Summarizing work - Statistical Theory

Causes of heart disease

By Noam Farhi and Yair Gilady

Abstract

Heart disease is a leading global cause of death, driven by lifestyle factors such as poor diet, lack of exercise, and smoking. The rising prevalence of cardiovascular conditions places significant pressure on healthcare systems, making it crucial to identify and manage risk factors. This research focuses on using statistical methods to analyze how variables like age, cholesterol and physical activity contribute to heart disease risk, aiming to provide insights for more effective prevention strategies. Ultimately, we seek to determine what actions can be taken to avoid heart diseases by identifying controllable factors versus those that are beyond our influence.

Our analysis identifies age, smoking, and cholesterol as significant predictors of heart disease.
These findings highlight the importance of lifestyle changes, particularly maintaining healthy cholesterol levels and avoiding smoking, to reduce heart disease risk.

Github: https://github.com/yairgilady/Statistical_Theory_Final_Project

Introduction

Heart disease, or cardiovascular disease (CVD), is a major global health issue, accounting for millions of deaths each year. It encompasses a range of conditions that affect the heart and blood vessels, such as coronary artery disease, heart attacks, and heart failure. The increasing prevalence of heart disease is closely linked to modern lifestyles, aging populations, and the rise of risk factors such as high blood pressure, diabetes, and poor dietary habits.

The goal of this study is to analyze heart disease risk using statistical tools and methods. By applying statistical theory to available data, we aim to identify key predictors of heart disease and evaluate how different factors, such as lifestyle choices, age, and health indicators, influence the likelihood of developing cardiovascular conditions. This research will contribute to understanding the statistical relationships between these variables and help identify potential strategies for heart disease prevention.

Results

In addition to whether a person has heart disease or not, we have another 15 features in our dataset. In order to calculate the correlations between them, we will classify them according to numerical and categorical variables:

Numerical: age, cholesterol, blood pressure, heart rate, exercise hours, stress level, blood sugar and heart disease (yes or no – so technically categorical)

Categorical: gender, smoking, alcohol intake, family history, diabetes, obesity, exercises induced angina and chest pain type

Correlation of numerical features with heart disease

In the numerical features, we will calculate 3 types of correlations: pearson, spearman and point-biserial, the features that their results are significant to us are:

feature	pearson	spearman	Point-biserial
Age	Pearson's r = 0.6469, p-value = 0.0000	Spearman's rho = 0.6514, p-value = 0.0000	Point-Biserial r = 0.646, p-value = 0.0000
Cholesterol	Cholesterol: Pearson's r = 0.3650, p-value = 0.0000	Spearman's rho = 0.3668, p-value = 0.0000	Point-Biserial r = 0.365, p-value = 0.0000

Correlations of categorical features with heart disease

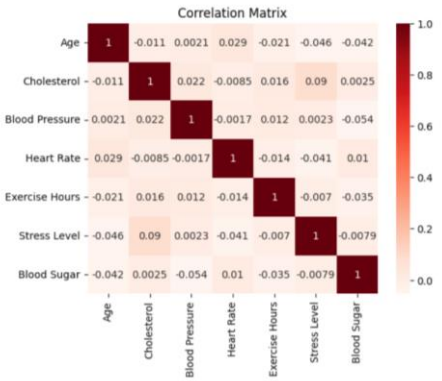
In the categorical features, we will calculate their Cramér's V correlations.
The results of the significant features are:

Feature	Cramér's V
Smoking	Cramer's V = 0.0909, p-value = 0.0161

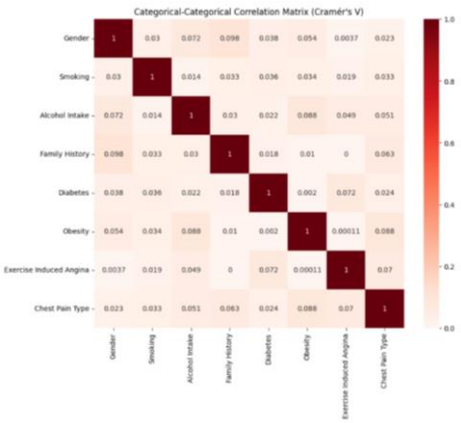
We saw that there is a significance correlation between getting a heart disease and the features Smoking, Age and Cholesterol. Later we will try to pinpoint how those three affect heart disease.

