Teaching statistics to future programmers using real data sets and R programming language

Liliia V. Pavlenko, Maksym P. Pavlenko and Vitalii H. Khomenko

Berdyansk State Pedagogical University, 4 Schmidta Str., Berdyansk, 71100, Ukraine

Abstract

This paper addresses the problem of teaching statistics to future programmers. It argues that the theoretical content of teaching statistics needs to be updated and oriented towards the practical field, even at the higher education level. It suggests that the teaching of statistics to students should move from theoretical methods to practical solutions of applied problems and emphasize the analysis and interpretation of results rather than the statistical calculations. The paper proposes a system of tasks based on real data sets obtained from statistical research as a way of improving the learning of statistics for future programmers. It shows that such tasks can increase the students' motivation compared to synthetic examples, which are commonly used in statistics courses. The paper also reviews the software tools for statistical data analysis and identifies their features and advantages for the learning process. It recommends using R, a specialized programming language, as the main tool for teaching statistics.

Keywords

statistics education, future programmers, real data sets, R programming language, applied problems

1. Introduction

The rapid growth of information in the modern world poses a challenge for the statistical education of society. Statistics is an essential component of the educational programs for training specialists in the field of IT [1, 2]. However, teaching statistics to students often encounters various problems, such as: different levels of prior knowledge, low level of motivation, lack of understanding of the relevance and applicability of statistics for their future profession [3].

The discipline of statistics has been undergoing significant changes and developments in recent years. Cox [4], Moore [5], Smith and Staetsky [6] raise many questions about the need to improve the objectives, content, methods and forms of teaching statistics.

Many researchers have investigated the issues and challenges of teaching statistics. They have provided recommendations for teaching statistics in different types of educational institutions [7, 8, 9, 10, 11, 12, 13]. Some of them have suggested moving from the theoretical learning to the practical application of statistical methods [10, 14].

Nicholl [15] notes that the theoretical content of teaching statistics has expanded significantly over the past 50 years, but this process has been uncoordinated, by adding new concepts without removing old ones. As a result, the content of teaching statistics has become overloaded with theoretical concepts that do not enhance the students' motivation and interest in learning statistics. Rumsey [16], Gal and Garfield [17] draw attention to the problems of teaching statistics and propose to change the paradigm of teaching and focus on the practical field, even at the higher education level.

Education is seen as an investment in human capital and production. The World Economic Forum predicts that the global demand for statistical data analysts will increase by almost six times in the next

CoSinE 2024: 11th Illia O. Teplytskyi Workshop on Computer Simulation in Education, co-located with the XVI International Conference on Mathematics, Science and Technology Education (ICon-MaSTEd 2024), May 15, 2024, Kryvyi Rih, Ukraine

✉ liliya.pavlenko@meta.ua (L. V. Pavlenko); pavlenko.2277@gmail.com (M. P. Pavlenko); v_g_homenko@ukr.net (V. H. Khomenko)
🌐 https://bdpu.org/en/faculties/fmkto/structure-fmkto/kaf-ktun/composition-ktun/pavlenkolv/ (L. V. Pavlenko); https://bdpu.org/en/faculties/fmkto/structure-fmkto/kaf-ktun/composition-ktun/pavlenko-2/ (M. P. Pavlenko); https://bdpu.org/en/faculties/fmkto/structure-fmkto/kaf-ktun/composition-ktun/homenko-2/ (V. H. Khomenko)
iD 0000-0001-7823-7399 (L. V. Pavlenko); 0000-0003-0091-696X (M. P. Pavlenko); 0000-0002-7361-2169 (V. H. Khomenko)

five years.

According to the Modis survey [18], 97.44% of respondents (representatives of banks and industry) consider data analysis as a promising skill for success in sales and marketing. However, they are more interested in interpreting the data than in performing the calculations. 42% of respondents complain about the shortage of qualified professionals who have skills in statistical data analysis in the labor market. 55% of respondents say that it is difficult to find specialists who can calculate and interpret the results.

Every day, large amounts of various data are generated and accumulated in the world [19]. Therefore, the labor market demand for data analysts and data scientists is constantly growing. Varian [20] notes that data analyst will become one of the most popular professions in the future.

Therefore, improving the education of students in statistics requires moving from the theoretical methods to the practical solutions of applied problems and shifting the emphasis from the statistical calculations to the analysis and interpretation of results.

To prepare an intellectually active, knowledgeable and skillful specialist, education needs to move from the reproductive to the innovative learning. Innovative learning is a creative combination of the traditional and new teaching methods, tailored to each discipline, based on its theoretical content and practical orientation [21]. Moreover, it should be considered that teaching students is not only about developing certain professional competencies, but also about aligning them with the current modern requirements [22]. It means that the future specialists should be able to express their thoughts and concepts verbally, understand the language of symbols, signs and schemes. This is not just the ability to think creatively, but also the ability to make original decisions and actions.

To organize innovative teaching of statistics in accordance with modern requirements, it is advisable to use special software tools for statistical data analysis. However, there are also specialized programming languages and environments that can be used to analyze data, interpret results and prepare conclusions and reports in various formats quickly and efficiently.

Therefore, there is a contradiction between the traditional approaches to teaching statistics and the society's expectations for the level of training of modern IT specialists in the field of statistical data analysis, as well as between the theoretical orientation of the content of statistics teaching and the need to train a specialist with applied tools and methods of statistical data analysis.

The aim of this paper is to justify the use of R programming language as a teaching method for learning statistics.

2. Results

The following main methods were used in the research process: content analysis of scientific and methodical literature, generalization and systematization to clarify the state of the problem development; questionnaire of those getting higher education and initial statistical processing of the obtained results to clarify the current state of the researched problem; generalization of theoretical and practical data to justify the introduction of innovative approaches to the study of statistics by students based on the use of programming language R.

The process of teaching statistics to the students is associated with certain difficulties: the study material in this course contains a large number of definitions and formulas. At the same time, students need not only to reproduce them, but also to understand the meaning and be able to apply in practice. However, with the traditional organization of the educational process, practical tasks are far from the real economic, social and other processes that occur in real life. The analyzed data are generalized and do not allow to fully form students' understanding of the need and expediency of studying this discipline and the opportunity to implement the acquired competencies in their further professional activities.

