STAT 240 Project Draft

Group

* 312 A

Names

* Colm Fitzsimmons
* Eric Choi
* Sanaya Kapadia
* Yingwei Song

**Introduction**

Cardiovascular disease is one of the most frequent diseases that threaten people’s health problems. Discussing the relationship between different factors which might cause cardiovascular disease might help us to prevent it. For example, whether there is a relationship between age, smoke, alcohol and cardiovascular disease? These questions could be extremely useful when examining the relationship between different pieces of information related to the disease.

In brief summary, this report will reveal the positive relationship between age, alcohol, tobacco intake, and cardiovascular disease. Also, the analysis will show it is a genetic disease, and 50% percent of patients have a family history. Finally, people with cardiovascular disease will have higher systolic blood pressure between the range of 4.07 mm Hg and 12.48 mm Hg compared to healthy people.

**Background**

For this project, we chose to work with the Cardiovascular Disease data set which was submitted by Yassine Hamdaoui on [Kaggle](https://www.kaggle.com/yassinehamdaoui1/cardiovascular-disease). It recorded detailed information collected through medical examinations from a larger data set in 1983. The larger original cardiovascular disease data set consists of 462 observations with 10 variables. The data set is a retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. And the full data set was described in Rousseauw et al, 1983, South African Medical Journal. The detailed source is: Rousseauw, J., du Plessis, J., Benade, A., Jordaan, P., Kotze, J. and Ferreira, J. (1983). Coronary risk factor screening in three rural communities, South African Medical Journal 64: 430–436. ElemStatLearn, R-Package.

The data set has 10 variables, but we will only focus on the following key variables:

* sbp: systolic blood pressure
* tobacco: cumulative tobacco (kg)
* famhist: family history of heart disease, a factor with levels “Absent” and “Present”
* alcohol: current alcohol consumption
* age: age at onset
* chd: response, coronary heart disease

Data limitations declaration:

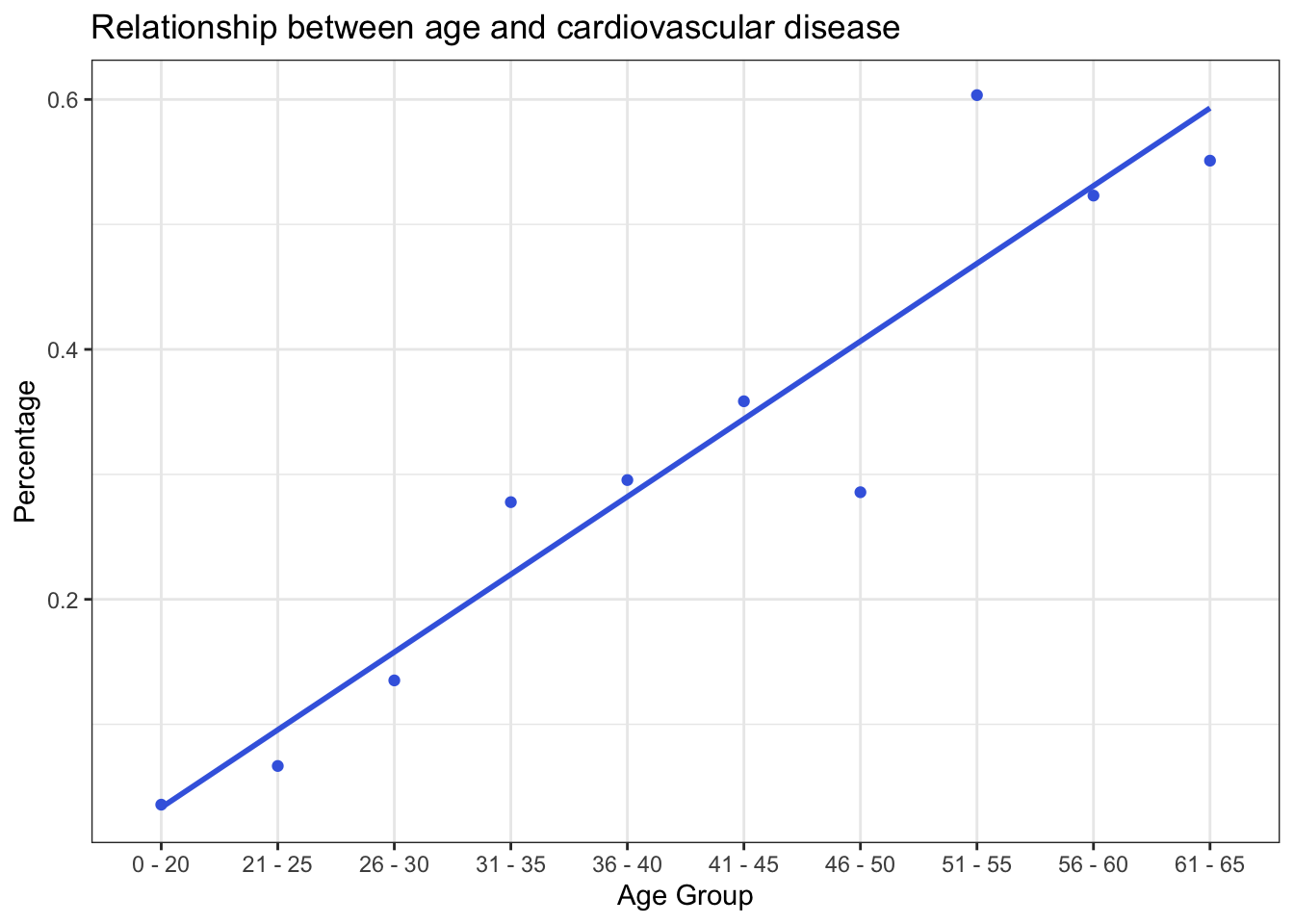
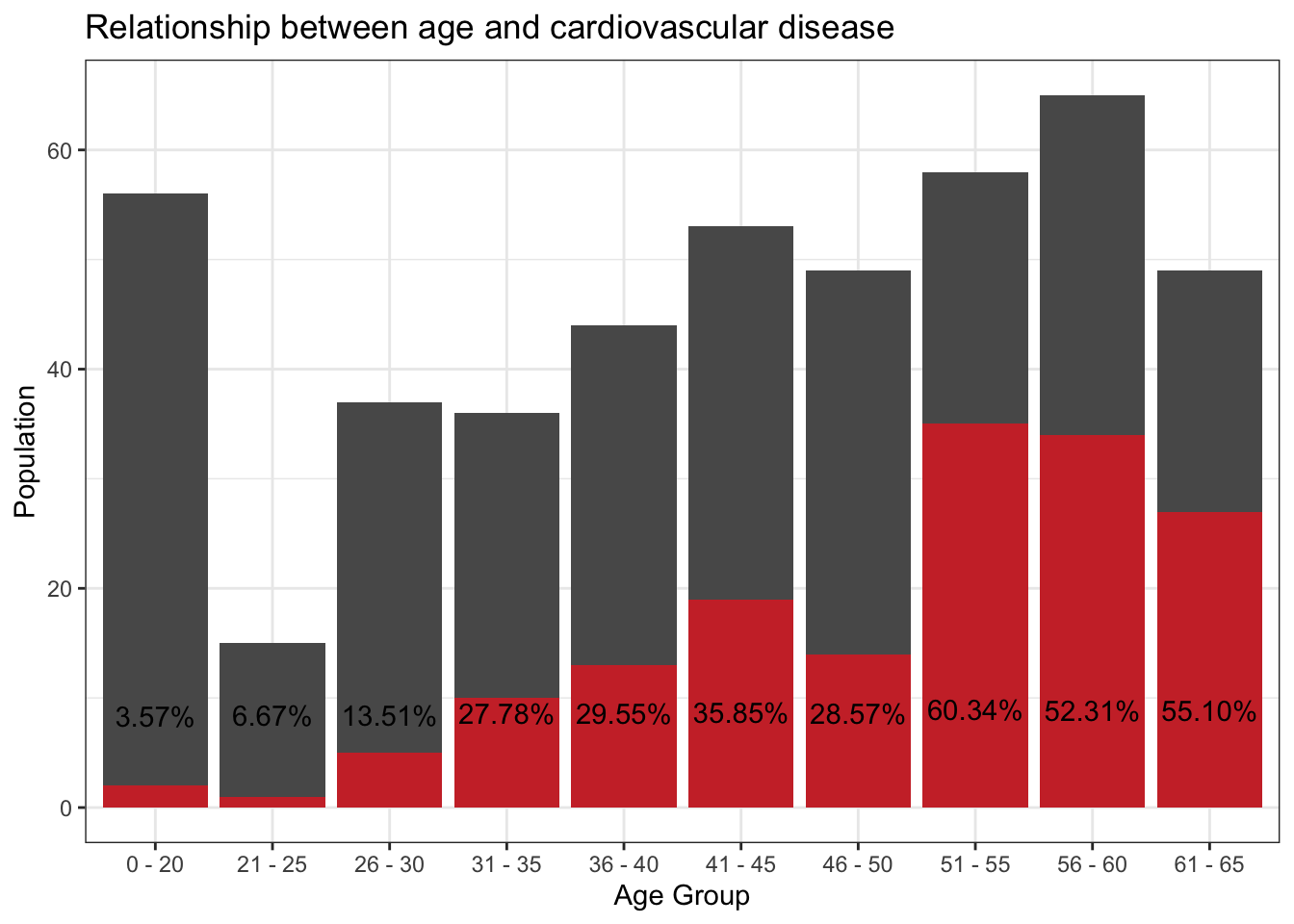
* Due to the original data being only retrieved from high-risk heart-disease areas in Africa, the geographical limitations may affect the interpretation of results. But by analyzing the data from the high-risk areas, the significant relationships between features can be found.

The remainder of the report will examine the following **questions**:

1. Is there a relationship between age and cardiovascular disease except all the other factors?
2. Is there a relationship between alcohol, tobacco intake and cardiovascular disease except all the other factors?
3. Is cardiovascular disease a genetic disease? If so, is there significant evidence showing that 50% of patients have a family heart disease history?
4. What is the difference between the average range of systolic blood pressure for a healthy person and a patient?

