Czech Technical University in Prague

Faculty of Electrical Engineering

Department of Computer Science

Master`s Thesis

STROKE MORTALITY PREDICTION

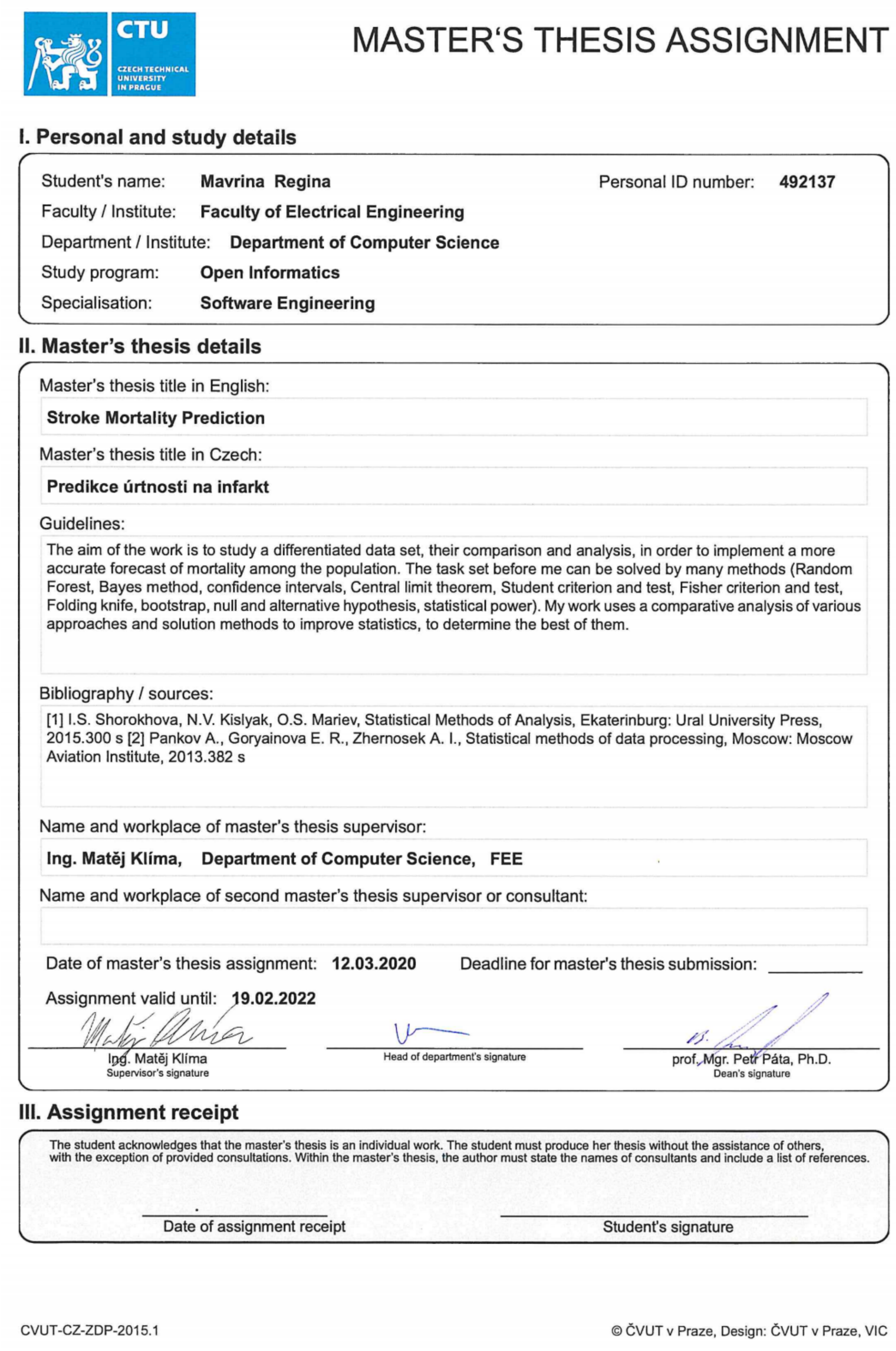
Regina Mavrina

Supervisor: **Ing. Matéj Klíma**

Study Program: Open Informatics

Field of Study: Software Engineering

May 22, 2020



CTU CZECH TECHNICAL UNIVERSITY IN PRAGUE

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, Department of Computer Science, 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: ___________

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

CVUT-CZ-ZDP-2015.1

© ČVUT v Praze, Design: ČVUT v Praze, VIC

Acknowledgements

I express my deep gratitude for the help in preparation of the diploma work to my supervisors, Ing. Matéj 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, May, 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. Matéj 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's a life-supporting system. If we take into account the world trend of mortality in medicine, cardiovascular diseases are steadily leading, along with oncology. However, despite the fact that 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. It motivated me to write this thesis, which can help prevent stroke deaths. The focus is on finding the best methods for predicting stroke. Analysis and specification using mathematical tests shows good results. The results of the tests are predicted data that help prevent the development of stroke. The developed software interacts with databases, which give the probability of stroke.

Differentiated data sets will be studied, tested, compared and analysed using different statistical methods, based on the information received and available, in order to be able to make more accurate predictions of stroke conditions from the onset of the disease to the fatal outcome in order to reduce them.

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, it will be searched for methods of symptom destructurization and key data in the study of pre-stroke conditions for universalization and optimization of this process, in order to predict stroke in accordance with the specific data.

The tasks of the master's thesis:

1. To analyze 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 assessment of reliability and accuracy to the predictability of mortality.

Thesis contents:

This thesis consists of an introduction, three chapters containing 4 paragraphs in chapter 1, 5 paragraphs in chapter 2 and 6 paragraphs in chapter 3, as well as conclusions and conclusions and a list of the literature used.

Scientific novelty of research:

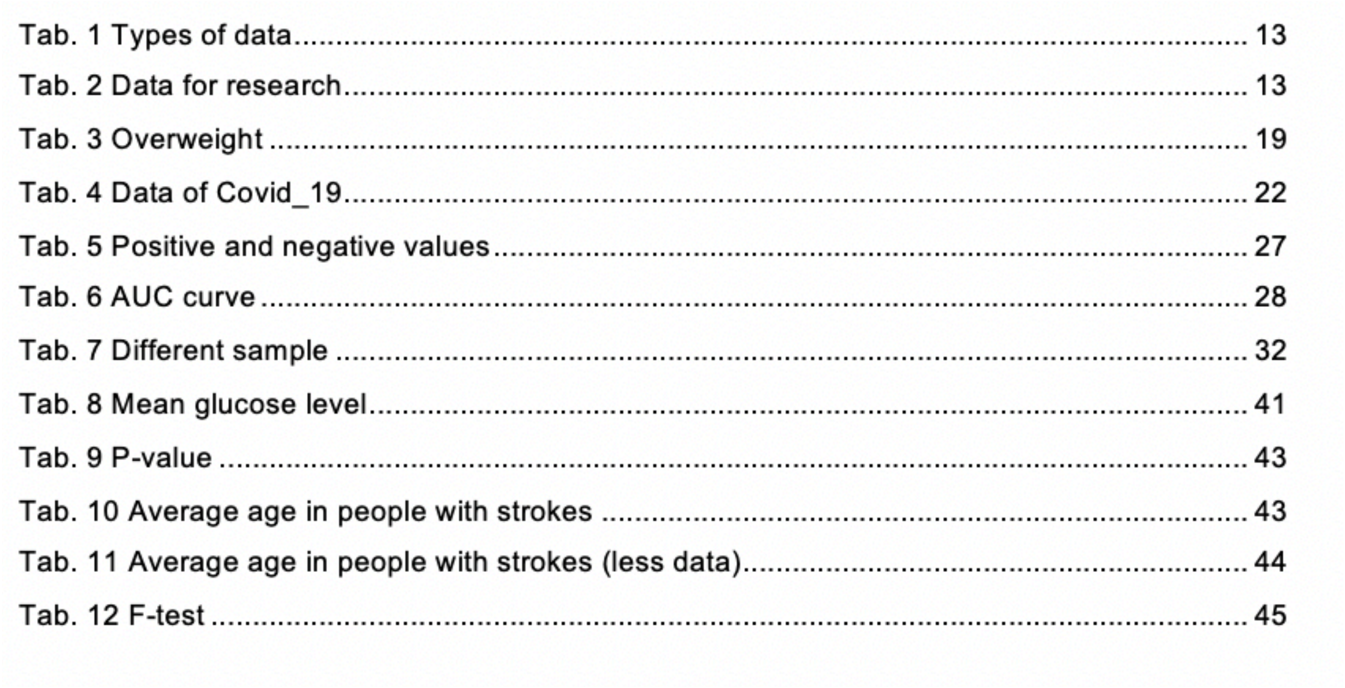
1. Prediction of stroke by the law of large numbers.
2. Prediction of stroke using the Bayesian method with control of the expected proportion of false positives.
3. Determine the relationship between COVID-19 and the onset of stroke.

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 in order 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 and 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, on the basis 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 of 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.

Chapter 1

Theoretical aspects. Main problems and approaches

1.1 General information on medical data analysis. Collection and processing

Data analysis is a process of inspecting, cleansing transforming and modeling data with the goal of discovering useful information, informing conclusions and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains. In today's business world, data analysis plays a role in making decisions more scientific and helping businesses operate more effectively. [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 actual current health status at the time. The analysis of medical indications, current symptomatology, predicted conditions and appropriate treatment is a study of the totality of all 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:

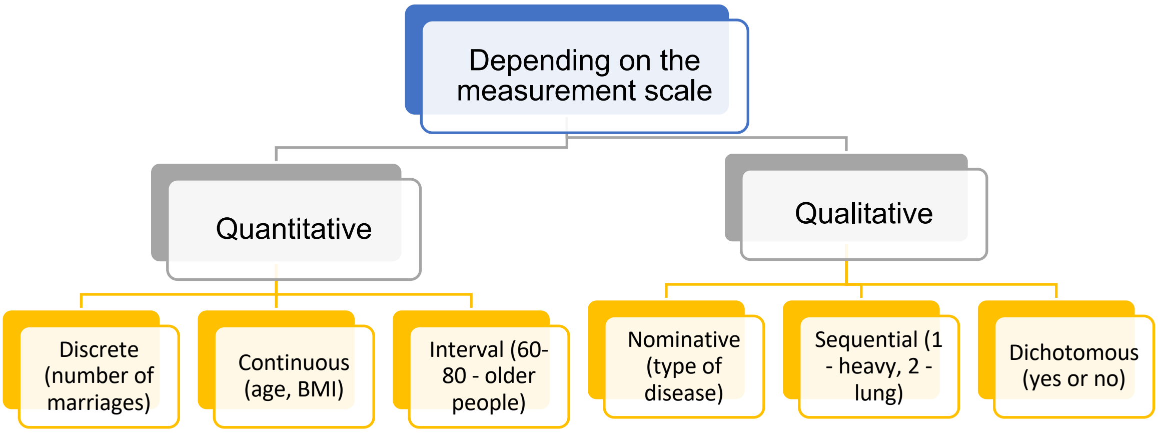
- Research concept
- Data collection and preparation
- Analysis
- Interpretation of results obtained
- Conclusion based on data obtained

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.

The means and methods of achievement of the result used in applied statistics, on the basis of clinical results and observations, allow to achieve the solution of classification tasks, their sequence, search for new ways of possibilities, for the solution of scientific hypotheses now and in the future. Statistics, in this regard, is often interpreted as the object language of describing the use of reality in sciences that use data that are not subject to strict formalization. [2]

Types of data (medical):

1.


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)
Dichotomous (yes or no)

Tab. 1 Types of data

2. Time parameters:
   - Dynamic - indicators that change over time. For example, electrocardiography.
   - Statistical - indicators that do not change in time. For example, X-rays.
3. Depends on the object of research:
   - Patient - his vital signs.
   - Population - health indicators of the population.

