**Wilcoxon's test**

**Question**

The wilcoxon’s test will be applied to a dataset which is a record of the amount of sick days taken annually between 2013 and 2014 from the department for business, innovation and skills in the United Kingdom.

*Question; Is there a significant difference in sick days taken between the two years? If so, this would validate the need of further testing to try to reduce the number of sick days taken.*

The Wilcox signed rank test is the non-parametric test version to the more familiar student's t test which used dependent variables. The wilcoxon is applied when the dataset is non-normally distributed (Statistics.laerd.com, 2019).

|  |  |
| --- | --- |
| ***Null Hypothesis*** | ***Alternative Hypothesis*** |
| H 0 -> μ1 = μ2 | H 1 -> μ1 ≠ μ2 |
| Alpha = 0.05 | |

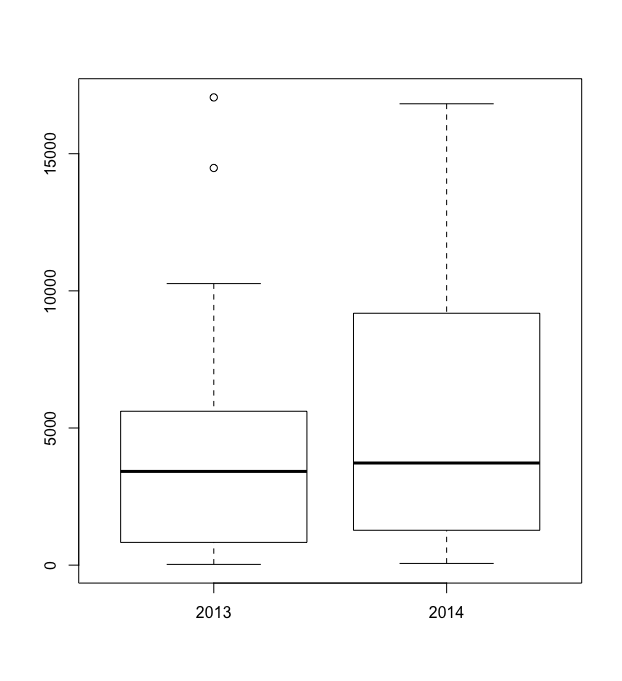
The null hypothesis is that there is no significant difference between the two means of the sick days taken between 2013 and 2014. The alternative hypothesis is that there is a significant difference between the amount of the sick days between the two years.

**Auxiliary Statistics**

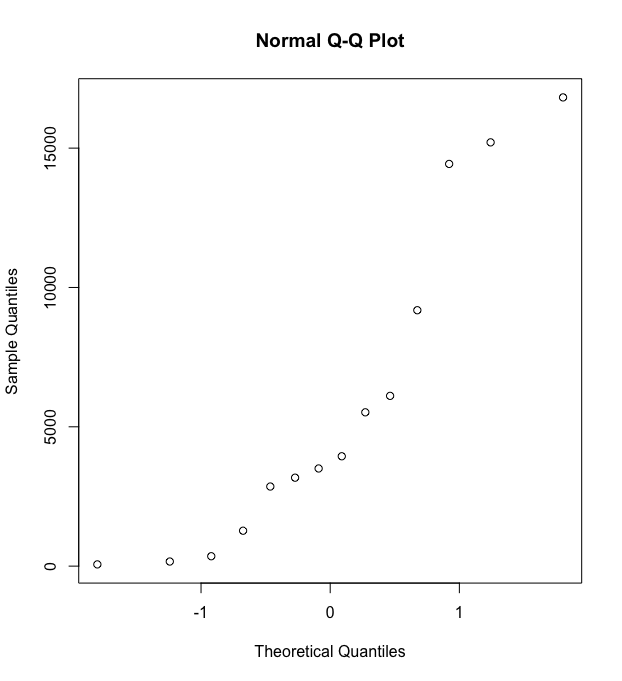
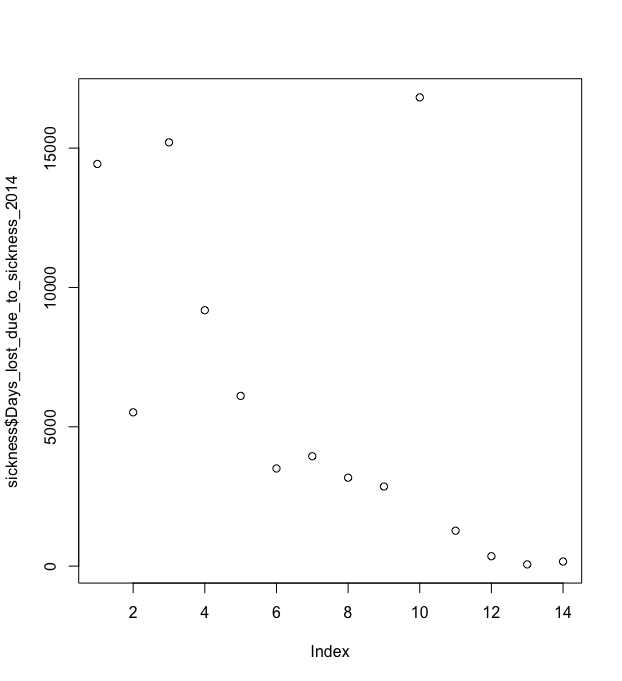
*Table 1.1*