**Analysis**

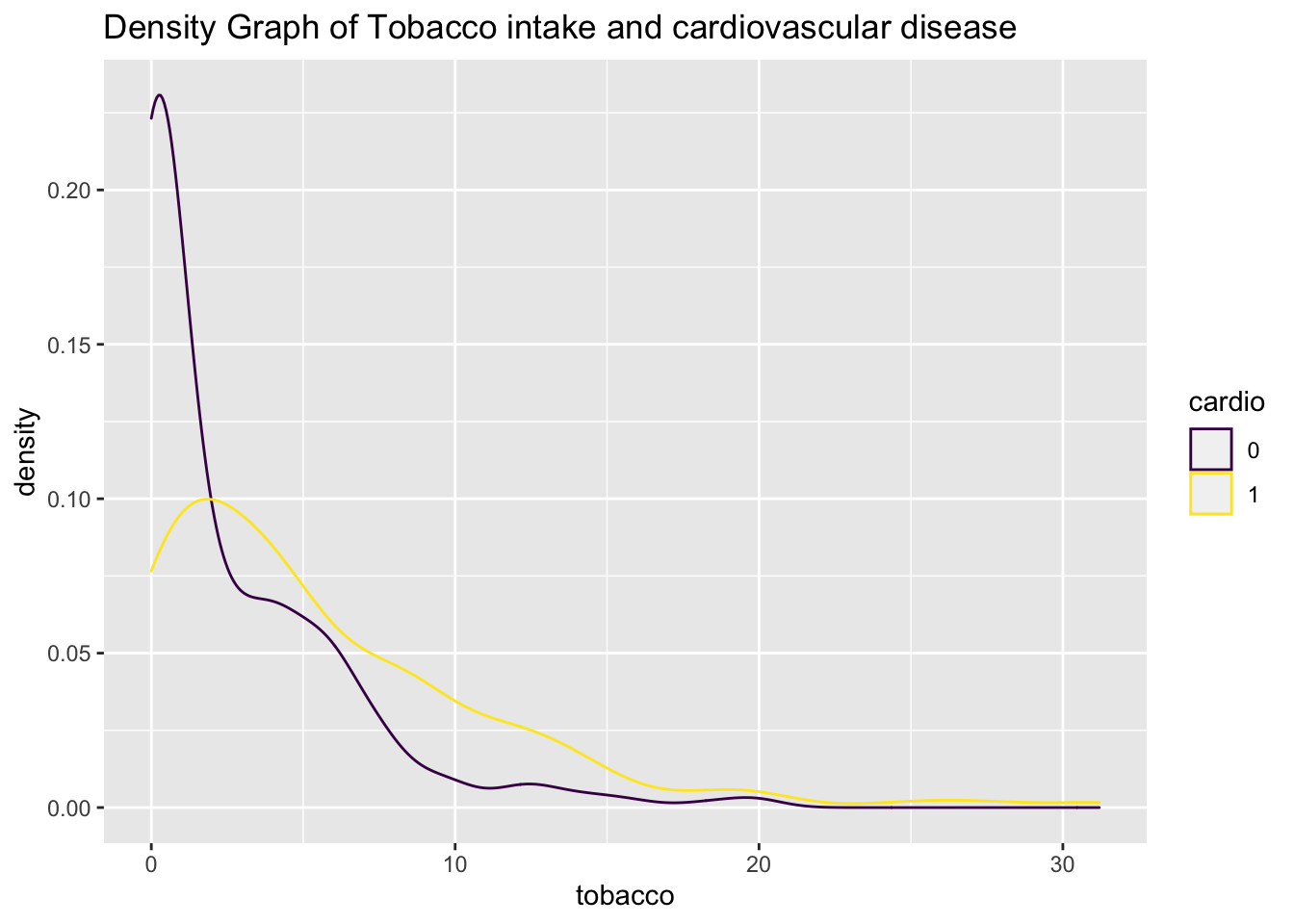
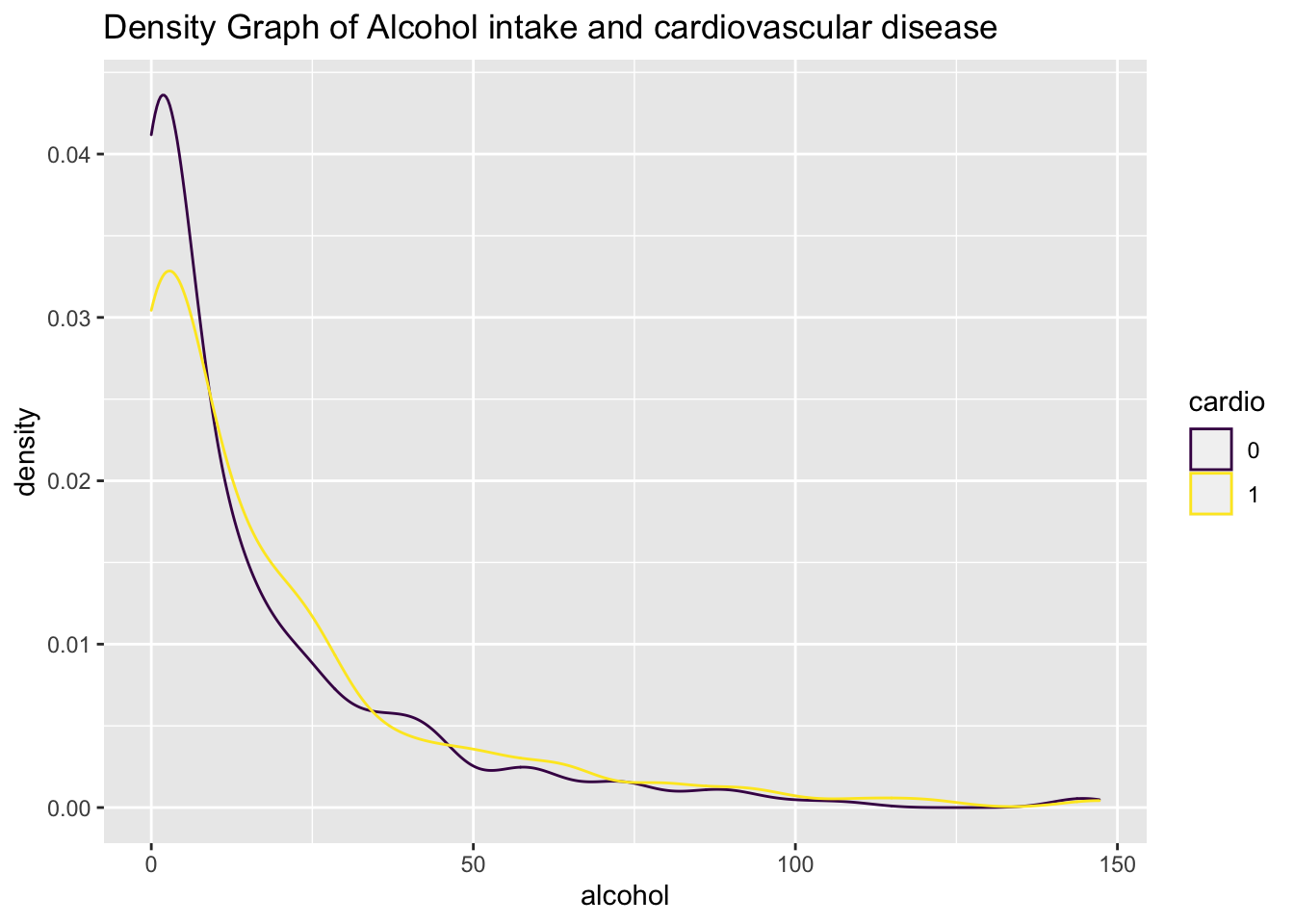
1. Relationship between age and cardiovascular disease except all the other factors