Materials and data for the study were taken from a public web platform in the form of open statistics for data processors and machine training engineers Kaggle.

id	gender	age	hypertension	heart_disease	ever_married	avg_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

Tab. 2 Data for research

However, a number of problems often arise when analysing medical data. And most of them require resources and time to solve and prevent.
Actual medical data are not publicly available and test data must be used to conduct the study. Because patient numbers are constantly changing due to the course of the disease, they need to be updated and adjusted periodically for more accurate and detailed statistics.

A fairly large data set should be used to predict stroke mortality in the population. Processing and correlation of such information should always be done on a computer with sufficient RAM. And, taking into account the fact that the number of lethal outcomes from stroke does not decrease every year, we can conclude that the health care system in our country, although it is developing every year, the infrastructure of many medical institutions is outdated and of inadequate 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 and to the health of its citizens with a view to reducing or preventing 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 which help to predict a stroke and to construct schedules on the basis of the available data.

Development environments:

- **RStudio** is a free software development environment with open source code for the R programming language, which is designed 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.
- **PyCharm** is an integrated development environment (IDE) used in computer programming, specifically for the Python language. It is developed by the Czech company JetBrains.It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCSes), and supports web development with Django as well as Data Science with Anaconda. [3]

Libraries:

- **RandomForest** is a library that corresponds to a number of decision tree classifiers on different subsets of the data set and uses averaging to improve forecasting and control accuracy.
- **Cowplot** - library for drawing charts.
- **AUC** - library for calculating ROC(AUC) score.

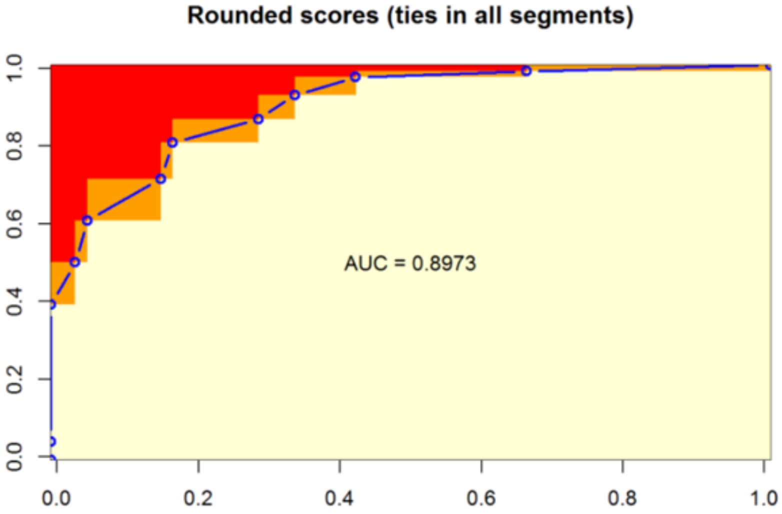


Fig. 1-1 Curve ROC

- **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 of 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 tails.

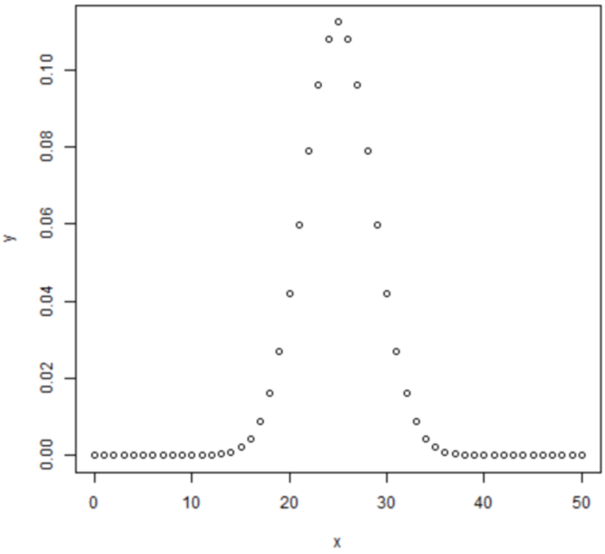


Fig. 1-2 Binomial distribution

- **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 Influencing 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.

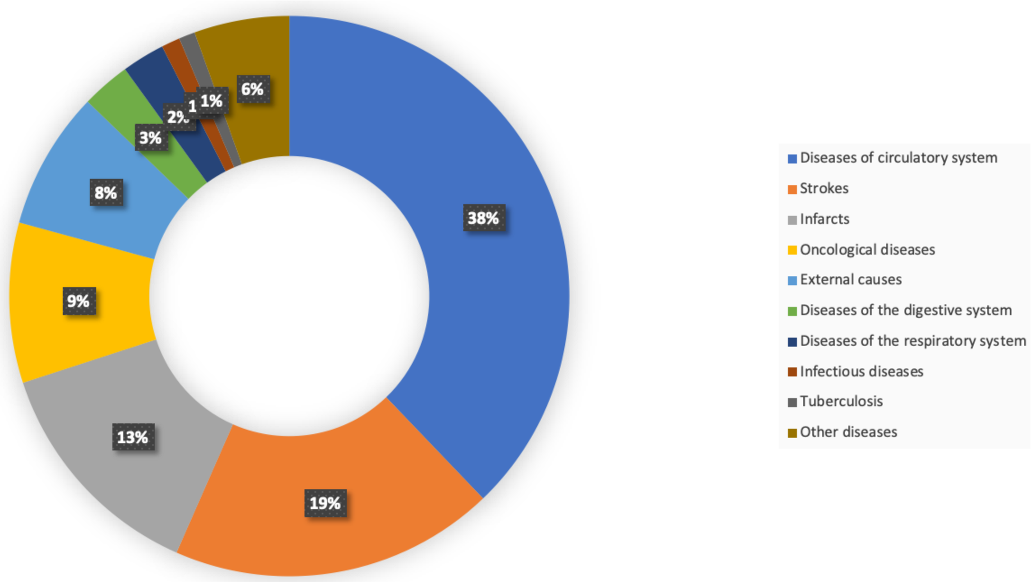


Fig. 1-3 The leading causes of death in the world

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

Let's compare the statistics on the number of deaths from stroke in Europe and Russia per 100 thousand inhabitants for 2019, according to the data of the Organization for Economic Cooperation and Development, which is freely available in the report Health at a Glance 2019.

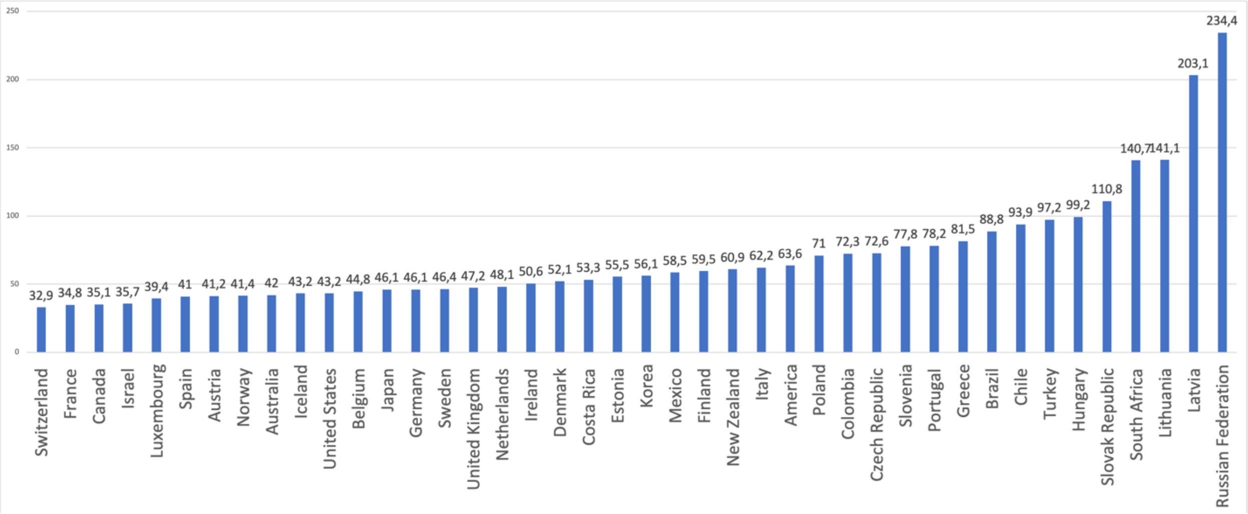


Fig. 1-4 Statistics on the number of deaths in the world

This graph clearly shows that Russia takes the leading position by the number of deaths from stroke among European countries.

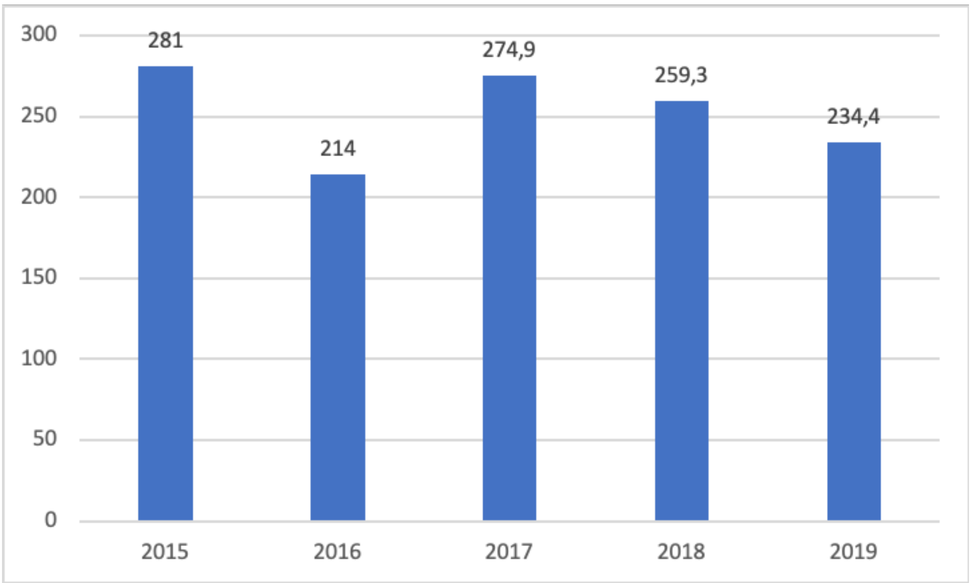


Fig. 1-5 Five-year statistics

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

That is why the purpose of this dissertation is to find methods for the destructurization of symptoms and key data in the study of pre-stroke conditions to universalize and optimize the process, in order to predict stroke taking into account the specific data.

Open databases of statistical data from the Kaggle website were taken as the initial data. The method of research data analysis (EDA) will be used as visual perception - such 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. [4]

Ten major fatalities worldwide have been identified and marked. 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:

- Abuse of alcohol
- Inheritance
- Smoking
- Obesity
- Unhealthy diet
- Hypercholesterolemia
- Hypertension. Hypercholesterolemia
- 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. Stroke is getting younger. In the American medical journal Annals of Neurology.

published in 2011, an article was presented 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, despite the fact that the age of stroke is decreasing, the percentage of disease among the younger generation, of the total number of cases and those who have suffered it, is on average 10%.

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. **Marital status and BMI**

Do people really tend to gain extra weight in marriage? Overweight is one of the points at which stroke occurs. The average BMI for people who have ever been married is 30.6. Obesity is BMI above 29. Overweight is between 25 and 29.

Ever Married	Median	Average	Variance	STD
Yes	29,5	30,6	50,4	7,1
No	23,6	25,1	58,4	7,64

Tab. 3 Overweight

A histogram was built which showed that people tend to gain extra weight in marriage.