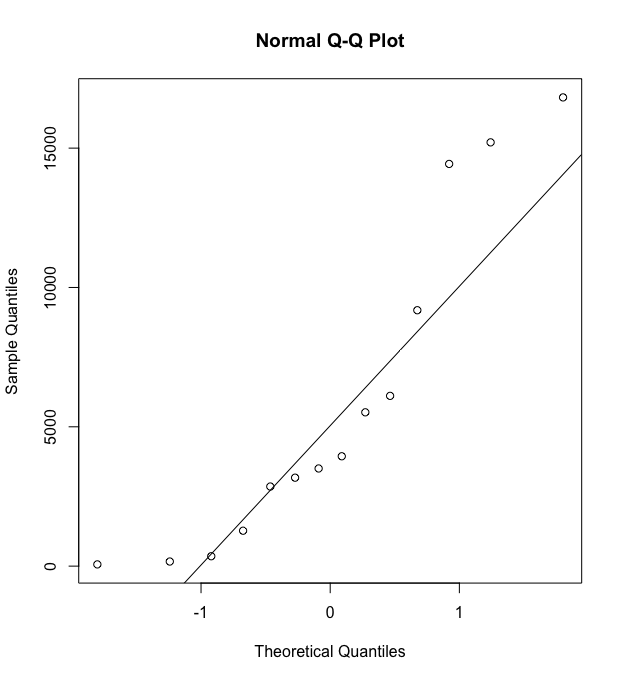
|  |  |  |
| --- | --- | --- |
| **Standard Deviation** | **Mean** | **Year** |
| 5265.69 | 5092.73 | 2013 |
| 5784.987 | 5900.227 | 2014 |

Table 1.1 shows the mean value and the standard deviation of the number of sick days taken in both 2013 and 2014. In 2014, the mean number of sick days taken increases by 15.85%. As we can see, there is a notable difference in standard deviation between the two years.

*Box plot 1.2*

Above in image 1.2, we have the boxplot for the number of sick days taken in 2013 and 2014. Evidently, there is a difference in the distribution of the number of sick days taken between the two years.

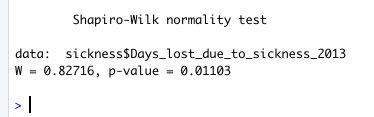
*Plot (un-normal) 1.3 Normalised 1.4*

*Qqline Normalised 1.5*

Graphs 1.3-1.5 show the process of normalisation for values 2014. From the first plot in table 1.3, it is evident that the distribution is not normal. This provides further assurance to use wilcoxon instead of the student T test for testing the two means. Please find the appendices for the remainder graphs.

**Shapiro-Wilk Test**

Apart from using more subjective means of testing for normality, we can implement the shapiro-Wilk test to test for normality by objective means.

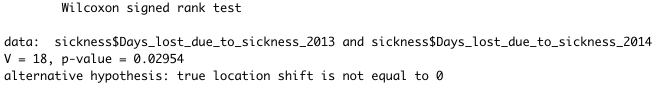
*Shapiro-Wilk Test 1.6* 

The Shapiro-Wilk test provided us with a value of 0.0113 for P. As the P value provided is less than our alpha 0.05, we can conclude that this data set is not normally distributed.

**Test**

The data has been tested for its normality and it can be concluded that the datasets are not normal. Therefore, applying a wilcoxon signed rank test to two variables is suitable.