After we saw the connection between the different features and heart disease, we will check the correlations between the features themselves. We will show the findings in 3 correlation matrixes – numerical & numerical, numerical & categorical and categorical & categorical:

Numerical features



Categorical features



Numerical & categorical features

For conclusion, it doesn't seems like there is a big correlation between any of the features. That is a good thing, because it should help our machine learning model in his prediction.

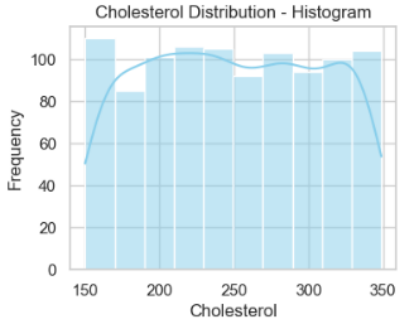
We will now focus on each of the three features that we have seen to have a strong connection to getting heart disease.

Prediction using machine learning

We split the data into train and test and then we run several models on the data and ended up choosing gradient boosted trees (GBT) as our model. With this model we were able to predict 100% of the data correctly.

Exploring cholesterol

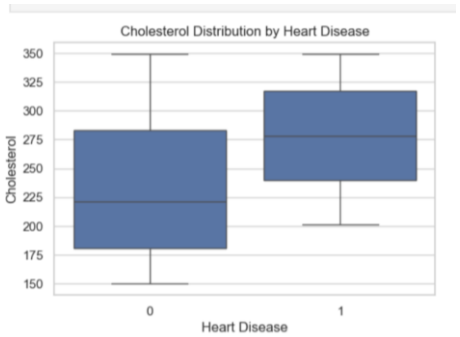
Let's start by looking at how the data distribute:



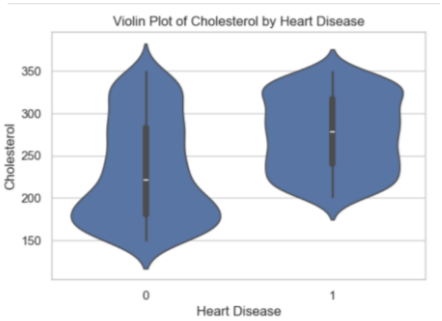
We can see from the graph that the distribution of cholesterol isn't normal, we also che cked this with the Shapiro test.

Now, if we will look at the minimum and the maximum cholesterol of people who don 't have heart disease compared to the people who do, we can see that for people with h eart disease the min: 201, max: 349. And for those that don't have heart disease - min: 150, max: 349.
To see it more clearly we will create plot box and violin graph:

Plot box

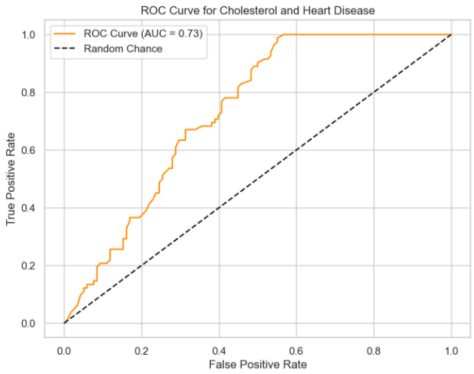


Violin plot



From looking at these graphs, it looks like most healthy people are in the 150-200 cholesterol range, while sick people are almost never there. Following this, we will try to check - whether it is possible to accurately find a cholesterol value in which there has been a change (the chance of getting sick).

We can use logistic regression to try and find this "turning point":



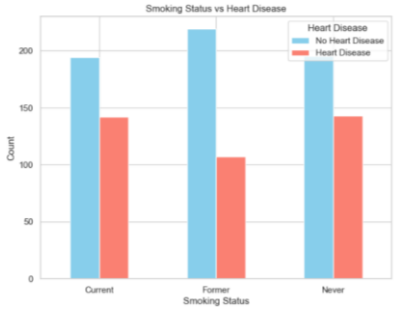
We see that the Optimal threshold for Cholesterol: 0.2405 and the
Cholesterol level corresponding to the probability threshold: 204.99 mg/dL

We tested our results statistically using two-proportion z test and found out that there is a significant difference from cholesterol above 205 mg to below it ($p = 0.000$). Meaning we should aim to keep out cholesterol lower than that. Just to be sure, we searched for another "turning point" where the chance for heart disease rises, just for cholesterol higher than 205 in jumps of 10 – we found out that there isn't any.