Therefore, most students learn statistics in fragments, and do not form systemic knowledge as a result. In addition, mainly verbal presentation of information increases fatigue, resulting reducing productivity of the learning process [23].

The number of statistically educated people is decreasing. It is difficult for potential employers to find a specialist who will be able to perform statistical calculations without prior training and explanation. Therefore, there is a need to improve the content of teaching this discipline through the introduction of practical tasks.

Improving the content of the statistics course requires the introduction of changes in the methods and means of its teaching using innovative technologies.

Scientific innovations that promote scientific progress cover all areas of knowledge. There are socio-economic, organizational and managerial, technical and technological innovations. One of the types of social innovations is pedagogical innovation.

Pedagogical innovation is an innovation in the field of pedagogy, purposeful progressive changes that make stable elements (innovations) in the educational environment that improve the characteristics of both – its individual components and the educational system on the whole [24].

Pedagogical innovations can be carried out both with the application of the educational system's own resources (intensive way of development) and with the involvement of additional capacities (investments) – new means, equipment, technologies, capital investments, etc. (extensive way of development).

Kazakov [25] notes that the combination of intensive and extensive ways of pedagogical systems development allows to carry out so-called "integrated innovations", which are built at the junction of various, multilevel pedagogical subsystems and their components.

The main ways and objects of innovative transformations in the teaching of statistics are:

- making concepts and strategies for the development of statistical education [26];
- updating the content of statistics training;
- change and development of new learning technologies;
- improving the training of IT specialists in the field of statistical data analysis;
- designing new models of the educational process for teaching statistics;
- improving the monitoring of the educational process and student learning;
- new generation electronic teaching aids development.

Innovation can take place at different levels. The highest level includes innovations that affect the entire pedagogical system.

Kulinenko [27] notes that while organizing the innovation, it should be considered that:

- innovative ideas must be clear, convincing and adequate to the real educational needs of man and society, they must be transformed into specific goals, objectives and technologies;
- innovation activity should be morally and materially stimulated, legal support of innovation activity is necessary;
- not only results are important in pedagogical activity, but also ways, means, methods of their achievement.

The current problems of teaching statistics in modern higher educational institutions include the review of experience associated with the intensification of learning. One of the main teacher's tasks is to teach students to obtain the necessary information independently, to teach them to consciously process the obtained information [28]. In order for them to be able to study the teaching materials on their own, the materials need to be designed primarily for students and not for teachers.

Possibilities of "Statistics" discipline for experts in the field of IT consists first of all of that knowing mathematical language and modeling that will allow the student to be better guided in forecasting of economic, social, technical and other processes; secondly, that statistics by its internal nature has rich opportunities for the formation of students algorithmic thinking.

Future IT professionals must not only know the theoretical foundations, but also be able to apply the means of automating statistical analysis. Such tools include specialized statistical software packages and programming languages.

Statistical packages on the basis of functionality can be divided into 3 main groups.

1. Universal statistical packages Statistica, SPSS, Statgraphics, STATA, Stadia, SYSTAT, S-PLUS and MS Excel. These packages are not targeted at a specific subject area and can be used to analyze data from different industries. Typically, they offer a wide range of statistical methods and have a relatively simple interface.
   It is recommended to work with such packages for starter users who have only basic knowledge in the field of statistics, as well as experienced users in the initial stages of working with data, when statistical methods that will be used to address a particular issue are not clearly defined yet. The versatility of the universal package allows holding a pilot analysis of different data types using a wide range of statistical methods. The vast majority of existing universal packages has much common functionality and is similar in the composition of the built-in statistical procedures.
2. Professional statistical packages such as SAS or BMDP. Professional packages, in the contrast to the universal ones, allow you to work with extremely large amounts of data, apply highly specialized methods of analysis and create your own data processing system. As a rule, such packages are complex and should not be used in the educational process.
3. Specialized statistical packages BioStat, Datastream, Datascope, etc. were designed for statistical analysis in specific areas of activity, which use special methods of statistical analysis, usually not presented in the universal packages.

Specialized packages allow analysis using a limited number of specialized statistical methods or are used in a specialized subject area. As a rule, such statistical packages are handled by specialists who are well acquainted with data analysis methods in the field to which the package is focused. For example, the BioStat statistical package was created to analyze data in the field of biology and medicine.

Most of the existing statistical packages have a flexible modular structure that can be supplemented and expanded owing to the custom modules that are optionally purchased or freely available on the Internet. Such flexibility allows you to adapt packages to the needs of a particular user.

Statistical packages are just the tools for an experienced professional. If the specialist does not have sufficient knowledge and competencies, then, even the most advanced software product will not allow holding quality data analysis. However, the wrong software, which does not contain the required set of statistical procedures, can make the work of even an experienced specialist more difficult.

Therefore, during the training of IT specialists it is necessary to acquaint those who get higher education with the available statistical packages and their characteristics, but the application of specialized programming languages is closer and more understandable for the students while conducting statistical data analysis.

For statistical data analysis it is possible and appropriate to use R and Python programming languages.

We will consider the features of the programming language R. The language R is a powerful high-level object-oriented programming language and environment for statistical calculations and visualization of source and calculation data, which allows you to solve many problems in the field of data processing. It's a free open source program under GNU GPL designed to run common operating systems (Windows, macOS, Linux).

Tens of thousands of specialized modules and utilities have been developed for this language. One of the most important features of the programming language R is the efficient implementation of vector operations, which allows the application of compact notation while processing large amounts of data. All this makes R an effective tool for obtaining useful information from large amounts of various statistics, including Big Data.

The R language is a convenient and effective tool for teaching statistical analysis, data processing and visualization.

It is also possible to use the Python programming language in the field of data analysis and interactive research calculations with results visualization. Python is an open source object-oriented programming language. The relatively recent advent of improved libraries for Python (primarily pandas) has made it a serious competitor to the R language for statistical data analysis. Combining with the benefits of Python as a universal programming language makes it an excellent choice for creating data processing applications.

So, the use of a specialized programming language as a learning tool contributes to the development of statistical data analysis skills as well as the development of algorithmic thinking of future IT professionals.

In order to study the relevance of the problem of scientific research, a ascertaining experiment was conducted among students of IT specialties. The issues that allow finding out the opinion of higher education students on the problem of improving the methods of teaching statistics to future IT professionals were studied.

