and business are the main demand trend in the national and international labor markets. This discipline is a response to the modern needs of society and provides students with an in-depth understanding of the business context of any socio-economic processes and will allow them to solve problems related to analytical work in the IT industry.

Economic systems studied by modern science are subject to investigation by conventional (verbal) theoretical methods with great difficulty. A direct experiment on them is impossible. The price of errors and miscalculations is high, therefore mathematical modeling is an inevitable component of scientific and technological progress.

Statistical thinking for data science is one of the optional disciplines of the Master's program "Computer Science".

The purpose of the course is to expand and deepen theoretical knowledge and acquire professional competences in forecasting socio-economic processes and modeling complex systems using statistical methods and models.

The tasks of the course are:

determination of the main features of modeling and forecasting of complex socio-economic systems;

familiarization with existing statistical methods and models;

research of socio-economic processes using econometric models, cluster and discriminant analysis models, analysis of weakly formalized situations using expert analysis.

The object of the course is the educational discipline is complex socio-economic systems.

The subject of the course is theoretical and practical issues related to the development of forecasts and the construction of models in the conditions of a market economy based on the use of modern economic and mathematical methods and models.

The learning outcomes and competencies formed by the course are defined in table 1.

Table 1

Learning outcomes and competencies formed by the course

Learning outcomes	Competencies
LO2. Have specialized computer science problem-solving skills necessary for conducting research and/or conducting innovative activities to develop new knowledge and procedures.	GC05. Ability to learn and master modern knowledge.
LO7. Develop and apply mathematical methods for the analysis of information models.	SC03. Ability to use mathematical methods to analyze formalized models of the subject area.
	SC06. Ability to apply existing and develop new algorithms for solving problems in the field of computer science.
LO8. Develop mathematical models and data analysis methods (including large ones)	SC04. The ability to collect and analyze data (including large data) to ensure the quality of project decision-making
LO18. Collect, formalize, systematize and analyze the needs and requirements for the information or computer system being developed, operated or supported.	SC12. Ability to develop, apply and integrate data processing and analysis technologies in high-performance systems and cloud platforms to ensure efficient use of computing resources of computer systems.
	SC05. Ability to develop, describe, analyze and optimize architectural solutions of information and computer systems for various purposes.
LO19. To analyze the current state and global trends in the development of computer sciences and information technologies.	SC05. Ability to develop, describe, analyze and optimize architectural solutions of information and computer systems for various purposes.
	SC12. Ability to develop, apply and integrate data processing and analysis technologies in high-performance systems and cloud platforms to ensure efficient use of computing resources of computer systems.

COURSE CONTENT

Content module 1. Methodological foundations of statistical modeling and forecasting

Topic 1. Categorical basis of statistical modeling and forecasting

1.1. The concept of socio-economic systems, their structure as an object of modeling
1.2. Economy as a subsystem of nature and society
1.3. Classification and stages of construction of economic and mathematical models
1.4. Requirements and principles for building models
1.5. Forecasting as a method of predicting socio-economic processes
1.6. Content of basic categories of forecasting

Topic 2. Regression models as a means of researching economic processes

2.1. The concept of univariate and multiple regression as a class of econometric models and approaches to its construction
2.2. Use of MNC to calculate model parameters
2.3. Checking the quality of built models
2.4. Multicollinearity, methods of its verification and exclusion

Topic 3. Modeling and forecasting of development trends

3.1. Concept of time series, its components and classification of time series analysis models

3.2. Methods for determining the presence of a trend in the variance and average: the Foster-Stewart method, the Fisher method, the method of averages
3.3. Time series decomposition models
3.4. Autocorrelation, methods of its determination, stationarity of time series
3.5. Criteria for checking the quality of built models

Topic 4. Adaptive forecasting models and integrated autoregression model

4.1. The concept of smoothing and its types
4.2. Adaptive smoothing according to Brown, Holt, Winter
4.3. Integrated autoregression model
4.4. Vector autoregressive model

Content module 2. Modeling and forecasting of multidimensional processes

Topic 5. Factor analysis of data

5.1. Basic concepts of factor analysis
5.2. Methods of factor analysis
5.3. Method of principal components

Topic 6. Cluster analysis as a means of forming homogeneous groups of data

6.1. The essence of cluster analysis
6.2. Standardization and rationing
6.3. The concept of distance
6.4. Methods of cluster analysis

Topic 7. Data recognition and discriminant analysis

7.1. Basic concepts of discriminant analysis
7.2. Discriminant functions
7.3. Criteria for assessing the quality of classification
7.4. The use of discriminant analysis in economics

The list of laboratory studies in the course is given in table 2.