Just to be sure, since our data of cholesterol isn't normal we run another Polynomial Logistic Regression and saw that the optimal threshold for Cholesterol based on polynomial model is: 0.2199 and that the cholesterol value corresponding to the optimal threshold is 205.00 mg/dL - We got pretty much the same result, which is quite nice.

Exploring smoking

If we will look specifically on smoking, in relation to whether the person has heart dis ease:



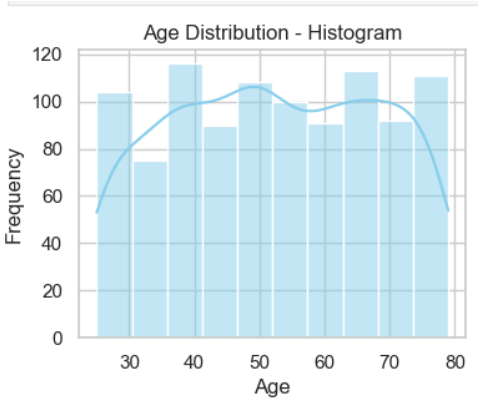
We also can see that the percentage of Heart Disease in each Smoking category:
Current: 42.2619, Former: 32.8220, Never: 42.3076, so according to this information, "Former" is the best option.

Exploring age

Now we will explore the key feature of heart disease: "Age". Starting with some graphs to get some information of how things are.

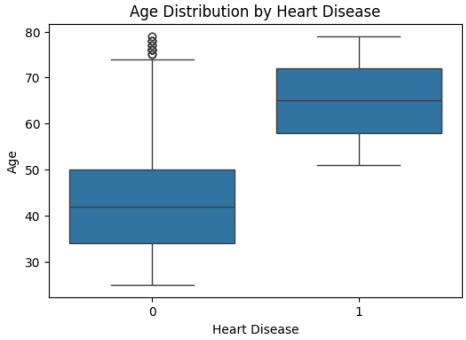
The distribution of "age":

Age	Number of people
0-10	0
11-20	0
21-30	92
31-40	165
41-50	19
51-60	181
61-70	184
71-80	182
81-90	0



After testing the feature we saw that the distribution of "age" isn't normal, just like our assumptions from the plot suggested.

Let's create a plotbox and violin plot for age:



Looking at the violin graph, we can see that from about 50, people chances of getting a heart disease spikes, and there isn't a single person with a heart disease below 40. We will focus on people above 50, and try to see if the chances rises up again using some statistical tests.

We split the data into 2 groups: age between 50 and 65, and age between 65 and 80. After running Two-proportion z-test and saw that there is no statistically significant difference in the proportion of heart disease between them.

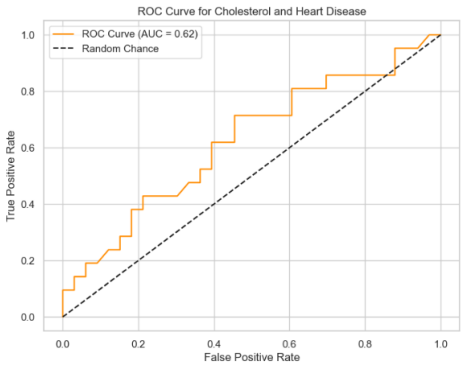
This result is very surprising, because you would except that the higher the age is, the more people have a chance of heart disease, but we have shown that if you are above 50, you chances of getter a heart disease are roughly the same as if you are 70.

Cross referencing our results

Now, in order to check the correctness of our results and improve them - we will cross our data set with another 2 data sets related to heart diseases.

Cholesterol

We took a second dataset to see if our results of 205 being the "turning point" is still true.
After running logistic regression:



We got that the Cholesterol level corresponding to the probability threshold is 249.00 mg/dL - a different result from our original 205, but not so far away.

We checked it statistically and saw that there is a significant difference in heart disease cases between the two groups (p = 0.000).

