STATISTICAL TESTING

In relation with the aim of this project we are trying to find out some of the biological traits of individuals who have heart disease. This data set consists of 14 attributes, however the attributes that were tested are in this research are ten as stated:

1. chest pain type (4 values)
2. resting blood pressure
3. serum cholesterol in mg/dl
4. fasting blood sugar > 120 mg/dl
5. resting electrocardiographic results (values 0,1,2)
6. maximum heart rate achieved
7. exercise induced angina
8. oldpeak = ST depression induced by exercise relative to rest
9. the slope of the peak exercise ST segment
10. thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

These attributes consist of continuous variable and categorical variables. The continuous variable were analyzed using the two sample t test for equal means these test have some assumptions that need to be verified to ensure that the test yields correct results. The t test assumptions are:

1. Independent observations
2. Normal distribution
3. Equal variances

Hypothesis Test Procedures and Assumption Checks for Continuous Variables

The procedures used to test the continuous variables are as follows:

1. The Shapiro-wilk test is run to check for normality of the distributions. However the t test normal distribution assumption is satisfied if the distribution of means from random samples of data forms a normal distribution.

2. Bootstrap method is used to plot the distribution of means from random samples to ensure the second criterion is met.

3. Levene's test for equal variances is used before the test is run to ensure the scipy code is set to the right variant to ensure accuracy of results.

4. The two sample t test for independence is run. With the null hypothesis stating there is no difference between the means of the two samples. The p value is compared to the t statistic, if less then null hypothesis is rejected.

Also in order to see how these variables relate to one another the Pearsons test for correlation was run on the continuous variables to see how the continuous variable interact with one another

Test for Categorical Variables

The Chi-Square test of independence is used to determine if there is a significant relationship between two nominal (categorical) variables. With the null hypothesis stating there is no independence between the variable in question and heart disease.

Results

There were 4 continuous variables : cholesterol, resting blood pressure, maximum heart rate, ST depression and 6 categorical attributes tested : fasting blood sugar, exercise induced aging, thalassemia, resting electrocardiograph, chest pain, and the slope of the peak exercise ST segment.

All three continuous variables cholesterol, resting blood pressure, maximum heart rate and ST depression are found to have non normal distributions hence the bootstrap method was used to find the distribution of the means, which were normally distributed allowing us to use the two sample t test. Using Levene's test we determined cholesterol and resting blood pressure have equal variances while maximum heart rate and ST depression do not, using this to apply the appropriate conditions in the t test. From the t test we determine that for all four continuous variables the absolute value of the t statistic is greater than the than the pvalue hence we reject the null hypothesis meaning that the mean values of the patients with and without heart disease are significantly different. From these studies all the continuous variables are significant in helping us to determine the presence of heart disease.
There was also correlation matrix between these four variable and age there was no real strong correlation between these variables as the strongest correlation was between age and maximum heart rate which was negative correlation of 0.4.

The chi square test for independence is carried on the categorical variables of the six categorical variables fasting blood sugar, exercise induced agina, thalassemia, resting electrocardiograph, chest pain, and the slope of the peak exercise ST segment, two attributes fasting blood sugar and resting electrocardiographic results have a pvalues greater than the alpha which is set to 0.01 hence we fail to reject the null hypothesis which means we cannot prove there any association or relationship between those two variable and heart disease. While all the other variables have pvalues lower than the 0.01 so there is a high probability that there is association between these variables and heart disease.