*Wilcoxon Signed Rank Test 1.7*

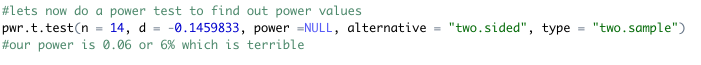
****

In image 1.7, the results of the signed rank test are presented. The p-value obtained is 0.02954. This is below the designated alpha level of 0.05. Therefore, the null hypothesis is rejected. There is indeed a significant difference between the mean number of taking sick days off work in 2013 and 2014.

**Power and Effect Size Testing**

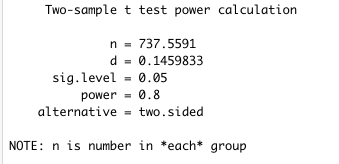
It is possible to test whether or not the sample size was big enough to determine whether or not it is plausible to reject the null hypothesis. For this, we must first understand the effect size between the two years. To accomplish this, we can use Cohen’s D formula to understand the effect size between the two years. Cohen’s D is obtained by taking the difference of the means of the two distributions and dividing the number by the pooled standard deviation (Bradburn, 2019). We have received a value of -0.1459833 for cohen’s D.

*Power Testing 1.8*

****

The power value obtained by inputting 14 for N as our sample size and 0.01459833 for D is 0.06 or simply put 6%. Typically, the higher the power, the less chance there is of committing a type II error (Statistics Solutions, 2019). Obtained power value is very low meaning the probability of committing a type II error is high.

*Desired Power Level Testing 1.9*



Ideally, our power level should be 80% in order to prevent a type II error. By inserting the received value for D and the given alpha, we receive a value for N which is 737.55. This means, in order to have a high probability that a type II error will not be committed given other variables, the sample size in each group should be 737.55.

By first testing for normality, by using qqline plotting techniques and Shapiro-Wilk’s test, we have been able to conclude that the distribution is not normal. After applying Wilcoxon’s test to test to see if there is a significant difference between the two means of the distributions, we can conclude that there is in fact a difference between the two means.

**Problem**

In the UK, sick days cost on average £77 billion annually due to lost productivity according to this report (Consultancy.uk, 2019). This is obviously a big problem. As we have seen that there is a significant difference between the means of the two years, we may produce the assumption that there are other external reasons that are causing people to have increasingly more sick days over the two years.

**Further Recommended Tests**

Reasons as to why more people are taking more sick days than before must be understood. The core problem that is causing the increase in the number of sick days must be understood in order to provide a suitable solution to the problem thus reducing the number of employee sick days.

A recommendation would be to hire a business analyst in the BIS department in the UK to further understand the core problem. Assumptively, the reasons for taking sick days will include many subjective reasons which may only be understood through close contact with the problem. For example, an increase in the number of sick days may occur because of an increase in the unhappiness index in the workplace.

The mean difference between the number of doctor visits by BIS employees should be tested for to see if there was also a significant difference between the two years. We will be able to test for similar patterns between sick days taken. If for example, there wasn’t a significant difference in doctor visits, this would suggest that there are other reasons for people taking more sick days off work which would suggest further testing internal to the organisation perhaps.

**Kruskal-Wallis test**

**Question**

The Kruskal–Wallis test will be applied to a dataset which has 3 independent and 1 dependenat variable. The experiment is looking at the weight gained by rats feeding on different meals, including pork, beef and cereal.

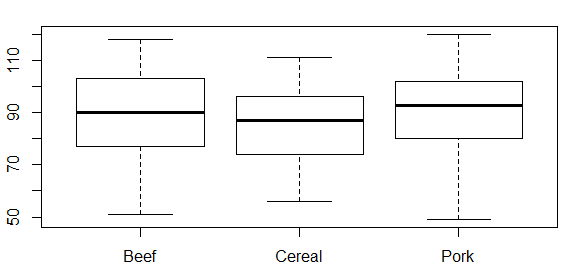
*Question; Does the food source have significant differential impact in weight gain on rats. If so, what can be deduced in relation to human physiology?*

The Kruskal-Wallis test, or sometimes referred to as the one-way anova test is a non-parametric test to determine if there are statistical differences between two or more independent variables and its effect on the dependent variable (Statistics.laerd.com, 2019).

|  |  |
| --- | --- |
| ***Null Hypothesis*** | ***Alternative Hypothesis*** |
| H 0 -> μ1 = μ2 = μ3 | H 1 -> μ1 ≠ μ2 ≠ μ3 |
| Alpha = 0.05 | |

The null hypothesis states that the mean ranks of the groups being weight gain by pork, cereal and beef is the same. The alternative hypothesis is that there is a significant difference between the mean ranks of the groups. In other words, a test to find out if the selected food has a significantly different impact on the weight gained by the rat in comparison to other variables.

