Czech Technical University in Prague

Faculty of Electrical Engineering

Department of Computer Science

Master`s Thesis

STROKE MORTALITY PREDICTION

Regina Mavrina

Supervisor: **Ing. Matěj Klíma**

Study Program: Open Informatics

Field of Study: Software Engineering

August 14, 2020

MASTER‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Mavrina Regina** Personal ID number: **492137**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Computer Science**

Study program: **Open Informatics**

Specialisation: **Software Engineering**

II. Master’s thesis details

Master’s thesis title in English:

Stroke Mortality Prediction

Master’s thesis title in Czech:

Predikce úrtnosti na infarkt

Guidelines:

The aim of the work is to study a differentiated data set, their comparison and analysis, in order to implement a more accurate forecast of mortality among the population. The task set before me can be solved by many methods (Random Forest, Bayes method, confidence intervals, Central limit theorem, Student criterion and test, Fisher criterion and test, Folding knife, bootstrap, null and alternative hypothesis, statistical power). My work uses a comparative analysis of various approaches and solution methods to improve statistics, to determine the best of them.

Bibliography / sources:

[1] I.S. Shorokhova, N.V. Kislyak, O.S. Mariev, Statistical Methods of Analysis, Ekaterinburg: Ural University Press, 2015.300 s [2] Pankov A., Goryainova E. R., Zhernosek A. I., Statistical methods of data processing, Moscow: Moscow Aviation Institute, 2013.382 s

Name and workplace of master’s thesis supervisor:

Ing. Matěj Klíma, Software Testing Intelligent Lab, FEE

Name and workplace of second master’s thesis supervisor or consultant:

Date of master’s thesis assignment: **12.03.2020** Deadline for master's thesis submission: **14.08.2020**

Assignment valid until: **19.02.2022**

Ing. Matěj Klíma Supervisor’s signature	Head of department’s signature	prof. Mgr. Petr Páta, Ph.D. Dean’s signature

III. Assignment receipt

The student acknowledges that the master’s thesis is an individual work. The student must produce her thesis without the assistance of others, with the exception of provided consultations. Within the master’s thesis, the author must state the names of consultants and include a list of references.

.

Date of assignment receipt	Student’s signature

Acknowledgements

I express my deep gratitude for the help in preparation of the diploma work to my supervisors, Ing. Matěj Klíma (CVUT), Ing. Artem Zaikin (KFU), and Galim Vakhitov (KFU). My heartfelt thanks to my family and friends for their support and assistance. Allow me to express my sincere gratitude to my reviewers for their work, recommendations and comments, also Ekaterina Turilova, Arslan Enikeev, Miroslav Bureš and all Czech team for the opportunity to participate in the Double Degree program (CVUT, KFU).

Declaration

I hereby declare that I have completed this thesis independently and that I have listed all the literature and publications used.

I have no objection to usage of this work in compliance with the act §60 Zákon č. 121/2000Sb. (copyright law), and with the rights connected with the copyright act including the changes in the act.

Prague, 14 August, 2020 ____________________________

ABSTRACT

MAVRINA, Regina: Stroke Mortality Prediction. [Master's Thesis] – Czech Technical University in Prague. Faculty of Electrical Engineering, Department of Computer Science. Supervisor: Ing. Matěj Klíma.

Relevance of the master's thesis:

Hundreds of thousands of strokes and pre-stroke conditions are reported each year in civilized and rapidly growing economies. The cardiovascular system of the body provides continuous blood circulation in the human body. It is the system that provides life. If we take into account the world trend of mortality in medicine, cardiovascular diseases are steadily leading, along with oncology. However, even though the optimization of the health care system is present and developing every year, the infrastructure of many medical institutions is an outdated system without proper quality. The main focus of work is to find the best method for predicting stroke with minimal error, which is also of interest to health professionals.

Based on the obtained open data on the Kaggle web platform, a systematic sampling and statistical analysis of indicators will be carried out to universalize and optimize the process of predicting stroke mortality, as well as an attempt to determine the relationship of COVID-19, as well as the influencing factors that somehow or other link it to the development of stroke. Statistical analysis methods are used to predict the data. The most accurate prediction of the available data depends largely on factors. Such as the study of the data obtained in the past, methods that most accurately determine a particular disease, the timing and cost of research, etc. In medical institutions, based on the available differentiating data, to reduce the frequency of stroke and identify pre-stroke conditions in patients, a comparative analysis is carried out for different groups of diseases.

Analysis and specification using mathematical tests show good results. The test results are predictive data that, based on the results, help predict and prevent a future stroke. The developed software interacts with the databases that give the probability of stroke. Statistical analysis of the data, based on the methods used in this thesis, during the decomposition of primary or secondary symptomatology, can help achieve more accurate predicted data that can directly affect the number of deaths.

The goal of the master's thesis:

Within the framework of existing software products and various methods for processing existing data, using the statistical data of the public web platform of open data Kaggle, search for methods for the structurization of symptoms and key data in the study of pre-stroke conditions to universalize and optimize the process, to predict stroke following the specific data.

The tasks of the master's thesis:

1. Carry out an analysis of scientific sources.
2. Develop and design mechanisms to work on stroke prediction, based on mathematical methods and models.
3. Get the results of comparable data and indicators preceding the stroke using different methods of solutions and choose the most optimal method of prediction.
4. Compare, analyze and evaluate the accuracy of the results obtained in the best decision method.

Objects of the research:

Determining the best mathematical method for predicting stroke based on pre-disease data.

Subjects of the research:

The use of probability theory methods, error estimation, systematic sampling, as well as the estimation of reliability and accuracy to the predicted mortality.

The scientific novelty of research:

1. Stroke prediction was made using the Bayes method with control of the expected proportion of false positives.
2. The relationship of COVID-19 with the onset of stroke was determined.

Keywords: statistics, stroke, covid, predict, coronavirus.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Based on the open data obtained on the Kaggle web platform, the objective was to perform a systematic sampling and statistical analysis of the indicators to universalize and optimize the forecasting of stroke mortality.

Statistical analysis methods are used to forecast the data. The most accurate prediction of the available data depends largely on factors. Such as the study of the data obtained in the past, the methods that most accurately determine this or that disease, the terms and costs of research, etc.

Predictions for symptom detection and current treatment may also vary in timing. They can be current, i.e. determining the disease and treatment in the process of detection here and now, determining and fixing for treatment the following incoming patients with the same symptoms and long-term, i.e. fixed, based on of available data, for a certain degree of repeated symptoms and course of the disease.

However, it is also necessary to take into account the fact that the result of the prognosis will be more accurate if the period time from detection to the receipt of indicators, for the beginning of appropriate treatment, is minimal.

In health facilities, a comparative analysis of different disease groups is carried out, based on available differentiation data, to reduce stroke and identify pre-stroke conditions in patients. Statistical analysis of the data, based on the methods used in this paper, in the decomposition of primary or secondary symptomatology can help to achieve more accurate predicted data that can directly affect the number of deaths.

1. Theoretical aspects, main problems and approaches

This chapter analyzed the data that affect the development of stroke. The data used in this chapter are publicly available and are currently the subject of extensive research. What is new in this paper is that it explores the relationship of the stroke to COVID-19. The results showed that the virus is very closely related to stroke, causing inflammation and blockage of blood vessels, disrupting heart function, disturbing its rhythm.

1.1 Basic information on medical data analysis, collection and processing

As part of this work, various databases will be analyzed to extract and obtain concrete and comparative results.

Data analysis - the field of mathematics and computer science, engaged in building and researching the most common mathematical methods and computational algorithms to extract knowledge from experimental (in the broad sense) data and indicators; the process of research, filtration, transformation and modeling, the studied and studied material to extract useful information and decision-making. Data analysis has many aspects and approaches and covers different methods in different fields of science and activities. [1]

Available information based on medical aggregated data and indicators is a good basis for research, as each patient has a medical record that shows and records his or her past, actual diseases, and in fact, current health status at the time.

Projections for symptom detection and current treatment may also vary in timing. They can be current, i.e. determining the disease and treatment in the process of detection here and now, determining and fixing for treatment the following incoming patients with the same symptoms and long-term, i.e. fixed, based on available data, for some degree of repeated symptoms and course of the disease.

For the prognosis result to be accurate, it is necessary that the period of time from detection to the receipt of indicators is minimal.

Analysis of medical indications, current symptomatology, predicted conditions and appropriate treatment is a study of the totality of all the data obtained, which provides a forecast for early detection and prevention of possible diseases in the future.

The study of medical data is based primarily on the following:

- Research Concept
- Data collection and preparation

- Analysis
- Interpretation of results obtained
- Conclusion based on data obtained

Since time immemorial, science such as medicine has worked and is working to find the right methods and solutions to improve diagnosis as well as treatment.

To achieve results, the methods used in applied observational-based statistics allow problem-solving. Statistics, in this regard, is very often interpreted as the object language of describing the use of reality in sciences that use data that are not subject to rigoric formalization. [9]

The first step determining the statistical analysis itself is first of all the study of the type of parameters, the main of which are quantitative and qualitative. Which, in turn, are ranked by indicators.

Types of medical parameters (see Figure 1.1):

1. Medical parametrs:

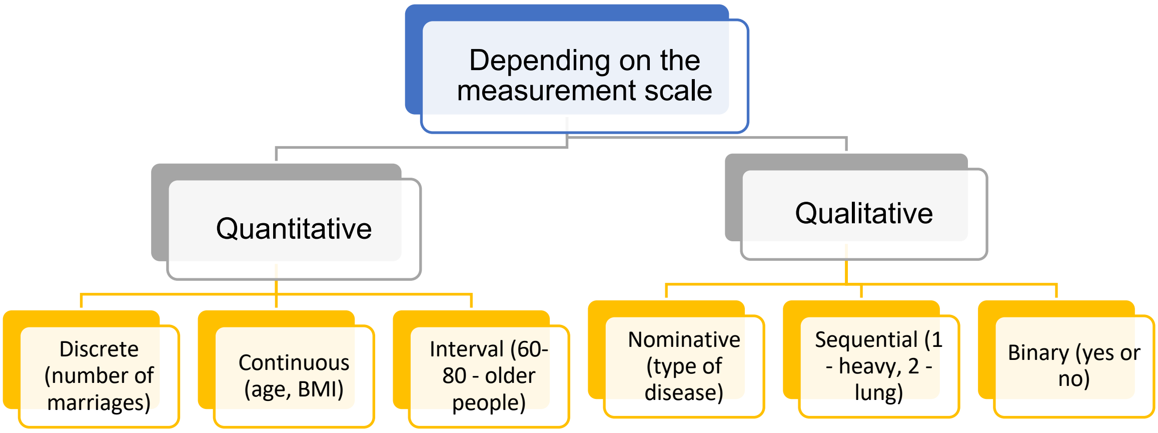

Depending on the measurement scale
Quantitative
Qualitative
Discrete (number of marriages)
Continuous (age, BMI)
Interval (60-80 - older people)
Nominative (type of disease)
Sequential (1 - heavy, 2 - lung)
Binary (yes or no)

Figure 1.1. Types of parameters

2. Time parameters:

- Dynamic - indicators that change over time. For example, electrocardiography. Electrocardiography - is a method of research and recording the electrical activity of the heart. The result of electrocardiography is the electrocardiogram (ECG) - a graphical representation of the difference in electrical potentials arising from the heart and projected on the body surface. [27]
- Statistical - indicators that do not change in time. For example, an X-ray. X-ray - examination of the internal structure of objects that are projected by X-rays on special film or paper. [28]

3. Depends on the object of research:

- Patient - his vital signs.
- Population - health indicators of the population.

The materials and data for the research were taken from a public web platform in the form of open statistics for Kaggle data processors and machine training engineers (see Table 1).

Table 1 - Data for the research [37]

id	gender	age	hypertension	heart_disease	never_married	glucose_level
36306	Male	80	0	0	Yes	83,84
61829	Female	74	0	1	Yes	179,5
14152	Female	14	0	0	No	95,16
12997	Male	28	0	0	No	94,76
40801	Female	63	0	0	Yes	83,57
9348	Female	63	1	0	Yes	219,98

However, some problems often arise when analyzing medical data. And most of them require resources and time to solve and prevent.

Real medical data is not publicly available and I used test data in my work to do research. Because patient numbers are constantly changing with the course of the disease, for more accurate and detailed statistics it is necessary to update and correct them periodically.

A fairly large data set should be used to predict stroke mortality in the population. The processing of such information should always be done on a computer with sufficient memory. It is necessary to take into account the fact that the number of deaths from strokes does not decrease with each year. This leads to the conclusion that the infrastructure of many medical institutions is outdated and of poor quality.

In countries where income levels are significantly high, systems exist to collect information on death and its causes. In countries where income levels are considered medium and low, such systems either do not exist or information on deaths is provided with incomplete or no specific cause data.

Improving reporting and statistics on a specific number of deaths and their associated data is essential to the health system to the health of its citizens to reduce or prevent deaths from specific causes or diseases in these countries.

1.2 Review of existing software products for medical data processing

Medical data analysis software makes it easier and more accurate to predict disease based on concomitant symptoms. Let's consider the most popular programs and libraries that help to predict the probability approach of a stroke and to build schedules based on the available data.

The development environment:

- RStudio - the free environment of software development with open source code for programming language R which is intended for statistical data processing and work with graphics. It is easy to learn and more convenient to use than a standard graphics shell for R. [11]
- PyCharm - is an integrated development environment for the Python programming language. It provides code analysis tools, a graphical debugger, a tool to run unit tests, and supports web development on Django. PyCharm was developed by JetBrains based on IntelliJ IDEA. [24]

Libraries [29]:

- RandomForest - is a library that builds the decision tree on different data packages and also averages for better results.
- Cowplot - a library for building charts.
- AUC - the library for calculation of ROC/AUC estimation (see Figure 1.2).

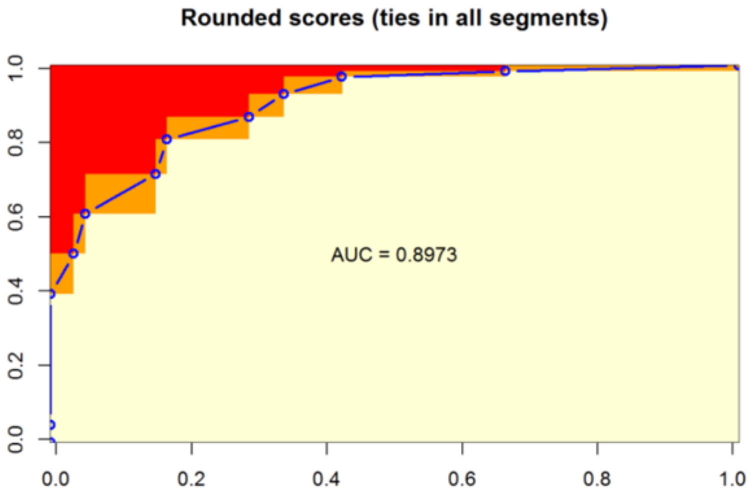


Figure 1.2. ROC curve

- Bootstrap - is a library for obtaining different types of confidence intervals of initial loading.
- Binom - a library. The model of binomial distribution deals with the search for the probability of success of an event that has only two possible results in a series of experiments. For example, flipping a coin always gives an eagle or a tailpiece (see Figure 1.3).

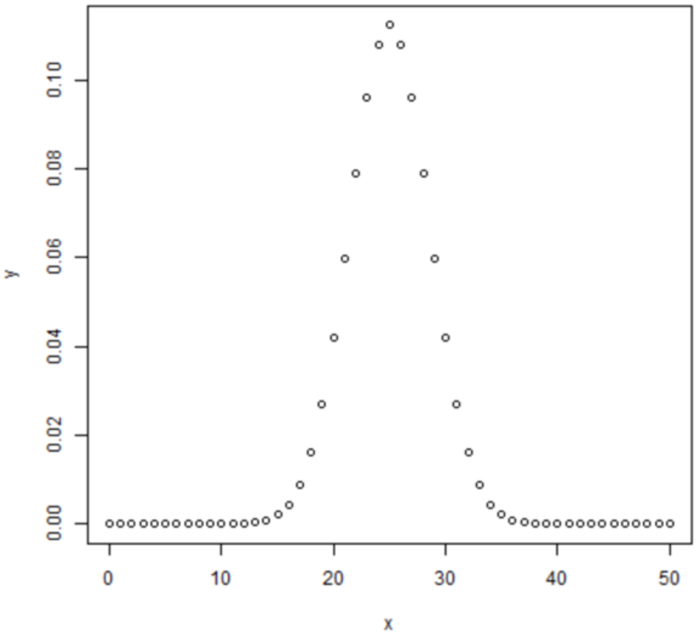


Figure 1.3. Binomial distribution

- Infer - The purpose of this package is to execute the output using expressive statistical grammar, which is consistent with a neat design structure.

1.3 Impacting factors and data research

The system of statistics on the number of deaths and their causes is one of the most important ways to assess the effectiveness and direction of health infrastructure measures in the country.

According to the World Health Organization, the 10 leading causes of death in the world are classified implicitly (see Figure 1.4) [30].

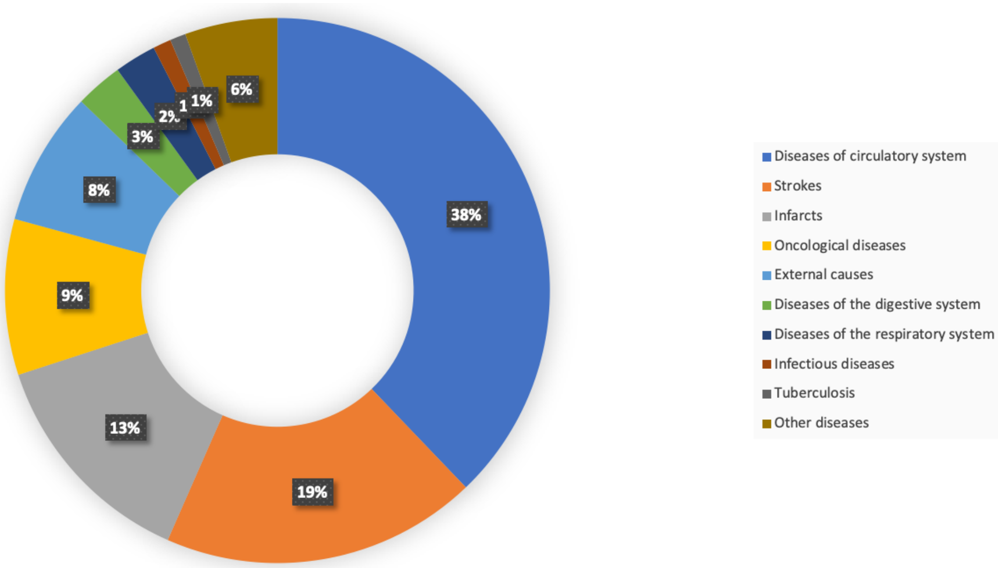


Figure 1.4. The leading causes of death in the world

You can see from the chart that coronary heart disease and stroke take the leading positions from the list.