The results of the ascertaining experiment are presented in percentages and indicate the number of positive answers to the questions. The survey was organized using Google Forms. 83 students majoring in 015 Professional Education (Digital Technology) and 015 Professional Education (Computer Technology) took part in ascertaining experiment.

2.1. Declared interest of students in studying the course of statistics

In this block students were asked two questions. You can see the results of the answers to the first question of the survey in figure 1. The analysis of answers allows establishing the level of awareness of students in the demand for specialists in the labor market who know how to analyze data.

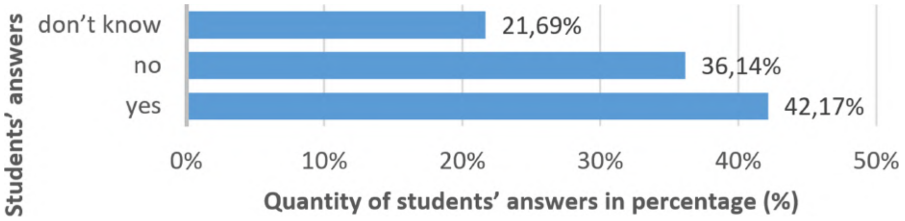

Figure 1: Results of answers to the question regarding students' awareness in the demand for the specialists on data analysis in the labor market.

Analysis of students' answers allows us to conclude that the majority of respondents, 42.17% believe that a data analysis specialist is in demand in the labor market. This confirms the relevance and need to study the course of statistics for IT professionals.

The second question clarified which specialties in data analysis students consider the most relevant today. The results of the student survey are shown in figure 2.

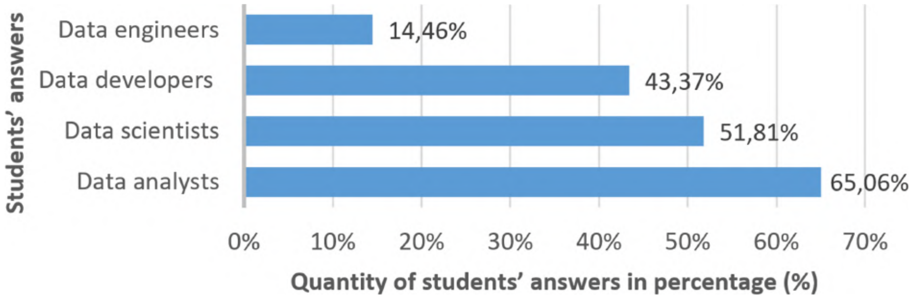

Figure 2: The results of questionnaire regarding students' awareness about modern professions on data analysis in the labor market.

The most famous profession among the future students programmers is the profession of data analysts (65.06%), in second place is the profession of data scientists (51.81%). These professions are known to more than 50% of students, which indicates their awareness and interest in this field.

So, based on the results of studying the answers to the questions of this block, we can draw the following conclusion. Training statistics of future IT professionals is relevant, because students are aware of the existence of professions in the field of data analysis and believe that they will need statistics in future professional activities.

2.2. Students' opinion about the need to fill the content with tasks of an applied nature

Students were asked to answer open-ended questions: "Which subject area data analysis you are interested conducting in?" The students' answers showed that the most popular data for processing are data from sociology, medicine, engineering, economics and biology.

Also, the idea of what data students are interested in working with in practice was studied. The results of answers to the questions are shown in figure 3.

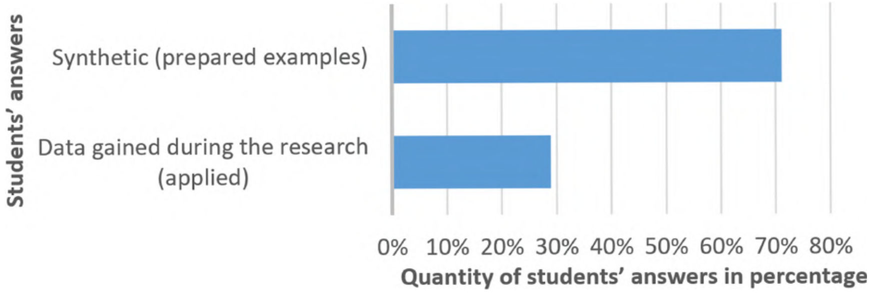

Figure 3: Students' opinion on data origin for practical tasks.

Among the surveyed respondents, 71.08% believe that data obtained as a result of practical research and having an applied nature are most attractive for them. This indicates the need to develop practical and laboratory work based on real data obtained from statistical studies.

2.3. Students' interest in using programming languages and software for statistical data analysis

The purpose of the third block of questions was to study the opinion of respondents about the need and feasibility of using software and programming languages for statistical data analysis.

Students were asked the following questions: "Do you know programming languages with which it is possible to perform statistical data analysis (enter)?", "Which software product interface is more user friendly for you?", "Are you more interested in data analysis using special software or using a programming language?"

According to the first question, the opinions of the respondents were divided as follows: 55.42% indicated the programming language R, 28.92% indicated the Python programming language. Programming languages such as C++ (9.64%) and Java (6.02%) were also indicated (figure 4).

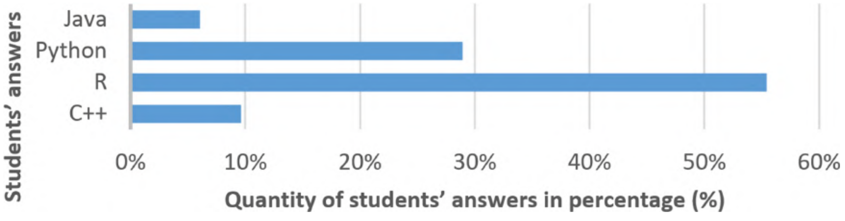

Figure 4: Respondents' answers to the question on convenience of program packages interface.

The obtained results allow us to state that the R language is the best known as a mean of statistical data analysis. So, we will use this programming language to solve application problems.

In choosing the convenience of the software package interface, respondents preferred MS Excel (56.63%), followed by Statistica software package (28.92%), followed by SPSS (14.46%) (figure 5).

So, the students will be asked to use MS Excel and Statistica for practical calculations.

According to the results of students' answers to the third question of this block, the programming language (57.83%) was chosen by the students as the main tool for organizing the training of statistical data analysis (figure 6).