**Auxiliary Statistics**

*Boxplot of each independent variable 2.1* 

Above are the boxplots of each independent variable and its respective weight gain impact. It appears that the mean and median values are all quite close to each other. We will do further normalisation tests to understand further.

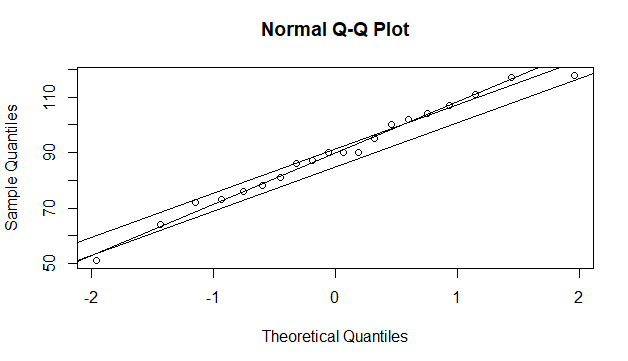
**Shapiro-Wilk Test**

*Shapiro-Wilk Test Results 2.2*



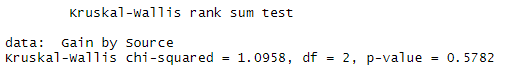
The P value of the independent variables are all clearly greater than our chosen alpha of 0.05. This would suggest that the independent variables are normally distributed.

*qqplot of beef, cereal and pork 2.3*



From the above qqline, we can see that all 3 slopes of the independent appear similar. Furthermore, the shapiro-wilk test result has proven that there is not a significant difference between the means of the independent variables. However, we will still go ahead and apply kruskal-wallis test.

**Test**

*Kruskal-Wallis Test 2.3*

The kruskal-wallis test has given us a p-value of 0.5782 which is greater than our alpha of 0.05. This means, we fail to reject the null hypothesis. After completing the dunn’s multi-comparisons test, we can see that there is a large similarity in means between the independent variables.

**Interpretation**

We can conclude that there is not a significant difference between the means of the independent variables being the weight gian impact on rats when fed on different food sources. Hence, we fail to reject the null hypothesis.

Rats and human biology are often compared for their similarities. Rats and human beings are both mammals that share many physiological traits such as sharing similar organs including hearts and livers and many other traits (Kalish, 2019). Therefore, without studying existing literature on the topic, one could make the assumption that the selected food sources, pork, beef and cereal would have little differential impact on weight gain on human beings. Pork is considered quite a high calorific food source with about 40% saturated fats (Seman, n.d.). With that being said, the test results show that the mean difference in weight gained between cereal and pork is minimal with a mean difference of 4.3 grams. One could plan their food habits based on the findings on this study and eat more high calorific food options without worrying about weight gain or obesity. However, this is definitely not recommended.

**Multi-Linear Regression**

**Question**

This dataset “house.csv”, is comprised of the dependant variable sale price of homes in the US and independent variables including but not limited to, age of home, lot size, squarefeet and other independent variables.

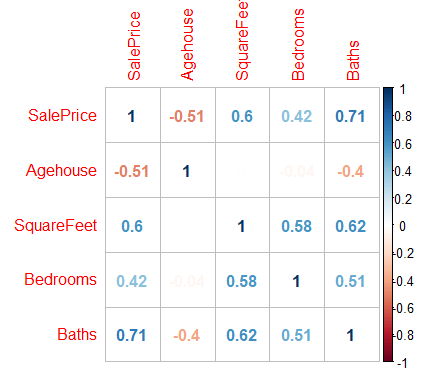
*Question; Do certain factors have significant impact on the value of a home. If so, which factors have the most impact and what would be the optimal size of these factors to better price homes?*

The 4 independent variables looked at in this study are number of bedrooms, number of baths, the total square feet of the home and the age of the house.

|  |  |
| --- | --- |
| ***Null Hypothesis*** | ***Alternative Hypothesis*** |
| H 0 -> B1 + B2 + B3 + B4 = 0 | H 1 -> B1 + B2 + B3 + B4 ≠ 0 |
| Alpha = 0.005 | |