Table 2

The list of laboratory studies

Name of the topic and/or task	Content
Topic 1.	Laboratory work 1 "Getting to know the Statistica 10.0 package. A study of the statistical characteristics of the variational series".
Topic 2.	Laboratory session on topic 2. "Construction of univariate and multiple regression models"
Topic 3.	Laboratory lesson on topic 3. "Building a time series decomposition model"
Topic 4.	Laboratory session on topic 4. "Assessment of autocorrelation of model residuals. Removing autocorrelation"
Topic 5.	Laboratory class on topic 5. "Building a factor analysis model"
Topic 6.	Laboratory session on topic 6. "Using cluster analysis for the study of economic processes"
Topic 7.	Laboratory lesson on topic 7. "Solving the problem of classification by the method of discriminant analysis"

The list of self-studies in the course is given in table 3.

Table 3

List of self-studies

Name of the topic and/or task	Content
Topic 1-7	Studying lecture material
Topic 1	Essay writing Solving a situational (case) task
Topic 2-7	Performing laboratory work

The number of hours of lecture and laboratory classes and hours of independent work is given in the work plan (technological map) for the academic discipline.

TEACHING METHODS

In the process of teaching an educational discipline, in order to acquire certain learning outcomes, to activate the educational process, it is envisaged to use such teaching methods as:

group work (Topic 1), case technologies (Topic 1, 7), problem lectures (Topic 1), situational tasks (Topic 2-7), creation of cognitive novelty situations (topics 2-7).

In person (demonstration (Topic 1-5)).

Practical (laboratory work (Topic 2-7), essay (Topic 1), case method (Topic 1-7), etc.).

FORMS AND METHODS OF ASSESSMENT

The University uses a 100-point cumulative system for assessing the learning outcomes of students.

Current control is carried out during lectures and laboratory classes and is aimed at checking the level of readiness of the student to perform a specific job and is evaluated by the amount of points scored:

− for disciplines with a form of semester control examination (exam): the maximum amount is 60 points; the minimum amount that allows a student of higher education to pass an exam is 35 points

The **final control** includes current control and an exam.

Semester control is conducted in the form of a semester exam (exam). The semester exam (exam) is taken during the exam session.

The maximum number of points that a student of higher education can receive during the examination (examination) is 40 points. The minimum amount for which the exam is considered passed is 25 points.

The **final grade** in the course is determined:

− for disciplines with a form of exam, the final grade is the amount of all points received during the current control and the exam grade.

During the teaching of the academic discipline, the following control measures are used:

Current control: Laboratory work (21 points), essay in the form of a presentation (5 points), homework in the form of a case study (5 points), test control (21 points), written control works (8 points).

Semester control:Grading including Exam (40 points).

More detailed information about the evaluation system is provided in the work plan (technological map) for the course.

An example of an examination paper

Semyon Kuznets Kharkiv National University of Economics
Second (master's) level of higher education
Specialty 122 "Computer sciences"
Educational and professional program "Computer sciences"
Course "Statistical thinking for data science"

EXAM CARD № 1

Stereotype task (tests). (20 points)

1	The model means: a) a material or imagined object, which in the process of research replaces the original object so that its direct study provides new knowledge about the original object; b) a system related to the analysis of statistical data on the object of management; c) a complex dynamic system covering the processes of production, exchange, distribution and consumption of material and other goods.
2	The data are: a) information processed in a special way for decision-making; b) numerically expressed characteristic of any property of an economic object, process or decision; c) the value of economic indicators, which are objects of storage
3	When implementing the Kaiser criterion, it is necessary to select: a) only those factors with eigenvalues greater than 1; b) only those factors with eigenvalues less than 1; c) only those factors whose eigenvalues are equal to 1.
4	Which type of analysis is not related to discriminant: a) step-by-step analysis with inclusion; b) step-by-step rank analysis; c) step-by-step analysis with exclusion.
5	The modeling process includes the following elements: a) system, analysis and synthesis; b) the subject and the object of research, a model that mediates the relationship between the knowing subject and the known object; c) validation and adequacy.
6	The main types of structures of socio-economic systems are the following: a) cognitive and pragmatic b) network, hierarchical, matrix and with arbitrary connections; c) informational, physical, analog and mathematical.
7	Multicollinearity means linear dependence: a) an endogenous variable with one or more regressors; b) two or more regressors; c) regressors with model disturbances.
8	The criterion of stony scree is: a) logical method; b) mathematical method; c) graphical method proposed by Keitel.