We will compare the statistics on the number of deaths from stroke in Europe and Russia per 100,000 inhabitants in 2019, according to the data provided by the Organization for Economic Cooperation and Development in the free access report Health at a Glance 2019 (see Figure 1.5). By mortality from strokes, Russia has 234 deaths per 100 thousand inhabitants [5]

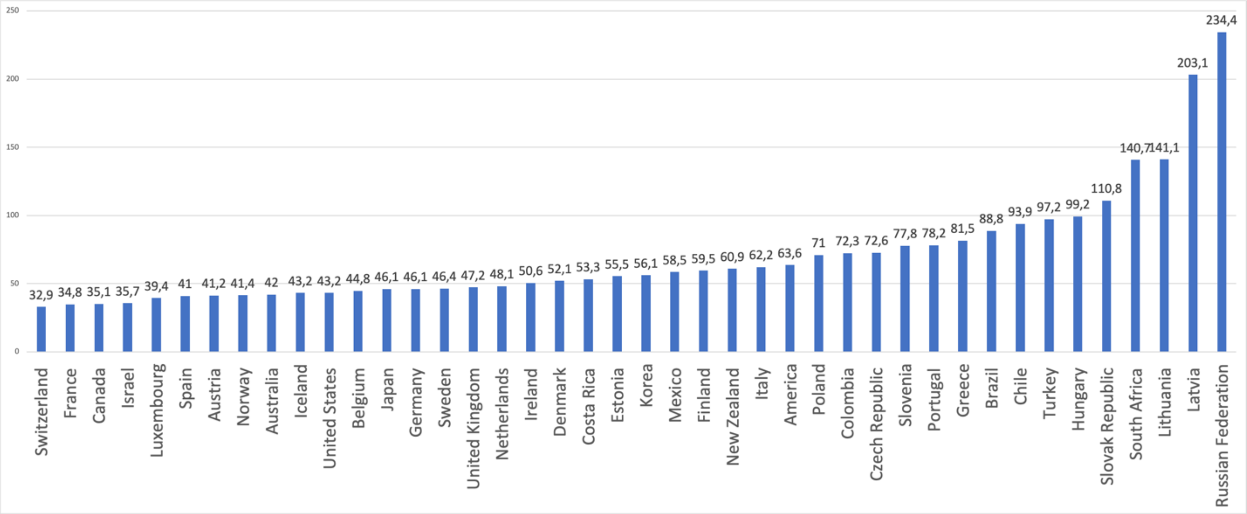


Figure 1.5. Statistics of deaths in the world

This graph clearly shows that Russia is the leading European country in terms of stroke deaths (see Figure 1.6). By mortality from strokes, Russia has 281 deaths per 100 thousand inhabitants in 2015, 214 deaths per 100 thousand inhabitants in 2016, 274.9 deaths per 100 thousand inhabitants in 2017, 259.3 deaths per 100 thousand inhabitants in 2018, and 234 deaths per 100 thousand inhabitants in 2019.

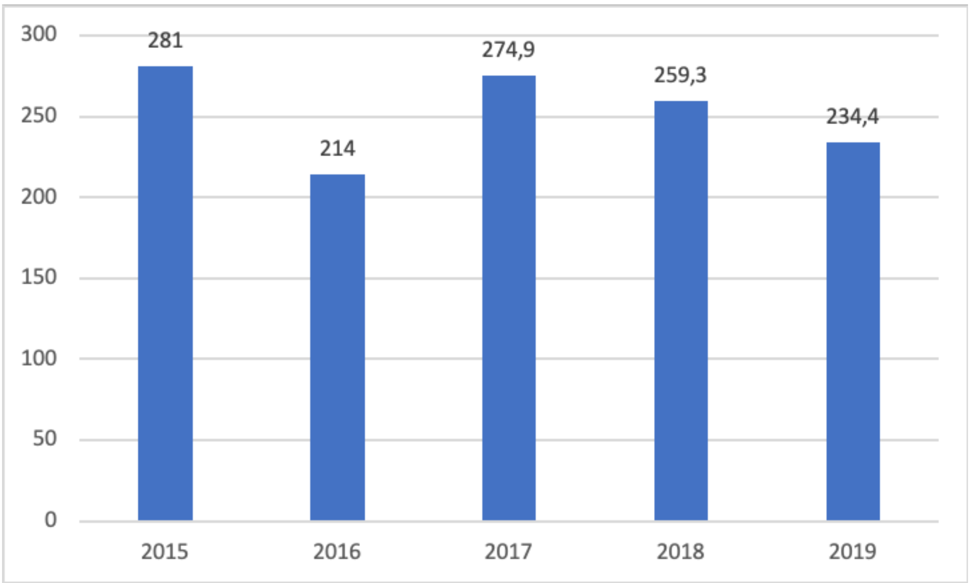


Figure 1.6. Statistics for 5 years

Having analyzed the statistics for 5 years per 100 thousand inhabitants, also, based on the data of the OECD report [21], we can see only a slight decline in the number of deaths from stroke.

Based on comparative statistics and disappointing indicators, the purpose of this thesis is to find methods to structure the symptoms and key data in the study of pre-stroke conditions to universalize and optimize the process, to predict stroke based on the specific data.

The source data are open statistical databases from the Kaggle website. The method of research data analysis (EDA) will be used as visual perception - a system of decomposition of a set of statistical data that combines their basic characteristics, often with visual methods, highlighting the main and useful aspects and decision-making. EDA is primarily designed to see what the data can tell us, beyond the formal task of modeling or hypothesis testing. [22]

Risk factors:

Previously, 10 of the world's leading deaths were identified and labeled, but the risk factors affecting future stroke should also be identified. If symptomatology occurs against these factors, the risk of stroke increases. These factors are referred to [31]:

- Inheritance.
- Abuse of alcohol.
- Smoking.
- Obesity.
- Unhealthy diet.

- Hypercholesterolemia.[1]
- Hypertension.
- Low physical activity.
- Age.

In the risk group in this case, according to world statistics and practice, are men over 40 (+/-) and women 55 (+). Today, however, more and more strokes are being recorded in 25-30-year-old young people worldwide. In the American medical journal Annals of Neurology [20] published in 2011, there was an article comparing the number of hospitalized young people with ischemic stroke aged 5-14, 15-34, and over 40 years. As the results of the study in the group of 15 to 40(+) years of age show, the number of patients increased by 30%.

However, even though the age of stroke is decreasing, the percentage of younger generation diseases of the total number of patients and those who have suffered from it is on average 10%, which is not a sufficient criterion for entering them into the study statistics.
Let's analyze several main factors against which the symptoms of the main causes of the risk of stroke in the future appear.

Variables in the data set:

1. Family status and BMI:

Do people tend to gain extra weight in marriage? Overweight and obesity tend to cause the excessive formation of fat deposits in the body, which in turn is harmful to human health. The Body Mass Index (BMI) is an indicator to determine whether a person's body weight is insufficient, corresponds to the norm, or is excessive and is calculated as a ratio of body weight to height.

BMI is a sufficient measure of body weight because it is calculated equally for both sexes, regardless of age and position.

According to WHO statistics [33], such conclusions about whether a person is overweight or obese are based on diagnostics:

- BMI above or equal to 25 is considered redundant.
- BMI is above or equal to 30 obesity.

The main reason for the weight gain of any category of citizens, of any age, is essentially excessive caloric intake of the diet, which is much higher than the needs of the body, consumption of food with a high content of fat, low physical activity, sedentary lifestyle, and so on.

- [1] Hypercholesterolemia - is an increase in blood cholesterol (a fat-like substance). [32]

Excess BMI, in turn, indicating the increased formation and deposition of fat in the body or already, as in the cycle of obesity, is not only an increased risk factor affecting the emergence of diseases such as cardiovascular disease (heart disease, atherosclerosis, thrombosis, hypercholesterolemia, stroke, etc.).), respiratory organs, nervous system, diabetes, oncology, disorders of the musculoskeletal system, but always affects all human organs as a whole, making their changes in their work, thus disrupting their usual rhythm. [33]

It is believed that married couples tend to gain extra weight. In the process of living together, partners often begin to adopt interests, lifestyles, sometimes quite sedentary, and hobbies of each other. Such interest can be and eating habits of the partner. However, there may also be such psychological techniques as no stress, the onset of confidence in stability and security.

Let us compare the indicators of people who have been or are being married to people who have never been married (see Table 2).

The average BMI for people who have ever been married is 30.9.

Table 2 – Overweight

Ever Married	Median	Average	Variance	STD
Yes	28,9	30,9	52,1	6,6
No	22,7	27,5	56,8	8,1

A histogram was constructed, which showed that people tend to gain extra weight in marriage (see Figure 1.7).

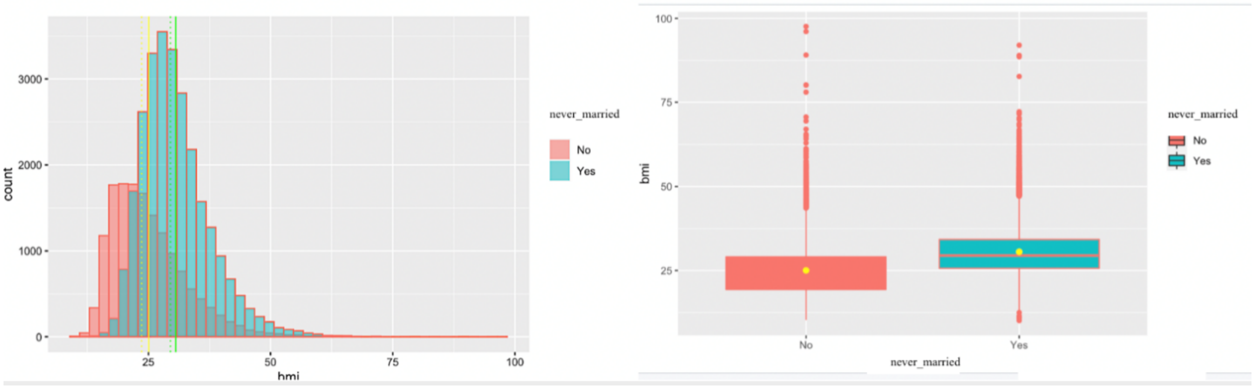


Figure 1.7. Overweight histogram

2. Age

The incidence of stroke increases with age. The risk of stroke is many times greater for older people than for others.

A histogram was constructed showing that people tend to stroke at ages 60-80 (see Figure 1.8).

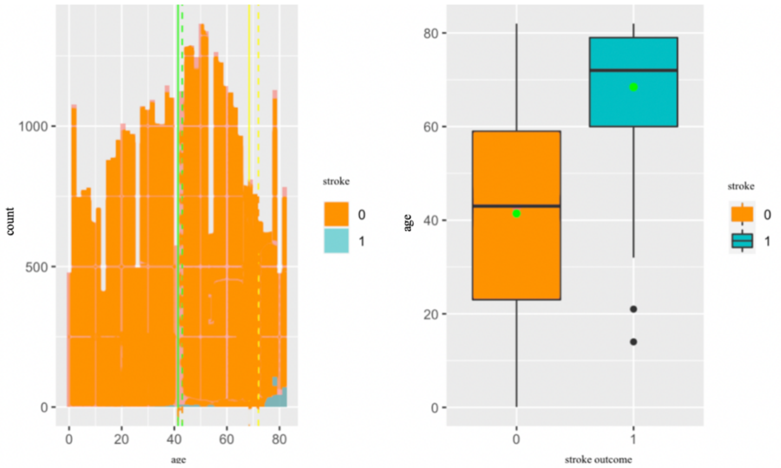


Figure 1.8. Histogram of age

3. Gender

A histogram was constructed which showed that stroke affects men before women. But according to statistics, 25% of men and 39% of women die from this disease. (see Figure 1.9)

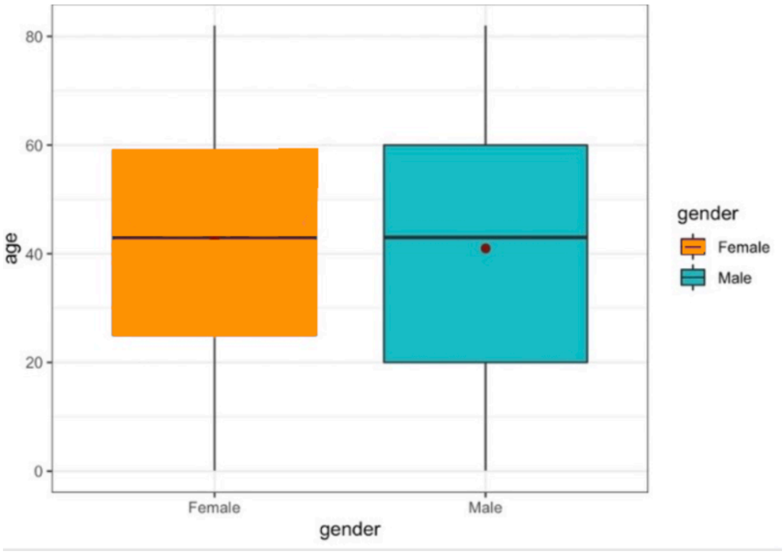


Figure 1.9. Gender histogram

4. Blood glucose levels

In the absence of diabetes, blood glucose levels range up to 140 mg/dL, in type 1 diabetes about 90-162 mg/dL, in type 2 diabetes about 90-153 mg/dL.

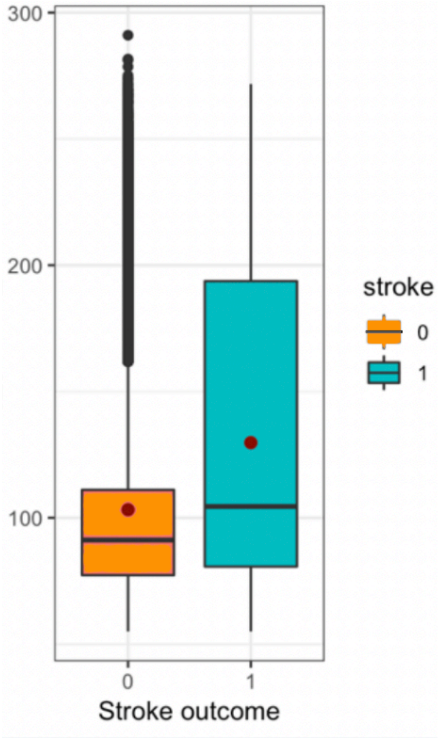


Figure 1.10. Histogram of glucose level

This histogram shows that people who have no stroke have normal blood glucose levels. People who have a stroke have elevated blood glucose levels.

The histogram shows not the best option for predicting the disease, as few people with a positive stroke result, to explore the correct distribution - it is impossible.

1.4 Possible relationship between COVID - 19 and stroke based on available data and statistics

The world has officially declared a pandemic against the background of the spread of COVID-19, an acute respiratory infection that affects all human organs during the disease caused by coronavirus SARS-CoV-2 (2019-nCoV).

Coronaviridae - a family of viruses currently consisting of 40 species and 2 subspecies and affecting both humans and animals. The name is formed in connection with the structure of the virus itself, the branches of which resemble a crown. Among the coronaviruses that affect humans in particular:

- HCoV-229E, an alphacoronavirus, first detected in the mid-1960s;
- HCoV-NL63, an alphacoronavirus, the pathogen was detected in the Netherlands in 2004;
- HCoV-OC43, betacoronavirus A, the agent was detected in 1967;
- HCoV-HKU1 - betacoronavirus A, the agent was detected in Hong Kong in 2005;
- SARS-CoV, betacoronavirus B, the causative agent of SARS pneumonia, the first case of which was registered in 2002;
- MERS-CoV, betacoronavirus C, a pathogen of Middle Eastern respiratory syndrome, which erupted in 2015;
- SARS-CoV-2, betacoronavirus B, responsible for a new type of pandemic pneumonia in 2020. [16]

Already a lot of months have passed since the SARS-CoV-2 virus began to spread, and scientists from all over the world still know very little about its impact on the body. The completeness of the picture of the disease and the course of the disease is composed of hundreds of articles in scientific journals, where doctors share their experience of symptomatology and treatment of patients.

The list of possible symptoms and related complications is constantly changing. According to WHO, most people with the virus, which is almost 80%, have sluggish symptoms. And only in one out of six cases does it develop into severe symptomatology with respiratory failure. Many medical articles indicate that the infection may affect the circulatory system and even the brain.[6]

The purpose of this section of the paper is to determine whether stroke can actually be a Covid-19 consequence and what complications in the human body caused by it can directly affect the signs of a stroke. Let's take as a basis and consider such common diseases affecting the appearance of stroke as lung disease, coronary heart disease, stroke experience, arrhythmia, hypertension, heart failure.

Calculations have been made that show how Covid-19 interacts with other diseases.