For the first graph, the average percentage of people having cardiovascular disease is around 10% in the younger age group 26-35 years old. For the age group 31-50, the percentage increased and remained around 28%. Then for the elder age group from 51-65, the figure for that increased even higher and reached 55% percentage.

Then using the linear method to draw the prediction line of the relationship between age group, and the percentage of having cardiovascular disease, as the the second graph shows above. The positive slope between the two variables indicates the positive relationship further. Therefore, it can be claimed that the age and the risk of having cardiovascular disease except all the other factors are positively related. Combined with the first and second graph, it is clear to figure out that while the age is increasing, a higher percentage of cardiovascular disease occurs.

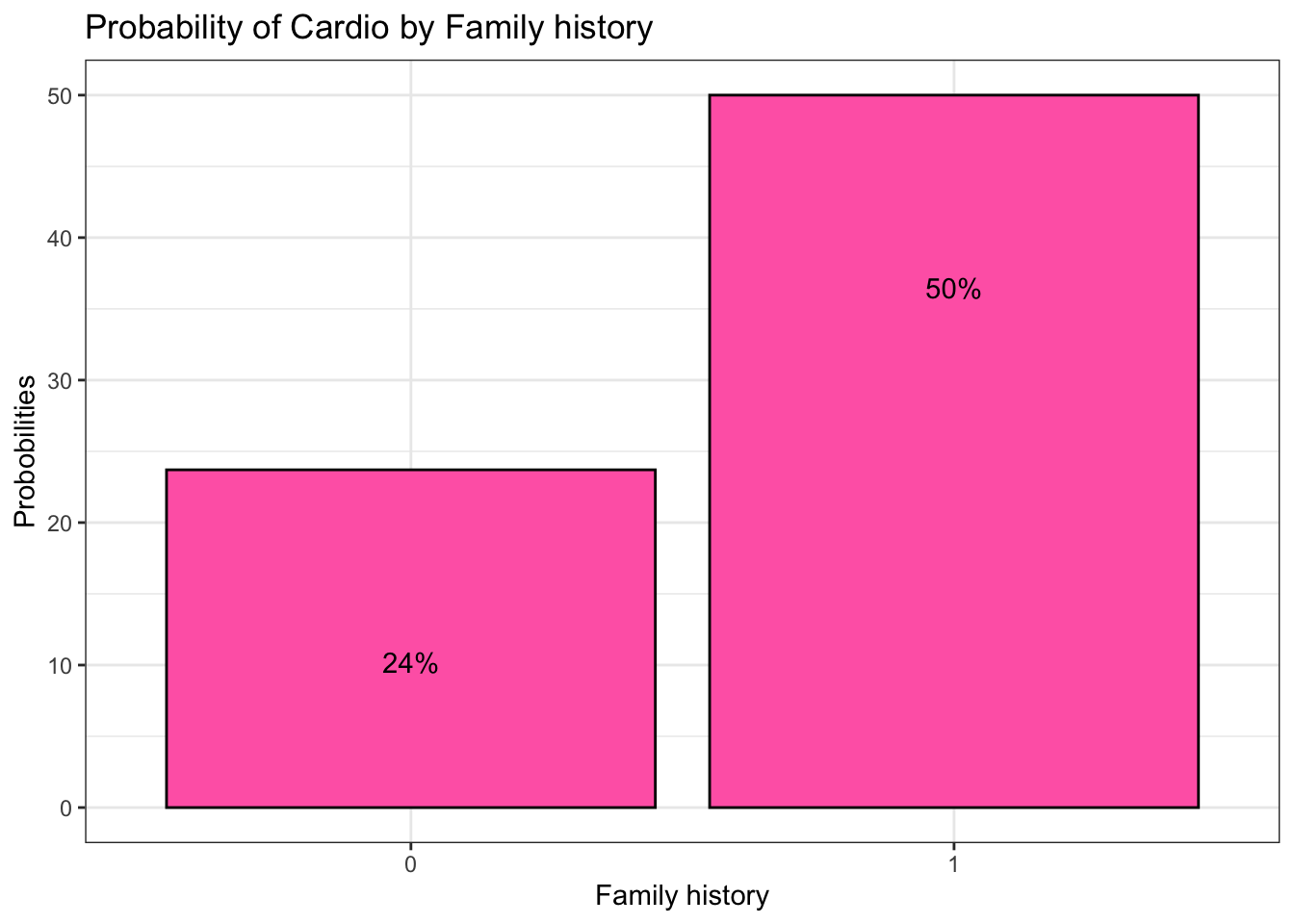
1. Relationship between alcohol, tobacco intake and cardiovascular disease except all the other factors