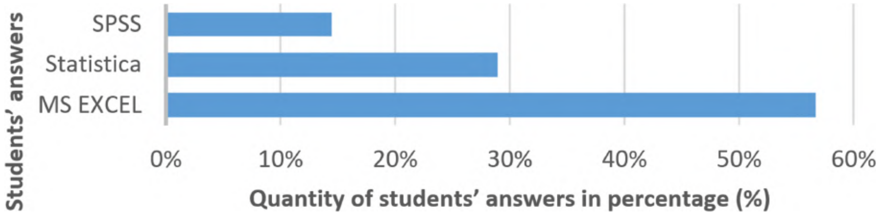

Figure 5: Choosing program packages for statistical data analysis.

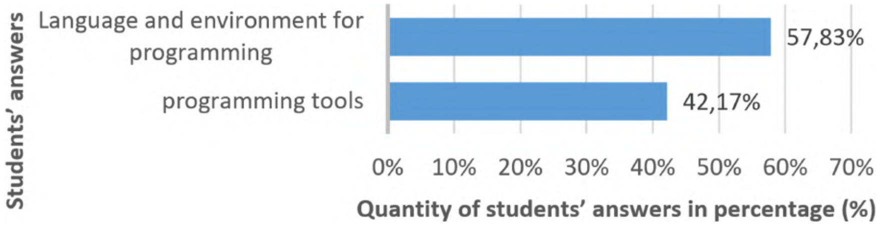

Figure 6: Respondents' answers regarding choosing the mean of solving the tasks of statistical data analysis.

So, students in the class will be asked to use the programming language R as the main tool for practical calculations. MS Excel and Statistica will be used as aids in statistical analysis.

3. Using applied tasks for teaching statistics

Taking into account and summarizing the results of the study, in our opinion, it is advisable to build the content and structure of the course considering the wishes of students. In practical classes, tasks that are of a real applied nature and based on real statistics should be considered. One of the main teaching methods should be a practical method of learning based on programming. The means of statistical data analysis in practical classes can be both software tools for data analysis (MS Excel and Statistica) and the language and programming environment R.

A system of tasks has been developed for the course. Let's consider an example for training of the statistical analysis in the R environment. For carrying out the analysis we will take data from the website https://abit-poisk.org.ua, namely data concerning entrants for 2017. This site contains large amounts of data, for our example we will take only entrants who entered the Faculty of Physical and Mathematical Computer and Technological Education of Berdyansk State Pedagogical University in the specialty "Professional Education (Computer Technology)" and "Professional Education (Digital Technology)", the level of "bachelor".

A total of 31 applications were submitted for these specialties. We will analyze these data, using descriptive statistics in R and present the results using the most common graphs in R when analyzing this data.

Step 1. We set the name, specialty, id, total score of the external evaluation, status (budget / contract), then enter the data into the table. We will set the value in the form of vectors with the command `<- c (’’vector_value1’’, ‘‘vector_value2’’,. . . )`. We build the table from the received vectors by means of the command > `studentdata`. Commands for a table creation with the information about applicants:

```
 > last_name <-c(’’Shvachko’’, ‘‘Dybiaga’’, ‘‘Kartashov’’, ‘‘Sytosenko’’,
 ‘‘Filipenko’’, ‘‘Klimenko’’, ‘‘Veretelnik’’, ‘‘Diakov’’,’’Salionov’’,
‘‘Bagnuk’’, ‘‘Kombarov’’, ‘‘Baranovsky’’, ‘‘Kiseliov’’, ‘‘Sakun’’, ‘‘Bova’’,
 ‘‘Potapova’’, ‘‘Kobzar’’, ‘‘Sementsov’’, ‘‘Cybulka’’, ‘‘Teplov’’,
 ‘‘Mitushkin’’, ‘‘Kartinik’’, ‘‘Gavrylenko’’, ‘‘Trotsenko’’,
 ‘‘Panchukov’’, ‘‘Kyslynsky’’, ‘‘Sagirov’’, ‘‘Korobov’’,
 ‘‘Shatalina’’,’’Tichovod’’,’’Popov’’)
 > specialty <-c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
 1,1,1,1,1,2,2,2,2,2)
 > id <-c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,
 21,22,23,24,25,26,27,28,29,30,31)
 > rating <-c(186,184,180,179,173,173,170,168,167,166,163,
 162,160,156,148,145,145,142,142,140,140,139,135,131,129,123,
 147,146,140,136,128)
 > status <-c(1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,
 0,0,0,1,1,1,1,0)
 > studentdata <- data.frame(id, last_name, rating, status)
 > studentdata
    id  last_name rating status
 1   1   Shvachko    186      1
 2   2    Dybiaga    184      1
 3   3  Kartashov    180      1
 4   4  Sytosenko    179      1
 5   5  Filipenko    173      1
 6   6   Klimenko    173      1
 7   7 Veretelnik    170      1
 8   8     Diakov    168      1
 9   9   Salionov    167      1
 10 10     Bagnuk    166      1
 11 11   Kombarov    163      1
 12 12 Baranovsky    162      1
 13 13   Kiseliov    160      0
 14 14      Sakun    156      0
 15 15       Bova    148      0
 16 16   Potapova    145      0
 17 17     Kobzar    145      0
 18 18  Sementsov    142      0
 19 19    Cybulka    142      0
 20 20     Teplov    140      0
 21 21  Mitushkin    140      0
 22 22   Kartinik    139      0
 23 23 Gavrylenko    135      0
 24 24  Trotsenko    131      0
 25 25  Panchukov    129      0
 26 26  Kyslynsky    123      0
 27 27    Sagirov    147      1
 28 28    Korobov    146      1
 29 29  Shatalina    140      1
 30 30   Tichovod    136      1
 31 31      Popov    128      0
```

Step 2. We will calculate the main statistical values: average, median, standard deviation, minimum and maximum value. The results of the main statistical values calculation:

```
> y <- mean(rating)
> y
[1] 153
> sd <-sd(rating)
> sd
[1] 18.03145
> var <-var(rating)
> var
[1] 325.1333
> mad <-mad(rating)
> mad
[1] 22.239
> min <-min(rating)
> min
[1] 123
> max <-max(rating)
> max
[1] 186
```

According to the results of the calculations, the following data were obtained: the average score of entrants with external evaluation is 153, the average difference between the scores of different entrants is 22 points, the lowest result (min) – 123 points, the best result (max) – 186 points.