The data from Kaggle "covid_19" were taken for analysis. This data was uploaded on April 19, 2020, by an Italian analyst who had data on covid_19 worldwide, I was looking at Russia [38]. In table region - is the place of the population of the patient, pathology - disease the patient, value_of_covid - whether the patient was sick with coronavirus, code - disease code (see Table 3):

Table 3 – Data of Covid_19 [38]

ID	Region	Pathology	Time	Value_of_Covid	Code
1	Moscow	Stroke	2017	0	378
2	Moscow	Cardiopathic	2018	1	37
3	Spb	Stroke	2018	0	378
4	Kazan	Arteriosa	2019	0	21
5	Moscow	Cardiac	2016	1	532
6	Spb	Stroke	2019	1	378

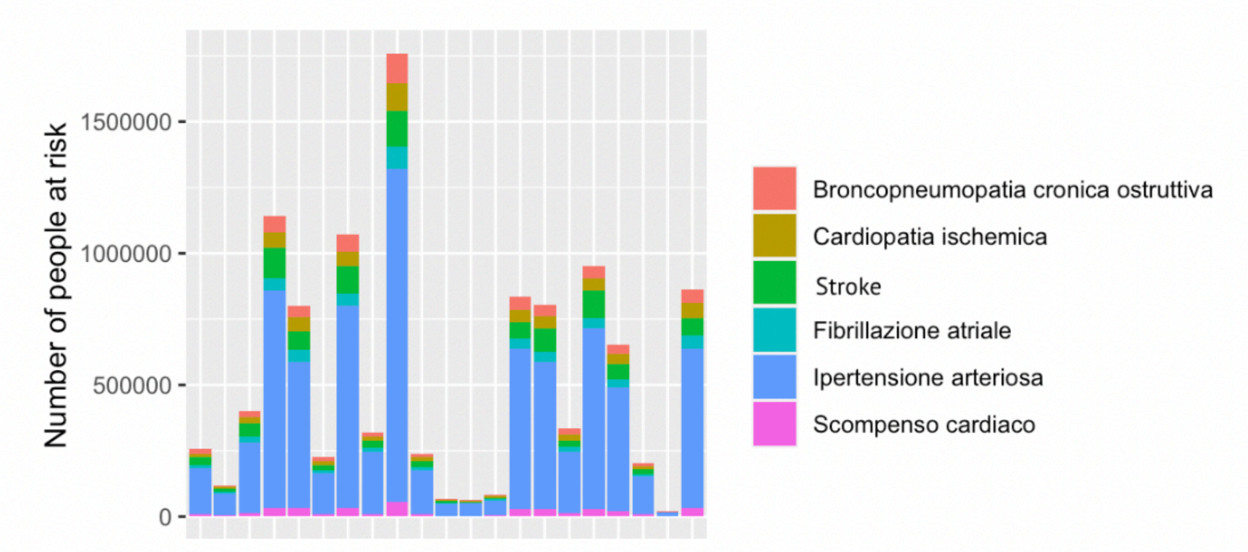


Figure 1.11. Death pathology from Covid_19

The resulting graph shows (see Figure 1.11) that hypertension, against the background of the virus, is the highest risk. Stroke is the second risk factor in this category.

The symptomatology and regional distribution of the disease have been considered: Moscow, St. Petersburg, Kazan.

In Moscow, the percentage of pathologies among the population is dominated by hypertension, with stroke coming second (see Figure 1.12).

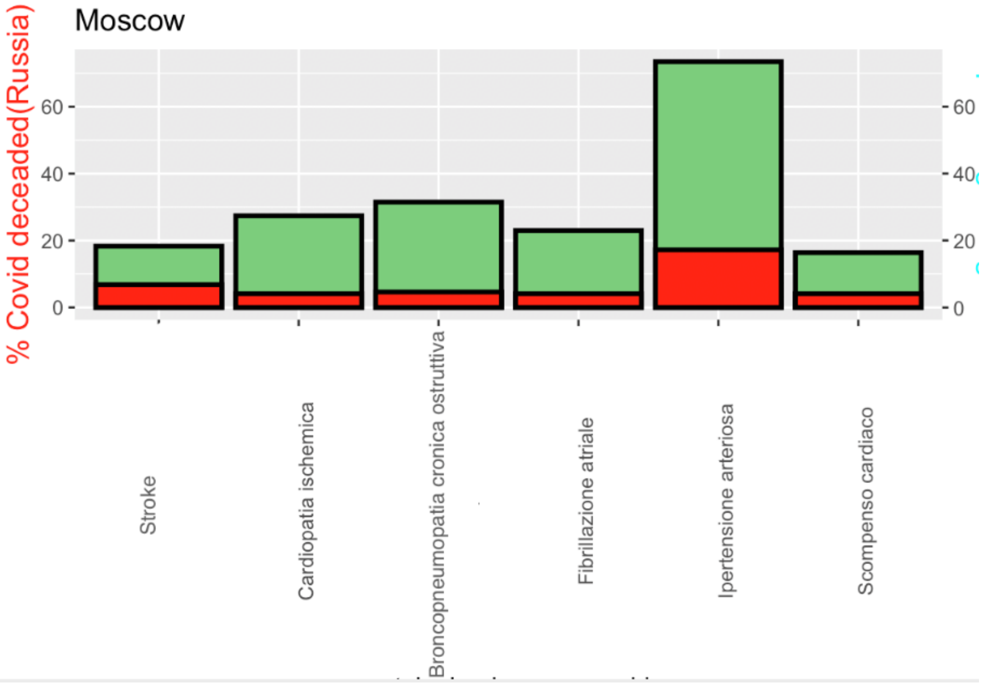


Figure 1.12 The result in the Moscow

In Kazan, the percentage of pathologies among the population is dominated by hypertension, followed by stroke (see Figure 1.13).

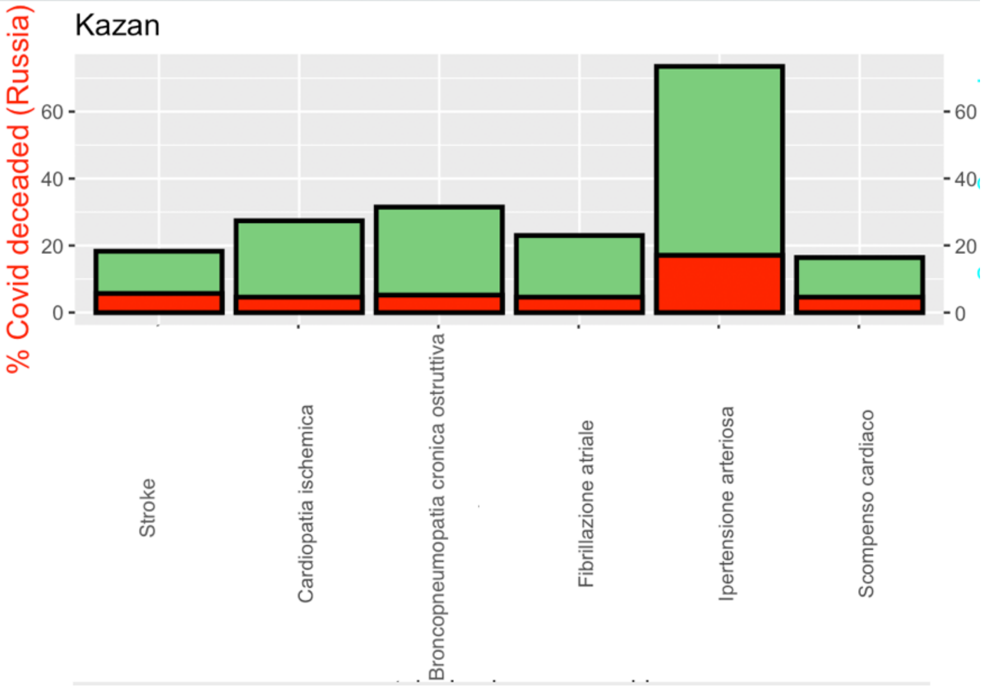


Figure 1.13 The result in the Kazan

In St. Petersburg, the percentage of pathologies among the population is dominated by hypertension, followed by stroke (see Figure 1.14).

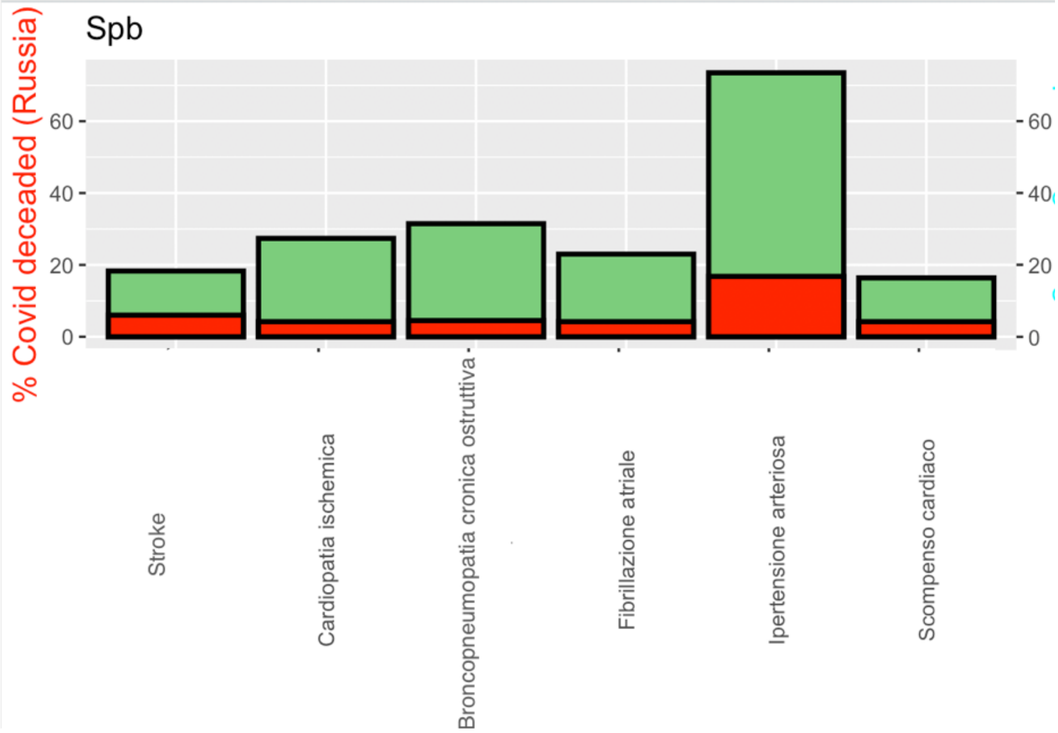


Figure 1.14 The result in the St. Petersburg

Covid_19, according to Neurosurgeon A. Kashcheeva, can cause serious changes in the entire blood system [18] and, against the background of coronavirus pneumonia, as the infection primarily originates as a viral infection affecting the human respiratory system, almost every third has thrombotic complications, which directly affects the formation of clots in large vessels and, as a consequence, affects the normal operation of almost any organ, as well as directly leads to a disturbance of cerebral blood circulation. Which, in turn, can lead to a stroke.

As can be seen from the results of tests, the virus is very closely related to the high load, which causes inflammation and blockage of blood vessels, disrupting the heart, disrupting its rhythm, leading to various myocarditis, arrhythmias, and related diseases.

Cardiovascular diseases and risk factors associated with them should be taken under special control, carefully regulated, and follow scientifically sound recommendations for remission or appropriate treatment.

2 Development of the software project

This chapter describes the analysis of requirements, which includes the collection of requirements for the software (software), their systematization, identification of relationships. The stages of activity with the help of the waterfall life cycle model are considered, and also the obtained results of calculations are tested.

2.1 Requirements analysis

Requirements analysis is a part of the software development process that includes the collection of software requirements, systematization, identification of relationships, and documentation. In the process of collection, it was important to take into account possible contradictions of the requirements of different stakeholders. [34]

The completeness and quality of the requirements analysis played a key role in the success of my project. Requirements to the software to be documented, executable, tested, with a level of detail sufficient to design the system.

Analysis of the requirements includes three types of activities:

- Collection of requirements - the analysis of the subject area was conducted, the requirements for stroke progression were defined.
- Requirements analysis - data processing was performed to ensure that the requirements were clear, complete; requirements interrelation was identified.
- Documentation of the requirements - simple description, scenarios for use.
- Requirements analysis can be a long and difficult process, involving many subtle psychological skills. New systems are changing the environment and relationships between people, so it is important to identify all stakeholders, take into account all their needs, and ensure that they understand the meaning of new systems.

Analysts can use several methods to identify the following requirements from a client: conducting interviews, using focus groups, or creating lists of requirements. More advanced methods include prototyping and scenario building. Where necessary, the analyst will use a combination of these methods to identify the exact requirements of stakeholders so that a system that meets business needs is created.

The process of analyzing information system requirements includes the following phases:

- Requirements development
- Identification of requirements
- Requirements analysis
- Specification of requirements
- Verification of requirements
- Requirements management

2.2 Description of a software project development

In this thesis, a cascade model of software development was chosen. Since in the cascade model each development stage corresponding to the software life cycle stage continues the previous one. That is, to move to a new stage, we must completely complete the current one. The structure of the cascade model (see Figure 2.1):

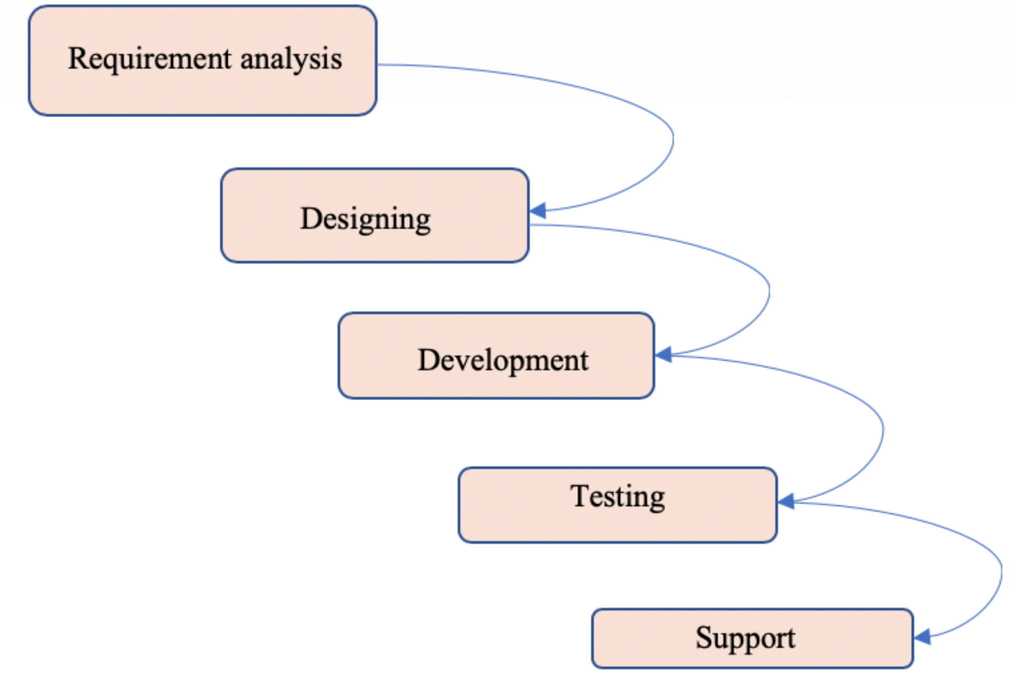


Figure 2.1. Waterfall model

In the thesis, the task is to determine the best mathematical method for predicting stroke.

At the system design stage, the development environment and programming language were chosen. The programming language R was chosen, and the development was carried out in RStudio.

At the development stage, all mathematical methods were developed and a structural approach to programming was used.

Testing was carried out on the test database of Kaggle. Mathematical methods were implemented independently, and ready-made functions were launched in RStudio. Test software demonstrates the correctness and accuracy of the algorithm and software in general.

Architecture**:**

This thesis uses monolithic architecture. A monolithic application is an application delivered through a single deployment (see Figure 2.2):

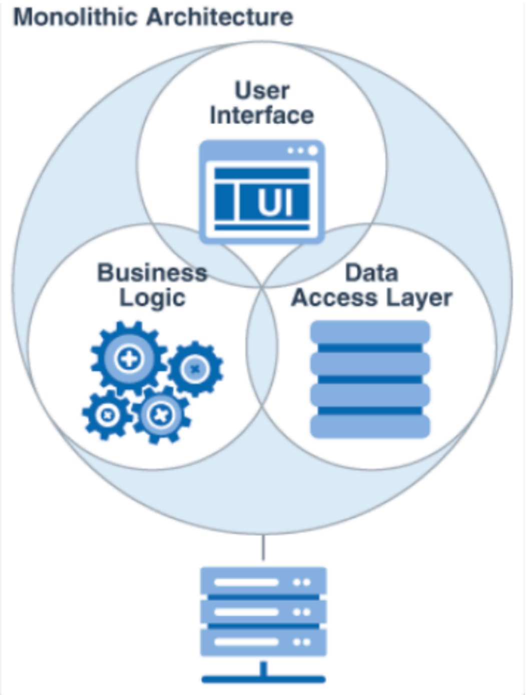


Figure 2.2. Monolithic Architecture [35]

The big advantage of the monolith is that it is easier to implement. In a monolithic architecture, you can quickly start implementing your business logic instead of wasting time thinking about interprocess communication.

Another thing is the end-to-end (E2E) tests. In a monolithic architecture, they are easier to perform.

2.3 Testing of the created program

Testing the program - checking the correspondence between the real and expected behavior of the program, carried out on the final set of tests chosen in a certain way. In a broader sense, testing is one of the quality control techniques that include Test Management, Test Design, Test Execution, and Test Analysis activities.

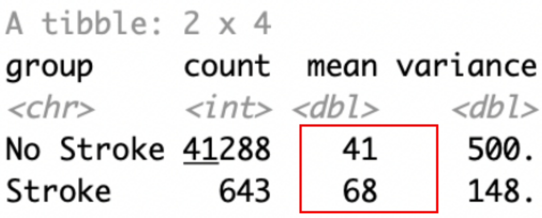
My work was testing functions, one of which was written by me, and the second function is built into RStudio. The essence of this test was to get the same values in the output. Mathematical methods were tested: T-test, F-test, null hypothesis. For each mathematical method (written and built-in) the same data were fed to the input.

Let's look at the T-test:

This test tested whether the age difference is significant for people who have already had a stroke and those who are at risk of experiencing it in the future. Let's take a look at a piece of code that I wrote.

```
library(dplyr)
group_by(mrtvice_Data, group) %>%
      summarise(
            count = n(stari),
            mean = mean(stari),
            sd = sd(stari)
            )
```

The result was as follows (see Figure 2.3):



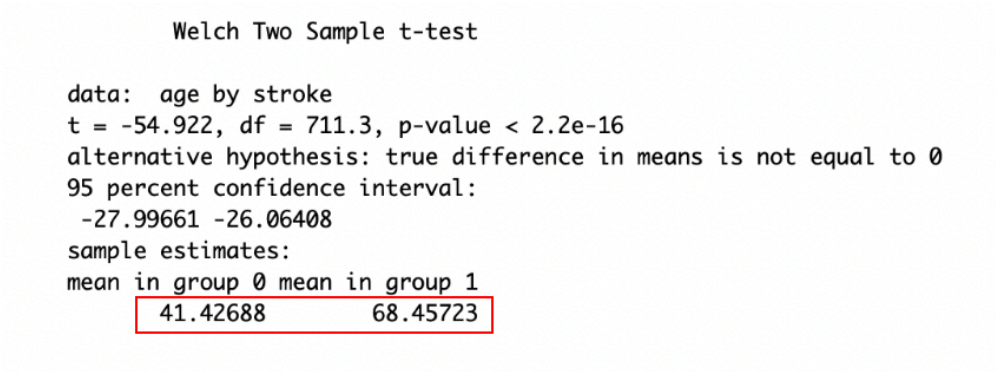
```
A tibble: 2 x 4
group      count  mean variance
<chr>      <int> <dbl>    <dbl>
No Stroke  41288    41     500.
Stroke       643    68     148.
```

Figure 2.3. Result T-test

Let's consider the result of the built-in function in RSudio

```
t.test(stari~stroke,
       data=mrtvice_Data,
       var.equal = FALSE,
       paired = FALSE,
       conf.level = 0.95) [26]
```

The result was as follows (see Figure 2.4):



```
	Welch Two Sample t-test

data:  age by stroke
t = -54.922, df = 711.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -27.99661 -26.06408
sample estimates:
mean in group 0 mean in group 1
       41.42688        68.45723
```

Figure 2.4. The result of the integrated function (t-test)

The results are the same, it means that the program code written by me works correctly.