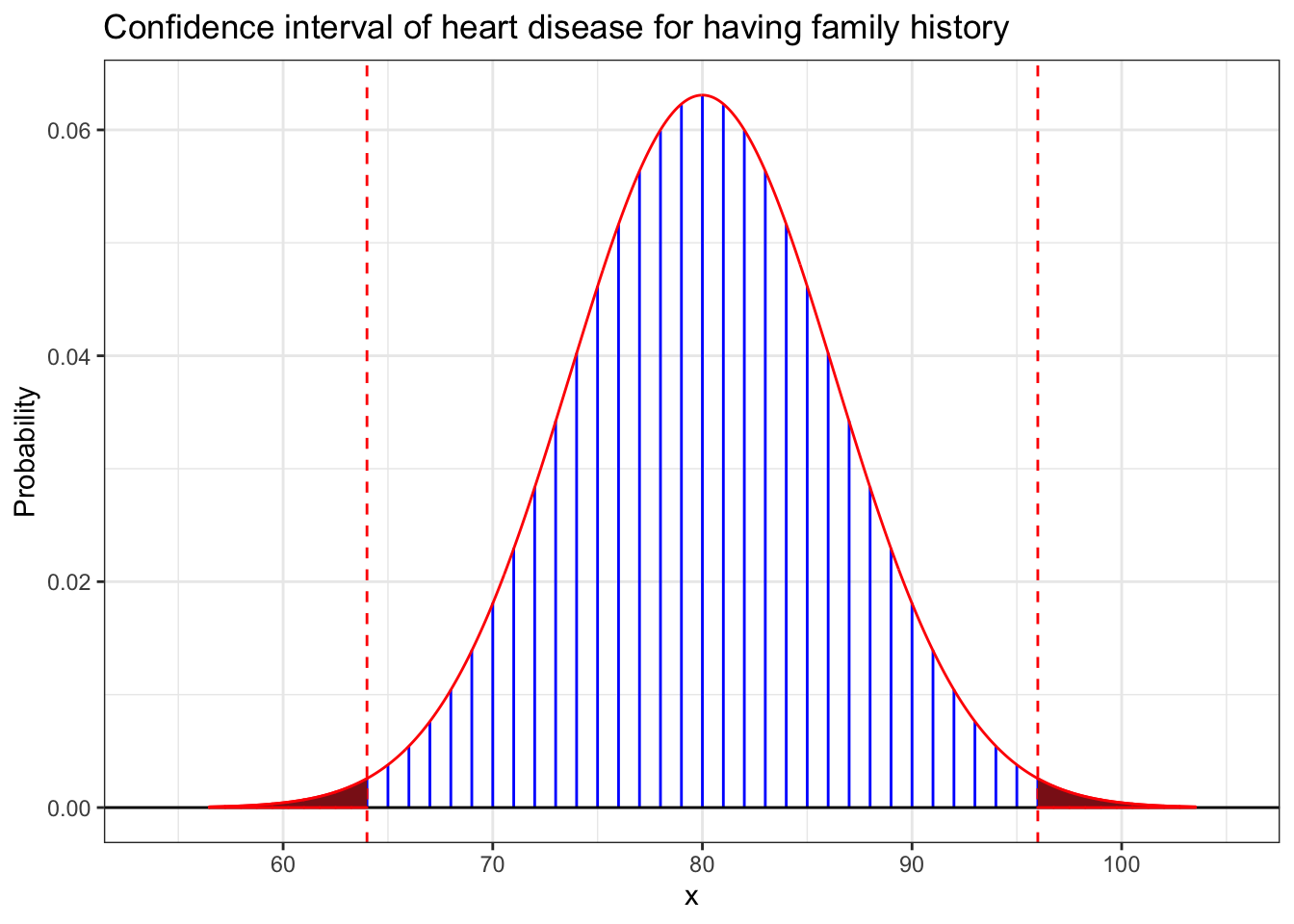
*The legend “cardio” represents whether you have cardiovascular disease; 0 - don’t have cardiovascular disease; 1 - have cardiovascular disease*

These two graphs above show the density distribution of alcohol intake and tobacco intake respectively. From the two graphs, the yellow lines, which represent the cardiovascular percentage, are almost always above the purple line. In addition, the density of healthy people who don’t intake alcohol and tobacco is extremely higher than those people with cardiovascular disease, which indicates that people in the population who don’t intake alcohol and tobacco have a low percentage of having cardiovascular disease. Also, for the second graph, the tobacco one, the density gap between healthy one and patient is significant. The method we used here is by comparing the density graph between healthy people and patients. Therefore, we can claim that with the alcohol and tobacco intake increasing, a higher percentage risk of having cardiovascular disease appears.