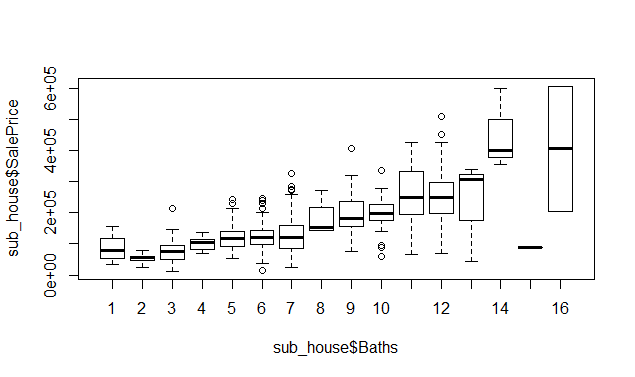
The null hypothesis states that there is no linear relationship between the sale price of the home and its independent variables. Hence, the slope equaling 0. The alternative hypothesis stats that there is a significant relationship between independent and dependent variables.

**Auxiliary Statistics**

*Correlation plot 3.1*

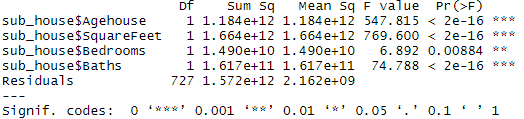
In the correlation plot in graph 3.1, we can see that there is a high both positive and negative correlation between sale price and the other independent variables. All have a positive correlation expected for the age of the house. The older the house, the lower the price the house is sold for.

*Box plot of Price of Home ~ Number of Baths 3.2*



In boxplot 3.1 above we can see that the more bathrooms a home has, the more expensive a home becomes.

**Anova Testing**

*Anova Test Results 3.3*

The anova test results show us that the interaction effect of each factor on the sale price of the home is quite significant. The P-value obtained tells us that the chance of obtaining these results again by chance is very low. The highest value for P is 0.0084, the interaction effect of the number of bedrooms on the sale price of the home, which means that there is 0.884% chance that we will arrive at these results by random chance. The obtained P-values are well below our alpha of 0.005.

**Test**

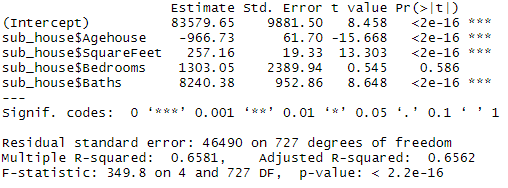
*Regression Testing 3.4*

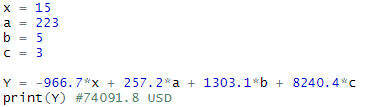
Table 3.4, is the results obtained from applying the multi-linear regression formula. As we can see, all factors except for age of the home, increases the price of the home as the factor itself increases in value. For every 1 year in age the home becomes older, the price of the home drops by 966.7 USD. For every 1 bathroom added to the home, the price of the home sold for increases by 8240.4 USD. The bull hypothesis can be rejected.

**Prediction Testing**

|  |
| --- |
| Multi-Linear Regression Formula for Y |
| Y = -966.7\*x + 257.2\*a + 1303.1\*b + 8240.4\*c |

Let's make a prediction. Let’s assume a new home is up for sale and we’re the estate agent we’re given the task of pricing this home. The home is 15 years old, has 3 bathrooms, 5 bedrooms, and is 223 sq ft in size. We use our multi-linear regression formula to arrive at a price for Y.

*Prediction Result of Multi-Linear Regression Formula Solved for Y 3.5*



According to the equation solved for Y, we receive a value of 740,91.80 USD. This gives us a Z score of -0.7703225, meaning we are -0.07703225 times away from the standard deviation of 79291.72 USD. The mean value is 135,172.00 USD.

**Value of Multi-Linear Regression**

As tested, using multi-linear regression for predicting the price of a home based on multiple factors can be quite valuable. We have been able to value the price of a home using multi-linear regressions and arrive at a price using previously collected data. Estate agents can benefit from this hugely using the estimated price as a first preliminary estimation. A clever application of this formula in this particular use case would be a stand-alone web application which enabled homeowners to quickly receive quotes by simply inputting information regarding their home, including all the different variables such as, number of bedrooms, bathrooms, size of home in square feet and even address of home for a more sophisticated result assuming that this variable would have a large effect size on the value of a home. This application could charge per quote enabling homeowners to potentially cut out fees estate agents charge homeowners for quotations.