Let's look at the null hypothesis:

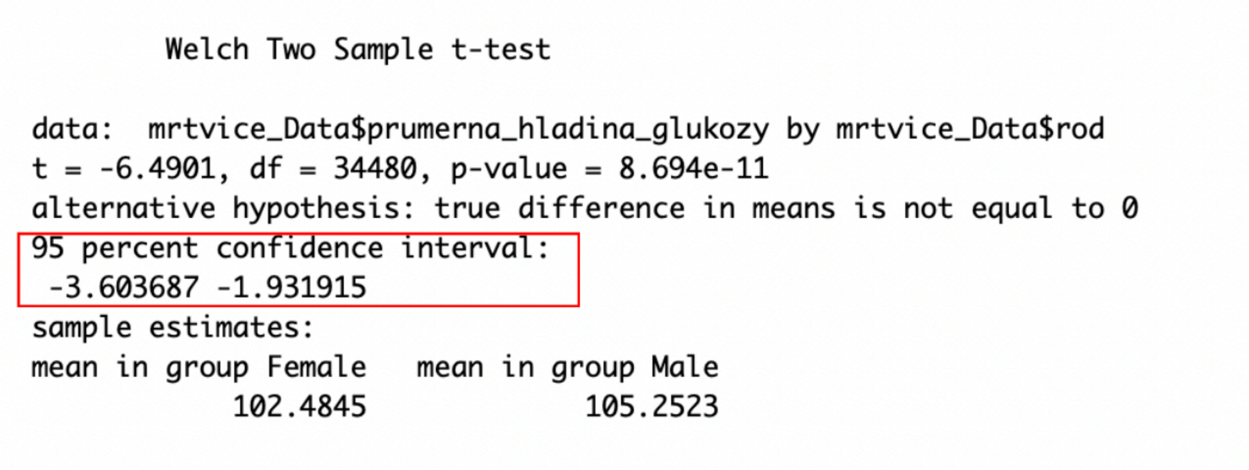
95% CI for difference in mean of average glucose level

```
round(estimate + c(-1, 1) * qnorm(.975) * se, 2)
```

The result was:

1. 1.931915
2. 3.603687

Let's consider the result of the built-in function in RSudio



```
	Welch Two Sample t-test

data:  mrtvice_Data$prumerna_hladina_glukozy by mrtvice_Data$rod
t = -6.4901, df = 34480, p-value = 8.694e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.603687 -1.931915
sample estimates:
mean in group Female   mean in group Male
            102.4845             105.2523
```

Figure 2.5. The result of the integrated function (null hypothesis)

The results are the same, it means that the program code written by me works correctly.

Let's look at the F-test:

This test tested whether the age difference is significant for people who have already had a stroke and those who are at risk of experiencing it in the future. Let's take a look at a piece of code that I wrote.

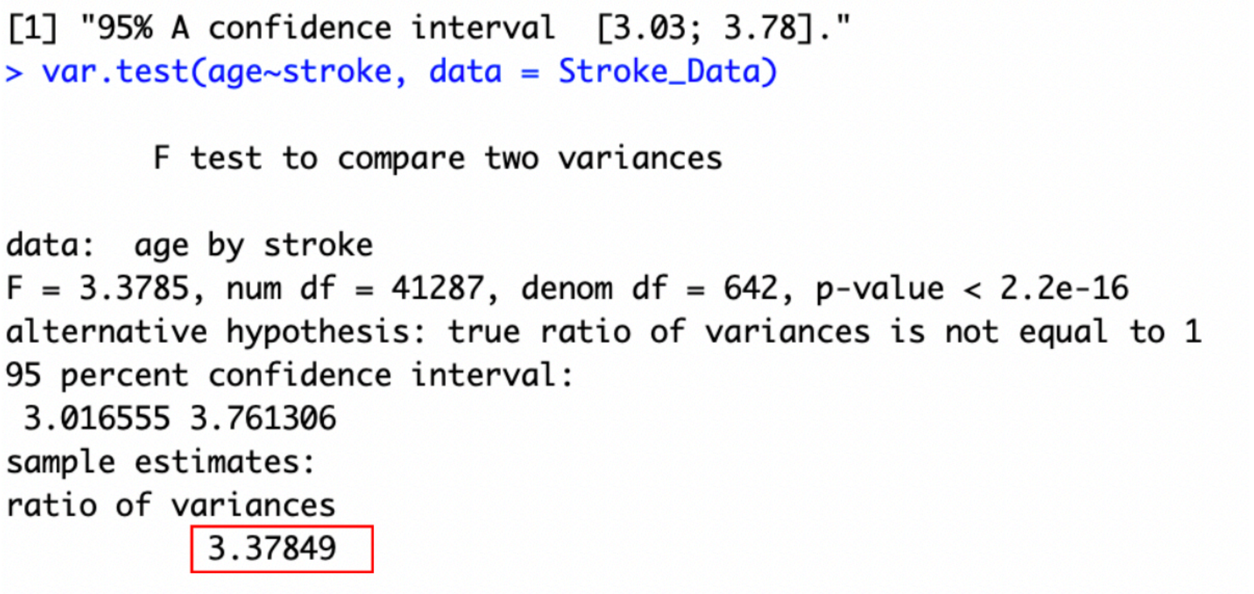
```
library(dplyr)
mutate(mrtvice_Data, group) %>%
     when (
                (mrtvice == 0) ~ “No stroke”,
                (mrtvice == 1) ~ “Stroke”,
           )
     summarise(
           count = n(stari),
           mean = mean(stari),
           sd = sd(stari)
           )
```

The result was:

```
F-test: 3.38
```

Let's consider the result of the built-in function in RSudio

```
f.test(stari~stroke,
       data= mrtvice_Data,
       var.equal = FALSE,
       paired = FALSE,
       conf.level = 0.95
```
[26]



```
[1] "95% A confidence interval  [3.03; 3.78]."
> var.test(age~stroke, data = Stroke_Data)

        F test to compare two variances

data:  age by stroke
F = 3.3785, num df = 41287, denom df = 642, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 3.016555 3.761306
sample estimates:
ratio of variances
          3.37849
```

Figure 2.6. The result of the integrated function (f-test)

The results are the same, it means that the program code written by me works correctly.

3 Practical aspects and identification of secondary factors influencing the onset of stroke

This chapter analyzes the auxiliary variables that affect the development of stroke. Stroke can be prevented if secondary factors are identified early. The average BMI of the population is found using a 95% confidence interval, the Central Limit Theorem is used, the Bootstrap method is used, and the Zero Hypothesis and Alternative Hypothesis are tested for a set of variables that influence the development of stroke.

3.1 Using graphical methods

The value of the graphical method in data analysis and synthesis is great. Graphic image first of all allows us to control the reliability of statistical indicators, because, presented on the graph, they more clearly show the existing inaccuracies associated with either the presence of observation errors or the essence of the phenomenon under study. With the help of the graphical image, I have studied the regularities of stroke development, the establishment of the existing links with other diseases and indicators. A simple comparison of the data does not always make it possible to catch the presence of causal dependencies, at the same time, their graphic representation helps to identify the causal relationships, especially in the case of initial hypotheses to be further developed. Graphs are widely used to study the structure of phenomena, their changes in time, and placement in space. The more expressively show comparative characteristics and distinctly identify the main development trends and relationships inherent in the phenomenon or process under study.

Nowadays, graphs have firmly entered the practice of medical activity in connection with the introduction of new mathematical methods and modern computer technology in the statistical work, using the packages of applied programs of computer graphics. These programs facilitate the task of a researcher in the practical application of graphics, if necessary, quickly changing some data in them, entering others, etc.

All my graphs are different in appearance, and the task is to find the most appropriate graphical image that shows the earliest detection of stroke in the population, which will help prevent deaths from stroke.

3.2 Finding the average BMI of the population using a 95 percent confidence interval

Stroke is the leading cause of acquired permanent disability in adults worldwide and the second-largest cause of death in patients over 60 years of age. The World Health Organization (WHO) estimates that one new patient worldwide suffers from stroke every 2 seconds. Every 6 seconds a person dies or remains disabled as a result of a stroke. Stroke causes 5.8 million deaths per year, more than all deaths from AIDS, tuberculosis, and malaria combined. Obesity has also reached epidemic proportions around the world. Given that obesity is an independent predictor of ischemic stroke that affects patients of all ages, overweight patients are a growing group of candidates for stroke. [4]

A confidence interval is an interval built using random sampling from a distribution with an unknown parameter, such that it contains this parameter with a given probability. [4]

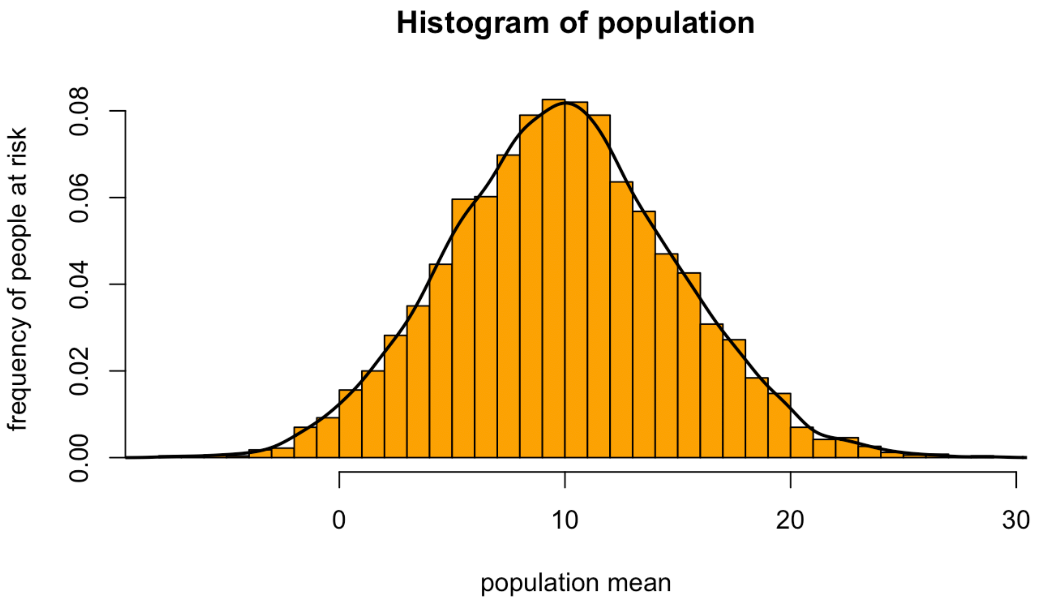
Construction of a confidence interval for the mathematical expectation of the general population with a known standard deviation. [5]

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}} \ (1),$$

where Z - is the value of a standardized normally distributed random value corresponding to an integral probability equal to 1 - α/2, σ - is the standard deviation of the general population, $\bar{x}$ - sample average, n - scope. [5]

Let us consider an example where we know the distribution of the population at risk of the total population. We show the average value of the population and the standard error (see Figure 3.1):

- Population mean: 10
- Population standard error: 5



–

Figure 3.1. Histogram of the population

Let us consider an example where we know the distribution of the population at risk of the total population. We show the average value of the population and the standard error at n = 250:

- Population mean: 10.2
- Population standard error: 4.9

A histogram has been constructed that shows the average of the population: (see Figure 3.2):

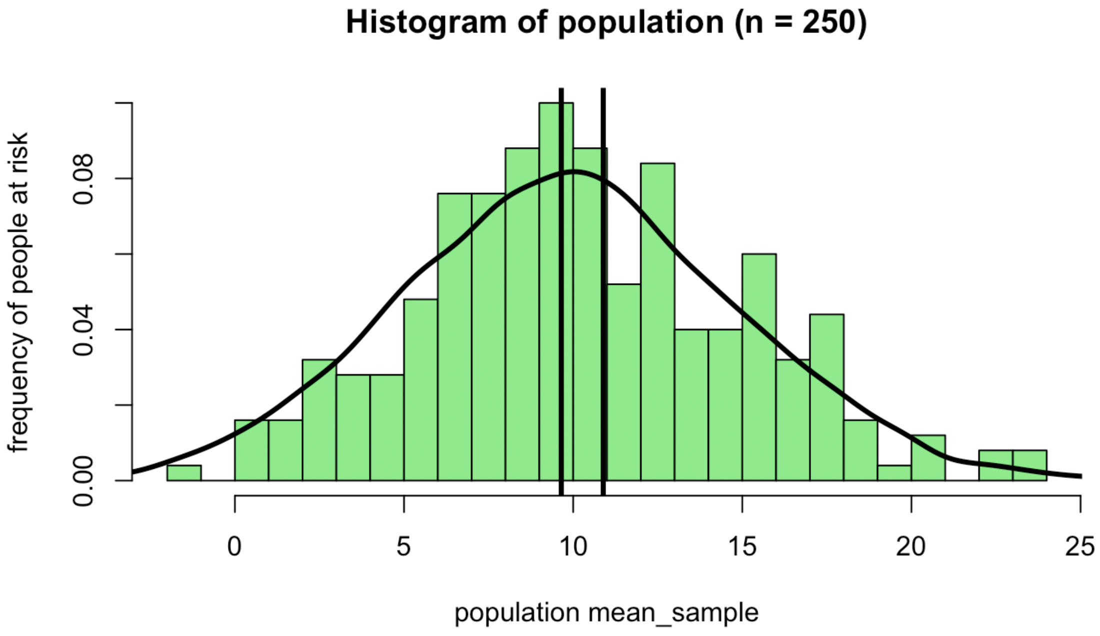


Figure 3.2. Histogram of the population (n = 250)

According to the results of the histogram, we can say that the average number of the population at risk is in the range from 9.4 to 10.6.

The confidence interval for mean value using t-distribution

If the data underlying the population is distributed abnormally and/or the total dispersion (population variance) is unknown, the average sample value depends on the t-distribution.

It provides a wider range than the normal distribution because it takes into account the additional uncertainty introduced by evaluating the standard deviation of the population and/or because of the small sample size. [11]

$$\left(\bar{x} \pm t_{n-1,\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right)(2),$$

Based on the sample, a confidence interval was found for the mean value of the human body mass index.

According to the results of the histogram, we can say that the average value of BMI of the population is in the range from 28.4 to 28.8 - it shows the tendency of excess weight, which further looks at the development of stroke. Where the upper limit of normal weight is 24.9. This is the benchmark (see Figure 3.3).

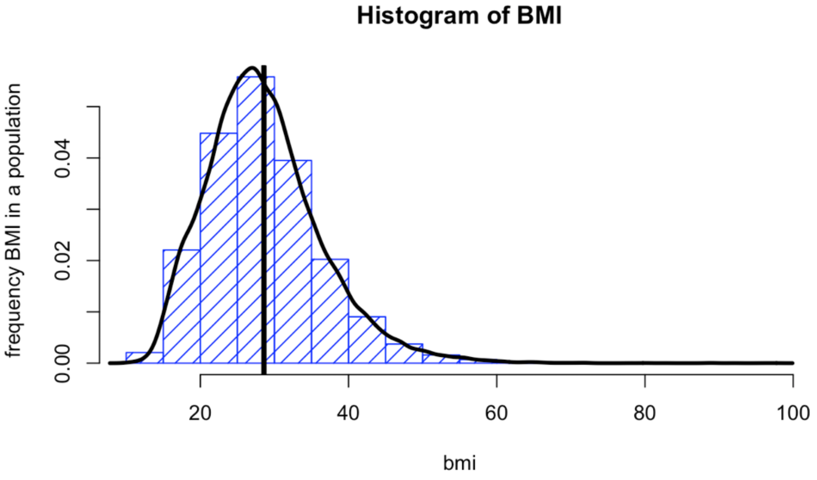


Figure 3.3. Average BMI value

The CI was very narrow. This is due to the large sample size.

Based on a sample “stroke” of 150 people, a confidence interval for the mean human mass index (BMI) was found.

Based on the results of the histogram, we can say that the average BMI of the population is in the range from 28.4 to 31.6. Where the upper boundary of obesity BMI is > 29.5 (see Figure 3.4).

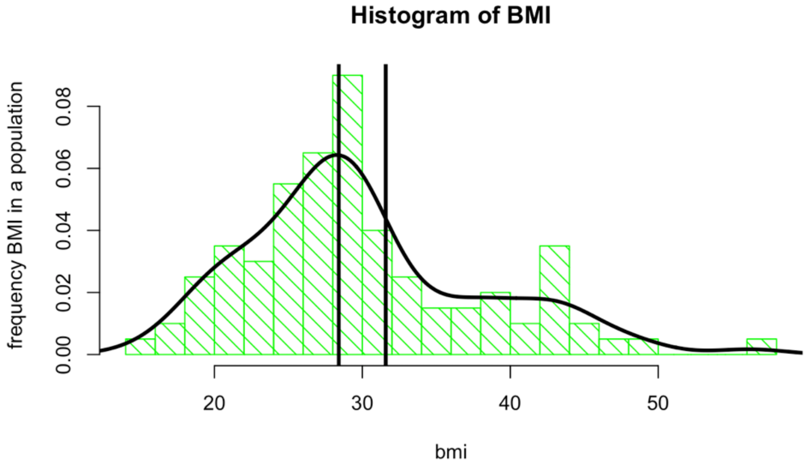


Figure 3.4. Average BMI value (n = 100)

3.3 Identification of BMI in the population influencing the development of stroke using the central limit theorem

CLT - a theorem in probability theory that states that the sum of a sufficiently large number of weakly dependent random variables, having approximately the same scale (none of the components dominates, does not contribute to the sum of the defining contribution), has a distribution close to normal. [15]

$$\sqrt{n}\ \frac{\overline{X_n}-\mu}{\sigma} \rightarrow N(0,1)\ (3),$$

where is $\overline{X_n}$ a sample average, μ is a mathematical expectation, σ is a dispersion.

