Introduction to Probability and Statistics

By: Kai Run Leong (S3862092)

Introduction and research questions

The dataset that my group and I have chosen is regarding heart diseases. The dataset contains a total of fourteen variables with thirteen of the variables known as features that were used to determine the label of this dataset called "target". The purpose of this research is to determine if age and cholesterol are a factor towards a person getting heart disease. For this report, we have chosen two continuous numerical variables which were known as "age" - age of the patient - and "chol" - cholesterol of the patient - and a binary variable called "target" which is the label of the dataset (0 = no heart disease, 1 = present of heart disease). The objective of using these three variables will help us determine if there is any relationship between the numerical variables and the binary variable that causes the chance of contracting heart disease. Thus, keeping the objective in mind, my group and I have constructed the following research questions:

1. Comparing the mean age of people with heart disease with respect to the binary variable.
2. Comparing the mean cholesterol of people with heart disease with respect to the binary variable.
3. What is the relationship between the cholesterol level and the age of a patient that could cause an individual to contract heart disease?

These research questions are important to understanding our dataset because it will help to determine three aspects. The first would be to determine if there is any relationship between the age variable and the target variable, the second would be to determine if there is any relationship between the cholesterol level variable and the target variable, and finally if there is any relationship between the age variable and the cholesterol level variable and if so, how strong is their relationship and what are its impact on the outcome – chance of contracting heart disease.

Data preparation

Before performing exploratory analysis on the heart disease dataset that was collected from Kaggle (Yasser 2022), the dataset must be cleansed first in order to prevent huge variability findings during the exploration process. Thus, two boxplots were plotted for each numerical variable to observe for any outliers – Minitab was the main tool that was used for this report.

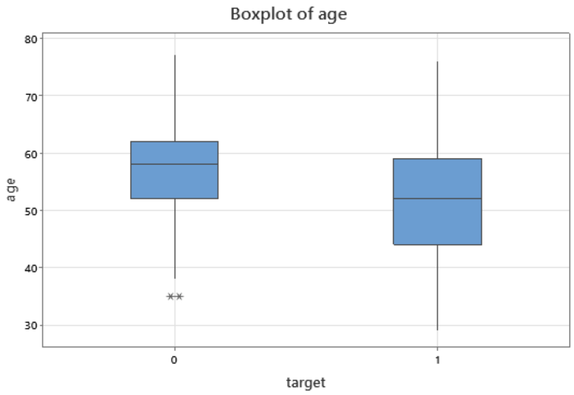


Figure 1

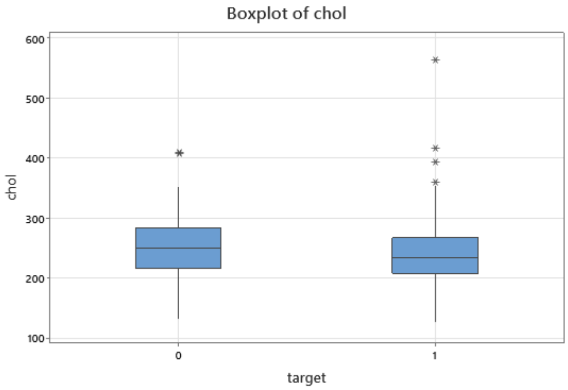


Figure 2

As can be seen from the two boxplot figures, there are a total of seven outliers. However, just because these data are beyond the upper or lower fence, this does not mean that all of these outliers are possible input errors or other form of errors in the dataset. For instance, in **Figure 1**, the boxplot claimed that the two outliers who had no heart disease and were of age thirty-five years old were odd. However, the two outliers can be considered realistic and true to the real world since age plays as a factor to contracting heart disease which is male who are the age of forty-five years old or older and female who are above the age of fifty-five years old or older (MedlinePlus n.d.). Hence, it can be concluded that these two outliers should not be removed and that it can be a result of insufficient data on patients in that age group that has caused these two subjects to be outliers.

The same can be said for the outlier in **Figure 2** that has no heart disease. The subject that is considered to be an outlier is aged sixty-three and has a cholesterol level of 407. This row of data should not be deleted because it is normal for humans to have a high cholesterol at an older age and that it is also possible for an individual to have a cholesterol level as high as 600 (Sarah Klein 2009). Furthermore, the subject's age is within the normal lifespan. Thus, this outlier should also be kept. As For the remaining outliers found in **Figure 2** that have heart diseases, their age and cholesterol level were also found to be in the realistic range. Therefore, none of the outliers were removed from the original dataset and the possible reason as to why these outliers were formed is probably due to insufficient data.