9	Using discriminant analysis, the task is solved: a) regressions; b) classifications; c) reductions.
10	Forecasting methods are: a) consecutive and mixed; b) logical-intuitive and formalized; c) functional and structured.
11	The system is: a) a complex of interconnected elements and their components that develop in the process of interaction; b) a set of components that are characterized by a common origin and are in the process of development; c) composition of elements that are connected to each other.
12	Which test can be used to determine the presence of multicollinearity: a) Durbin-Watson; b) Ferrara-Glober; c) Geyser.
13	The characteristics of the cluster can be called: a) internal homogeneity; b) external uniformity; c) efficiency.
14	With the help of factor analysis, the task is solved: a) regressions; b) classifications; c) reductions.
15	According to the degree of determinism of the object, forecasts are: a) deterministic, stochastic and mixed; b) quantitative and qualitative; c) search and target.
16	Properties of complex systems include: a) emergency; b) complex structure; c) robustness.
17	The time series is this a) sequence of values of observations of the economic process during a certain number of periods; b) sequence of values of observations of an economic process that depends on time; c) set of points on the time line.
18	The initial step in cluster analysis is: a) formation of a matrix of observations; b) calculation of distances between objects; c) rationing.
19	It is considered that the recognition ability of the discriminant function is high if the Wilks lambda is equal to: a) 1; b) 0; c) -1.
20	According to the scale of the object, the forecasts are: a) deterministic, stochastic and mixed; b) sublocal, local, superlocal and global; c) search and target.

Diagnostic task 1 (5 points)

Based on the data in Table 4, it is necessary to reduce the factor space. Justify the number of factors. Evaluate the quality of the obtained results. Give economic conclusions about the obtained results.

Diagnostic task 2. (5 points)

According to the data given in the table 4 to group the regions of Ukraine. Carry out natural and artificial clustering. Give economic conclusions about the obtained results.

Heuristic task (calculation). (10 points)

Check the quality of the clustering carried out in diagnostic task 2 using discriminant analysis. Build a discrimination function. Conduct a comparative characterization of the resulting clusters and determine which group the Kharkiv region will fall into.

Table 4

Development indicators of the regions of Ukraine

Regions	Industrial production index,%	Capital investments, million hryvnias	Consumer price index, %	Export, thousands of dollars . USA	Import, thousands of dollars . USA	Unemployed, thousands of people	Economically active population, persons
Vinnytsia	109.4	840	99.7	149757.9	89281.8	64.3	27.9
Volynsk	95.8	503	99.7	160493.9	220719.4	64.1	12.7
Dnipropetrovsk	103.2	4352.2	99.6	2430643.2	1513161.9	71.1	30.3
Donetsk	99.3	4650.6	100	3561360.3	1088254.3	68.7	28.4
Zhytomyr	137.1	474.4	99.8	119178.8	89009.1	65.3	22.7
Zakarpattia	115.4	489.8	100	324461.3	467413.4	65.1	13
Zaporizhzhia	99.7	1248.3	99.3	936886.4	463283.7	69.4	21.4
Ivano -Frankivsk	106.6	922.5	99.8	231943.1	243368.5	60.1	16.7
Kyivska	92.2	4140.6	100	432208.2	944093.3	66.7	16.4
Kirovohradsk	104.6	555.8	99.7	120244.3	63450	66.1	17.5
Luhansk	97.4	1375.1	99.3	1440403	576522.4	65.7	17.9
Lviv	102.2	1623.9	99.6	322177.3	654652.2	64.6	24.2
Mykolayivska	108.2	509.1	99.2	578173.9	181629.5	68.1	17.7
Odesa	89.4	1578.1	99.7	277857.4	680277	66.6	16.5
Poltava	101.1	2171.4	99.6	772974.5	320086.2	68.1	25.6
Rivne	99	522.1	99.5	119491.4	154471.8	62.3	19
Sumy	99.6	365.9	99.5	259055.9	134355	65.2	17.5
Ternopilsk	102.4	548	99.4	57689.2	65503.9	60.2	16.3
Kharkivska	101.6	2076.6	99.5	463099.2	621686.8	69.4	29.9
Khersonsk	104.1	386.1	99.8	61184.3	51553.7	67	11.6
Khmelnytska	101	561.9	99.6	93313.7	150481.8	65.6	17.9
Cherkasy	100.5	389.4	99.7	244360.3	120390.3	69	25.2
Chernivtsi	94.9	353.4	99.8	27179.5	29751.6	58.2	11
Chernihivska	101.1	307.1	99.6	117419	113481.2	65.5	16.5
Kyiv	98.6	13611.5	99.9	2587839.7	5355211	70.3	7.3