Let us consider the central limit theorem, which will show with what probability the results of the experiment are close to the true goal. As the sample size becomes larger, the average sample BMI values are distributed evenly among the population.

The central limit theorem applies to the distribution of values of all random samples. The distribution of all outcomes converges to those normally distributed as the number of responses increases according to the central limit theorem.

The biggest problem is that each sample must be independent of all samples before and after. It should also have the same distribution.

Let us look at the distribution of BMI for people who have never had a stroke (see Figure 3.5). Average BMI in the population: 28.6

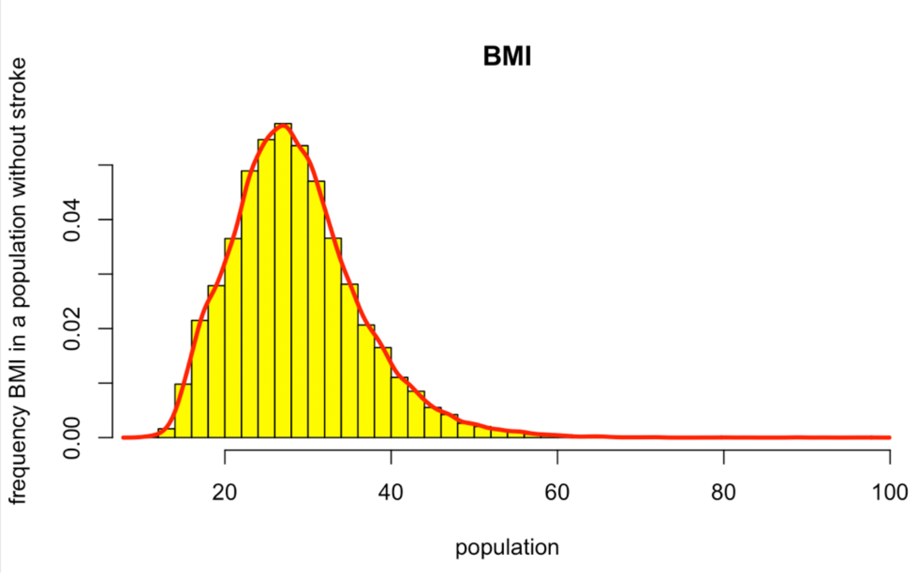


Figure 3.5. Average BMI in the population

The larger the sample, the more it will strive for normal distribution of values.

Let's look at the average value of the sample n = 40:

Average population BMI value: 31.6, which suggests a tendency towards overweight (see Figure 3.6).

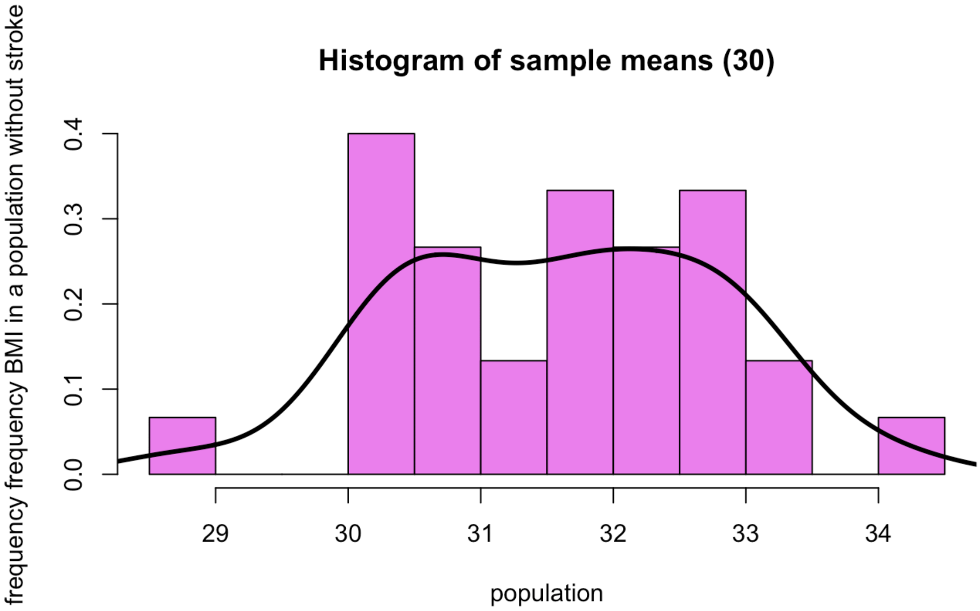


Figure 3.6. The average value (n = 30)

Let's look at the average value of the sample n = 300:

Average population BMI value: 28.1, which implies an upper bound of the value aspiring to obesity (see Figure 3.7).

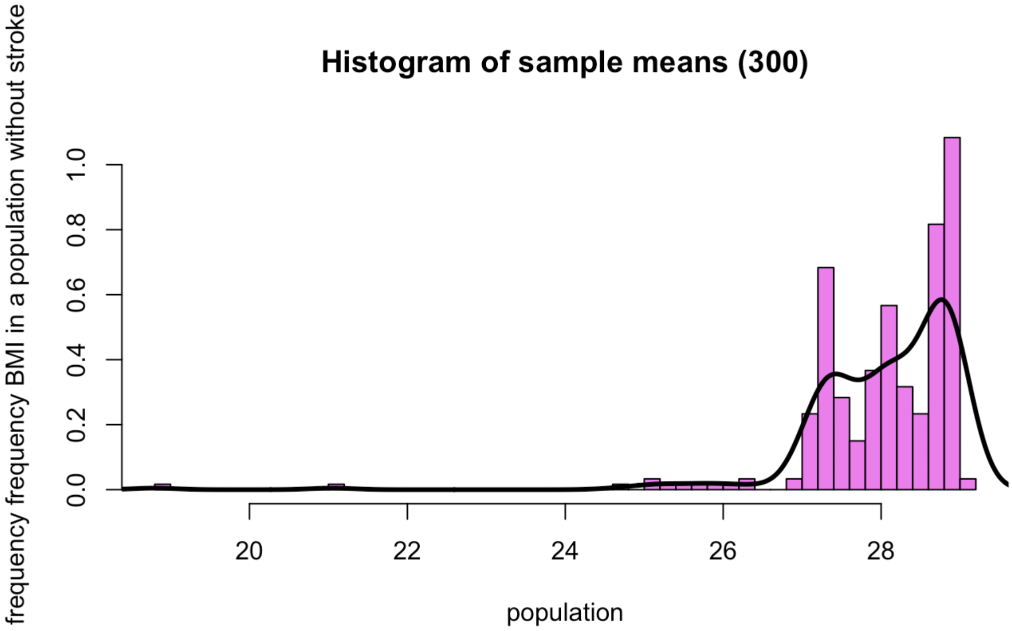


Figure 3.7. The average value (n = 300)

Thus, we can see that the distribution of sample averages looks more likely to be normal as the sample size increases.

3.4 Using the Bootstrap method to determine the average BMI value of a population

Bootstrap is a practical computer-based method for studying the distribution of probability distributions statistics based on multiple Monte Carlo sampling generation based on an available sample. It allows us to easily and quickly estimate a variety of statistics (confidence intervals, variance, correlation, etc.) for complex models. [3]

The idea of initial sampling is to use the results of sample calculations as a "fictitious set" to determine the statistical distribution of the sample. It analyses a large number of "phantom" samples called bootstraps. The bootstrap is randomly selected with a return, the selected elements in the original sample are returned to the sample and can be selected again. With the bootstrap we do not get any new information but use the available data sensibly based on the task at hand. Bootstrap is best used for small samples, for median estimates, correlations, confidence intervals, and other situations.

$$y_i = \theta_{x_i} + \epsilon_i \; (4),$$

where θ_{x_i} - parameter estimation, ϵ_i - empirical distribution function

Bootstrap is mainly used as a method to evaluate statistical precision, i.e., SE, offsets, and CI. This thesis uses the observed value of test statistics and parameter estimation as the best guess with the true value of an unknown parameter or statistic. For example, if we are interested in estimating the average BMI value of the population, it may seem that the best estimate of the average BMI value of the population is the average of all bootstrap estimates. The result showed that this is not the case, as the average of all bootstrap estimates is biased. The baseline average observed in the sample is always the best estimate of the population's average BMI. The same result applies to other statistics such as median and regression coefficients (see Figure 3.8). The bandwidth is a measure of how close you want the density to match the distribution. Bandwidth the smoothing bandwidth to be used. The kernels are scaled such that this is the standard deviation of the smoothing kernel.

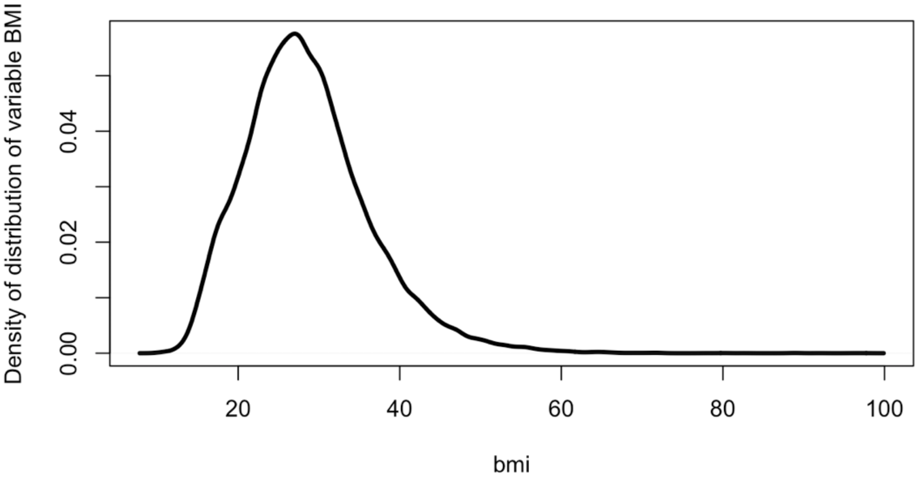


Figure 3.8. Pop_ average BMI value of the population

Both charts look the same as it should be, because the right chart was effectively created by simulating from the left chart (see Figure 3.9).

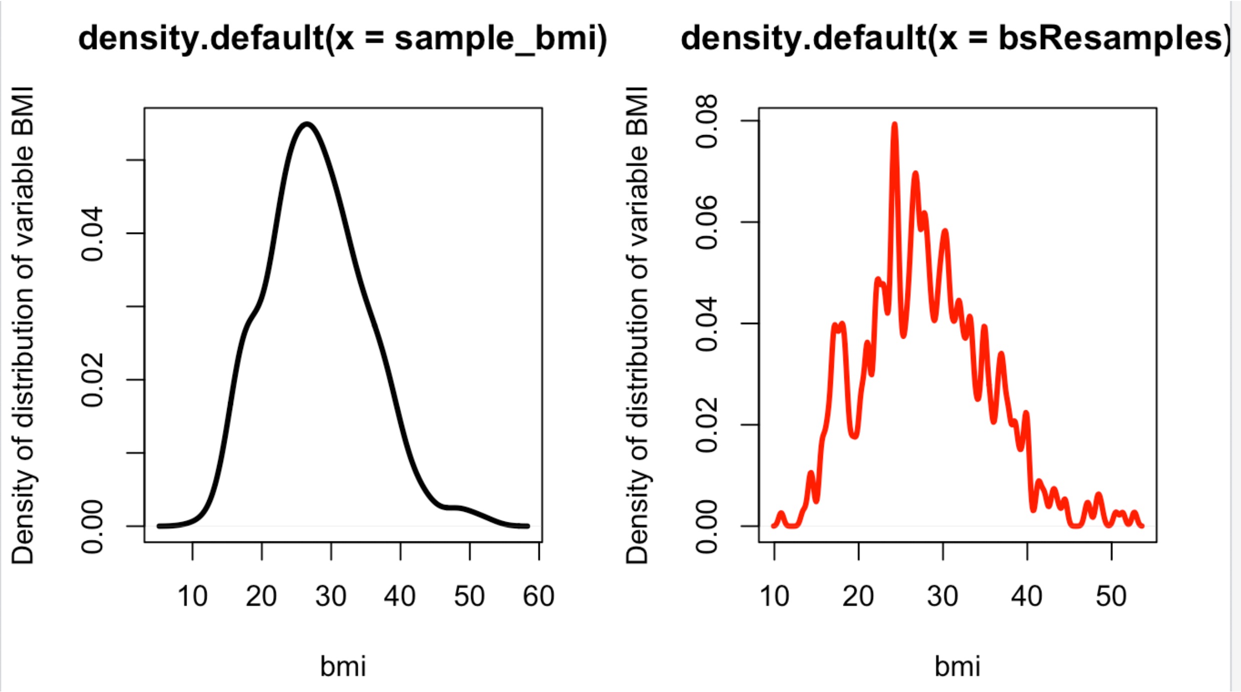


Figure 3.9. Re-sample Pop_ average BMI value of the population

- Calculation of average value for each recalculated data
- Building a histogram with a modified sample
- Standard error: 0.348

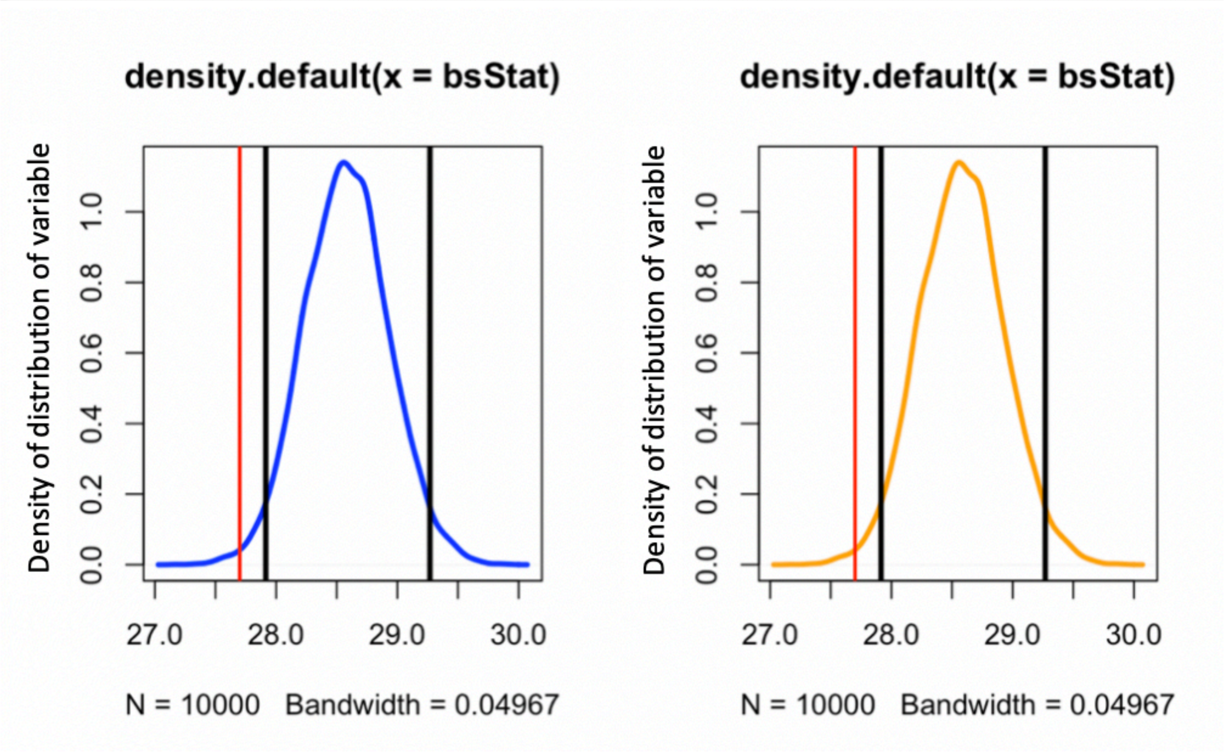


Figure 3.10. Charts with a modified sample

With the built-in function in R, the following result was obtained (see Figure 3.11):

- Standard error: 0.342

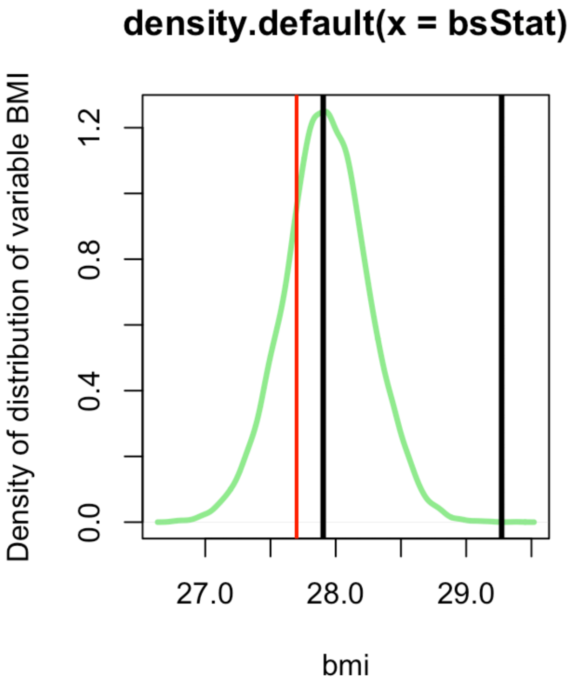


Figure 3.11. The bootstrap function

At repeated sampling, a false result was given because the sample was small and had some deviations. When the code was run frequently, you could see (Figure 3.9) that the average value of the re-sample proportions was different from the average population BMI.

3.5 Testing the significance of the Zero and Alternative Hypothesis for a set of variables affecting stroke development

The testing of hypotheses is one of the most fundamental methods in statistics. For example, we want to know the answer to the question: "Is the prevalence of the disease higher among men than women?

The most popular method, called the "Zero Hypothesis", was considered. According to this method, we will always declare one hypothesis as a status hypothesis, also called the Zero Hypothesis, H0 - there is no evidence to suggest that the prevalence of the disease is different by sex. The alternative hypothesis is usually referred to as Na, and that is exactly the hypothesis we want to test.

