Student no: 20493252

Final Project

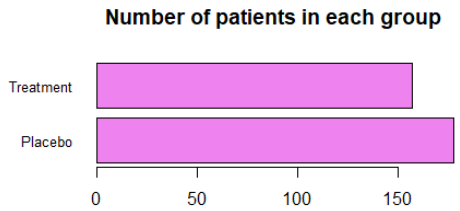
STAT10060

1) Introduction

Throughout this report we want to answer whether a new drug could reduce a patient's risk of cardiovascular. This report is also going to analyse the data we have received from the Irish pharmaceutical company. Researchers have split a sample group into two groups, treatment and placebo. After 6 months on this new drug each patients' risk is recorded. By the end of this report, I want to summarise our data, compare both groups, comment on the effectiveness of the drug in reducing cholesterol level, explore the relationship between cholesterol level and BMI and finally comment on the relationship between height and blood group.

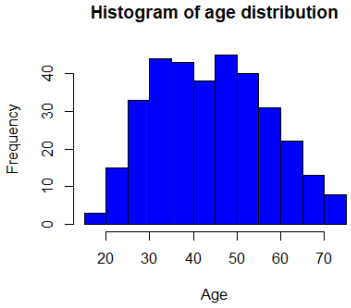
2) Explanatory Analysis

1)Number of patients in each of Treatment and Placebo



We obtain from R that there are 178 patients in the placebo group and there are 157 patients in the treatment group. This information is represented by the horizontal bar chart.

2)Age distribution

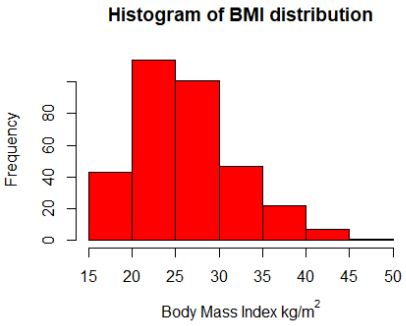


We want to investigate the age distribution within our data.

- This is a roughly symmetrical distribution meaning that the mean, mode and median are very close in value if not equal.
- The mode seems to lie in the interval 46-50 inclusive.
- Using R, I have calculated that there are 45 people in the modal interval that is 46-50 in age.
- We have very few people aged above 70 and below 20. Precisely there are 3 people younger than or equal to 20 and 11 people aged 70 and over.

3)BMI distribution

Next, we will look at the distribution of body mass index in kg/m^2.



- This histogram is skewed right which implies a positive skew.
- It seems to suggest that very few people have a High BMI. A lower BMI would be significantly more common. There are only 8 people with a BMI higher than 40 kg/m^2.
- The mode seems to lie in the interval 20-25 inclusive.
- Using R, I have calculated that there are 114 people in this modal interval.

We also noticed a slight difference in the means between male and female BMI. The male BMI average was 24.84059 kg/m^2 and the female BMI average is 27.76996 kg/m^2.

4)Counties with the highest and lowest mean weights

Monaghan has the highest mean weight with an average of 77.95455 kilograms and Wicklow has the lowest mean weight with an average of 61.26364 kilograms. The reasoning for this difference could be related to a variety of factors such as sports facilities within the respective counties or even the number of fast-food restaurants.

5)Percentage of men belonging to blood group B

23.52941% of men are in blood group B. Given that there are 4 blood groups we would expect about a 25% distribution for each blood group so our percentage of blood group B could indicate that there is a roughly even distribution of blood group type amongst men. However, we would need to investigate this further to give any concrete conclusion about whether more men have one blood type as opposed to another.

6)Mean difference in each subject's cholesterol.

I found that there is a mean difference of 0.2107463 mol/l in each subject's cholesterol from before participating in the study and after participating in the study. This positive value represents a decrease in the cholesterol level from before to after.

3) Hypothesis Testing

1)Is there a difference in the risk of cardiovascular disease between males and females?

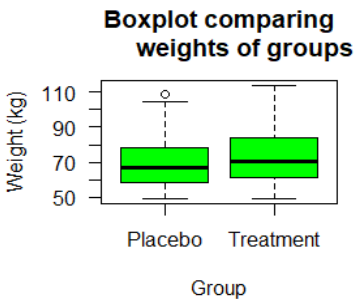
Two-way Contingency Table

	Female	Male	Sum
High	56	41	97
Low	48	59	107
Medium	61	70	131
Sum	165	170	335

I used a chi-squared test with a two-way contingency table (shown to the right) to answer this question as we are measuring 2 categorical characteristics for each item in the sample. The categorical characteristics we are measuring are gender (male or female) and risk of cardiovascular diseases (Low, Medium or high). When conducting this test, we assume that each patient fits into either male or female and, we assume that each patient fits into one of low, medium or high-level risk of cholesterol **[1]**. I decided to use a significance level of $\alpha=0.01$ because we want to be very sure that our conclusion is right as there is a potential to lose lives due to cardiovascular diseases. From R we obtain an X-squared of 3.995 with 2 degrees of freedom and a p-value of 0.1357. At this significance level we would fail to reject our null hypothesis because the significance level is smaller than our p-value and this test is upper tailed. Thus, we can conclude there is not enough evidence to suggest that there is a difference in the risk of cardiovascular disease between male and females.