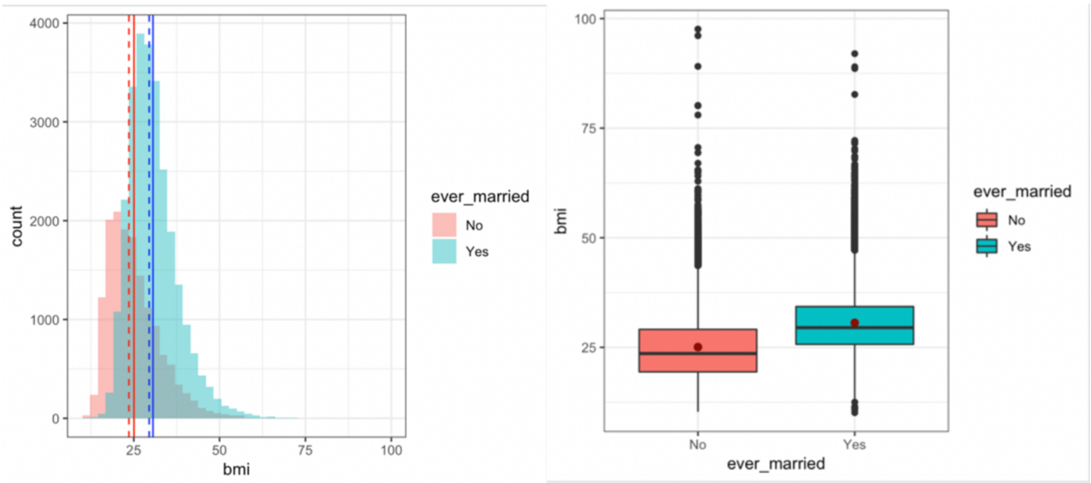


Fig. 1-6 Histogram of overweight

2. **Age**

A histogram was constructed which showed that people tend to stroke at the age of 60-80 years.

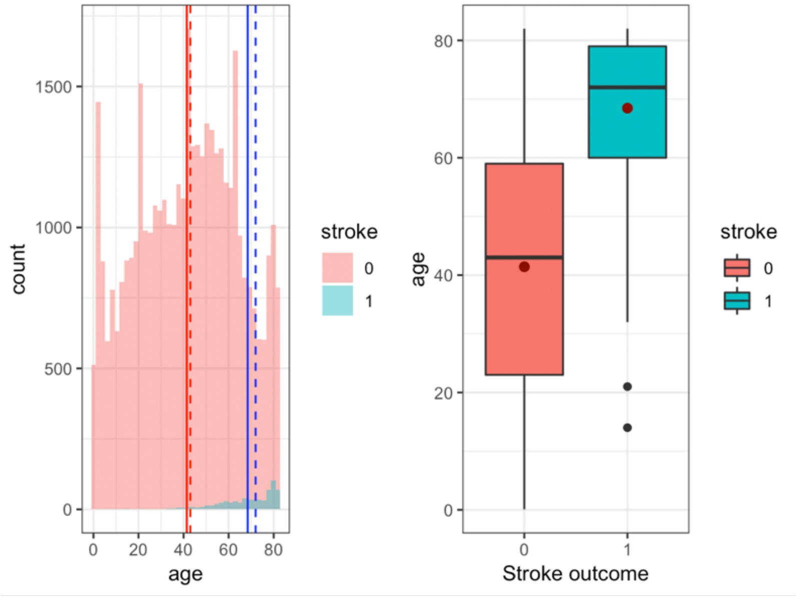


Fig. 1-7 Histogram of age

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.

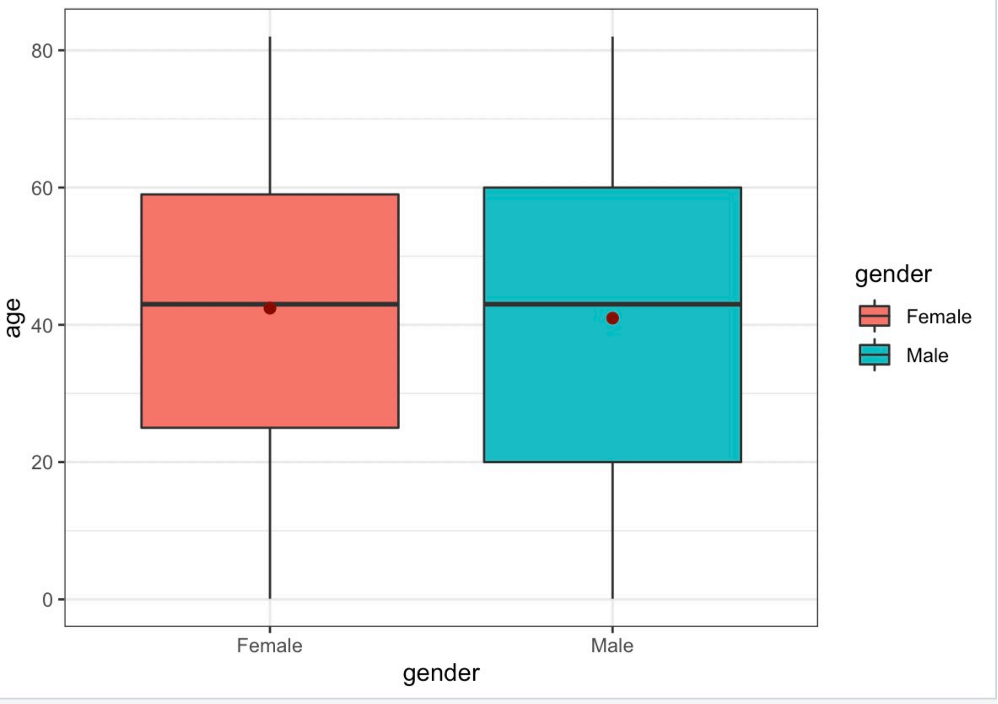


Fig. 1-8 Histogram of gender

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.

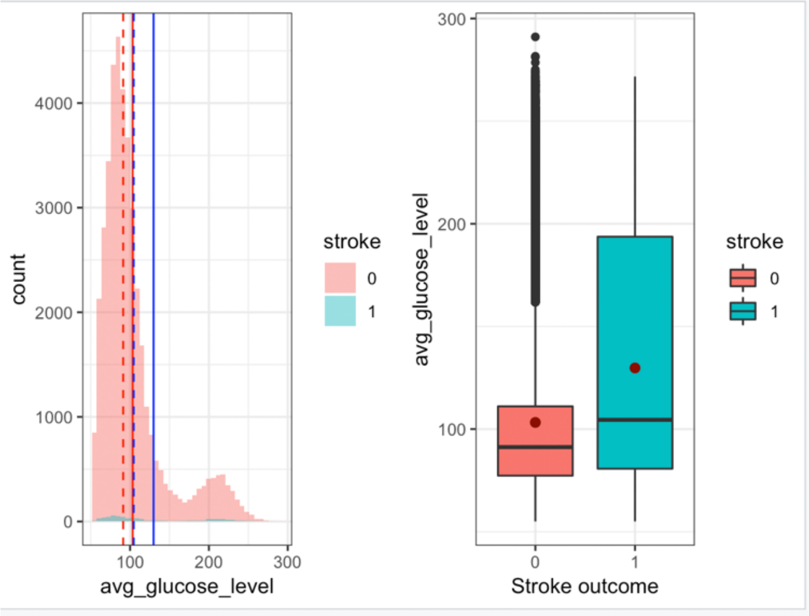


Fig. 1-9 Histogram of glucose level

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 of COVID-19 to 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).

Coronaviruses are 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, 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. [5]

It has been 6 months 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 can affect the blood system and even the brain.[6]

The purpose of this section 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 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. Data from Kaggle were taken for analysis (covid_19):

ITTER	Region	Pathology	Time	Value	Code
1	Moscow	Stroke	2017	0	37.2
1	Moscow	Cardiopathic	2018	1	37
2	Spb	Stroke	2018	0	456
3	Kazan	Arteriosa	2019	0	21
1	Moscow	Cardiac	2016	0	532
2	Spb	Stroke	2019	0	12

Tab. 4 Data of Covid_19

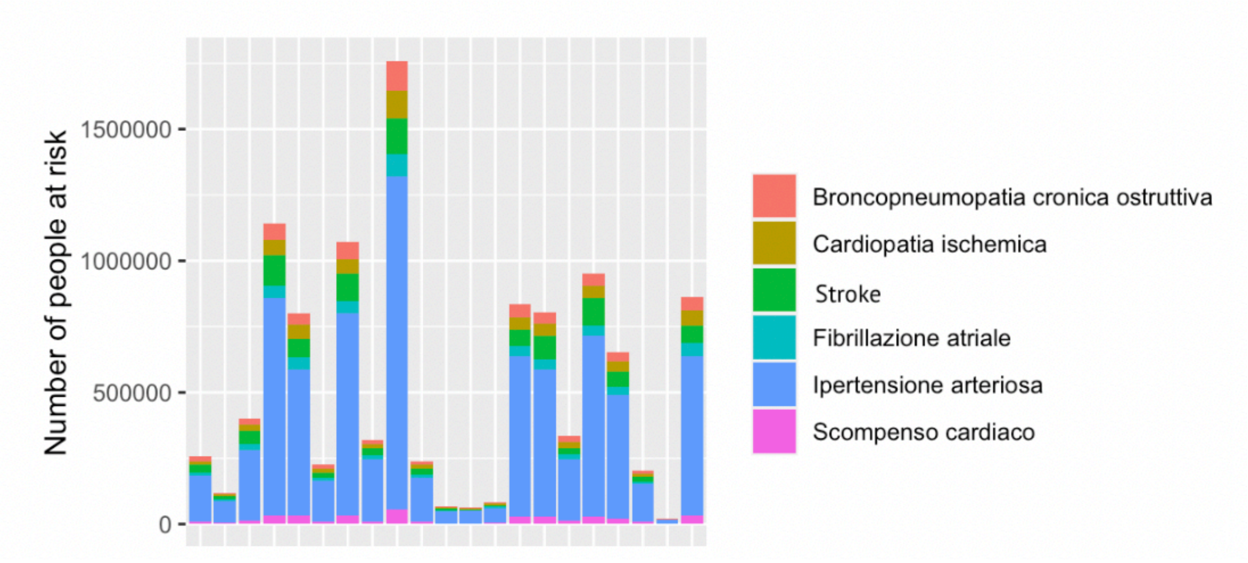


Fig. 1-10 Covid death pathologies

The resulting graph shows that hypertension, against the background of the virus disease, is the greatest 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.

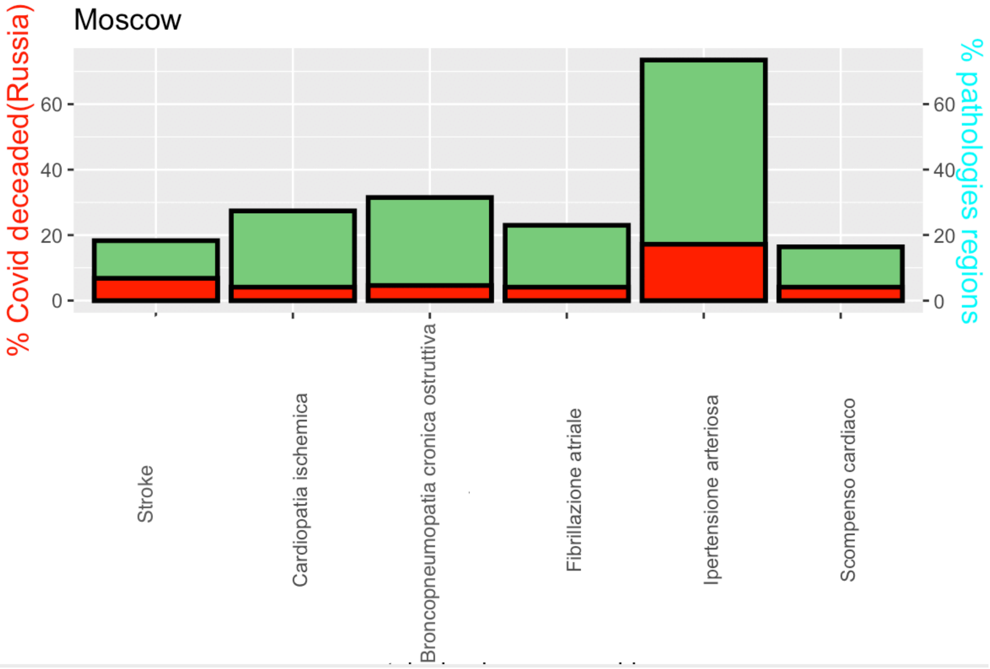


Fig. 1-11 The result in the Moscow

In Kazan, the percentage of pathologies among the population is dominated by hypertension, followed by stroke.