The calculations are made under the assumption that the H0 hypothesis is correct, and evidence is needed to reject the H0 hypothesis in favor of the Alternative. There can be 4 different types of outcomes:

- If H_0 is right, and we couldn't reject $H_{0,}$ we took zero correctly;
- If H_0 is true, and we rejected H_0, then we made a Type I error;
- If H_a is true, and we rejected H_0, we correctly rejected zero;
- if H_a is true, and we rejected H_0, we made an error of type I;
- if H_a is true, and we rejected H_0, we correctly rejected zero;
- If H_a is correct, and we couldn't reject H_0, we made a Type II error.

Zero hypotheses:

Testing whether the average glucose level in men and women is different.

- Z score result = $\frac{\sqrt{n}(\bar{x}-\mu 0)}{S} = 6.47$ (5)
- qnorm = 1,96 > quintile

Table 4 – Result: mean glucose level

Gender	Average	STD	N
Male	105	44	17000
Female	102	41	25000

We see that the resulting test statistics are much larger than the 0.975 normal distribution quantile (qnorm = 1.96 > 0.975), so we reject the null hypothesis.

One-sided test for one mean. We believe that the average BMI (BMI) for the population exceeds 30 (regardless of gender).

- Z score result = $\frac{\sqrt{n}(\bar{x}-\mu 0)}{S} =$-36.76 (5)
- qnorm = 1,64 > quintile
- Average = 28,7
- STD = 7,8
- N = 42000

The calculated score is much lower than the 0.95 normal distribution quantile (qnorm = 1.64 > 0.95), so we do not have enough evidence to reject the null hypothesis.

Connection to CI:

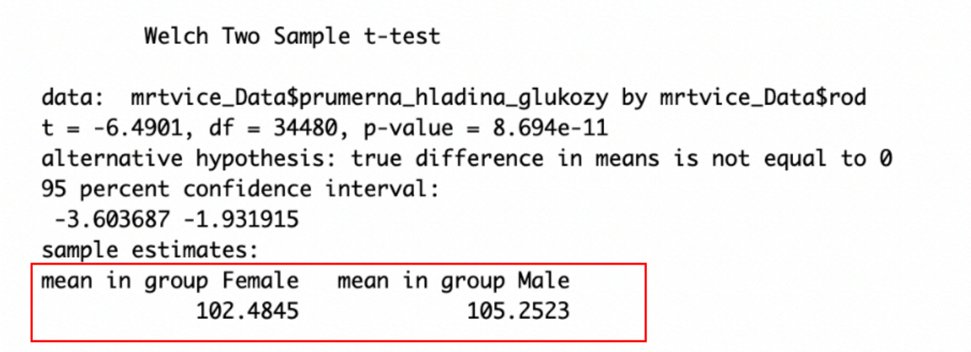
This interval does not include 0, so we can reject $H_{0,}$ which gives some idea of the probability distribution that led to the observed data sampling.

$H0\colon\ \mu1 - \ \mu2 = 0$

$Ha\colon\ \mu1 - \ \mu2 \neq 0$

The result of CI and Hypothesis: 1.93 и 3.6

In addition, you can call the t.test function. We can see that the results are the same (mean in group Female = 102, mean in group Male = 105) (see Figure 3.12):

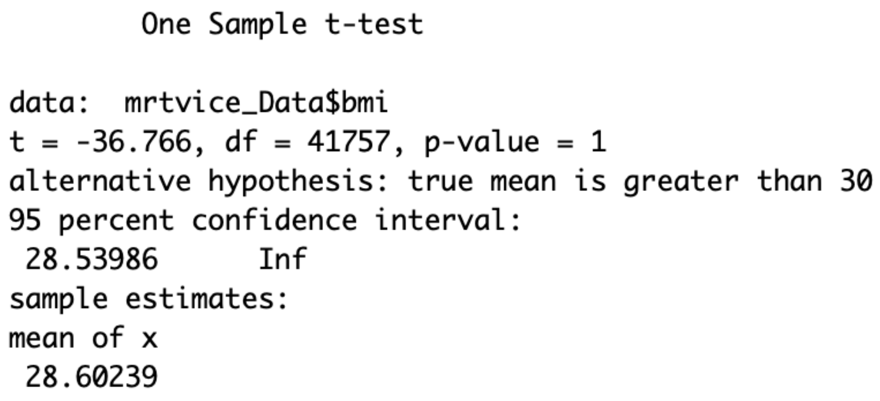


```
	Welch Two Sample t-test

data:  mrtvice_Data$prumerna_hladina_glukozy by mrtvice_Data$rod
t = -6.4901, df = 34480, p-value = 8.694e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.603687 -1.931915
sample estimates:
mean in group Female   mean in group Male 
            102.4845             105.2523 
```

Figure 3.12. T-test hypothesis

Mean BMI result in groups Female and Male (see Figure 3.13):



```
	One Sample t-test

data:  mrtvice_Data$bmi
t = -36.766, df = 41757, p-value = 1
alternative hypothesis: true mean is greater than 30
95 percent confidence interval:
 28.53986      Inf
sample estimates:
mean of x 
 28.60239 
```

Figure 3.13. BMI

Let:

H_0 - stroke Male and Female are the same.

H_a - stroke Male and Female are different.

α = 0,05

P-Value = 0,043

Level = 0,05

Table 5 – P-value

Gender	N	X[2]
Male	17000	286
Female	25000	357

It can be seen that p.value < α, so we reject the null hypothesis - there is not enough evidence to suggest that the ratio of strokes in men and women is the same.

[2] X - total stroke rate

4 Practical aspects, systematic sampling, forecasting and standard error estimation

This chapter analyzed the data on stroke. These data are publicly available and are currently undergoing extensive research. What is new about this paper is that it attempts to analyze unusual classification errors about these data, namely False Discovery Rate (FDR) and False Omission Rate (FOR). Procedures for which some functionality from FDR and FOR is being optimized are obtained. Considering these values can be considered as a response to the imbalance in the outcome of a stroke (stroke is quite common if we consider its frequency within a population).

4.1 Naive Bayesian method for predicting stroke

The Bayesian theorem (or Bayesian formula) is one of the basic theorems of elementary probability theory, which allows us to determining the probability of an event provided that another statistically related event has occurred. In other words, using the Bayesian formula, we can more accurately recalculate the probability by taking both previously known information and new observations. The Bayesian formula can be derived from the basic axioms of probability theory, in particular from conditional probability. The peculiarity of the Bayesian theorem is that its practical application requires a large number of calculations, calculations, so Bayesian estimates began to be actively used only after the revolution in computer and network technologies. [14].

$$P\ (A|B) = \frac{P(B|A)\ P(A)}{P(B)}\ (7),$$

- P(A) is the a priori probability of hypothesis A;
- P(A│B) is the probability of hypothesis A when event B occurs (a posteriori probability);
- P(B│A) - probability of event B when hypothesis A is true;
- P(B) is the full probability of event B. [14]

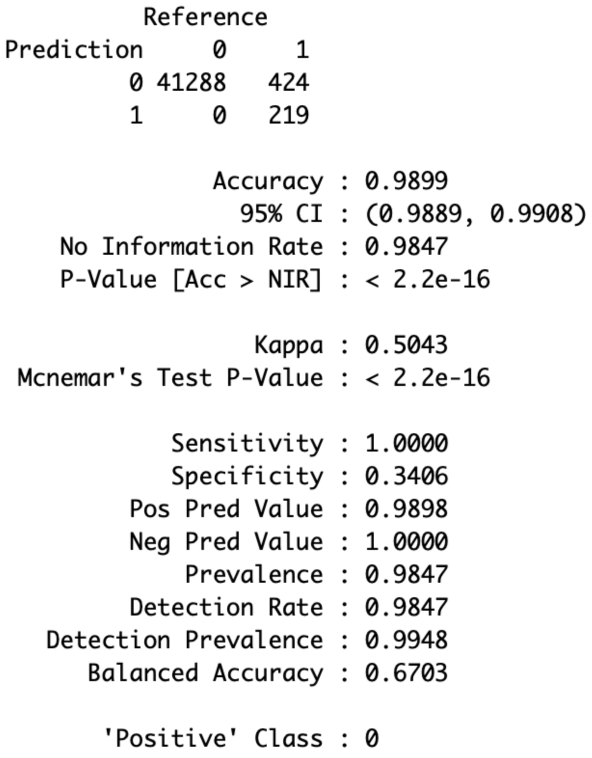
Take as an example a sample of 1,862 people who in their disease have a total of 4,128 symptomatic phenomena such as gender, age, hypertension, heart disease, ever married, work type, residence type, avg glucose level, BMI related to the symptomatology of suspected stroke patients.

The naive Bayesian method, using a formula for predicting stroke, showed that a population with a positive stroke test result corresponds to a 99% probability that the subject is having a stroke.

Let's assume that a subject from a population with a 99% prevalence of stroke would get a positive stroke test result. We are interested in calculating the probability of this subject having a stroke. Mathematically, we would like to calculate P(D+|T+), the probability of being a stroke-positive (D+) provided that one stroke test was positive (T+). The sensitivity of the P(T+|D+) test is known to be 0.99, the specificity of the P(T-|D-) test is known to be 0.35, and the prevalence of the disease in the P(D+) target group is known to be 0.99. Using the Bayes formula, we can calculate the amount we are interested in. P(T+|D+) = 0.99.

The Mcnemars stroke test is visible (see Figure 4.1):

- (TPR) That the stroke test is positive given that the person had a stroke - 1.0
- (TNR) That the stroke test is negative given that the person did not have a stroke - 0.35
- (PPV) That the person has a stroke and the stroke test result is positive, given that the person had a stroke - 0.99
- (NPV) That the person has no stroke and the result of the stroke test is negative - 1.0.



```
          Reference
Prediction     0     1
         0 41288   424
         1     0   219

               Accuracy : 0.9899
                 95% CI : (0.9889, 0.9908)
    No Information Rate : 0.9847
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5043
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 1.0000
            Specificity : 0.3406
         Pos Pred Value : 0.9898
         Neg Pred Value : 1.0000
             Prevalence : 0.9847
         Detection Rate : 0.9847
   Detection Prevalence : 0.9948
      Balanced Accuracy : 0.6703

       'Positive' Class : 0
```

Figure 4.1. Bayes Formula

Diagnostic ratio that the test is positive:

$$Res(LR)\ (+) = \frac{TPR}{1-TNR} = \frac{1}{1-0.35} = 1.5\ (8)$$

Diagnostic ratio that the test is negative:

$$Res\ (LR)(-) = \frac{1-TPR}{TNR} = \frac{1-1}{0.35} = 0\ (9)$$

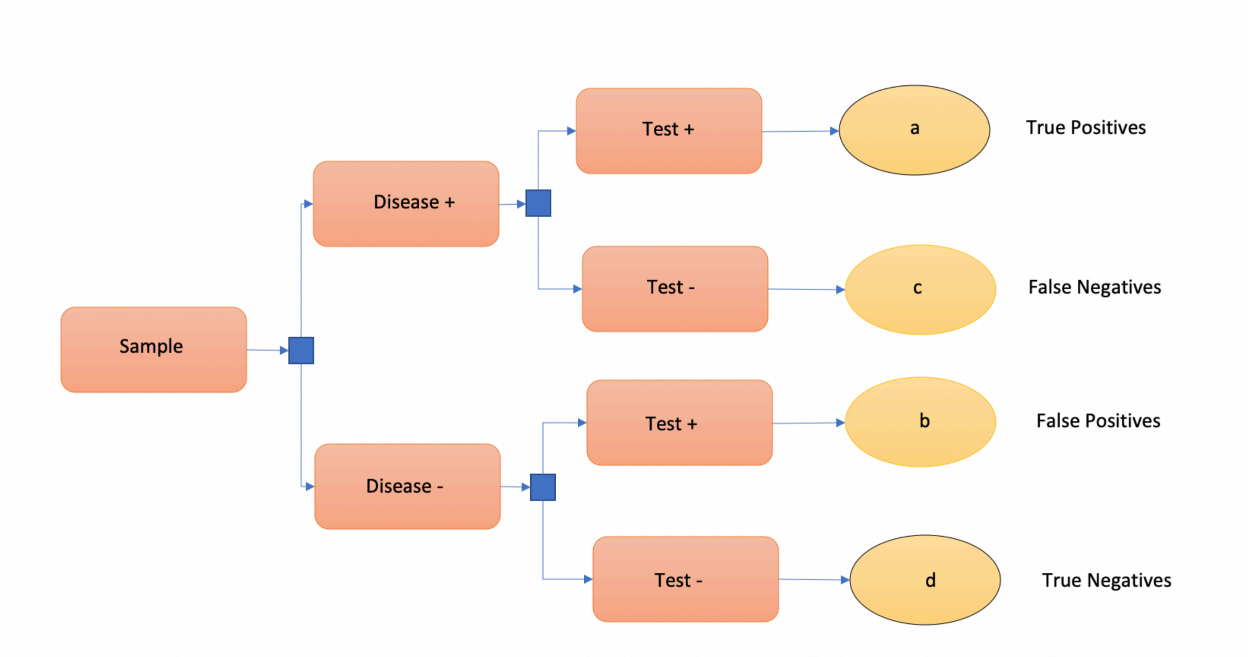


Figure 4.2. Test result [25]

A table of positive and negative prognostic values was constructed (see Table 6):

Table 6 - Positive and negative values

	Condition positive	Condition negative		
Test Pos	TP = 41288	FP = 393	PPV = 0,9900	FDR = 0,01
Test Neg	FN = 0	TN = 250	FOR = 0	NPV = 1
ACC = 0,9906	TPR = 1	FPR = 0,6112	LR+ = 1,5	DOR = 0
Preval = 0,9946	FNR = 0	TNR = 0,3530	LR- = 0	F1 Score = 0,9953

Where:

- TP - True positive: Sick people correctly identified as sick,
- FP - False positive: Healthy people incorrectly identified as sick,
- TN - True negative: Healthy people correctly identified as healthy,
- FN - False negative: Sick people incorrectly identified as healthy,
- TPR – True positive rate: The percentage of sick people who are correctly identified as having the condition,
- TNR – True negative rate: The percentage of healthy people who are correctly identified as not having the condition,
- FPR – False positive rate: The level of significance that is used to test each hypothesis is set based on the form of inference,
- FNR - False negative rate: The result of the test corresponds with reality, then a correct decision has been made,
- PPV and NPV – Positive and Negative predictive values: Are the proportions of positive and negative results in statistics and diagnostic tests that are true positive and true negative results, respectively,
- FOR and FDR - False omission rate and False Discovery Rate: Unusual classification errors. [36]

The credibility of a belief depends on how well the facts are explained. The more different the explanation of precedents, the less true belief is.

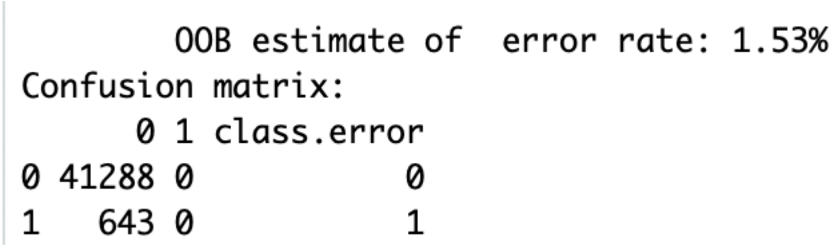
The results show that from the classifier of the selected symptomatology TP is equal to 41288, FP is equal to 393, which means that these symptoms have the wrong attitude to the class in question. Thus, TN equal to 250 shows that the object cannot be classified as a stroke patient, which is 1% of the total sample. In this situation, a test error of 1% is a truly random value related to the trend of changing symptoms in the disease, which is a problem in evaluating the accuracy of the test, not the Bayes method itself. This indicates that this percentage can be diagnosed by another method to get the best result.

Thus, a positive test result corresponds to a 99% probability that the person is a stroke. This is a positive predictive value for the test (PPV). The high positive prognostic value is due to the high prevalence of the disease and somewhat high specificity about the prevalence of the disease.

AUC and ROC

Random Forest is a universal machine learning method capable of performing both regression and classification tasks. It also uses resizing methods, handles missed values, deviation values, and other important steps in data mining, and does a pretty good job. This is a type of ensemble learning method where a group of weak models is combined into a powerful model (see Figure 4.3).

- Type of random forest: classification
- Number of trees: 500
- No. of variables tried at each split: 3.



```
        OOB estimate of  error rate: 1.53%
Confusion matrix:
      0 1 class.error
0 41288 0           0
1   643 0           1
```

Figure 4.3. OOB error

As you can see, it lists the call used to build the classifier, the number of trees (500), the variables at each interval (3), displays the matrix and OOB error rate estimate. This estimate is calculated by calculating how many points in the training set would be incorrectly classified and dividing this number by the total number of observations (~= 1.53%).

Estimation of OOB error rate is a useful measure to distinguish different Random Forest classifiers. We can, for example, vary the number of trees or the number of variables considered and choose the combination that gives the smallest value for this error rate. For more complex data sets, i.e. when more functions are present, it is a good idea to use cross-checking to perform function selection using the OOB error rate.

Let us consider the value that our classifier has assigned to each variable (see Figure 4.4):

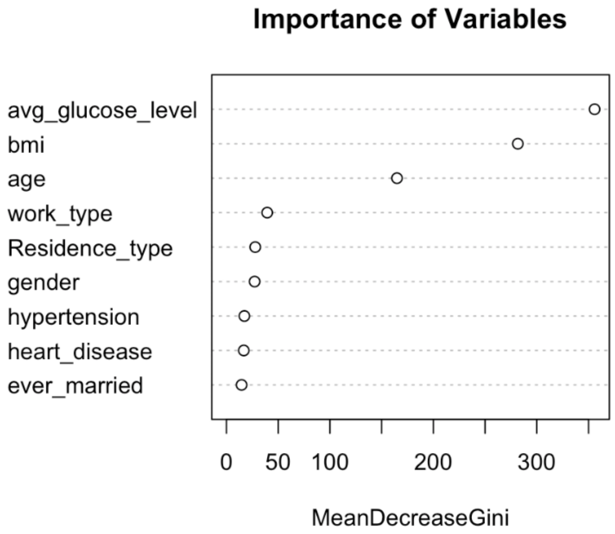


Figure 4.4. Mean Decrease Gini

MeanDecreaseGini: GINI is a measure of delamination in nodes. The highest purity means that each node contains only elements of one class. Estimating the GINI reduction when this function is omitted leads to an understanding of how important it is to correctly separate the data.