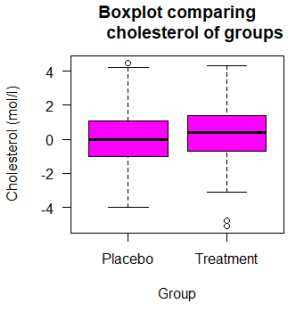
2)Is there a difference in weight between the Treatment and Placebo groups?

I used a z-test on the normal distribution because the treatment and placebo groups are independent, and the sample is large (bigger than 30). To conduct this test, I assumed that the two groups are independent and that the sample standard deviations for each group is equal to the sample standard deviation of the population for each group. The non-threatening nature of the weight data would make me choose a significance level of $\alpha=0.05$. From R I found a p-value of 0.04977. The p-value is below the significance which would mean I reject the null hypothesis meaning there is evidence to suggest that there is a difference in weight between the treatment and placebo group at our chosen significance level. However, the difference between our p-value and alpha level is minimal, it is so close to not being significant thus this needs further investigation. The 95% confidence interval for the mean is (0.0032,6.3801) rounded to 4 decimal places. Thus, we are highly confident that the true value for the difference in average weight between the two groups lies in this interval. I decided to do a boxplot to try get an informal impression of this result due to our p-value being so close to our significance level. We see from the boxplot that the median line of our placebo group boxplot is slightly different to the median line of our treatment group. However, the boxplots of both groups overlap so there is no strong evidence of a difference between weight in the two groups.



3)Do the data suggest that the new drug reduces cholesterol level compared to the placebo?

Again, I used a z-test on the normal distribution because the treatment and placebo groups are independent, and the sample is large. We have the same assumptions in the previous section above. Cholesterol can cause death, so I decided to use a significance level of $\alpha=0.01$ to decrease the chances of falsely accepting the alternative hypothesis. From R I found a p-value of 0.07078 which is bigger than the significance level hence we fail to reject the null hypothesis. There is not enough evidence to suggest that the new drug reduces cholesterol level more than the placebo at our chosen significance level.



We use boxplots to compare both distributions. From the boxplots we can see that the median cholesterol level for both placebo and treatment are roughly the same value, but the treatment group's median is slightly higher. The box plots overlap which suggests that there is is not enough evidence to suggest that the new drug reduces cholesterol level

compared to the placebo. This mimics the results from our z-test. We then conduct a 99% confidence interval for the difference. We are highly confident that the interval (-0.1949007,0.7130643) contains the actual average difference in cholesterol levels between the two groups. Given that zero lies in the interval, it is a plausible value for the difference – indicating no difference in cholesterol level.

4) *Model Fitting*

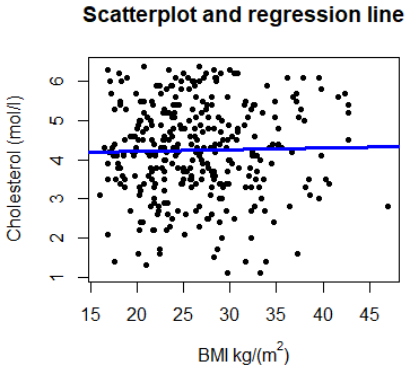
1)Correlation coefficient between BMI and cholesterol level

For our correlation coefficient we assume that both BMI and cholesterol level are continuous data, and that each patient has a value for both variables **[2]**. Our corelation coefficient is 0.02301364. This value indicates a very weak positive linear relationship between BMI and cholesterol level. It is so close to 0 that we could hypothesise that there is little to no relationship between BMI and cholesterol level.

2)Regression line between BMI and cholesterol level

For our regression line we assume that there are no significant outliers and that our residuals of our regression line are approximately normally distributed **[3]**.

The value of our y-intercept for our linear model is 4.112817 and the value for our slope of the linear model is 0.004634.

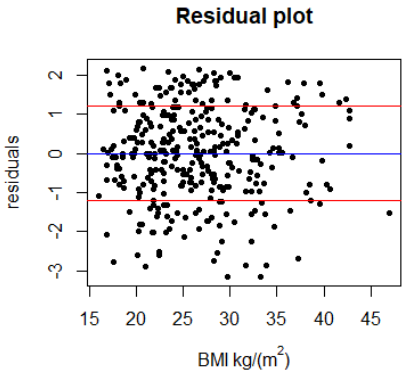


This means that the equation of our regression line is $\hat{y}$ = 0.0046x+ 4.1128 rounded to 4 decimal places which is shown in blue in our scatter plot. 0.004634 is the increase in cholesterol level in mol/l for one additional kg/m^2 in BMI. 4.112817 mol/l is the average cholesterol level for when BMI is 0, however we must remember that it is impossible for BMI to be equal to 0.