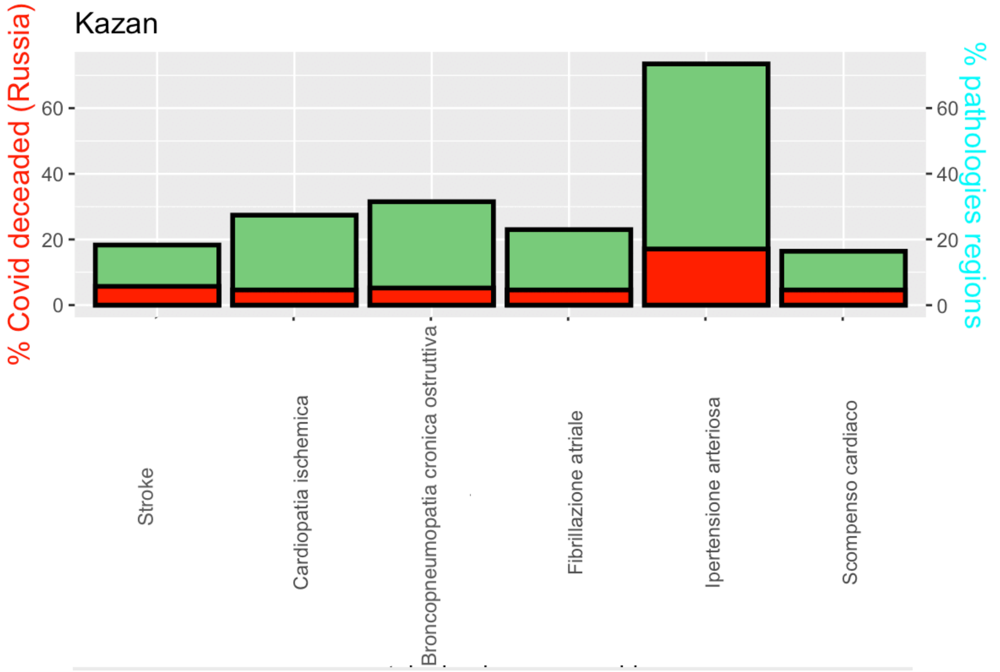


Fig. 1-12 The result in the Kazan

In St. Petersburg, the percentage of pathologies among the population is dominated by hypertension, followed by stroke.

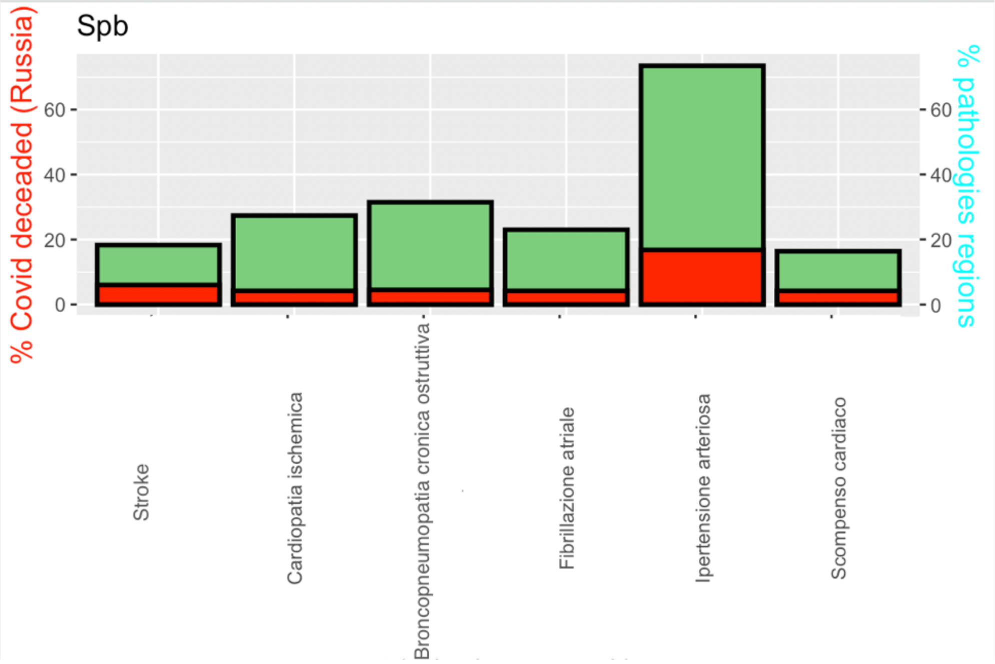


Fig. 1-13 The result in the St. Petersburg

Covid_19, according to Neurosurgeon A. Kashcheeva, can cause serious changes in the entire blood system [7] and, against the background of coronavirus pneumonia, as the infection primarily originates as a viral infection that affects the respiratory system, almost every third recorded thrombotic complications, which directly affects the formation of blood 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 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.

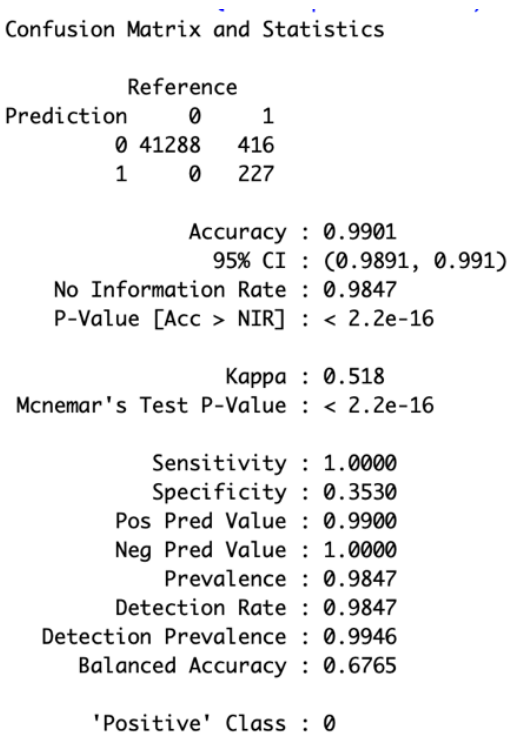
Chapter 2

Practical aspects. Theory of Probability

2.1 Prediction of stroke using the Bayesian Rule

In probability theory and statistics, Bayes' theorem (alternatively Bayes's theorem, Bayes's law or Bayes's rule) describes the probability of an event, based on prior knowledge of conditions that might be related to the event. For example, if the risk of developing health problems is known to increase with age, Bayes’s theorem allows the risk to an individual of a known age to be assessed more accurately than simply assuming that the individual is typical of the population as a whole. [8]

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



```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 41288   416
         1     0   227

               Accuracy : 0.9901
                 95% CI : (0.9891, 0.991)
    No Information Rate : 0.9847
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.518
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 1.0000
            Specificity : 0.3530
         Pos Pred Value : 0.9900
         Neg Pred Value : 1.0000
             Prevalence : 0.9847
         Detection Rate : 0.9847
   Detection Prevalence : 0.9946
      Balanced Accuracy : 0.6765

       'Positive' Class : 0
```

Fig. 2-1 Bayes Rule

You can see from this test:

- True positive rate (TPR) - 1.0
- True negative rate (TNR) - 0.35
- Positive predictive value (PPV) - 0.99
- Negative predictive value (NPV) - 1.0

Diagnostic ratio that the (PT) – positive test:

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

Diagnostic ratio that the (TN) – test negative:

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

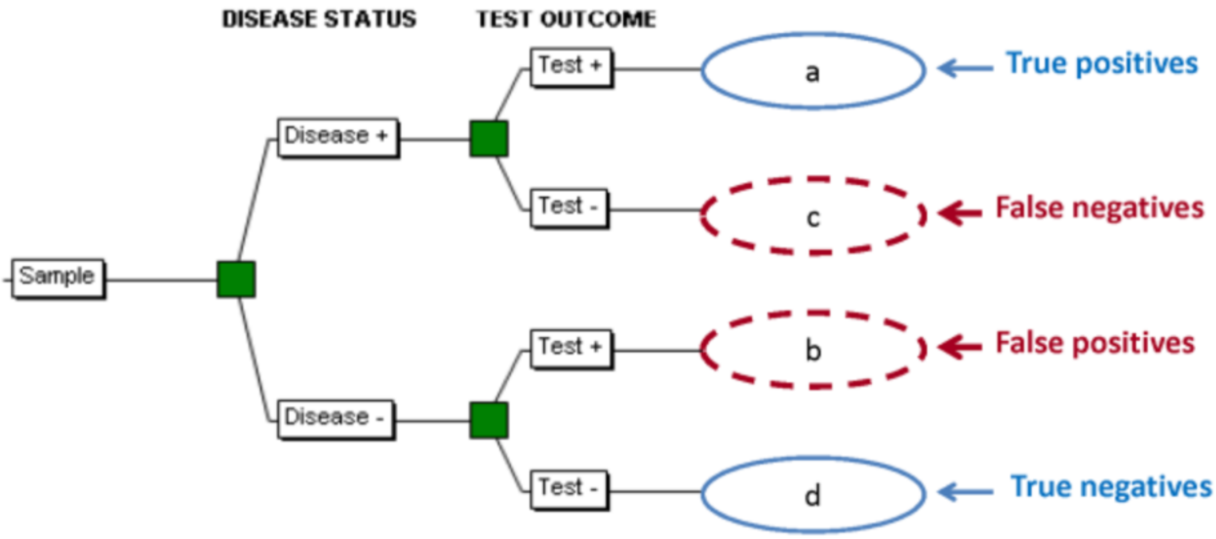


Fig. 2-2 Test outcome

A table of positive and negative prognostic values was constructed:

	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

Tab. 5 Positive and negative values

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

Thus, in this population, a positive test result corresponds to only 99% of the probability that the test subject is actually a stroke. The high positive predictive value is due to the high prevalence of the disease.

AUC и ROC

ROC is a curve that represents a graph showing the diagnostic capabilities of a binary classifier system.