Exploratory analysis, and discussion

Now that the data has been cleansed, data exploration can now begin. For the difference in mean age and cholesterol level for the first and second research questions, it was found that both the mean values were odd as younger people generally had a higher chance of getting heart disease based on the dataset. For instance, the difference in mean value for age of people who had heart disease against those who did not have heart disease was –4.1044, while the difference in mean value cholesterol level of those who had heart disease against those who did not have heart disease was –7.857. The result appeared to be not in line with the real-world data, as it is generally expected that younger people of both sexes who also have a lower cholesterol level should have a lower chance of contracting heart disease. Thus, in order to investigate this phenomenon and determine why younger people are more prone to getting heart disease, a dot-plot graph was plotted by age group.

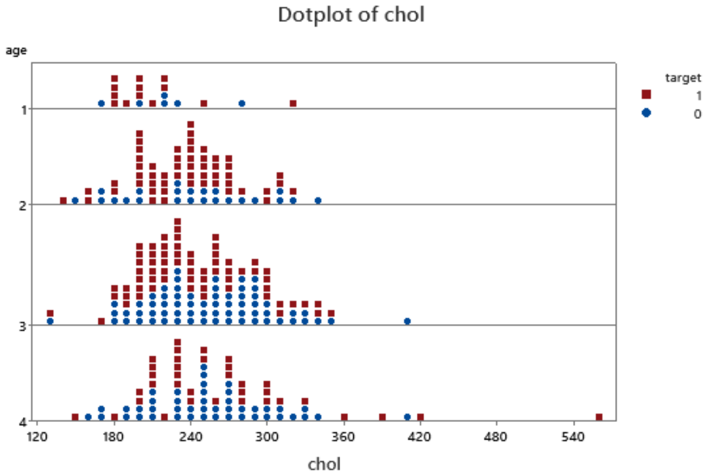


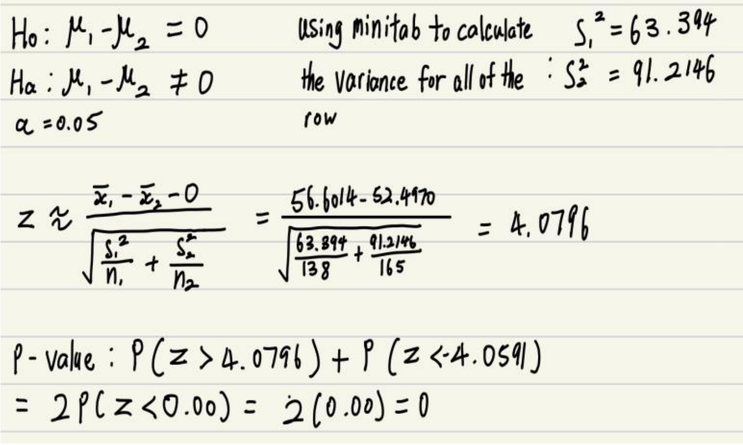
Figure 3

According to the dot-plot, it was found that 69.737% of the total population in age group 2 (41-50 years old) contracted heart disease while only 49.612% of the total population in age group 3 (51-60 years old) contracted heart disease. This showed that age group 2 in the dataset was highly populated with subjects who have extremely high cholesterol levels of above 200 which is above the healthy cholesterol level threshold a person should have (Cleveland Clinic 2020). Which could explain the phenomenon of an increased number of younger people having heart disease as compared to older people in the dataset. However, when Pearson's correlation model was used to determine the relationship between the numerical values, it scored a coefficient of 0.214 which showed a weak positive correlation between the two variables which could also mean that even though the dataset was biased towards younger people having heart disease, this could have also been a random event of an unintentional biased dataset. Therefore, two possible conclusions can be made based on the dataset. The first possible conclusion is that the dataset has insufficient data on healthier younger people and older morbidity people, thus skewing the data to show younger people to be more prone to having heart disease. The second possible conclusion is that the trend has changed over the past few years and younger people are taking less good care of their health than older people. However, based on an article posted by MedlinePlus which is a scientifically based and peer-reviewed health information provider, it can be concluded the first possible conclusion is more likely true. Thus, the data can be deemed as insufficient, and it will be able to provide more accurate information once more data is collected and added into the dataset.

Hypothesis testing and regression analysis

In order to prove that either one of the conclusions that was mentioned in the exploratory analysis is true, hypothesis testing will be performed to prove that the average age and cholesterol for both targets will never be the same. In order to justify that there is a difference in the means for both numerical features between the two targets, sampling distribution and hypothesis testing will be performed. The sampling distribution will first be performed to determine if the difference in average age and cholesterol are within the confidence interval. If the difference between the two targets (µ1 - µ2) are within the confidence interval, it can be said that regardless of how many times a random sample is drawn, the dataset should contain the true population for both numerical features and therefore the difference in average age and cholesterol that was found between the two groups can be used for hypothesis testing. The method that would be used for hypothesis testing are the six steps for performing a hypothesis testing.