1. Whether cardiovascular disease is a genetic disease. And if it is, is there significant evidence showing that 50% of patients have a family heart disease history?



*Figure: The x-axis Famhist represents whether there is a family history of heart disease; 0 - don’t have; 1 - have*

****

## [1] 0.01400102

From the first graph, the probability of having a cardiovascular disease with family heart disease history is 50%, for those who don’t have reached 24%. It is clear that people with a family heart disease history have a higher risk of having cardiovascular disease.

From the second graph, to verify the hypothesis that whether 50% of patients have a family heart disease history, we make the ***hypothesis***:

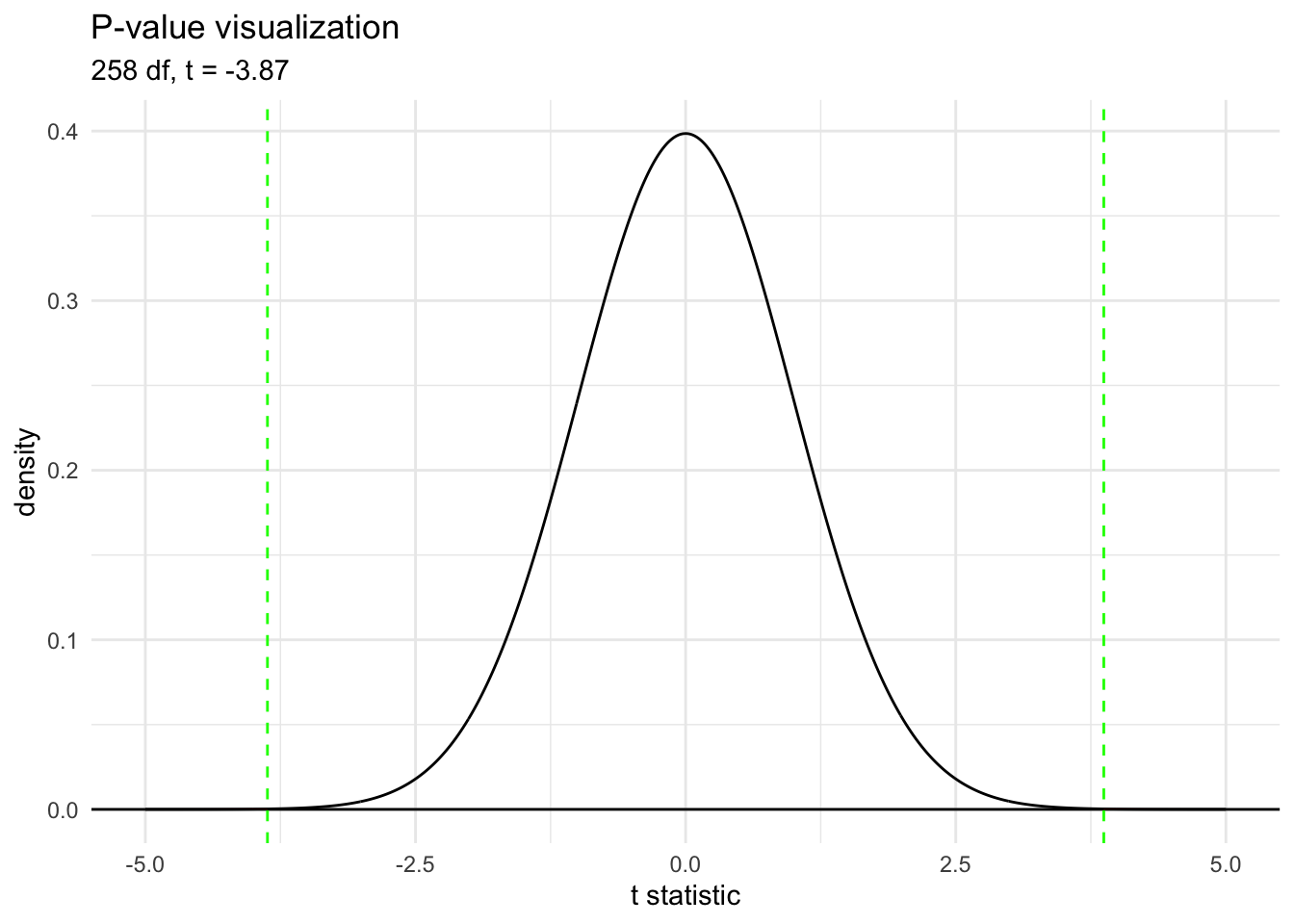
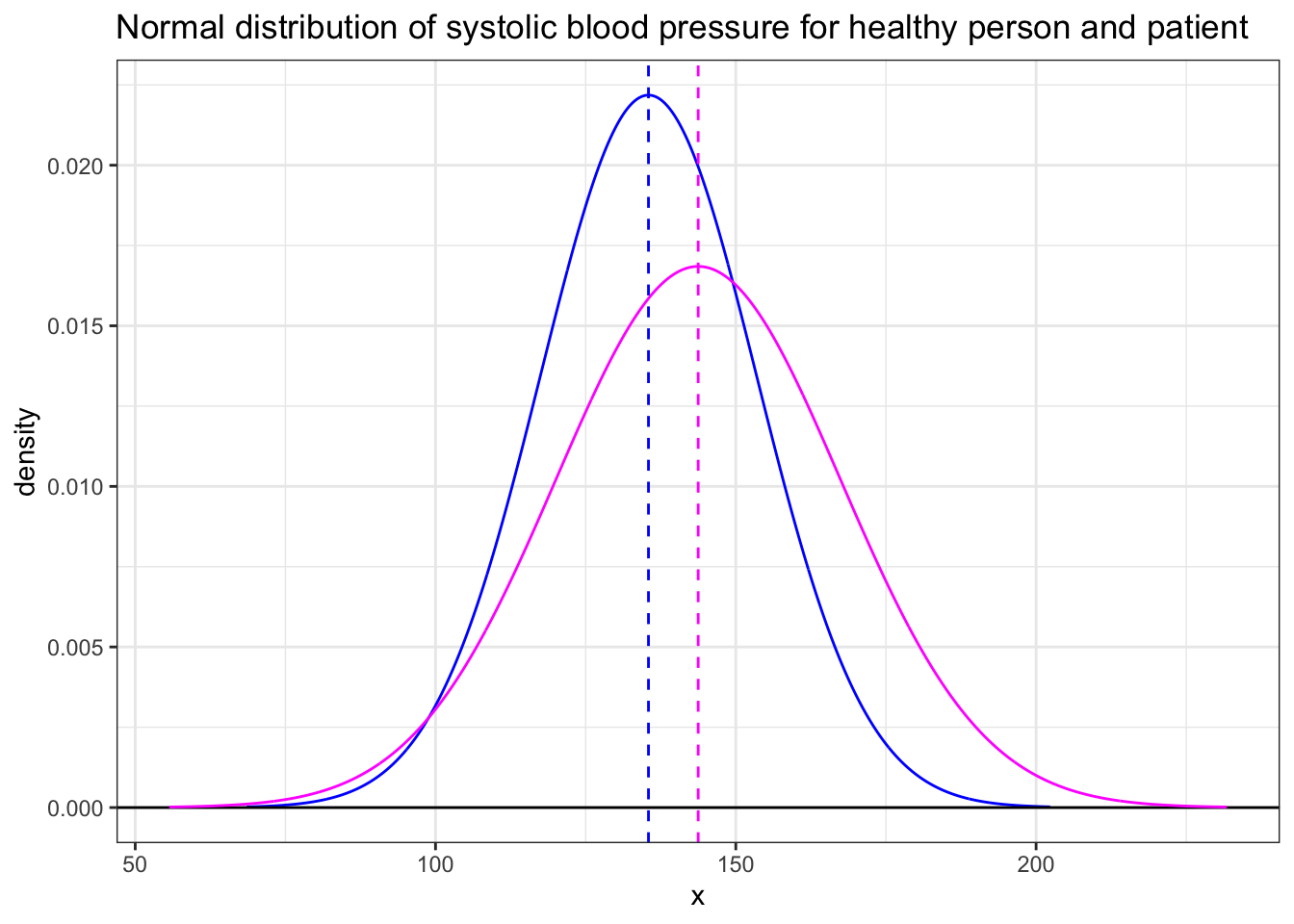
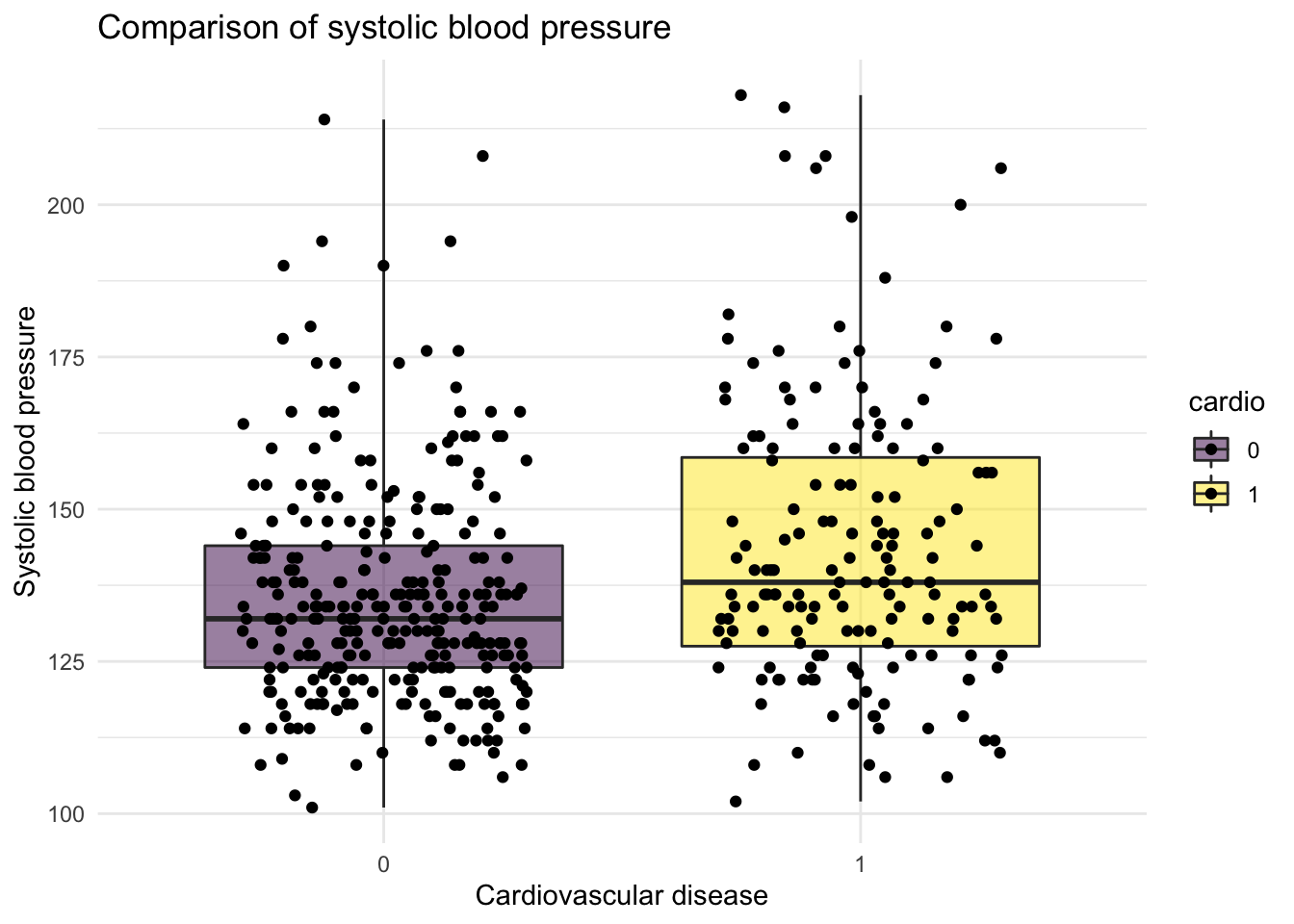
h0: p = 0.5, h1: p != 0.5.