The ROC curve is determined by a formula:

$$ROC = 1 - TNR = 1 - 0.35 = 0.65$$

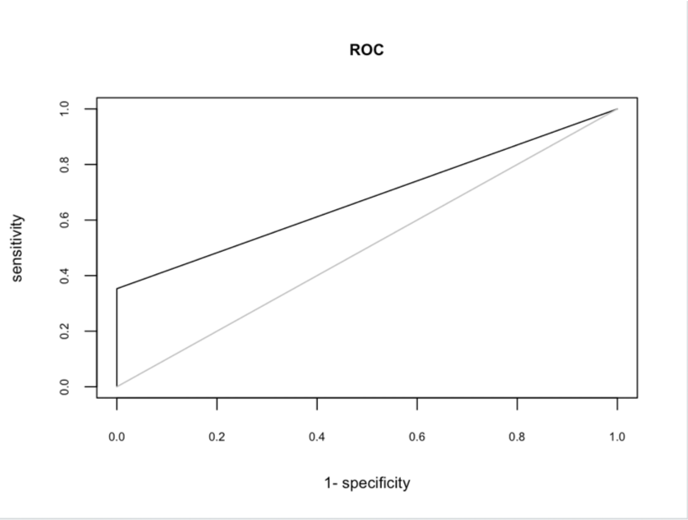


Fig. 2-3 ROC

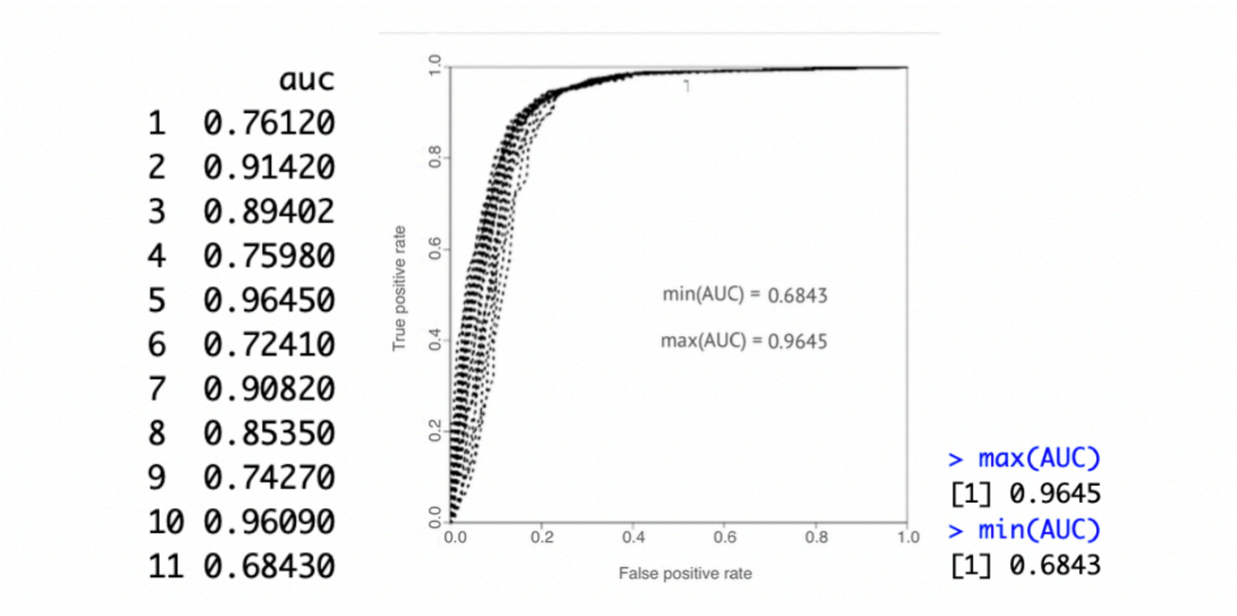
AUC:

AUC Curve Interpretation

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

Tab. 6 AUC curve

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 between a negative and a positive class.



```
       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
```

Fig. 2-4 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 ROC = 0.68

2.2 Prediction of stroke using the Law of Large Numbers (LLN)

The law of large numbers in probability theory asserts that the empirical mean (arithmetic mean) of a final sample from a fixed distribution is close to the theoretical mean (mathematical expectation) of this distribution. [9]

Checking the ratio of people to stroke and without it.

Ratio of people stroke =0,088

- Size of samples - 500
- Converting the stroke column to (0.1)
- We take X_1, X_2, X_3 random samples (size n)
- Create 3 vectors that will contain correlation information (from 1 to n)

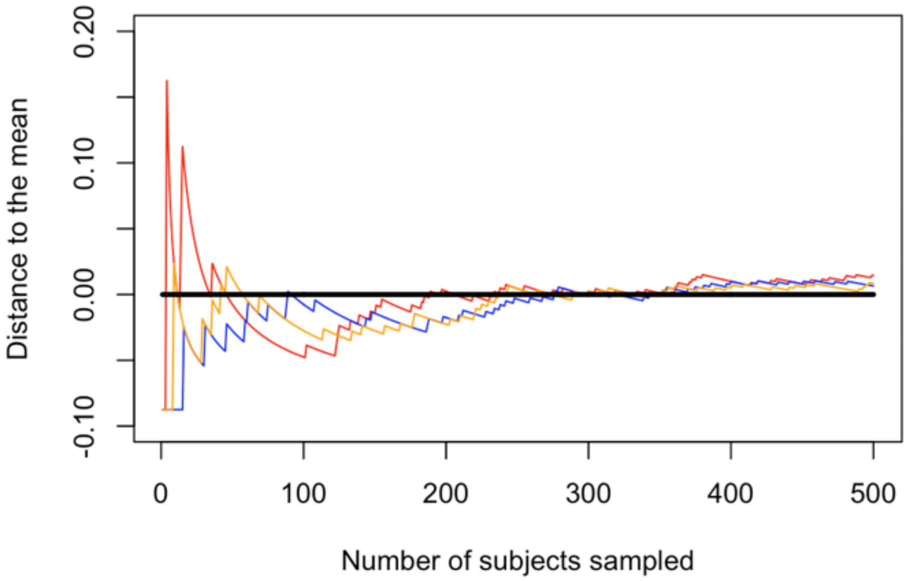


Fig. 2-5 LLN

From all of the above it can be seen that the greater the number of objects used in the sample, the higher the factor of the intermediate variant of the sample will be close to the average given indicator.

The law of large numbers states that when n is increased, the mean values are close to the target.

2.3 Central Limit Theorem (CLT)

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. [10]

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

Let us consider a Central Limit Theorem, which will tell us how close and with what probability the results of the experiment are to the true purpose.

A histogram of the body mass index in people who've never had this disease was constructed.

Average number of populations = 28,6

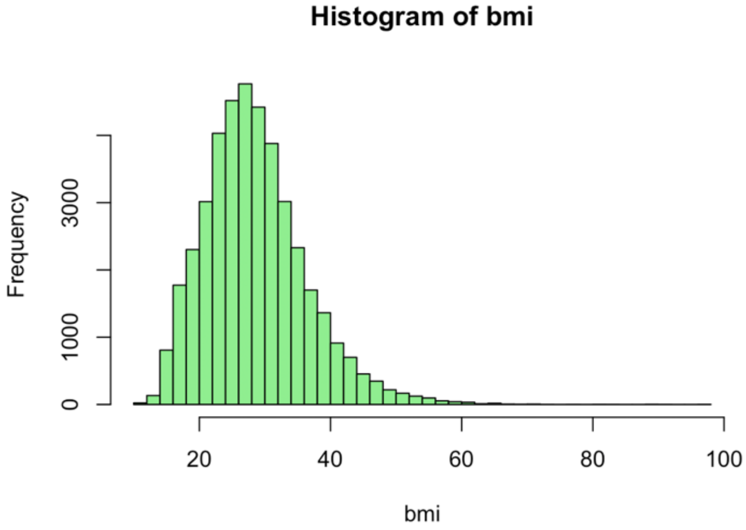


Fig. 2-6 BMI

The larger the sample, the more it will resemble a normal distribution.

Let's look at the average value of the sample n = 30:

Average number of populations = 28,7

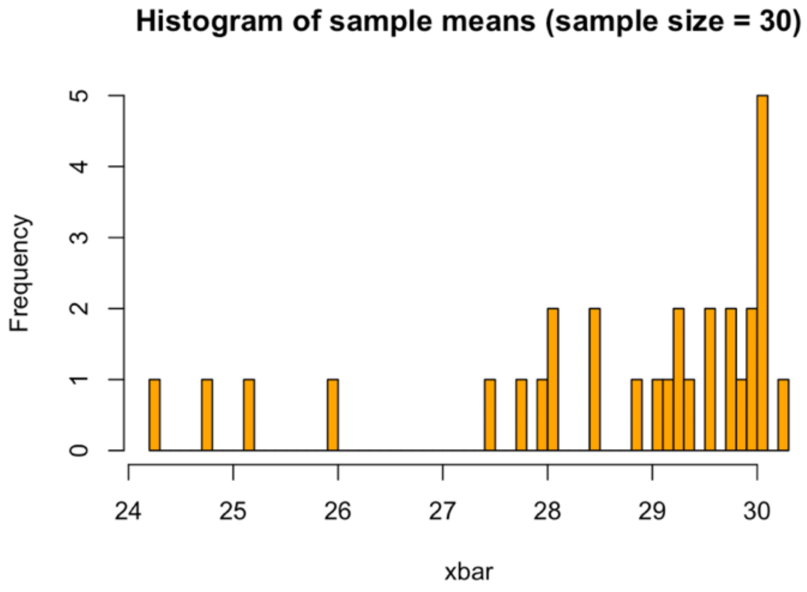


Fig. 2-7 Sample mean (n = 30)

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

Average number of populations = 29,2

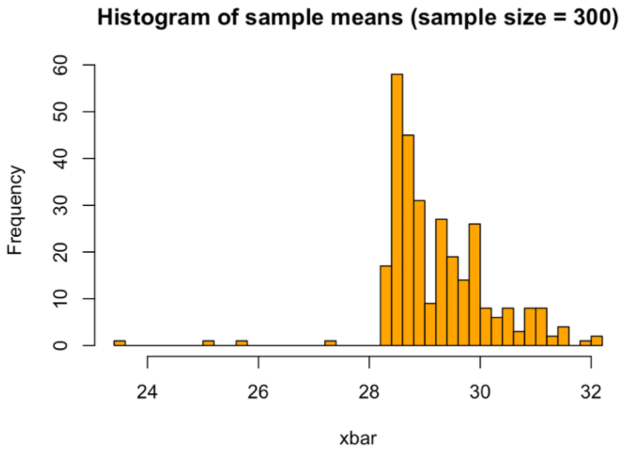


Fig. 2-8 Sample mean (n = 300)

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

2.4 Prediction of stroke with Statistical Power

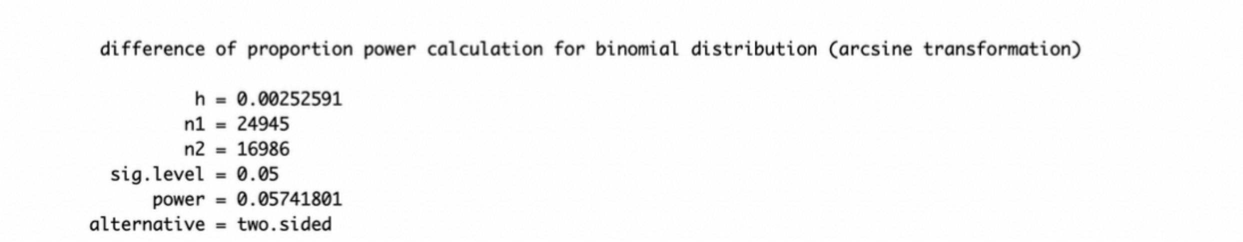
Statistical power in mathematical statistics - the probability of the main (or zero) hypothesis rejection when testing statistical hypotheses in the case when the competing (or alternative) hypothesis is correct. The higher the power of a statistical test, the less likely it is to make a second type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

With a known standard deviation of the general population and a given level of significance α = 0,05, the power $1 - \beta$ can be calculated using the Z-criterion by the formula:

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

Where $\mu 0$ there is an average in the null hypothesis, $\mu 1$ - the average in the alternative hypothesis,1,64 - the value of the critical value of Z-statistics in the one-way Z-test, and SE - the standard error. [11]

Power = $1 - \beta$
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
```

Fig. 2-9 Power Calculating

The power is only 5.741%. In other words, the probability of correct deviation of the null hypothesis is 5.741%.

For the two groups, it is necessary to find the best effect when running the test so that the difference is about 0.2.

	n	$1 - \beta$
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

Tab. 7 Different sample

To achieve a 51,60 % probability of correct deviation of the null hypothesis, it is necessary to collect information from two groups of 100 people.

2.5 Prediction of stroke using the Intervals for Binomial Probabilities

The binomial interval is a measure of uncertainty for a proportion in a statistical population. It takes a proportion from a sample and adjusts for sampling error. [12]

The formula for the CI on parameter p is:

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

Where $\hat{p}$:

$$\hat{p} = \frac{x}{n}$$

In order to estimate how hypertension is spreading, which is one of the factors of stroke development in the population, we will consider a sample of 50 people.

- The result of P-Hat = 0,04
- The result of P-Tilde = 0,97

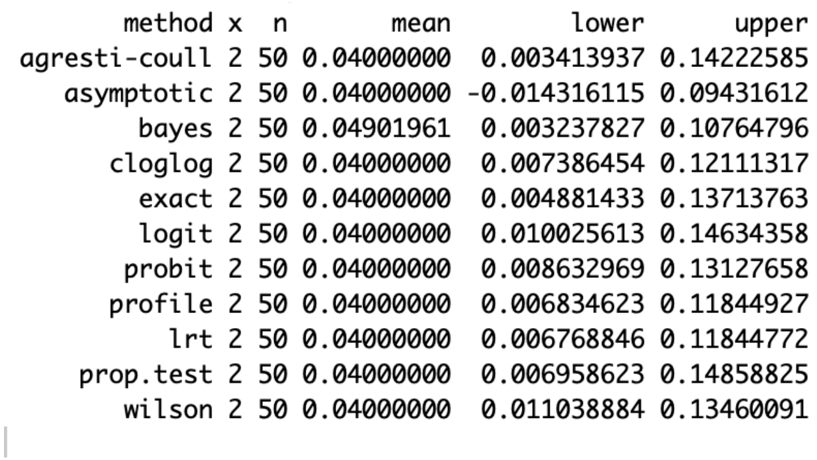
Result of 95% interval Wald:

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

Result of 95% interval Agresti/Coull:

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

Result of integrated function:



```
       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
```

Fig. 2-10 Integrated function

To calculate P-hat you need two numbers, the first number is the sample size (n), and the second number is the number of events or parameters in question (X). **P-hat** = 0.04 found, 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.

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 for cases where consists in adding two successes and two failures. That is, let $\hat{p} = (\mathrm{X}+2)/(\mathrm{n}+4)$ and $\hat{n} = n+4$. Then there is the so-called Agresti-Coull interval.

Chapter 3

Practical aspects. Systematic sampling and statistical criterion

3.1 Prediction of stroke with Confidence interval

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

Construction of a confidence interval for the mathematical expectation of the general population at a known standard deviation, where Z is the value of a standardized normally distributed random value corresponding to an integral probability equal to 1 - α/2, σ - standard deviation of the general population.

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

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:

Average value of the population: 10

Standard Error: 5

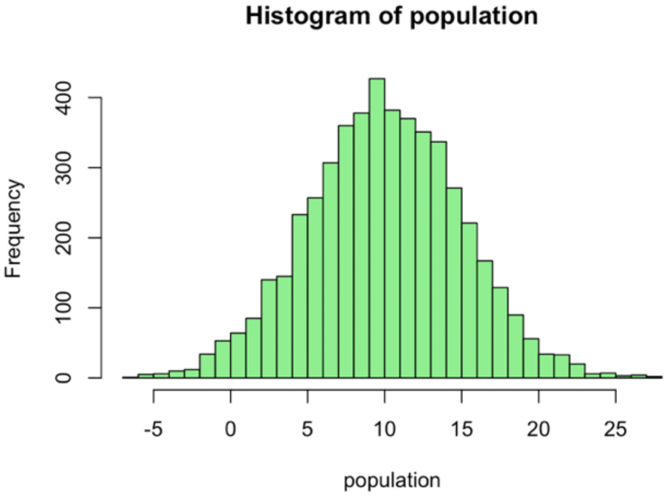


Fig. 3-1 Mean BMI of Population

At n = 250:

Average value of the population: 10.2

Standard Error: 4.9

A histogram has been constructed that shows the average value of the population:

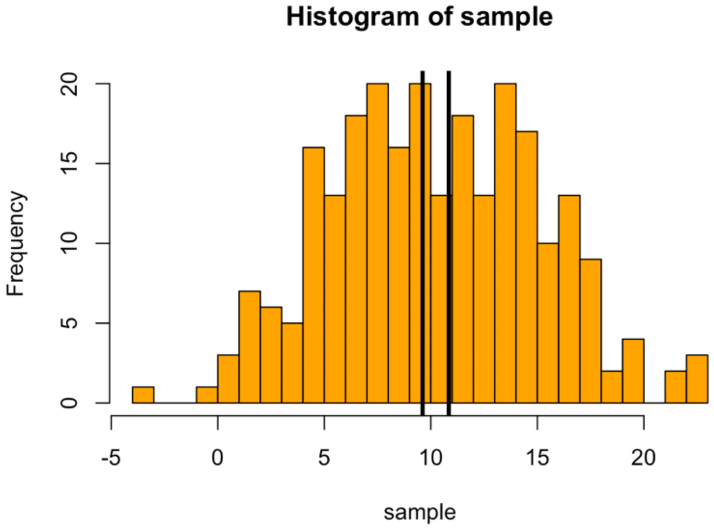


Fig. 3-2 Average BMI of population (n = 250)

According to the results of the histogram, we can say that the average BMI of the population at risk ranges from 9.6 to 10.8.

Confidence interval with the t-distributions for the average value

If the data underlying the population is distributed abnormally and/or the general dispersion (population variance) is unknown, the sample average is subject to the t-distribution.

Provides a wider interval than normal distribution because it takes into account the additional uncertainty introduced by estimating the standard deviation of the population and/or because of the small sample size.

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

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 28.5 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 indicator is a benchmark.

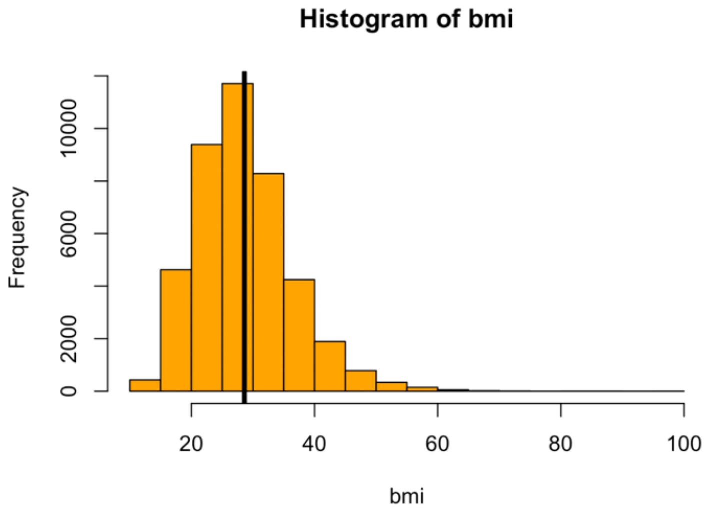


Fig. 3-3 Average of BMI

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

On the basis of the sample for 100, was found the confidence interval for the average value of the human mass index.

According to the results of the histogram, we can say that the average IMT of the population is in the range from 26.8 to 29.6. Where the upper limit of obesity BMI is > 29.5.

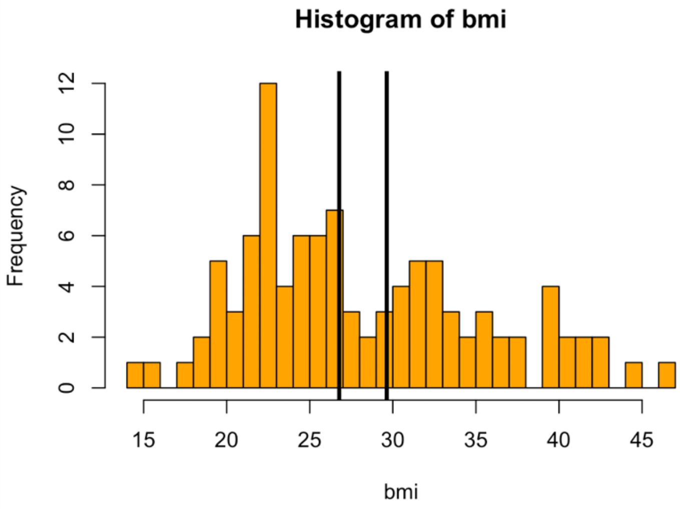


Fig. 3-4 Average of BMI (100)

3.2 Prediction of a stroke using the Bootstrapping

Bootstrapping 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 to easily and quickly estimate various statistics (confidence intervals, variance, correlation, etc.) for complex models.

The idea of the bootstrapping is to use the results of sample calculations as a "dummy population" in order to determine the sampling distribution of statistics. In fact, it analyses a large number of "phantom" samples called bootstraps. The bootstrapping is randomly selected with a return, the selected items in the original sample are returned to the sample and can be selected again. In bootstrap we do not get new information, but we use the available data reasonably based on the task at hand. Bootstrapping is better used for small samples, for median estimates, correlations, confidence intervals and other situations.

$$y_i = \theta_{x_i} + \epsilon_i$$

where θ_{x_i} - parameter estimation obtained using the least squares method.

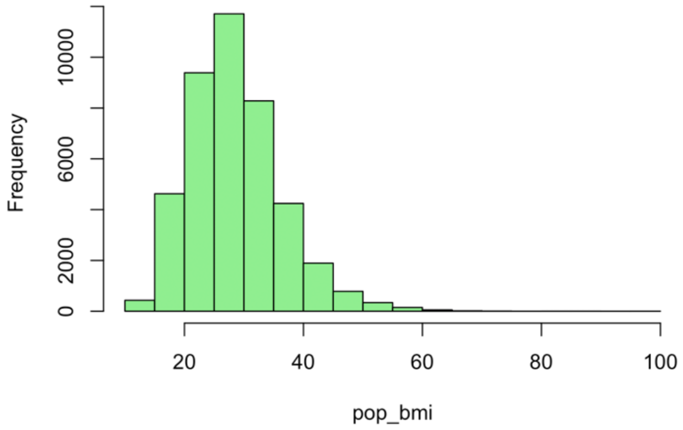


Fig. 3-5 Pop_bmi

Two histograms were constructed, with a sample of 500 pieces, as well as re-sampling. Both histograms look the same as they should, as the right histogram was obtained by an image from the left histogram.

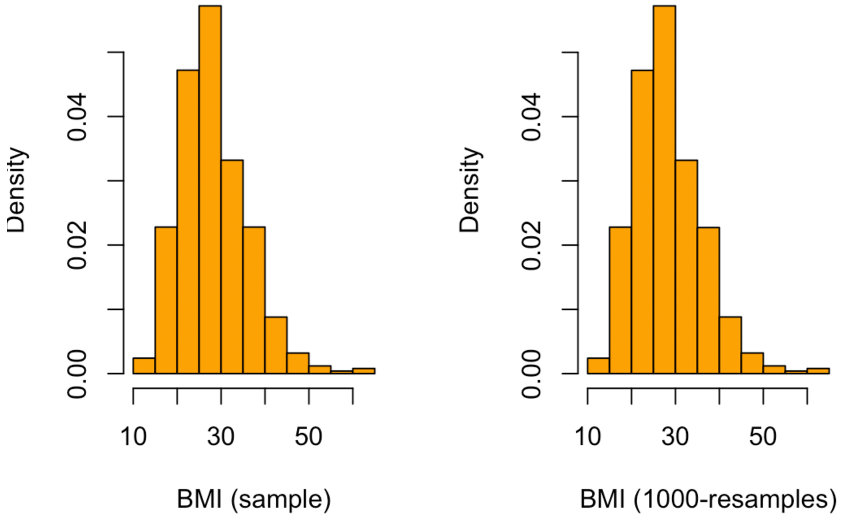


Fig. 3-6 Re-sampling

1. Calculation of average value for each recalculated data
2. Building a histogram with a modified sample
3. Standard Error: 0.346

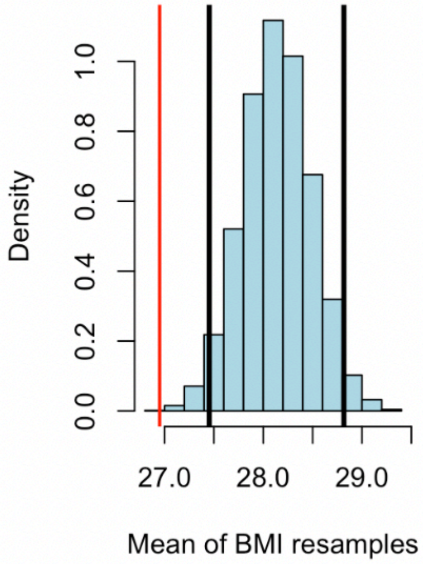


Fig. 3-7 Histogram with a modified sample

With the function in R bootstrap, the following result was obtained:

Standard Error: 0.343

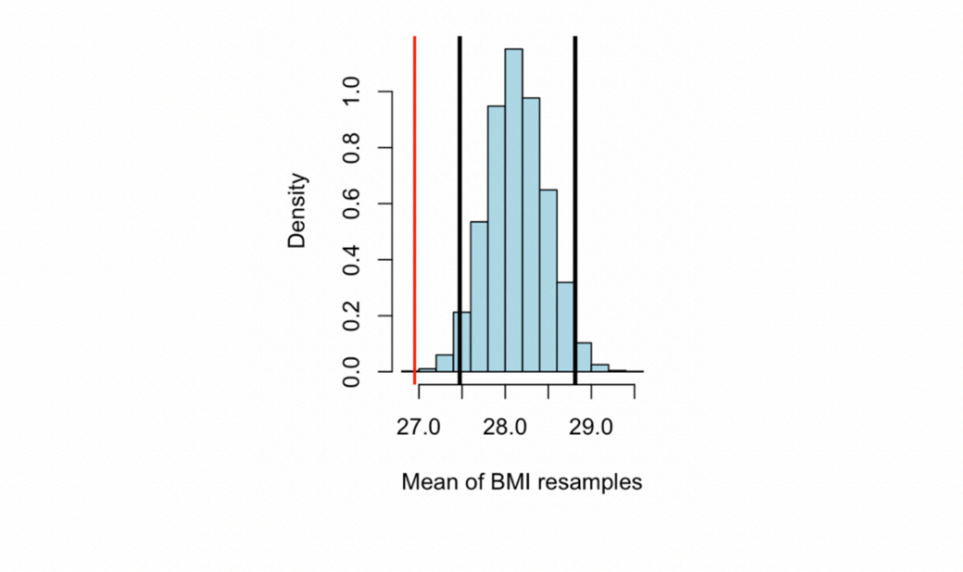


Fig. 3-8 Histogram with bootstrap function

The re-sampling gave a false result. Due to the fact that the sample was small and had some deviations. When rerunning the code often, you could see that the average value of the re-sample proportions was different from the average value of the population.

3.3 Prediction of stroke using the Null and Alternative Hypothesis

When testing the significance of a hypothesis, it should be formulated independently of the data used in its testing. In this case you can get a really productive result.

A null hypothesis (h_0) is always checked, which rejects an effect (for example, the difference of averages is equal to zero) in a population.

For example, when comparing glucose rates in the male and female populations, the null hypothesis h0 would mean that glucose rates are the same in the female and male populations.

An alternative hypothesis (h_1) is then determined and accepted if the null hypothesis is incorrect. The alternative hypothesis is more related to the theory that is going to be investigated. So, on this example, the alternative hypothesis (h_1) is to claim that glucose rates are different for women and men in a population.

The glucose difference has not been clarified, i.e. it has not been determined whether the male population has higher or lower glucose rates than the female population. This approach is known as a two-way approach because it takes every opportunity into account and is recommended insofar as there is rarely any certainty in advance that any difference, if any, will occur.

In some cases, a unilateral criterion can be used for hypothesis (h_{1}), in which the direction of the effect is given. It can be applied, for example, if we consider a disease from which all patients who have not received treatment have died; a new drug would not make things worse.

The result of Z score = $\frac{\sqrt{n}(\bar{x}-\mu 0)}{S} =$ 6.47

qnorm = 1,96 > quintile

Gender	Average	STD	N
Male	105	44	17000
Female	103	41	25000

Tab. 8 Mean glucose level

Let's assume that the average BMI for the population is over 30. now let's run a one-way test.).

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

This test has yielded a result, a score > 0.95 quantile, so we cannot refute the null hypothesis as in the previous test.

CI and Hypothesis

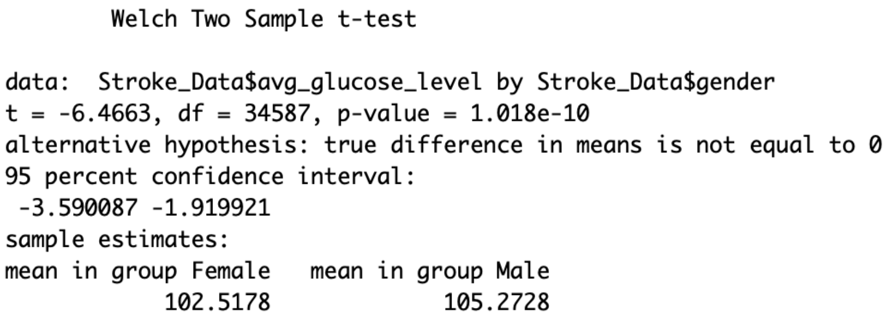
This interval doesn’t include 0, of which we can reject H0, which gives some idea of the probability distribution that gave rise to the observed data sample.

$$H0\colon\ \mu 1 - \ \mu 2 = 0$$

$$Ha\colon\ \mu 1 - \ \mu 2 \neq 0$$

Result CI and Hypothesis: 1,92 and 3,59

Result of T-test:

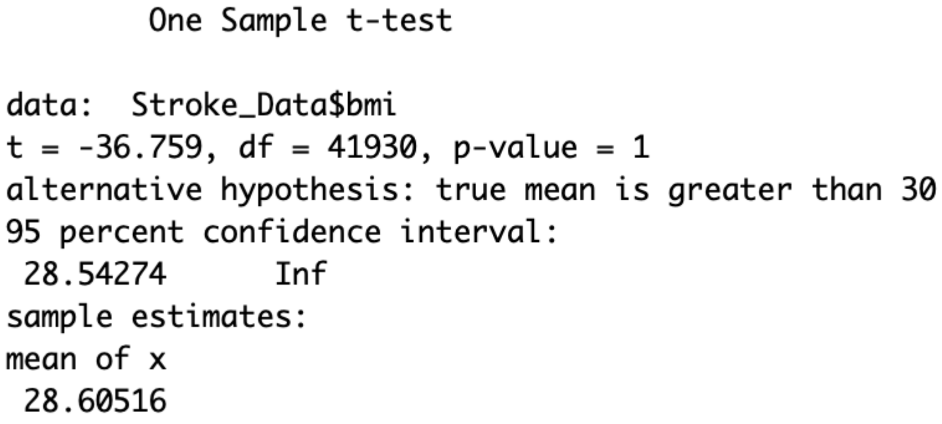