Starting with age, in order to justify that the difference in the mean is within the confidence interval, a 95% confidence interval for the average age for both targets was found to be (2.09, 6.12). Thus, since the difference in means for age for both targets was found to be 4.1044 (µ1 - µ2 = 56.6014 - 52.4970 = 4.1044), it can be said that the actual difference in the mean of age will lie within the confidence interval with 95% confidence. Furthermore, because the confidence interval does not contain µ1 - µ2 = 0, it is possible to conclude that there is a difference in average age between two of targets. Therefore, hypothesis testing can now be performed to further determine if there is a difference.



$H_0: \mu_1 - \mu_2 = 0$ using minitab to calculate the variance for all of the row : $S_1^2 = 63.394$

$H_a: \mu_1 - \mu_2 \neq 0$ $S_2^2 = 91.2146$

$\alpha = 0.05$

$$z \approx \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{56.6014 - 52.4970}{\sqrt{\frac{63.894}{138} + \frac{91.2146}{165}}} = 4.0796$$

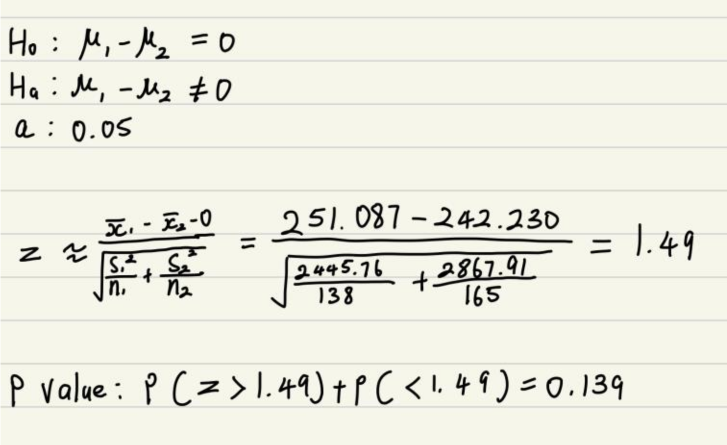
$$p\text{-value}: P(z > 4.0796) + P(z < -4.0591)$$

$$= 2P(z < 0.00) = 2(0.00) = 0$$

Figure 4: *Hypothesis testing for difference in mean age*

After performing hypothesis testing for age, since our *p-value* is less than the pre-assigned significant level, the test is statistically significant, and it can be concluded that average age for getting heart disease has changed. Thus, the null hypothesis is rejected, and it can be concluded that the average age between the two target groups will never be the same.

When another hypothesis test was performed for the difference in the average cholesterol – refer to **Figure 5 -**, it achieved similar results where the difference in mean is within the confidence interval (-2.89, 2.60). However, the null hypothesis was not rejected when hypothesis testing was done for the difference in average cholesterol since it got a *p-value* 0.139, which is greater than the pre-defined significant level.



$H_0 : \mu_1 - \mu_2 = 0$

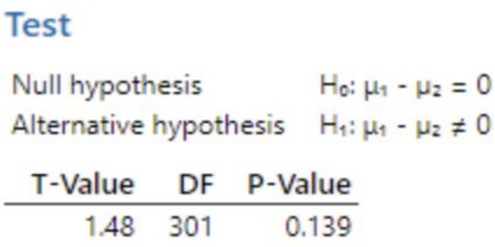
$H_a : \mu_1 - \mu_2 \neq 0$

$a : 0.05$

$$z \approx \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{251.087 - 242.230}{\sqrt{\frac{2445.76}{138} + \frac{2867.91}{165}}} = 1.49$$

P value : $P(z > 1.49) + P(< 1.49) = 0.139$

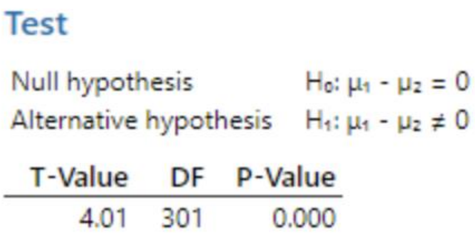
Figure 5: *Hypothesis testing for difference in mean cholesterol*



```
Test

Null hypothesis          H₀: μ₁ - μ₂ = 0
Alternative hypothesis   H₁: μ₁ - μ₂ ≠ 0

T-Value   DF   P-Value
   1.48  301     0.139
```



```
Test

Null hypothesis          H₀: μ₁ - μ₂ = 0
Alternative hypothesis   H₁: μ₁ - μ₂ ≠ 0

T-Value   DF   P-Value
   4.01  301     0.000
```

Figure 6: *2 Sample T-test in Minitab for Chol* **Figure 7:** *2 Sample T-test in Minitab for age*

Lastly, regression analysis was performed to determine if there is a relationship between age and cholesterol. As seen from **Figure 8**, the normal probability plot follows the correct pattern for the assumption of a normally distributed dataset and that since majority of the data points appears to be randomly scattered around the residual line, a linear assumption is reasonable. Furthermore, all outliers have already been removed during data preparation, therefore the data is now prepared to find Pearson's correlation.