Step 3. Let's construct a histogram of frequencies for external evaluation points using the command `> barplot` (figure 7):

```
> counts <- table(studentdata$rating)
> barplot(counts, main=''Frequency diagram'', xlab=''Rating'',
ylab=''Frequency'')
```

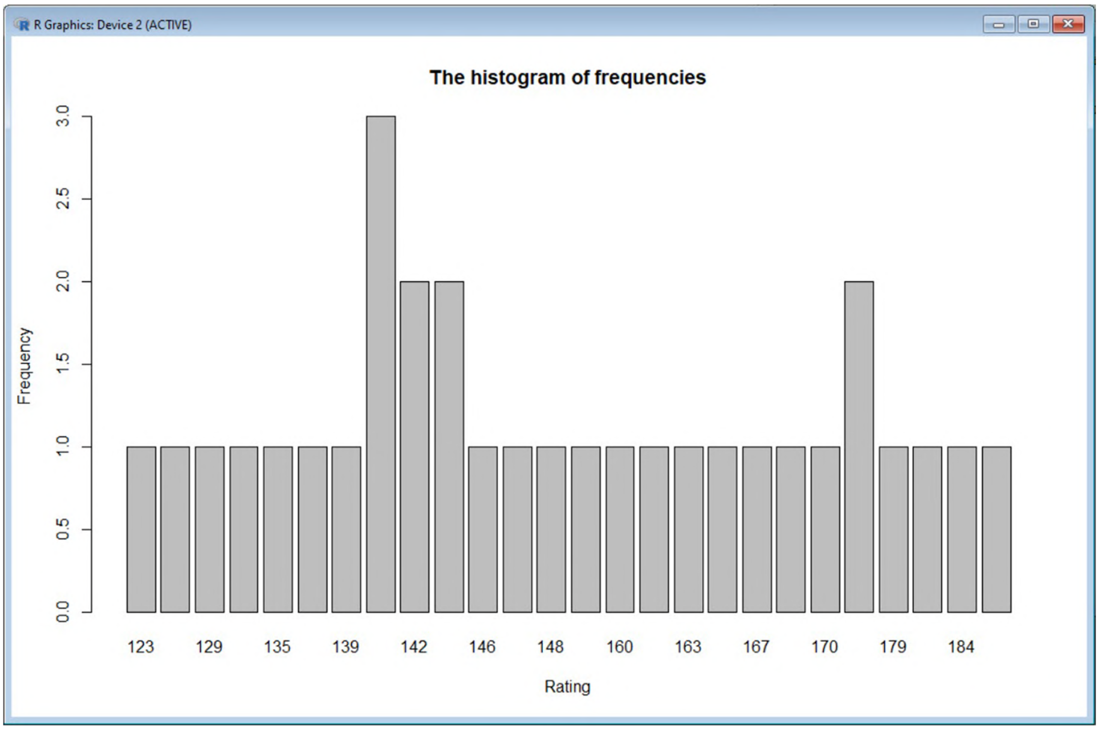

Figure 7: The histogram of frequencies for external evaluation points.

The histogram of frequencies shows that the largest number of entrants has a score from 139 to 142 points, as well as the fact that the vast majority has a unique score with EIT, which is no longer repeated.

Step 4. We construct histograms of points / frequencies with a normal distribution curve. With this purpose we use the command > box. We will build: on the x-axis – the parameter rating, and on the y-axis – the frequency of the score in the table (figure 8):

```
> box()
> library(plotrix)
> x <-studentdata$rating
> h <-hist(x, breaks=12, col=’’darkblue’’, xlab=’’ZNO score’’,
main=’’Frequencies histogram with the curve of distribution ‘‘)
> xfit <-seq(min(x), max(x), length=40)
> yfit <-dnorm(xfit, mean=mean(x), sd=sd(x))
> yfit <-yfit * diff(h$mids[1:2] * length(x))
> lines(xfit, yfit, col=’’red’’, lwd=3)
```

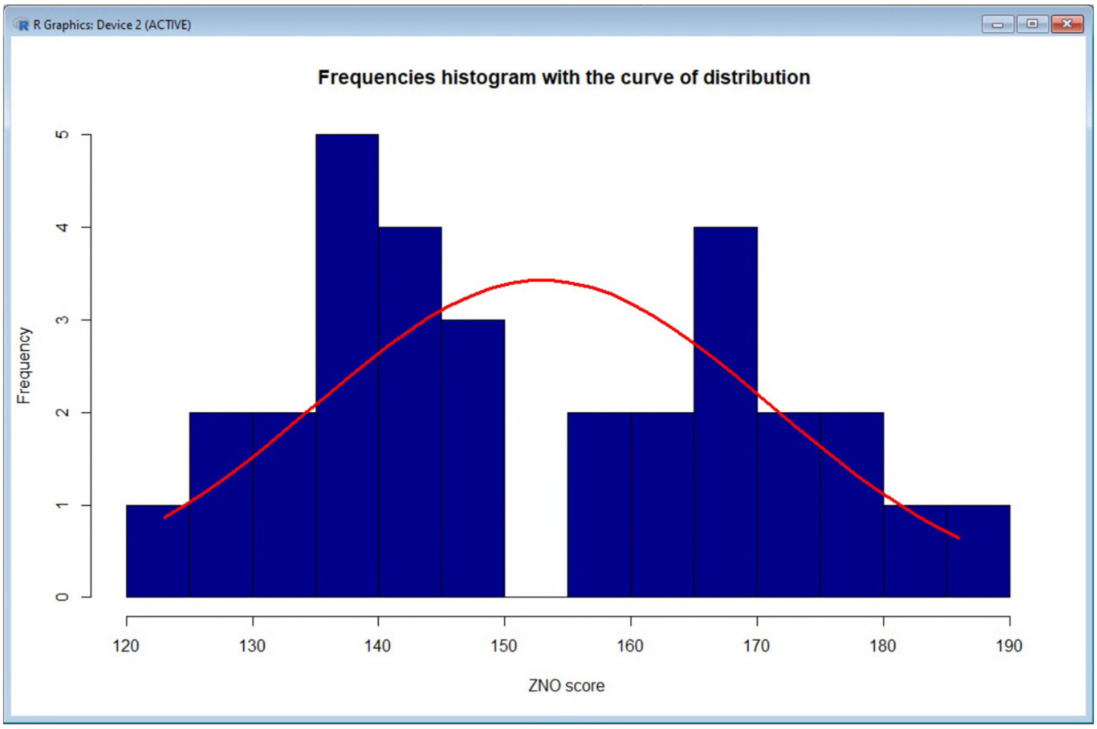

Figure 8: Frequencies histogram with the curve of distribution.