Another way to evaluate the performance of our classifier is to generate a ROC curve and calculate the area under the curve.

A ROC curve is a graph that shows the diagnostic capabilities of a binary classifier system.

The ROC curve is determined by a formula:

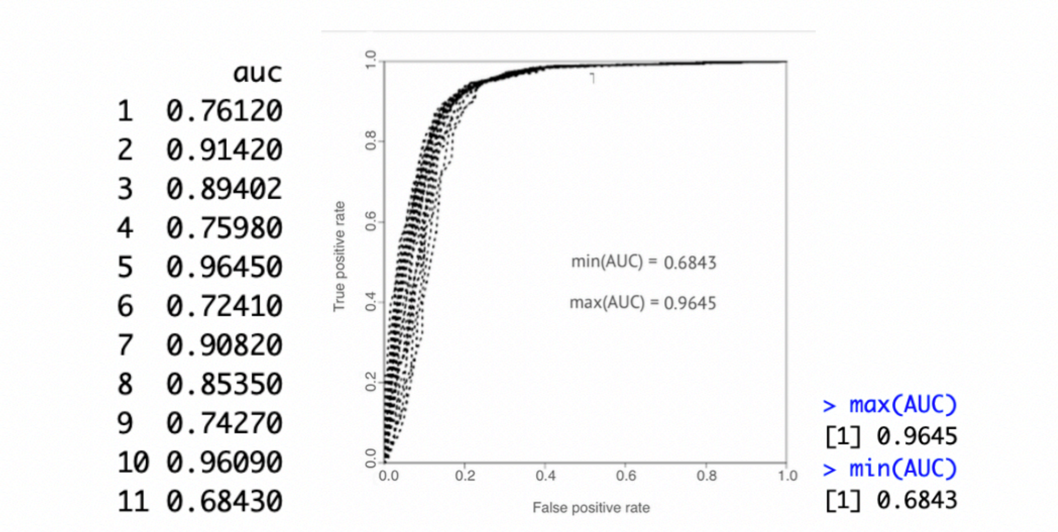
$$ROC = \ 1 - TNR = 1 - 0.35 = 0.65 \ (10)$$

AUC: Interpretation of AUC curves (see Table 7):

Table 7 – AUC curves

AUC	Diagnostic accuracy
0.9-1.0	Perfect
0.8-0.9	Very good
0.7-0.8	Ok
0.6-0.7	Normal
0.5-0.6	Bad
<0.5	Very bad

Depending on the threshold, it can be maximized or minimized. If AUC = 0.7 and above, it means that our model will probably be able to distinguish the negative from the positive class (see Figure 3.5).



```
       auc
1  0.76120
2  0.91420
3  0.89402
4  0.75980
5  0.96450
6  0.72410
7  0.90820
8  0.85350
9  0.74270
10 0.96090
11 0.68430
```

```
> max(AUC)
[1] 0.9645
> min(AUC)
[1] 0.6843
```

Figure 4.5. AUC

The higher the AUC, the better the classifier. The closer the curve follows the upper left corner and the larger the area under the curve, the better the test differentiates between those with and without a disease. In our case the AUC threshold (max) = 0.96 and the AUC threshold (min) = 0.68. With the lowest AUC (min), the principle cannot be used.

To determine the optimal threshold, it is necessary to establish a criterion for its determination.

Minimum total FDR and FOR

$$M = min(FDR + FOR) \quad (11)$$

Minimum FDR and FOR

$$N = \; min\,|FDR - FOR| \quad (12)$$

Let's consider the comparison criterion - FOR at the level of FDR 0.09. When using this method, FDR 0.09 and FOR 0.01 are reached. The maximum FOR reaches 0.04 - this indicates an improvement in the forecast. The threshold value affects the FDR to FOR ratio. We can talk about the task of finding the optimal cutoff value. The balance point value is 0.0582 (see Figure 4.6).

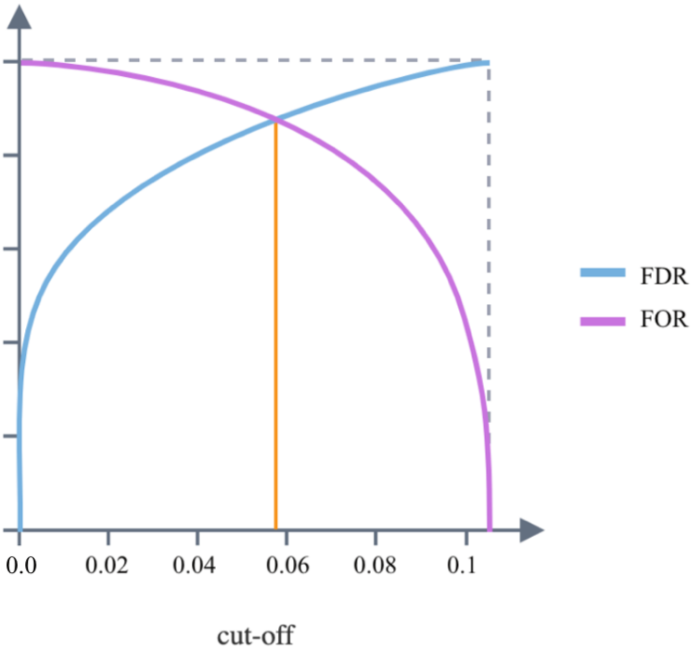


Figure 4.6. Balance point

Considering these values can be considered as a response to the imbalance in the outcome of a stroke (stroke is quite common, if we consider its frequency within a population).

4.2 Using the principle of Binomial Probability Intervals when trying to predict stroke

The binomial interval is a measure of uncertainty for a part of the statistical population. It takes the proportion from the sample and is adjusted for sample error.

$$\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \,(13),$$

where

- $\hat{p} = \frac{x}{n}$,
- p = proportion of interest
- n = sample size
- α = desired confidence
- $z_{1-\alpha/2}$ = “z value” for desired level of confidence
- $z_{1-\alpha/2}$ = 1.96 for 95% confidence
- $z_{1-\alpha/2}$ = 2.57 for 99% confidence
- $z_{1-\alpha/2}$ = 3 for 99.73% confidence

The binomial distribution model allows to calculate the probability of observing a certain number of "successful" outcomes when the process is repeated a certain number of times (for example, in a group of patients) and the outcome for a given patient is either successful or unsuccessful. First, you must enter some notation required for the binomial distribution model.

Using binomial distribution requires three assumptions:

- Each process response leads to one of two possible outcomes (success or failure).
- The probability of success is the same for each answer.
- The answers are independent, i.e. here success in one patient does not affect the probability of success in the other.

Consider a sample of 50 people.

For calculation P-hat we will need two numbers, the first number is the sample size (n), and the second number is the number of considered events or parameters (X). P-hat = 0.04. It is found by dividing the number of occurrences of the required event by the sample size.
P-tilde = 0.07. CI is more reliable than those based on p-hat.

- Result P-Hat = 0.04
- Result P-Tilde = 0.97

The Wald based confidence interval has good coverage (i.e. the confidence interval includes the true value 95% of the time) when the log-likelihood function, on the scale on which the Wald interval is constructed, is close to being a quadratic function.

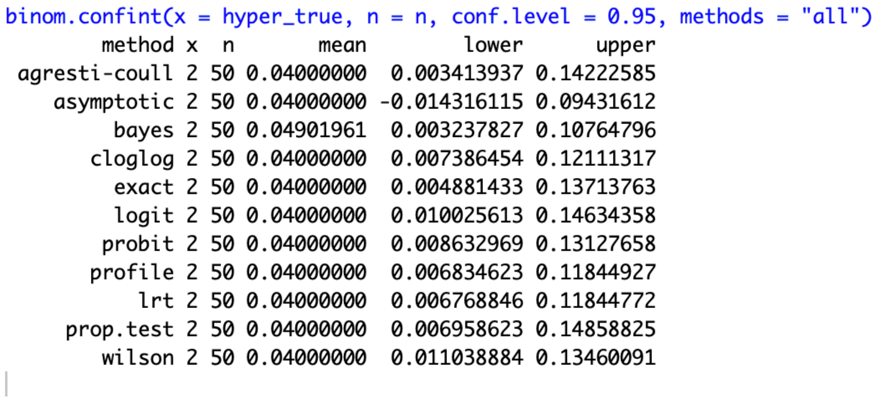
The result is 95% interval Wald:

$$Ci\ Wald = \frac{P{-}Hat+Z*Error(Walid)}{n} = 0{,}094 \quad (14)$$

The result is 95% interval Agresti/Coull, which the binomial proportion is defined as the number of successes divided by the number of trials:

$$Ci\ AC = \frac{P{-}Tilde+Z*Error(AC)}{n} = 0{,}147 \quad (15)$$

The result of the built-in function (see Figure 4.7):



```
binom.confint(x = hyper_true, n = n, conf.level = 0.95, methods = "all")
         method x  n       mean        lower      upper
  agresti-coull 2 50 0.04000000  0.003413937 0.14222585
     asymptotic 2 50 0.04000000 -0.014316115 0.09431612
          bayes 2 50 0.04901961  0.003237827 0.10764796
        cloglog 2 50 0.04000000  0.007386454 0.12111317
          exact 2 50 0.04000000  0.004881433 0.13713763
          logit 2 50 0.04000000  0.010025613 0.14634358
         probit 2 50 0.04000000  0.008632969 0.13127658
        profile 2 50 0.04000000  0.006834623 0.11844927
            lrt 2 50 0.04000000  0.006768846 0.11844772
      prop.test 2 50 0.04000000  0.006958623 0.14858825
         wilson 2 50 0.04000000  0.011038884 0.13460091
```

Figure 4.7. Built-in function “Binom.confict”

The probability of coverage varies widely, and when p is small or large, coverage can be quite poor even for a very large number of n. In practice, a good rule is that the coverage probability and asymptotic approximation to work with binomial probability p in cases where $\mathrm{np}(1-\mathrm{p}) \geqslant 5$. We have got less than 5.

A simple fix in cases when two successes and two failures must be added. That is, let $\hat{p} = (\mathrm{X}+2)/(\mathrm{n}+4)$ and $\hat{n} = n+4$. Then there will be the so-called Agresti-Coull interval. It follows that this principle cannot be used for the predictability of stroke.

4.3 Evaluation of Statistical Power for Stroke Prediction

Statistical power in mathematical statistics - the probability of the main (or zero) hypothesis rejection when testing statistical hypotheses in the case when the competing (or alternative) hypothesis is correct. The higher the power of a statistical test, the less likely it is to make a second type error. The power value is also used to calculate the sample size necessary to confirm the hypothesis with the necessary effect force. [13]

With the known standard deviation of the general population and the given level of significance α = 0,05, the power can be calculated using the Z-criterion by the formula:

$$1-\ \beta = P(Z > \frac{\mu 0+1{,}64(SE)-\ \mu 1}{SE}) \ (16),$$

where μ0 is the mean in the null hypothesis, μ1 is the mean in the alternative hypothesis, 1.64 is the critical value of Z-statistics in the one-way Z test, and SE is the standard error. [13]

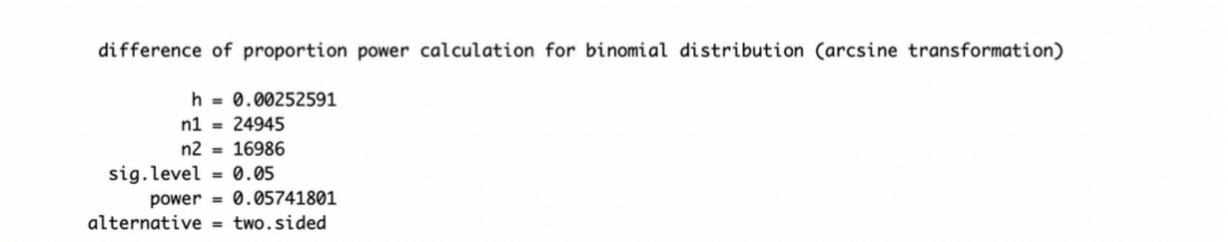
A population study on stroke demonstrated a causal relationship between the two variables. The resulting 5.7% power in a clinical study means that when tested statistically (i.e. statistically significant treatment effect), the study has a 5.7% chance of getting a p value of less than 5% if there was indeed a significant difference between the treatments. The results are questionable (the result is too small to detect any differences). As a general rule, 80% is an acceptable power level.

Power = 1 - β
Level - 0,05

1 proportion - 0.1

2 proportions - 0.2

Each group consists of 100 observations.



```
     difference of proportion power calculation for binomial distribution (arcsine transformation)

              h = 0.00252591
             n1 = 24945
             n2 = 16986
      sig.level = 0.05
          power = 0.05741801
    alternative = two.sided
```

Figure 4.8. Power calculation

The power is only 5.7%. In other words, the probability of correct deviation of the null hypothesis (no difference between groups) is 5.7% (see Table 8).

Now let's run the test to find the "small" effect between two groups of the proportion (difference ~ 0.2).

Table 8 – Different sample

	n	
1	10	9,7
2	20	14,5
3	30	19,4
4	40	23,41
5	50	28,30
6	60	33,08
7	70	38,7
8	80	43
9	90	47,5
10	100	51,60

To achieve a 51% probability of correct deviation of the null hypothesis, it is necessary to collect information from 2 groups of 100 people each.

This study has shown that the principle of using Statistical Power in Stroke Prognosis displays too scattered data, which cannot lead to the correct definition of the disease and the construction of appropriate treatment.

4.4 Determination of mean age of stroke in population T and F - tests

T-test is a statistical method that allows you to compare the mean values of two samples and, based on the results of the test, to conclude whether they statistically differ from each other or not.

$$t = \frac{Z}{s} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \ (17),$$

where:

- **x̄** = Observed Mean of the Sample
- **μ** = Theoretical Mean of the Population
- ***σ*** = Standard Deviation of the Sample
- **n** = Sample Size

Let's see if the age difference is significant for people who have already had a stroke and those who are at risk of having one in the future. To do that, we can run a t-test to find out. Let's visually examine the age distribution of the stroke result.

Table 9 – Average age in people with strokes

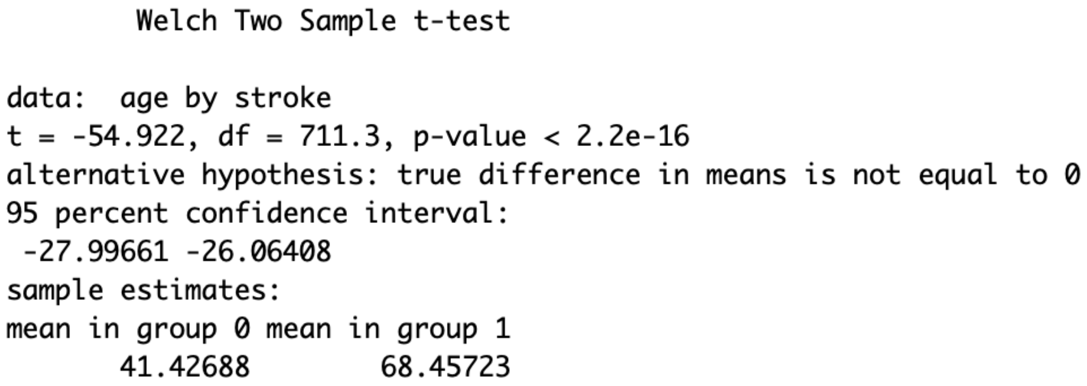
Group	N	Average	Variance
Stroke	643	68	148
No Stroke	41000	41	500

We can see that the mean age of stroke is 41-68. If we compare the mean values of the two groups, we'll get it:

$$\overline{X_1} - \overline{X_2} = -27$$

The confidence interval of average differences is -28 to -26. This does not include 0, so we reject the null hypothesis and say that there is a difference in significance between the average groups with a level of confidence of 95%. In other words, we are 95% sure that strokes are more likely to occur in older people.

With the function in the R T-test the following result was obtained:
The test with the built-in function in R gave the same result.
T-test (see Figure 4.9):



```
          Welch Two Sample t-test

data:  age by stroke
t = -54.922, df = 711.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -27.99661 -26.06408
sample estimates:
mean in group 0 mean in group 1
       41.42688        68.45723
```

Figure 4.9. T-test

In order to check on which data this test works best, I took a sample with less data.

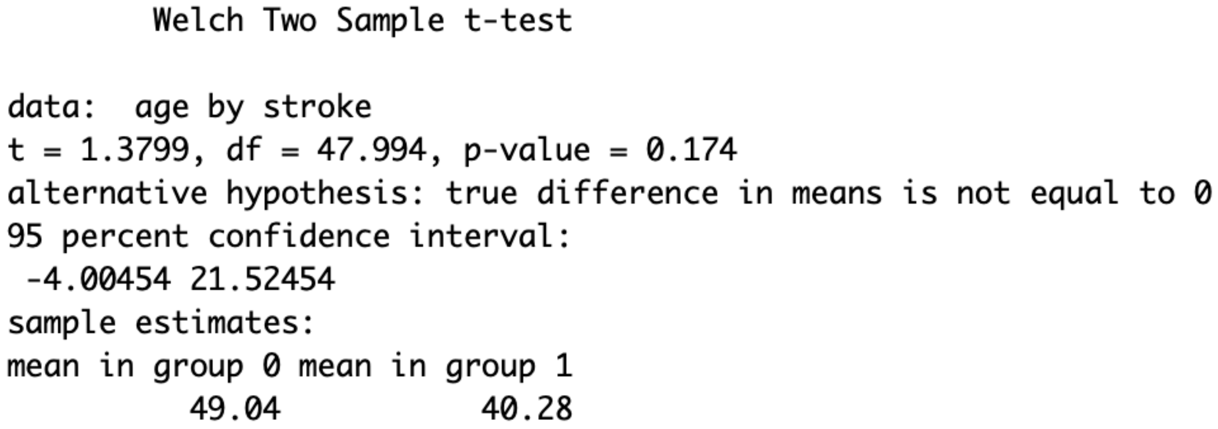
Table 10 – Average age in people with strokes (less data)

Group	N	Average	Variance
Stroke	25	49	510
No Stroke	25	40	498

We can see that the average age of stroke is between 40 and 49 years. In order to get a result, it is necessary to compare the obtained mean values between the two groups, get a result:

$$\overline{X_1} - \overline{X_2} = 9$$

The following result was obtained using the function in R T-test (see Figure 4.10):



```
        Welch Two Sample t-test

data:  age by stroke
t = 1.3799, df = 47.994, p-value = 0.174
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.00454 21.52454
sample estimates:
mean in group 0 mean in group 1
          49.04           40.28
```

Figure 4.10. T-test (less data)

In this case, the difference between groups is 9. The resulting difference is far from 0. After performing the test, the result of the confidence interval was obtained, which includes too wide a range from -4 to 21. This interval includes zero, which does not reject the null hypothesis. Therefore, we can conclude that in this sample, age is not related to the stroke.