However, based on the results, the regression analysis returned an R-square value of 4.57%. This meant that only 4.57% of the total variance for the response variable can be explained on the independent variable, which meant that the fit of the linear model is very weak. Thus, the coefficient of correlation had a score of *r* = 0.214 which shows that there is a very weak positive correlation between the age variable and cholesterol variable.

Figure 8 *Four in one regression residual plot graph*

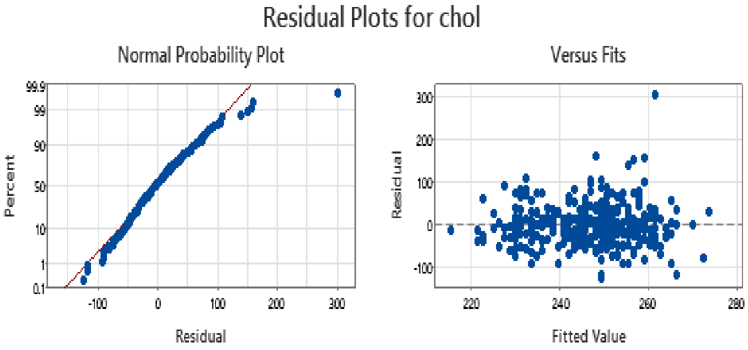


Figure 9: *Four in one regression residual plot graph*

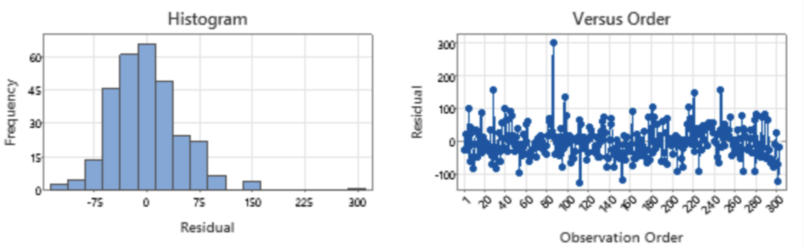
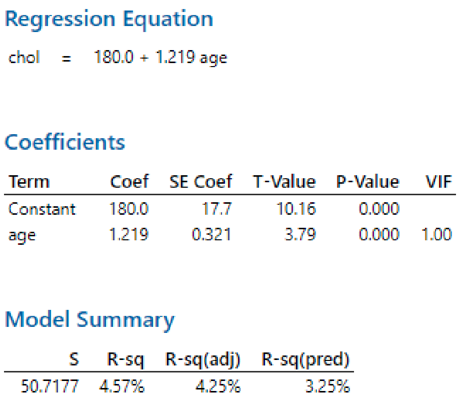


Figure 10: *Results of Regression Analysis in Minitab*



Regression Equation

chol = 180.0 + 1.219 age

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	180.0	17.7	10.16	0.000	
age	1.219	0.321	3.79	0.000	1.00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
50.7177	4.57%	4.25%	3.25%

Discussion and conclusion

In summary, it was found that based on the hypothesis testing and regression analysis, age and cholesterol do not have a strong relationship. Thus, this explains why it is difficult to determine if a person is prone to getting heart disease just based on two of these numerical features. For instance, even though the average age for both target groups should be different – based on the hypothesis testing -, the difference in average cholesterol for both groups could still be almost the same which meant that younger people could have a very high cholesterol but still have no presence of heart disease. Therefore, predicting if a person will have heart disease is not accurate or possible just by limiting the use of the two numerical values that were discussed throughout the report. For future improvements, it would be best to ensure that the data collected will not be skewed towards younger morbid patients as this has caused the data to provide extremely biased information. Furthermore, the numerical feature age should be replaced by other more relevant numerical features such as the patient's resting electrocardiogram – although this would have to be tested again. In conclusion, despite the data being skewed, it was found that age and cholesterol had a very low correlation and thus using just age and cholesterol to determine the probability of a patient having heart disease is insufficient.

References

M Yasser H. (2022). *Heart Disease Dataset*

https://www.kaggle.com/datasets/yasserh/heart-disease-dataset

MedLinePlus. (n.d.). *How to Prevent Heart Disease*

https://medlineplus.gov/howtopreventheartdisease.html

Sarah Klein. (2009). *10 surprising facts about cholesterol*

http://edition.cnn.com/2009/HEALTH/11/24/moh.healthmag.cholesterol.surprises/index.html

Cleveland Clinic. (2020). *Cholesterol Numbers: What Do They Mean*

https://my.clevelandclinic.org/health/articles/11920-cholesterol-numbers-what-do-they-mean