```
        Welch Two Sample t-test

data:  Stroke_Data$avg_glucose_level by Stroke_Data$gender
t = -6.4663, df = 34587, p-value = 1.018e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.590087 -1.919921
sample estimates:
mean in group Female   mean in group Male
            102.5178             105.2728
```

Fig. 3-9 T-test (hypothesis)

Result of BMI:



```
        One Sample t-test

data:  Stroke_Data$bmi
t = -36.759, df = 41930, p-value = 1
alternative hypothesis: true mean is greater than 30
95 percent confidence interval:
 28.54274      Inf
sample estimates:
mean of x
 28.60516
```

Fig. 3-10 BMI (hypothesis)

P – Value

Let:

H_0 - stroke Male and Female are the same.

H_a - stroke Male and Female are different.

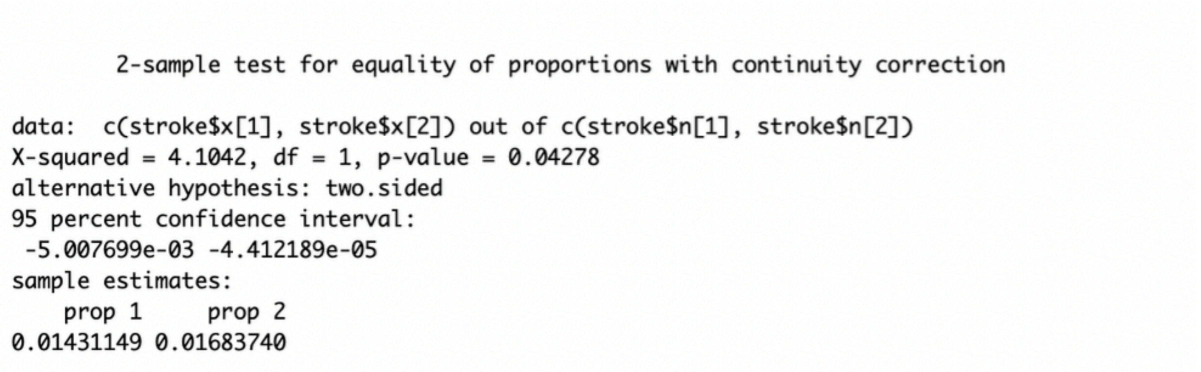
α = 0,05

P-Value = 0,043

Level α = 0,05

Gender	N	X
Male	17000	286
Female	25000	357

Tab. 9 P-value



```
        2-sample test for equality of proportions with continuity correction

data:  c(stroke$x[1], stroke$x[2]) out of c(stroke$n[1], stroke$n[2])
X-squared = 4.1042, df = 1, p-value = 0.04278
alternative hypothesis: two.sided
95 percent confidence interval:
 -5.007699e-03 -4.412189e-05
sample estimates:
    prop 1     prop 2 
0.01431149 0.01683740 
```

Fig. 3-11 P-value (hypothesis)

From the result, we can see that the p-value is less than α, we can reject H0, which gives some idea of the probability distribution that gave rise to the observed data sample. This leads to the conclusion that it cannot be said that the ratio of strokes in Male and Female is the same.

3.4 Prediction of stroke T and F - tests

T-test

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

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

Let's see if the age difference is significant for people who have not a stroke and who have. Let's visually examine the age distribution of the stroke result and run a T-test.

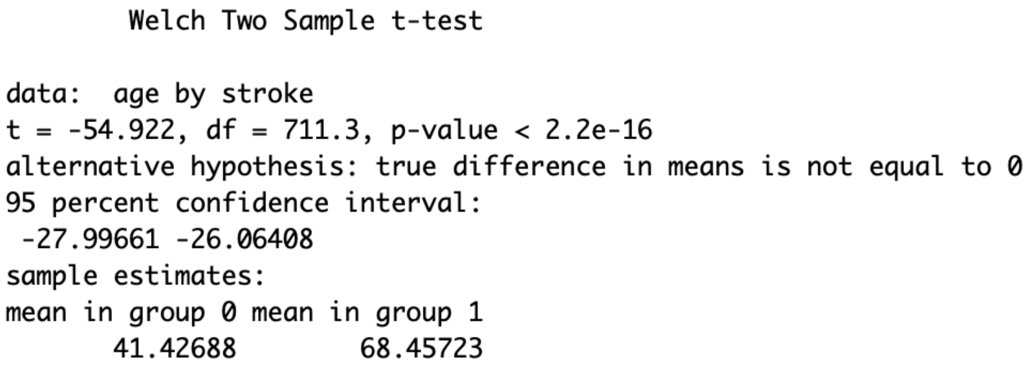
Group	N	Average	Variance
Stroke	643	68	148
No Stroke	41000	41	500

Tab. 10 Average age in people with strokes

We can see that the mean age of stroke from 41 to 68. If In order to get a result, it is necessary to compare the obtained average values between two groups, we'll get it:

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

The confidence interval of the overage differences is between -28 and -26. This interval doesn't include 0, of which we can reject H_0, which gives some idea of the probability distribution that gave rise to the observed data sample. So, there is a 95% difference between groups. From this we can conclude that stroke is much more common in older people than in younger people.
With the function in R T-test, the following result was obtained:
The test carried out, with built-in function in R, gave the same result.