The distribution histogram shows that the data on the scores of applicants are not the subject to the normal law of distribution. We have a lot of “average” entrants, i.e. those who passed the external examination from 135 to 145 points. There are also those who passed 165 points, i.e. entrants with a “sufficient” level. There are very few who scored more than 180 points.

Step 5. We construct a diagram of the nuclear estimation of the density of values for external evaluation points using the command > box (figure 9):

```
> box()
> par(mfrow=c(2,1))
> d <- density(studentdata$rating)
> plot(d)
```

The nuclear density estimation diagram shows that the highest density is observed in the range from 130 to 155 points. That is, in this interval, based on the graph, the values differ by 25 points, then, if you take the full table, they differ by 22 (see standard deviation).

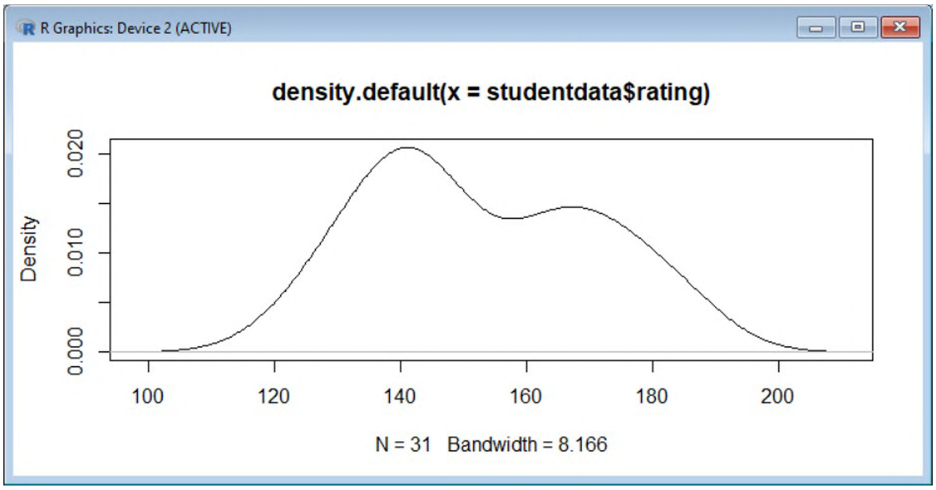

Figure 9: The diagram of nuclear density estimation.

As a result of solving applied problems using theoretical knowledge from different sections of statistics, students will not only master the skills of using statistical methods, but also develop the ability to interpret the results and predict the studied processes. It should be emphasized that the use of programming as a practical teaching method will allow students to improve their knowledge and skills in the field of programming as well as the use of algorithms and design patterns.

Using real data for statistical analysis, students will be able to understand the need and feasibility of statistical research in future professional activities.

One of the problems of using application tasks with real data is the selection and use of data sets. Much of the datasets are closed and inaccessible for free research and use. However, there are organizations that provide free access to data:

- World Bank Open Data (https://data.worldbank.org/) provides more than 3,000 sets of economic and social data on various indicators. Data can be downloaded in csv and xml formats. The service supports API access, which allows you to automate data downloads using the programming language R.
- The unified state web portal of open data (https://data.gov.ua/) contains 15 categories of data sets that are constantly updated. Datasets are available for download in Excel, csv, json and xml formats. All data are available from Creative Commons Attribution 4.0 International license.
- The official page of the All-Ukrainian Population Census (http://database.ukrcensus.gov.ua) provides access to information on the population living in the country, socio-economic characteristics, and demographic indicators, level of education, national composition and language characteristics. Datasets can be downloaded in txt, csv, html formats.
- Open World Health Organization data repository (https://www.who.int/data/gho/). The site provides datasets on the health status of citizens of World Health Organization member states. Datasets are divided into over 100 categories. Data can be downloaded in Excel format or use the API for direct access to data.
- UNICEF Dataset (https://data.unicef.org/) collected relevant data on education, child labor, child disability, infant mortality, maternal mortality, water and sanitation, pneumonia, malaria and more. Datasets are available in Excel and csv formats.
- Registry of Open Data on AWS (RODA) (https://registry.opendata.aws/) contains data located on AWS servers. The service offers access to over 200 datasets. There is a page with additional information, usage examples, license information, and more for each data set. Using the wide range of computing products offered by AWS (Amazon EC2, Amazon Athena, AWS Lambda and Amazon EMR), it is possible to share data in the cloud. This allows users to spend more time analyzing data rather than collecting data. When using data sets hosted on AWS, it is necessary

to consider the type of license of each specific data set, as they belong to different agencies, government organizations, researchers, businesses and individuals.

- Data.gov (https://www.data.gov/) provides open data sets of the US government. The resource contains more than 200,000 data sets from various sources: federal agencies, states, counties, cities, etc. Data can be obtained in various formats, including Excel, csv, json, xml.
- The GroupLens Research (https://grouplens.org/) provides several sets of movie ratings data provided by MovieLens users. The kits contain movie ratings, movie metadata (genre and release year), and user demographics (age, gender, and occupation). Such data can be used to develop a recommendation system based on regression analysis.
- Open data sets Yelp (https://www.yelp.com/dataset) is a subset of our businesses, reviews, and user data for application in personal, educational, and academic purposes. Available as JSON files, use it to teach students about databases, to learn NLP, or for sample production data while you learn statistics.
- Kaggle (https://www.kaggle.com/datasets) a social network for researchers, which provides access to various data sets for analysis and research. The convenience of Kaggle is that it is not just a data warehouse. Each data set brings together a community of researchers in which data are discussed and approaches to data processing are elucidated.
- Google Public Data Explorer (https://www.google.com/publicdata/directory) provides access to more than 130 datasets submitted by World Bank, U. S. Bureau of Labor Statistics, OECD, IMF and other organizations.