F - test:

The F-test is used to compare the dispersions of two general normally distributed assemblies, i.e. the following null hypothesis is tested:

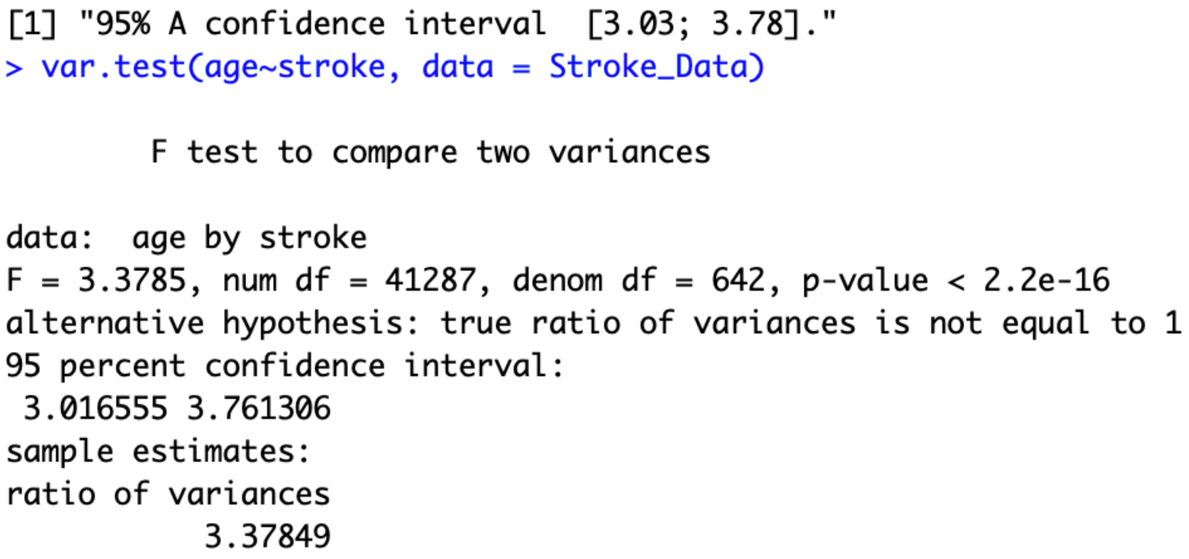
$$H_0: \sigma_1^2 = \sigma_1^2$$

$$F = \frac{S_1^2}{S_2^2}$$

A sample was considered which showed the outcome of the stroke with an age distribution of the population:

- F = 3.38

A confidence interval was constructed (see Figure 4.11):



```
[1] "95% A confidence interval  [3.03; 3.78]."
> var.test(age~stroke, data = Stroke_Data)

	F test to compare two variances

data:  age by stroke
F = 3.3785, num df = 41287, denom df = 642, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 3.016555 3.761306
sample estimates:
ratio of variances
           3.37849
```

Figure 4.11. F-test (confidence interval)

The confidence interval of average differences is from 3.03 to 3.78. This interval does not include 0, so we can reject H0, which gives some idea about the probability distribution that led to the observed data sampling. Thus, the difference between the groups is 95%.

T and F - tests have been used to check whether the age difference is significant for people who have already had a stroke and those who are at risk of future stroke. It has also been shown that strokes are more common in older people.

4.5 Estimation of the standard error using the folding Jackknife

The Jackknife is one of the resampling methods (a linear approximation of the statistical bootstrap) used to estimate the error in the statistical output. The method consists of the following: for each element, the average sample value is calculated without taking into account this element, and then the average of all such values is calculated. For a sample of N elements, the estimation is obtained by calculating the average value of the other N-1 elements. [12]

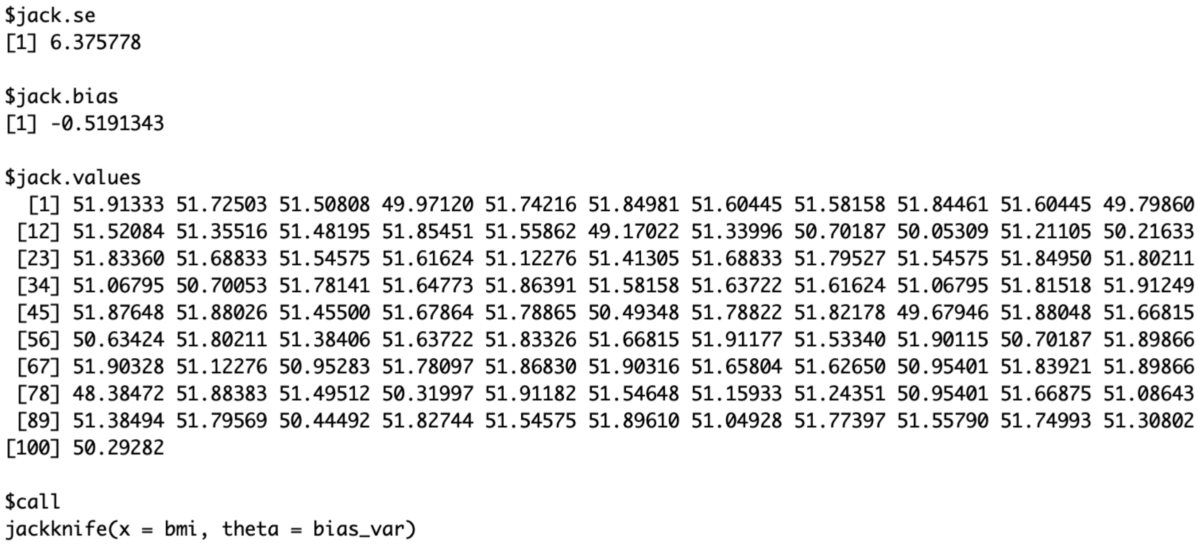
$$Var_{(jackknife)} = \frac{n-1}{n}\sum_{i=1}^{n}(\overline{x_\iota}\overline{-x_{(.)}})^2 \; (18),$$

where x_i - is the evaluation parameter, $x_{(.)} = \frac{1}{n}\sum_{i}^{n} x_i$ - is the evaluation based on all elements.

The Jackknife removes each observation and calculates an estimated base for the remaining (n-1) of them. It uses this collection of estimates for things like offset and st.error estimates.

Offset and st.error estimates are not needed for things like sampling tools, which we know are impartial estimates of BMI in the population.

Let us consider the BMI data set (see Figure 4.12):



```
$jack.se
[1] 6.375778

$jack.bias
[1] -0.5191343

$jack.values
  [1] 51.91333 51.72503 51.50808 49.97120 51.74216 51.84981 51.60445 51.58158 51.84461 51.60445 49.79860
 [12] 51.52084 51.35516 51.48195 51.85451 51.55862 49.17022 51.33996 50.70187 50.05309 51.21105 50.21633
 [23] 51.83360 51.68833 51.54575 51.61624 51.12276 51.41305 51.68833 51.79527 51.54575 51.84950 51.80211
 [34] 51.06795 50.70053 51.78141 51.64773 51.86391 51.58158 51.63722 51.61624 51.06795 51.81518 51.91249
 [45] 51.87648 51.88026 51.45500 51.67864 51.78865 50.49348 51.78822 51.82178 49.67946 51.88048 51.66815
 [56] 50.63424 51.80211 51.38406 51.63722 51.83326 51.66815 51.91177 51.53340 51.90115 50.70187 51.89866
 [67] 51.90328 51.12276 50.95283 51.78097 51.86830 51.90316 51.65804 51.62650 50.95401 51.83921 51.89866
 [78] 48.38472 51.88383 51.49512 50.31997 51.91182 51.54648 51.15933 51.24351 50.95401 51.66875 51.08643
 [89] 51.38494 51.79569 50.44492 51.82744 51.54575 51.89610 51.04928 51.77397 51.55790 51.74993 51.30802
[100] 50.29282

$call
jackknife(x = bmi, theta = bias_var)
```

Figure 4.12. Jackknife

Standard Jackknife Estimation θ= 6.37.

Bias is an offset to the correct answer.

Theta θ applies to BMI with the removal of 1st observation, Theta θ applies to BMI with the removal of 2 observations ... theta θ applies to BMI with the removal of n observation = -0,5.

The Jackknife method is more conservative than the Bootstrap method, i.e. its calculated standard error was greater. And, as a rule, it is less computationally intensive in comparison with the Bootstrap method.

4.6 Analysis of the data obtained based on the forecasting methods used

The process of extracting knowledge from the data is carried out according to the same scheme as the establishment of physical laws: a collection of experimental data, their organization in the form of tables, and the search for such a scheme of reasoning, which, firstly, makes the results obtained obviously and, secondly, makes it possible to predict new facts. At the same time, there is a clear understanding that our knowledge about the analyzed process, as well as any physical phenomenon, to some extent approximation. In general, any system of reasoning about the real world involves different kinds of approximations. In fact, the term Data Mining is an attempt to legitimize the physical approach, as opposed to the mathematical one, to solving problems of data analysis.

The work in this thesis is aimed at the earlier prediction of stroke, resulting in reduced mortality.

After conducting mathematical experiments to determine the best prediction of stroke mortality, I obtained percentage accuracy in predicting different methods and displayed them in a histogram (see Figure 4.13).

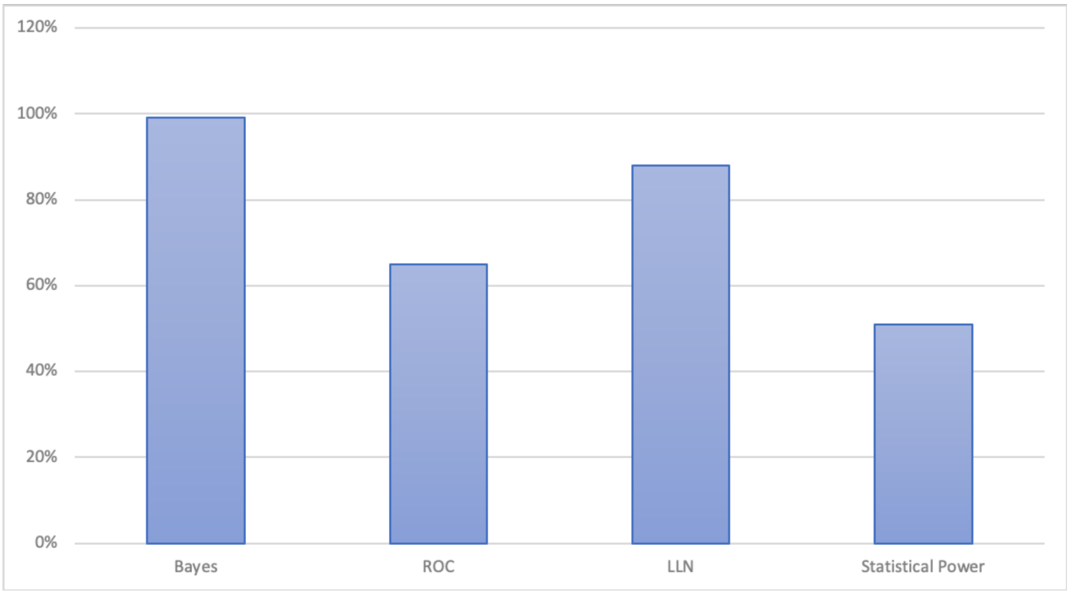


Figure 4.13. Best forecasting methods

From the histogram, we can conclude that the best approach to prevent stroke in my work is to predict using the Bayes method. For this method, a large number of tests are implied.

From this we can conclude: analyzing the available information, which is inherent in the property to update that it can predict future events, in our case, to predict the onset of stroke, based on the symptoms and available information in general on the specifics of the disease.

The best method for detecting a standard error in the methods I use - Jackknife. It is a relatively simple method for determining accuracy and error in calculations. The Jackknife method is less expensive and more intensive in its estimates.

CONCLUSION

In my master's thesis, various methods of probabilistic stroke prognosis based on different age groups and corresponding diseases were used. When comparing the accuracy of the prognosis, it was proved that in most of the methods used there is a so-called significant percentage of error, that in a disease such as a stroke, should be minimal.

Based on the data obtained and comparative results, we can say that the best method and approach to the diagnosis of stroke in my work is the Bayes method. The Bayesian method, based on the available information and new evidence, is the most accurate method to determine the likelihood that an event will occur. In this case, predicting and identifying early signs of morbidity.

However, it is important to understand and take into account the fact that the earlier a disease is diagnosed before a future pre-stroke condition, the more likely it is to be successfully treated to prevent symptoms and the disease itself. Current work and comparative statistics, based on available and repetitive data, make it possible to identify a disease much faster and to exclude or prevent the possibility of the signs of a stroke soon or at a more mature age.

Thus, the research project uses differentiated data, based on different indicators and statistics, which, being investigated in the material, allows to better diagnose early signs of the disease in practice. Which in this case can help various professionals, doctors, or diagnosticians to reduce the percentage of deaths from stroke.

Based on studies and tests of various methods capable of diagnosing the course of the disease, the software code can be provided to physicians as well as specialists in various fields.

BIBLIOGRAPHY

1. Bilich, G.L. Popular medical encyclopedia. - Moscow: Veche, 2012. – 399 p. [14]
2. Kruk, I.V. Explanatory dictionary of psychiatric terms/ I.V. Kruk, V.M. Bleicher. - Voronezh: NPO "Modec", 1995. – 221 p. [13]
3. Kurochkina, A.I. Modern methods of analysis of medical data: article in the journal - scientific article / A.I. Kurochkina, E.N. Timin / Annals of surgical hepatology. - Moscow: "Vidar" Ltd., 1998. – 131 p. [2]
4. Dan Morris, "Bayes Theorem: A visual introduction for beginners" - USA, 2016. - p.324. [8]
5. Jacob Cohen, "Statistical Power Analysis for the Behavioral Sciences" - USA, 1988. – p.428. [11]
6. Mar G. George MD, "Annals of Neurology" - USA, 2011. – p.309. [15]
7. OECD INDICATORS, "Health at a Glance 2019" - England, 2019. - p.243. [16]
8. William J. Adams, "The Life and Times of the Central Limit Theorem (History of Mathematics)" - England, 2009. - p.195. [10]
9. Data analysis [Electronic resource]. Access mode: https://clck.ru/JLo9d (circulation date: 21.01.2020). [1]
10. Boothstrap [Electronic Resource]. Access mode: https://clck.ru/NUJUV (circulation date: 1.02.2020). [3]
11. Confidence intervals [Electronic resource]. Access mode: https://clck.ru/NUJPi (circulation date: 5.03.2020). [4]
12. Confidence interval [Electronic resource]. Access mode: https://clck.ru/NUHpQ (circulation date: 5.03.2020). [5]
13. Disease caused by coronavirus (COVID-19) [Electronic resource]. Access mode: https://clck.ru/MRe7C (circulation date: 18.04.2020). [6]
14. Changes in the circulatory system [Electronic resource]. Access mode: https://clck.ru/NUGuSx1 (circulation date: 20.02.2020).
15. Construction of confidence interval for mathematical expectation of general population [Electronic resource]. Access mode: https://clck.ru/NUHvQ (circulation date: 5.03.2020). [12]
16. Setting process R [Electronic resource]. Access mode: https://clck.ru/NSm7m (circulation date: 13.01.2020).
17. Jackknife [Electronic resource]. Access mode: https://clck.ru/NUJZ6 (circulation date: 16.03.2020).
18. Statistical power [Electronic resource]. Access mode: https://clck.ru/NUHit (circulation date: 24.02.2020).

19. The Bayes Theorem [Electronic Resource]. Access mode: https://clck.ru/Gg4hf (circulation date: 7.12.2019).
20. Central limit theorem [Electronic resource]. Access mode: https://clck.ru/DBJa8 (circulation date: 4.01.2020).
21. Coronaviridae [Electronic resource]. Access mode: https://clck.ru/M7HVc (circulation date: 21.04.2020).
22. EDA [Electronic Resource]. Access mode: https://clck.ru/NSnCy (circulation date: 8.04.2020).
23. OECD [Electronic Resource]. Access mode: https://clck.ru/NSn6P (circulation date: 20.04.2020).
24. PyCharm [Electronic resource]. Access mode: https://clck.ru/NSmbu (circulation date: 2.02.2020).
25. Prevalence [Electronic resource]. Access mode: https://clck.ru/PsdnX (circulation date: 2.05.2020). [25]
26. T-test [Electronic resource]. Access mode: https://clck.ru/PsknY (circulation date: 15.05.2020). [26]
27. Electrocardiography [Electronic resource]. Access mode: https://clck.ru/Q5GCg (circulation date: 4.08.2020). [27]
28. X-ray [Electronic resource]. Access mode https://clck.ru/GbjFV (circulation date: 4.08.2020). [28]
29. Libraries in RStudio [Electronic resource]. Access mode https://clck.ru/Q5e5F (circulation date: 12.01.2020). [29]
30. Causes of death in the world [Electronic resource]. Access mode https://clck.ru/JkdHz (circulation date: 13.04.2020). [30]
31. Risk factors [Electronic resource]. Access mode https://clck.ru/Q5kVy (circulation date: 14.04.2020). [31]
32. Hypercholesterolemia [Electronic resource]. Access mode https://clck.ru/Q5kec (circulation date: 14.04.2020). [32]
33. Overweight and obesity Hypercholesterolemia [Electronic resource]. Access mode https://clck.ru/G6e9S (circulation date: 14.04.2020). [33]
34. Requirements analysis [Electronic resource]. Access mode https://clck.ru/Q5osC circulation date: 14.07.2020). [34]
35. Monolithic Architecture [Electronic resource]. Access mode https://clck.ru/Q5p8W circulation date: 15.07.2020). [35]
36. Sensitive and specificity [Electronic resource]. Access mode https://clck.ru/Q896N circulation date: 16.07.2020). [36]
37. Data of Stroke [Electronic resource]. Access mode https://clck.ru/Q8BiU
38. Data of Covid_19 [Electronic resource]. Access mode https://clck.ru/Q8DxW