3)Coefficient of determination

We get a value of 0.0005296 for the coefficient of determination. This tells us that about 0.05% of the variation in cholesterol level is explained by BMI in this linear model that we have fitted.

Residuals



The residuals seem to be on average around the 0 mark. There is a random scatter within our residuals which would satisfy our assumptions of constant variance. Within our red bands there seems to be most of our residuals lying between these lines. For a good fit we would expect 68% of our residuals lie between our two red lines. This seems to be true so we can say our linear model is a good fit for the data. Meaning that our analysis of a very weak positive linear relationship between BMI and cholesterol based off our slope of our regression line is accurate.

4)Model utility test

Our null hypothesis is that the standard error of the slope of the linear model is 0. The y-intercept parameter is 0.297342. The standard error on the slope is 0.011033. The p-value of β is 0.675 which is much higher than our given significance level of 0.05. Hence this is not an important variable in explaining the relationship between BMI and cholesterol in a linear model. We also get a confidence interval of (-0.017, 0.027) rounded to 3 decimal places. This interval contains 0 which means that 0 is a plausible number for our standard error on the slope which furthers our conclusion that BMI is not significant in explaining cholesterol level.

5) Secondary Analysis

1)Average height for each level factor variable denoting the blood group.

Blood Group	Height Mean (cm)
A	165.6779
AB	165.7023
B	164.5973
O	164.0488

This table describes the average height for each blood group.

Clearly from the table we can see that each average height is within one centimetre of 165 cm.

2 & 3) ANOVA to test that the mean height differs significantly different across the different Blood groups.

Blood Group	Height Variance(cm)
A	56.27861
AB	58.97835
B	54.98185
O	46.63228

Before we conduct ANOVA, we must look at the sample variance in height for each blood group. During an ANOVA test we must assume that we have equal variances across groups. Here we see that the variance of blood group O is about 10 smaller than the rest which could mean our conclusions may be affected in the ANOVA test.

From the ANOVA test we get (3,331) degrees of freedom. We are given a significance value of 0.05. We also calculated our F statistic to be F=1.057 and our p-value is 0.368. The p-value is much greater than our significance level so we fail to reject the null hypothesis hence we can say that there is no evidence to suggest that one of the means is significantly different from the others. We can check whether any of the means differ significantly using the Tukey-Kramer method.

Using this method, we get the following table.

	Difference	Lower	Upper	P adj
AB-A	0.02436233	-2.932636	2.981361	0.9999965
B-A	-1.08060976	-4.046556	1.885336	0.7828903
O-A	-1.62917683	-4.544128	1.285775	0.4733508
B-AB	-1.10497208	-4.061971	1.852027	0.7694944
O-AB	-1.65353916	-4.559387	1.252308	0.4572629
O-B	-0.54856707	-3.463519	2.366384	0.9621909

Zero lies comfortably in each interval which would suggest that 0 is a plausible value for the difference in mean height between each blood group. Also, each p-value is much greater than our significance level with would further my claim.

6) Conclusion

In hypothesis testing we concluded that there was not enough evidence to suggest that there is a difference in the risk of cardiovascular disease between males and females. We also found that there is a difference in weight between the treatment and placebo groups, but this would need further investigation. From our test in hypothesis testing part 2 we found that there was not enough evidence to suggest that the new drug reduces cholesterol level compared to placebo at our chosen significance level of 0.01. This could suggest a lack of effectiveness of the new drug that the Irish pharmaceutical company have made.

We would expect there to be a relatively strong positive linear relationship as both BMI and cholesterol level are both closely related to the health of a person. However, we found that for this sample of people there is a very weak positive linear relationship between BMI and initial cholesterol level. Based off our residuals this analysis seems accurate. Thus, we may be able to infer that BMI and cholesterol are not related for people with high blood pressure. Further analysis could be conducted by doing a similar test with people who instead do not have a history of high blood pressure. We then may be able to see if cholesterol level and BMI are related for the population, we picked our sample from.

In our secondary analysis section, we found that there is not enough evidence to suggest that there is a difference between the mean height in each blood group. Thus, we can hypothesis that there is no relationship between height and blood group at our significance level.

References

[1] McHugh M. L. (2013). The chi-square test of independence. Biochemia medica, 23(2), 143–149. https://doi.org/10.11613/bm.2013.018

[2] Laerd Statistics (2020) Pearson's product moment correlation. Statistical tutorials and software guides. Retrieved April 16, 2021 from https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

[3] Laerd statistics (2020) Linear regression analysis using SPSS statistics. Retrieved April 17,2021 from https://statistics.laerd.com/spss-tutorials/linear-regression-using-spss-statistics.php