```
	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 
```

Fig. 3-12 T-test

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

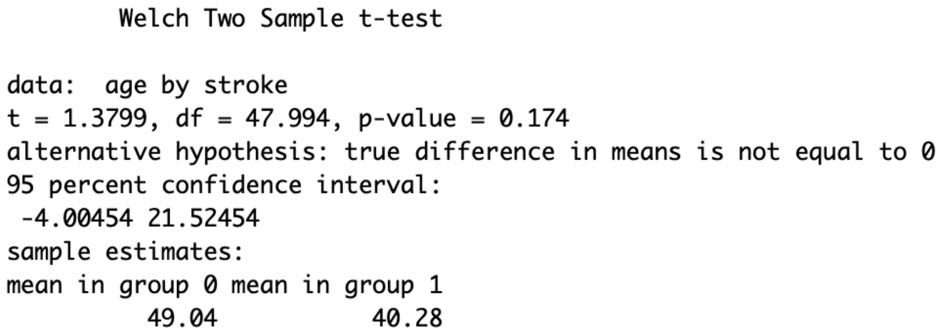
Stroke	N	Age	Variance
1	25	40	498
0	25	49	510

Tab. 11 Average age in people with strokes (less data)

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

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

With the function in R T-test, the following result was obtained:



```
            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
```

Fig. 3-13 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

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

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

General dispersions are assessed on the basis of samples, and the criterion itself is directly calculated as a ratio of one sample variance to another:

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

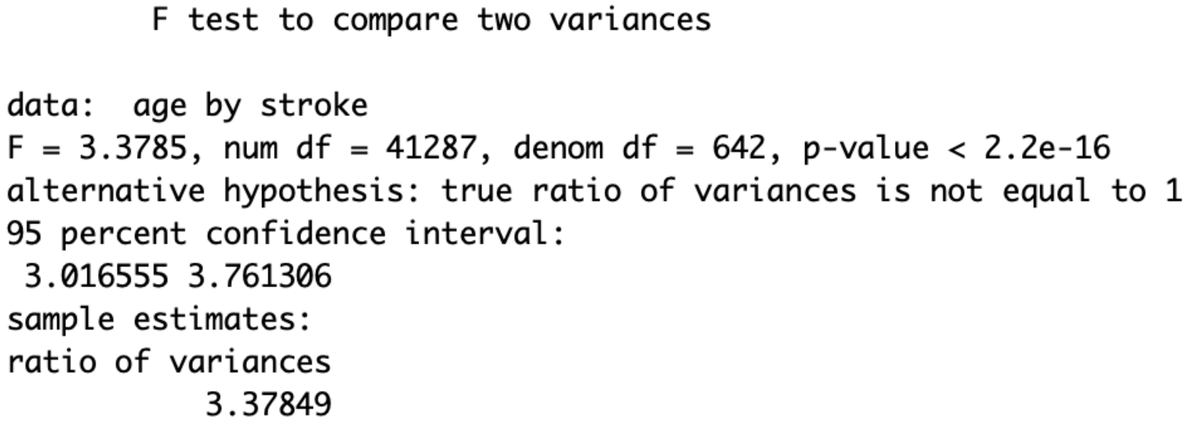
We considered a sample that showed the outcome of stroke with the age distribution of the population:

The result of ratio = 3,38

Group	N	Variance
Stroke	645	148
No stroke	41000	500

Tab. 12 F-test

A confidence interval was built



```
        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
```

Fig. 3-14 F-test (confidence interval)

The confidence interval of the overage differences is between 3,03 and 3,78. This interval doesn’t include 0, of which we can reject H0, which gives some idea of the probability distribution that gave rise to the observed data sample. So, there is a 95% difference between groups.

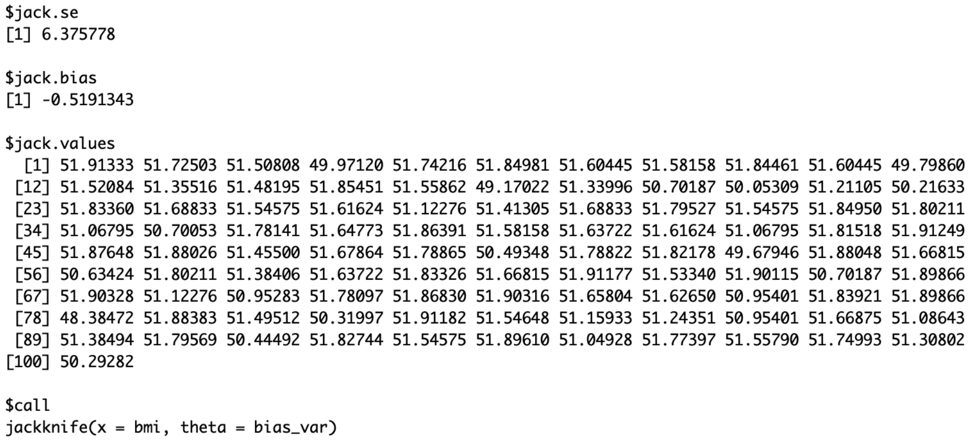
3.5 Prediction of stroke with Jackknife

The **jackknife** is one of the resampling methods (linear approximation of the statistical bootstrap) used to estimate the error in the statistical output. The method consists in 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.

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

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

There was a review of the data oversight linked to the IMT:



```
$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)
```

Fig. 3-15 Jackknife

Folding jackknife assessment standard theta error = 6.37.

Bias is an offset relative to the correct answer.

Theta applies to BMI with removal of 1 observation, theta applies to BMI with removal of 2 observation ... theta applies to BMI with removal of n observation = -0.5.

3.6 Analysis of data obtained based on forecasting methods used

After conducting experiments to determine the best prediction of stroke mortality, I obtained the percentage accuracy of predictions for different methods and displayed them in a histogram.

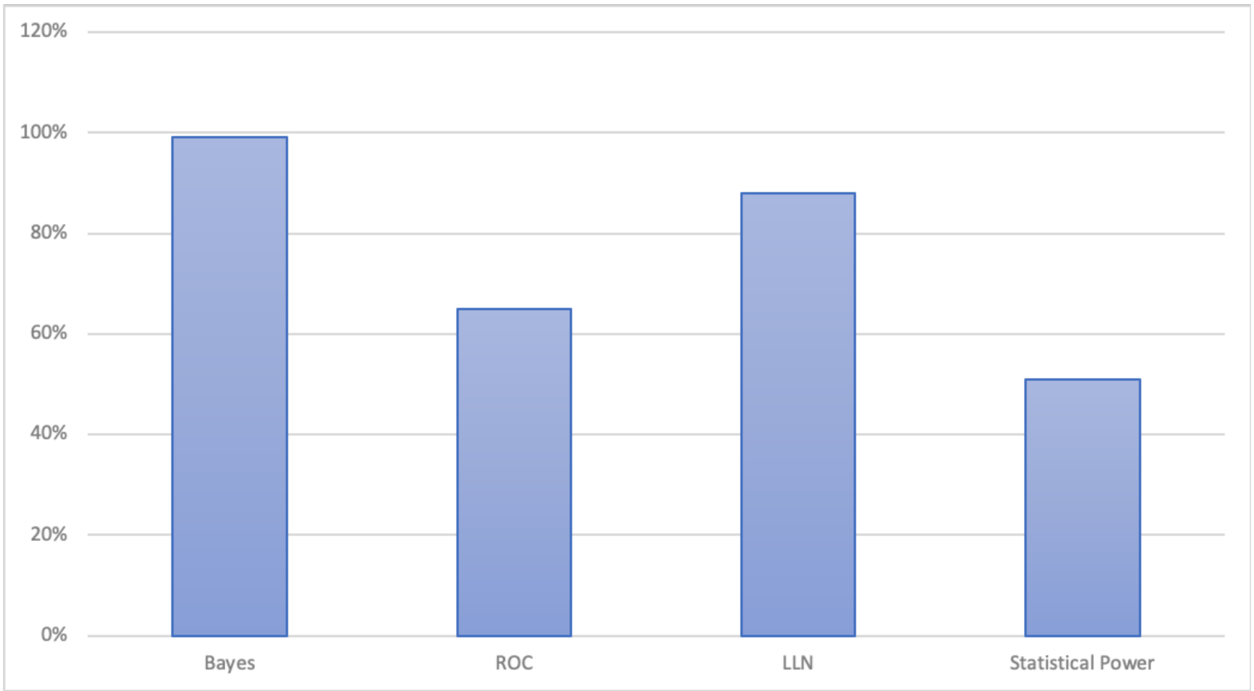


Fig. 3-1 Methods

From the histogram we can conclude that the best approaches to prevent stroke in my work are the Law of Large Numbers (LLN) prediction and the Bayes method.

Both methods involve a large number of tests.

From this we can conclude: the first method identifies the most significant recurring factors and fixes their position, the second, analyzing the available information or information that has the property of being updated, is able to 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.

Conclusion

In this way, the research project studies and explores differentiated data based on different indicators and statistics, which, being investigated in the material, will make it possible to diagnose early signs of disease and reduce the percentage of deaths from stroke.

The best approaches to stroke prevention in my work are the Law of Large Numbers (LLN) forecasting and the Bayes method. However, the Law of LLN is inherently more constructive only when a large number of trials are implied. It in turn guarantees a stable position for average events in a chain of trials. In this way, it highlights and records individual, consistently repetitive test readings. The Bayes method, in turn, determines the probability of an event based on available information and new incoming facts.

Based on general indicators and the specific repetitive symptomatology of most subjects, it is important to understand that the earlier a disease is diagnosed prior to a future pre-stroke condition, the more likely it is to be successfully treated to prevent the symptomatology and the specific case. The work being done and the comparative statistics being conducted, based on the available data, make it possible to identify a disease much faster and to eliminate or prevent the possibility of the signs of stroke in the near future or at a more mature age.

Bibliography

1. Data analysis at: [https://www.igi-global.com/dictionary/default-probability-prediction-of-credit-applicants-using-a-new-fuzzy-knn-method-with-optimal-weights/6677]
2. 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. - book. 3, № 1. – pp. 127-131
3. PyCharm at: [https://www.zdnet.com/article/pycharm-heres-what-python-programming-language-developers-get-in-new-ide-update/]
4. Data research analysis, characterization design and modelling using data at: [https://www.machinelearningmastery.ru/exploratory-data-analysis-feature-engineering-and-modelling-using-supermarket-sales-data-part-1-228140f89298/]
5. Coronavirus at: [http://kvd8.spb.ru/p59/l82/index.html](in Russian)
6. Disease caused by coronavirus (COVID-19) at: [(https://www.who.int/ru/emergencies/diseases/novel-coronavirus-2019/advice-for-public/q-a-coronaviruses)] (in Russian)
7. Coronavirus research at: [https://www.google.com/amp/s/www.mk.ru/amp/science/2020/04/29/neyrokhirurg-raskryl-podrobnosti-issledovaniy-koronavirusa-porazhaet-ne-tolko-legkie.html)] (in Russian)
8. Dan Morris. «Bayes Theorem: A visual Introduction for beginners», USA, 2016, pp. 65-80
9. Pal Revesz, Z. W. Birnbaum E. Lukacs, «The Laws of Large Numbers», USA, 1967
10. William J. Adams, «The Life and Times of the Central Limit Theorem (History of Mathematics)», 2009, pp.180-195
11. Jacob Cohen «Statistical Power Analysis for the Behavioral Sciences», USA, 1988
12. Binomial Confidence Interval at: [https://www.statisticshowto.com/binomial-confidence-interval/]
13. Kruk, I.V. «Explanatory dictionary of psychiatric terms»/ I.V. Kruk, V.M. Bleicher. - Voronezh: NPO "Modec", 1995. - – 221 pp.
14. Bilich, G.L. «Popular medical encyclopedia». - Moscow: Veche, 2012. - – 399 pp.
15. Mar G. George MD «Annals of Neurology», USA, 2011
16. OCED INDICATORS «Health at a Glance 2019», England, 2019, pp.243
17. I.S. Shorokhova, N.V. Kislyak, O.S. Mariev, Statistical Methods of Analysis, Ekaterinburg: Ural University Press, 2015.300 pp
18. Pankov A., Goryainova E. R., Zhernosek A. I., Statistical methods of data processing, Moscow: Moscow Aviation Institute, 2013.382 pp