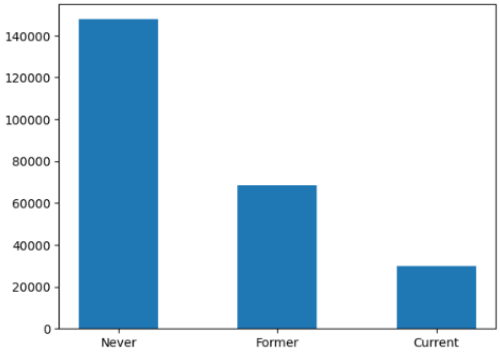
And it does seems like the correct point, since our p-value is low.

We got our result, so after comparing the two datasets and looking at the result for where it's more likely to get a heart disease, we'll just take the mean as our final answer: 227 mg/dL (but just to be certain, our recommendation will be to not go over 205 mg\dL).

Smoking

We took the only features that were relevant to us in a third data set and left "HadHear tAttack" and "SmokerStatus". After fixing them, they look like this:

Didn’t get heart attack: 232587
Got heart attack: 13435



We checked in percentage, and tried to see if Former is really the best.

Smoker status	Percentage of getting a heart attack
Former	0.081866
Current	0.079105
Never	0.037032

We did get a completely different result, and now we see that the best is not being a smoker at all, and there isn't a really big difference in former and never.

After we checked this using Chi2 test we got that there is a statistically significant difference between smoking categories and heart attacks (reject null hypothesis), and with Z-test we got that there is no statistically significant difference between 'Former' and 'Current' smokers (fail to reject null hypothesis).

After checking is statistically we can be assure that someone who never smoked has a much better at not getting a heart disease, while there isn't a difference between a curr ent smoker and a former one.
That means that either this dataset has a bias which prevent it from getting the right results, or our original data have bias.

Age

By taking the second data set again, which contain additional data on heart disease and age here, our goal is to check if our original data set was derived from an i.i.d patient list or not.
After we split the data again into the 2 age groups we created earlier and run proportio n z test again we got again a high p-value meaning we don't reject H_0 (there isn't any difference between the groups), and any difference between the groups is probably coi ncidence.

This bring us to our first result, which is that from age 50 till 70, you have the same chance of dying from a heart disease every moment, and it's not rising or slowing down in any meaningful manner. Moreover, being below 50 pretty much mean that you have close to zero chance of having a heart disease.

Methods

Correlations:

Pearson - measures the linear relationship between two continuous variables, providing a value between -1 and 1, where 1 indicates a perfect positive correlation and -1 indicates a perfect negative correlation.
Spearman - assesses the strength and direction of the association between two ranked variables. It is a non-parametric measure, making it suitable for data that do not meet the assumptions of normality.

Point-baserial - quantifies the relationship between a binary categorical variable and a numerical variable. It is similar to Pearson correlation but specifically used when one variable is with 2 categories.
Cramér's V – a measurement of association between two categorical variables. It's giving out a value between 0 and 1 (higher value - the variables are more correlated).

Parametric hypothesis tests:
Z test – a statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution

Non parametric hypothesis tests:

Chi square test - a statistical test used to determine if there is a significant relationship between two categorical variables.
Shapiro-wilk test - a statistical test used to assess whether a dataset follows a normal distribution.

Graphs:

Plot box - a method for demonstrating graphically the locality, spread and skewness groups of numerical data through their quartiles.
Violin graph - a statistical graphic for comparing probability distributions. It is similar to a box plot, with the addition of a rotated kernel density plot on each side
Roc graph - A receiver operating characteristic curve - a graphical tool that evaluates the performance of a binary classification model. It helps determine the optimal threshold for classification.

Logistic regression - a statistical method used to model the relationship between a binary dependent variable and one or more independent variables.

Discussion

Our analysis indicates that smoking, age, and cholesterol are significant factors in heart disease. The risk can be reduced by maintaining cholesterol levels below 205 mg/dL and avoiding smoking, with non-smokers being at the lowest risk. However, we observed no significant distinction between current and former smokers. Regarding age, we found that the risk of heart disease increases significantly after age 50, but there was no further increase beyond that age.

It's important to note that smoking and cholesterol are key factors that can be influenced by lifestyle changes. On the other hand, age is beyond our control but still impacts the risk, that’s why taking early prevention and regular checkups for those over 50 is crucial.

Despite these findings, there are several limitations to the analysis. The data may not represent the entire population, and measurement errors in variables like smoking and cholesterol are possible. Additionally, the sample may be biased in certain age groups, and important genetic influences were not considered in the analysis.