Approved at the meeting of the Department of Statistics and Economic Forecasting protocol No._____dated "___"_____20___year.

Examiner Doctor of Economics, Prof. K. STRYZHYCHENKO

Chief Department Doctor of Economics, Prof. O. Rayevnyeva

Assessment criteria

The final marks for the exam consist of the sum of the marks for the completion of all tasks, rounded to a whole number according to the rules of mathematics.

The algorithm for solving each task includes separate stages that differ in complexity, time-consumingness, and importance for solving the task. Therefore, individual tasks and stages of their solution are evaluated separately from each other as follows:

The stereotypical task is valued at 20 points (each correct answer to the test is 1 point).

The diagnostic task is evaluated in 5 points on a scale:

1 point - the task is solved incorrectly, but some stages are given correctly or the task is solved with gross errors that affect the final result;

2 points – the task was completed half correctly: only part of the calculations were carried out;

3 points - the task is completely completed, but there are insignificant inaccuracies in the calculations or there are no comments to the calculations and conclusions;

4 points - the task is completely completed, but the expediency of using one or another statistical toolkit is not substantiated or there are no conclusions based on the results of the calculations;

5 points - the task was completed correctly, it was well designed, a full justification of the calculations was provided, and a thorough interpretation of the results was provided.

The heuristic task is evaluated in 10 points on the scale:

1 point – the student created only a file with raw data;

2 points - the task is solved incorrectly, but some stages are given correctly;

3 points – the task was solved with gross errors affecting the final result of the calculations;

4 points - the task was completed half correctly: only part of the calculations were carried out;

5 points - the task is completely completed, but there are insignificant inaccuracies in the calculations and there are no comments on the calculations and conclusions;

6 points - the task is completely completed, but no justification is provided for the feasibility of using one or another statistical toolkit;

7 points - the task is correctly completed, qualitatively designed, the expediency of using one or another statistical toolkit in the analysis of the proposed situation is substantiated, but there is no economic interpretation of the results;

8 points - the task is completed correctly, qualitatively designed, the expediency of using one or another statistical toolkit in the analysis of the proposed situation is substantiated, but the conclusions are incomplete;

9 points - the task was solved flawlessly, knowledge of the software and

statistical apparatus was demonstrated, a full justification of the calculations and economic conclusions was given;

10 points - the task was performed flawlessly, without a single error, it was qualitatively designed, a comparative analysis of one or another statistical toolkit for solving practical situations was carried out, based on the results of calculations, reasoned analytical conclusions and generalizations were made.

The maximum number of points that a student can receive based on the results of the exam is 40 points. A student should be considered certified if the minimum number of points obtained for the exam is 25 points.