All considered services provide access to open data sets. This allows you to fill the content of teaching statistics for future programmers with the tasks of applied direction.

4. Experimental verification of the effectiveness in the use of applied tasks to teach statistics to the future programmers

Using programming language R and tasks of applied direction while training statistics with future IT specialists.

The main purpose of the pedagogical experiment is to test the hypothesis that the use of programming language R and applied problems in teaching statistics to the future IT professionals will help increase the educational motivation of students.

According to the hypothesis of the study, the experiment involved checking the level of motivation of students of IT specialties in the field of statistics based on the results of implementation of applied problems and programming language R. The experiment was conducted on the basis of Berdyansk State Pedagogical University. Students majoring in 015 Professional Education (Digital Technology) and 015 Professional Education (Computer Technology) took part in it.

Control and experimental groups were organized. In the control group, the educational process was carried out according to the traditional methods. This technique involved the use of specialized software (Microsoft Excel, Statistica, etc.) and synthetic tasks, the content of which did not take into account the specifics of future professional activities of students of IT specialties. The control group (CG) consisted of 42 students. The experimental group used application problems and the programming language R to solve them. The experimental group (EG) included 32 students.

During the formation of control and experimental groups, their alignment was carried out taking into account the initial level of educational motivation of students.

The success of the pedagogical research was ensured by the application of the standardized methods. This guaranteed the reliability of the results.

Experimental methods of teaching statistics of future programmers using professional tasks and programming language R was based on their application at all stages of learning: in learning new material as a motivating task, at the stage of consolidation, in independent work of students as a professionally oriented project.

An electronic learning tool has been developed for students programmers to provide information and methodological support for the statistics course. The development of an electronic tool takes into account students age and preparation level. The developed learning tool contains theoretical materials, tasks for practical implementation, visual materials with examples of the application of the programming language R, a guide to the commands of the R language and a list of recommended reading. The e-learning tool is available on the Internet at the link https://r.ktuni.bdpu.org/.

In order to test the effectiveness of the implemented experimental training, the level of educational motivation was chosen as a criterion. To assess the dynamics of changes in motivation to study statistics, future IT specialists used the method of Rean and Yakunin [29] aimed at diagnosing educational motivation in general in order to identify the predominant types of motives for learning. The technique allows identifying the predominant type of motives and to trace the dynamics of changes in the structure of educational motivation. The methodology is standardized and involves the study of 16 types of educational motives of students.

Positive motivation for learning ensures the successful formation of knowledge and skills. High positive motivation can compensate for insufficiently high abilities of students. With the right choice of means of motivation for learning, there is a positive pedagogical influence. Focusing only on "negative" motives (avoidance, fear of failure, fear) is always less effective than "positive" ones. In our study, we will determine the impact of the developed system of tasks on the level of educational motivation of students.

Table 1 presents the results of calculating the average scores for each type of educational motives on the scale of Rean and Yakunin [29]. Comparative analysis of table 1 allows us to conclude that before the experiment the levels of educational motives of students in the control and experimental groups did not differ. After the experiment in the experimental group there is an increase in the levels of the internal educational motives of students. In general, the level of educational motivation in the experimental group is higher than in the control group, except for the motives of avoiding failure and punishment.

Table 1
The results of students' questionnaire according to the methods of Rean and Yakunin [29].

Educational motivation	Before the experiment		After the experiment	
	CG	EG	CG	EG
1. To become a qualified specialist	6.6	6.6	6.7	6.8
2. To get the diploma	6.7	6.6	6.2	6.8
3. To continue successful studies at further courses	5.6	6.3	6.0	6.2
4. To study successfully, to pass exams for "good" and "excellent" marks	6.0	5.3	4.5	6.2
5. To get constant scholarship	5.5	5.2	4.9	5.5
6. To gain deep and profound knowledge	6.0	6.3	6.3	6.8
7. To be always ready for classes	4.5	4.5	5.0	5.2
8. Not to give up learning the subjects of the educational cycle	5.5	5.6	5.5	6.5
9. Not to lag behind the classmates	6.0	5.6	5.5	5.8
10. To provide future successful professional activity	6.8	6.6	6.5	6.9
11. To execute pedagogical requirements	5.0	4.7	5.2	5.5
12. To get teachers' respect	4.8	5.2	3.6	4.9
13. To be an example for the classmates	3.2	4.7	3.5	4.3
14. To gain parents' and relatives' respect	4.5	4.8	5.0	6.6
15. To avoid condemnation and punishment for bad studying	4.1	4.9	5.0	4.5
16. To get intellectual satisfaction	4.9	4.91	4.5	6.6

Table 2 shows the results of statistical comparison of the control and experimental groups before

Table 2
Statistical comparison of the students of control and experimental groups educational motivation levels before and after the experiment.

Before the experiment			After the experiment		
W_{emp}	W_{crit}	Taken hypothesis	W_{emp}	W_{crit}	Taken hypothesis
0.1508	1.96	H_0	2.186	1.96	H_1

and after the experiment. The following statements were formulated as working hypotheses: H_0 – levels of learning motivation in the compared groups do not differ; H_1 – levels of motivation to learn in the compared groups differ. The Mann-Whitney U-test was used to determine the difference between the samples. This is a non-parametric statistical criterion used to estimate the difference between two samples at the level of any qualitatively measured trait. It allows you to detect differences in the value of the parameter between small samples.

Statistical analysis allows us to conclude that at the level of significance $\alpha = 0.05$ the initial states of the experimental and control groups (before the experiment) coincide. At the end of the experiment, the levels of educational motivation differ.

So, the results of the study indicate that the hypothesis of the study was confirmed, namely the introduction of statistics of the R programming language and applied problems in the learning process helps to increase the level of educational motivation of future IT professionals.

5. Conclusions

This paper has provided a theoretical justification for the introduction of innovative approaches to teaching statistics. It has shown that the teaching of statistics to future programmers should be based on the use of applied tasks developed with real data sets obtained from statistical research. Such tasks can increase the students' motivation and interest compared to synthetic examples, which are often used in statistics courses.

Real data sets for statistical analysis are a rich source of applied tasks. They are freely available on the Internet and cover various subject areas, such as sociology, medicine, engineering, economics and biology. Therefore, the development of practical and laboratory work for future IT professionals should include tasks that involve real data from these domains.