Then set the ***model***:

X | p ~ Binomial(160, p), n = 160, x = 96

Then we get the p-value is 0.0140, a two-sided test. Therefore, we can claim that cardiovascular disease is a genetic disease, and there is significant evidence that 50% of patients have a family heart disease history(p=0.014, two-sided binomial test).

1. Difference between the average range of systolic blood pressure for a healthy person and a patient



*The blue line represent the distribution for healthy person, and the red line represent the patients*

The first box plot shows the distribution of systolic blood pressure for healthy people and cardiovascular disease patients respectively. It is clear that the figure for health people cluster at the lower level compared to the patients. The mean blood pressure for healthy people is around 130 and the patient is around 135. The second graph draws the normal distribution of systolic blood pressure for healthy people and patients. The sigma for patients is higher than the healthy one. To figure out the statistical average range of systolic blood pressure relationship between healthy people and patients, we compared the two Independent means:

Xi∼F1(μ1,σ1),i=1,…,n1  
Yi∼F2(μ2,σ2),i=1,…,n2

The third graph shows the p-value line on the density graph, which indicates p-value= 0.00013. Therefore, we can claim that after doing the hypothesis test, we are 95% confident that the mean systolic blood pressure of patients is between 4.07 mm Hg and 12.48 mm Hg higher than the mean pressure of healthy people who do not have cardiovascular disease.

**Discussion**

For the relationship between age and cardiovascular disease, we can strongly recognize that with higher age, people have a higher percentage of risk to have cardiovascular disease. It follows the normal sense because the physiology of the elderly will deteriorate. One significant point to mention is the age groups 40s and 50s. The risk of having heart disease increases dramatically between the two age groups. Therefore, future research might be discussed to figure out what causes such a gap.

For the relationship between drinking, smoking, and cardiovascular disease. We can infer that drinking and smoking are highly possible to lead to cardiovascular disease. Also, the larger gap shown in the tobacco taking graph indicates that tobacco might have a higher influence on cardiovascular disease.

For the question of whether or not cardiovascular disease is a genetic disease, we can conclude from the chart above that 50% of people who have a family heart disease history will have cardiovascular disease, which shows a strong relationship to inheritance.

For the last question we analyzed in this report, the proofs are more accurate and reliable due to the use of the hypothetical test and confidence interval. We compared the two independent means then discovered that the systolic blood pressure for patients is between 4.07 mm Hg and 12.48 mm Hg, which is higher than the mean pressure of healthy people who do not have a cardiovascular disease with a confidence of 95%.

Another important factor of cardiovascular disease is weight, but we didn’t cover it in this report. We intuitively believe that there is a positive relationship between weight and heart disease. However, we still need a new data set and analysis to figure out how strong the relationship is in the future.

Finally, for potential short-comings, it is worth noting that the method we analyze the relationship between drinking, smoking, and cardiovascular disease can be improved and become more precise. The density graph to make a visual interpretation may involve subjectiveness and lack accuracy. Also, the outcomes generated from the data set selected may not be universal, due to the reason that it was collected in a high-heart disease area in Africa. Other variables might also pose threats to heart health significantly, which we don’t include in this report.

In conclusion, the report shows a strong positive relationship between age and cardiovascular disease. Also, people who drink and smoke would have a higher risk of having such a disease, and smoking affects them more. And it is a genetic disease, about 50% of patients have a family heart disease history. Finally, the mean systolic blood pressure of patients is between 4.07 mm Hg and 12.48 mm Hg higher than the mean pressure of healthy people who do not have cardiovascular disease.

**References**

Data set source: [Cardiovascular disease dataset](https://www.kaggle.com/yassinehamdaoui1/cardiovascular-disease)