RECOMMENDED LITERATURE

Main

1. Andrew Bruce, Peter Bruce, Peter Gedeck. Practical Statistics for Data Scientists. 2nd Ed. - Scale, 2020. - 350 p.
2. Marco Peixeiro.Time Series Forecasting in Python. - Scale, 2021-286 p.
3. Howard F. Horton, Megan H. Quirk, Thomas J. Quirk. Excel 2019 for Physical Sciences Statistics: A Guide to Solving Practical Problems. - Springer., 2021. - 242 p.
4. Scott E. Page. The Model Thinker: What You Need to Know to Make Data Work for You., 2019. - 448 p.
5. Alfonso Zamora Sais, Carlos Quesada Gonzalez, Diego Mondejar Ruiz, Luis Hurtado Gil. An Introduction to Data Analysis in R: Hands-on Coding, Data Mining, Visualization and Statistics from Scratch. - Springer, 2020. - 276 p.
6. Marcus du Sautois. Thinking Better. The Art of the Shortcut. - HarperCollins Publishers, 2022. - 352 p.
7. Прийняття рішень: теорія та практика : підручник / А. В. Катренко, В. В. Пасічник. – Львів : «Новий Світ – 2000», 2020. – 447 с.
8. Статистичний аналіз даних : навчальний посібник / Т. М. Паянок, Т. М. Задорожня. – Ірпінь : Університет державної фіскальної служби України, 2020. – 312 с.
9. Статистичне моделювання та прогнозування: навчальний посібник / під ред. д-ра екон. наук, проф. Раєвнєвої О.В. – Х.: ВД „ІНЖЕК”, 2013. – 537 с.
10. Rayevnyeva O. Statistics [Electronic resource]: textbook / O. Rayevnyeva, I. Aksonova, O. Brovko [et al.]; Simon Kuznets Kharkiv National University of economics. - E-text data (3,53 МБ). - Kharkiv : S. Kuznets KhNUE, 2020. - 376 p.: il. - The title screen. - Referenc.: p. 356-362. http://repository.hneu.edu.ua/handle/123456789/25678
11. Шабельник Т. В. Математичне моделювання соціально-економічних систем : навч. посібник / Т. В. Шабельник. – Маріуполь : МДУ, 2019. – 135 с. http://repository.hneu.edu.ua/handle/123456789/28090

Additional

12. Гороховатський В. О. Методи інтелектуального аналізу та оброблення даних : навч. посіб. / В. О. Гороховатський, І. С. Творошенко ; М-во освіти і науки України, Харків. нац. ун-т радіоелектроніки. – Харків : ХНУРЕ, 2021. – 92 с.

https://openarchive.nure.ua/server/api/core/bitstreams/2e55d639-52fd-48d9-b7b7-14989f49f291/content

13. Чаговець Л. О. Моделі ідентифікації та прогнозування стану цифровізації країн у світовому просторі / Л. О. Чаговець, В. В. Чаговець // Комунальне господарство міст. – 2023. – № 1(175). – С. 2–12. http://repository.hneu.edu.ua/bitstream/123456789/29885/1/%D0%A7%D0%B0%D0%B3%D0%BE%D0%B2%D0%B5%D1%86%D1%8C.pdf

14. Donoho, D. (2024). Data science at the singularity. Harvard Data Science Review, 6(1).

15. White, H. (2006). oApproximate Nonlinear Forecasting Methods, pin G Elliott, CWJ Granger, and A Timmermann, eds, Handbook of Economics Forecasting. New York: Else% vier, 460, 512.

16. Rayevnyeva O. V. The diagnostic model for assessing the state of stability of an industrial enterprise/ O. V. Rayevnyeva, M. P. Karpinski, O. I. Brovko et al // 13th PLAIS EuroSymposiumon Digital Transformation, PLAIS EuroSymposium 2021. Sopot, Poland, September 23/ – 2021 – P. 51-70. Режим доступу: https://doi.org/10.1007/978-3-030-85893-3

17. Rayevnyeva O. Computer-Mathematical Modeling of the Influence of the Macro-Environment on the Economic Behavior of the Enterprise / O. Rayevnyeva, O. Brovko, Su Rui // 7th International Symposium on Multidisciplinary Studies and Innovative Technologies. – 2023. http://repository.hneu.edu.ua/handle/123456789/30966

Information resources

18. Єдиний державний веб-портал відкритих даних // [Електронний ресурс]. Режим доступу: https://data.gov.ua/

19. Офіційний сайт Світового банку. – Режим доступу: http://www.worldbank.org/

20. Офіційний сайт Державної служби статистики України. – Режим доступу: http://www.ukrstat.gov.ua

21. Офіційний сайт Головного управління статистики в Харківській області [Електронний ресурс]. – Режим доступу: http:// uprstat. kharkov. ukrtel.net/,http: /uprstat. kharkov. ukrtel.net/.

22. Course page on the Moodle platform (personal learning system). - Access mode: https://pns.hneu.edu.ua/course/view.php?id=4771