Using the R programming language to teach statistics to future programmers allows the use of a practical training method based on programming. This method engages students in familiar and relevant activities and develops their programming skills. Therefore, we propose to use R as the main tool for teaching statistics. MS Excel and Statistica software packages can be used as supplementary tools.

In further research, we plan to develop a methodology for implementing and applying R and Python programming languages for statistical data analysis.

References

[1] O. V. Bondarenko, O. V. Hanchuk, O. V. Pakhomova, G. Tsutsunashvili, A. Zagórski, Visualization of demographic statistical data, IOP Conference Series: Earth and Environmental Science 1049 (2022) 012076. doi:10.1088/1755-1315/1049/1/012076.

[2] L. F. Panchenko, V. Y. Velychko, Unveiling the potential of structural equation modelling in educational research: a comparative analysis of Ukrainian teachers' self-efficacy, Educational Technology Quarterly 2023 (2023) 157–172. doi:10.55056/etq.601.

[3] A. Zieffler, J. Garfield, S. Alt, D. Dupuis, K. Holleque, B. Chang, What Does Research Suggest About the Teaching and Learning of Introductory Statistics at the College Level? A Review of the Literature, Journal of Statistics Education 16 (2008) 8. URL: https://www.tandfonline.com/doi/full/10.1080/10691898.2008.11889566. doi:10.1080/10691898.2008.11889566.

[4] D. R. Cox, The current position of statistics: a personal view, International statistical review 65 (1997) 261–276.

[5] D. S. Moore, New pedagogy and new content: The case of statistics, International statistical review 65 (1997) 123–137.

[6] T. M. F. Smith, L. Staetsky, The teaching of statistics in UK universities, Journal of the Royal Statistical Society: Series A (Statistics in Society) 170 (2007) 581–622. doi:10.1111/j.1467-985X.2007.00482.x.

[7] D. Ben-Zvi, J. B. Garfield, The challenge of developing statistical literacy, reasoning and thinking, Springer, 2004.

[8] R. Biehler, D. Frischemeier, C. Reading, J. M. Shaughnessy, Reasoning about data, in: International handbook of research in statistics education, Springer, 2018, pp. 139–192.

[9] A. J. Bishop, M. A. K. Clements, K. Clements, C. Keitel, J. Kilpatrick, C. Laborde, International Handbook of Mathematics Education, Springer Science & Business Media, 1996.

[10] J. Garfield, D. Ben-Zvi, Developing students' statistical reasoning: Connecting research and teaching practice, Springer Science & Business Media, 2008.

[11] C. W. Langrall, K. Makar, P. Nilsson, J. M. Shaughnessy, Teaching and learning probability and statistics: An integrated perspective, 2017.

[12] J. M. Shaughnessy, Research in Probability and Statistics : Reflections and Directions, in: Handbook on Research in Mathematics Education, 1992, pp. 465–494. URL: https://ci.nii.ac.jp/naid/10029959707/.

[13] J. M. Shaughnessy, Research on Statistics' Reasoning and Learning, in: Second Handbook of Research on Mathematics Teaching and Learning, 2007, pp. 957–1009. URL: https://ci.nii.ac.jp/naid/10029959708/.

[14] J. M. Watson, N. E. Fitzallen, P. Carter, Top drawer teachers: Statistics, 2013. URL: http://ecite.utas.edu.au/87993.

[15] D. F. Nicholl, Future directions for the teaching and learning of statistics at the tertiary level, International Statistical Review 69 (2001) 11–15.

[16] D. J. Rumsey, Statistical literacy as a goal for introductory statistics courses, Journal of Statistics Education 10 (2002).

[17] I. Gal, J. Garfield, Curricular goals and assessment challenges in statistics education, The assessment challenge in statistics education (1997) 1–13.

[18] Modis, STEM IQ Survey Results 2018, 2018. URL: https://www.modis.com/en-us/resources/employers/stem-iq-survey-2018/.

[19] V. H. Khomenko, L. V. Pavlenko, M. P. Pavlenko, S. V. Khomenko, Cloud technologies in informational and methodological support of university students' independent study, Information Technologies and Learning Tools 77 (2020) 223–239. URL: https://journal.iitta.gov.ua/index.php/itlt/article/view/2941. doi:10.33407/itlt.v77i3.2941.

[20] H. R. Varian, Nel 2020 il data analyst sarà la professione più ricercata, 2017. URL: https://www.giornaledibrescia.it/rubriche/impresa-4-0/nel-2020-il-data-analyst-sar%C3%A0-la-professione-pi%C3%B9-ricercata-1.3182021.

[21] A. V. Kaminskaya, Forming of readiness of future teachers to innovative activity in higher educational establishment, Scientific Bulletin of Donbass (2011). URL: http://nvd.luguniv.edu.ua/archiv/NN13/11kavvnz.pdf.

[22] A. M. Striuk, S. O. Semerikov, Professional competencies of future software engineers in the software design: teaching techniques, Journal of Physics: Conference Series 2288 (2022) 012012. doi:10.1088/1742-6596/2288/1/012012.

[23] M. M. Fitsula, Pedagogy, 2000.

[24] E. S. Rapatsevych, Psychological and pedagogical dictionary, Minsk, 2006.

[25] V. H. Kazakov, New times - new technologies of professional training, Professional education (2006) 12.

[26] S. Tishkovskaya, G. A. Lancaster, Statistical Education in the 21 st Century: A Review of Challenges, Teaching Innovations and Strategies for Reform, Journal of Statistics Education 20 (2012) 4.

doi:10.1080/10691898.2012.11889641.

[27] L. Kulinenko, Technologies of innovative educational space, Naukovyi chasopys Natsionalnoho pedahohichnoho universytetu imeni M. P. Drahomanova. Seriia 07. Relihiieznavstvo. Kulturolohiia. Filosofiia (2013). URL: http://enpuir.npu.edu.ua/handle/123456789/12492.

[28] M. Pavlenko, L. Pavlenko, Formation of communication and teamwork skills of future IT-specialists using project technology, Journal of Physics: Conference Series 1840 (2021) 012031. doi:10.1088/1742-6596/1840/1/012031.

[29] E. P. Il'in, Human motives: theory and methods of study, High school, 1998. URL: https://www.elibrary.ru/item.asp?id=21748410.