Multi-linear regression formulas are simple yet effective. The multitude of use cases of this formula are presumably largely unexplored.

**Two-way ANOVA**

**Question**

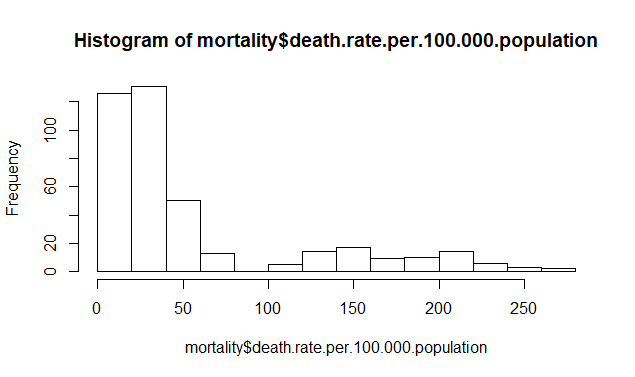
The two-way anova test will be applied to a dataset which is a record of deaths and its causes. It will be applied to a dataset which has more than two variables. In retrospect, a Kruskall-Wallis test would have been more appropriate given the number of independent variables and shapiro testing results for P. However, it is evident that valuable results are obtained using two-way anova test as a means for testing the significance of different independent variables.

*Question; Is there a significant difference in the means of the independent variables and its impact on the number of mortality. In other words, does different causes of deaths and gender cause more or less deaths? If so, how could the medicine industry benefit from this?*

|  |  |
| --- | --- |
| ***Null Hypothesis*** | ***Alternative Hypothesis*** |
| H 0 -> μ1 = μ2 = μ3 = μ4 = μ5 = μ6 = μ7 = μ8 = μ9 | H 1 -> μ1 ≠ μ2 ≠ μ3 ≠ μ4 ≠ μ5 ≠ μ6 ≠ μ7 ≠ μ8 ≠ μ9 |
| Alpha = 0.05 | |

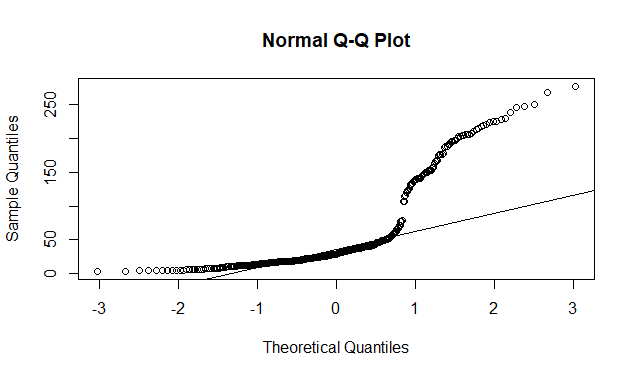
The null hypothesis is that there is no significant difference between the means of the 9 different causes of death and that there is no interaction between the variables. The alternative hypothesis states that there is indeed variance in the means of the independent variables and they do have varying interaction effects between them on mortality rates.

**Auxiliary Statistics**

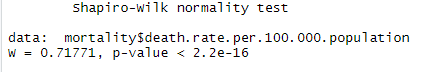
*Histogram of Mortality Rates 4.1*

From the barplot, we can see that the distribution of mortality rates are not normally distributed. Let’s do some further normalization tests to further analyse this distribution.

*Q-qline and Q-q plot of Number of Deaths 4.2*

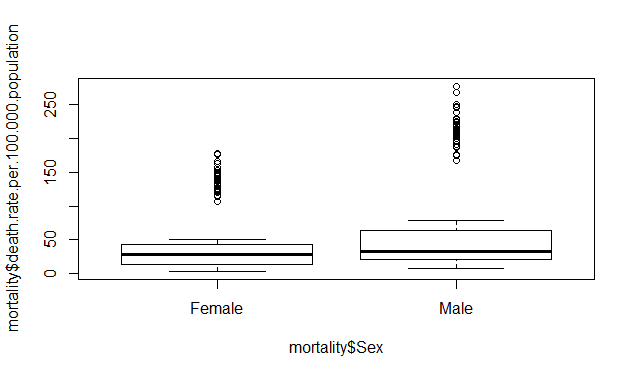


As we can see in the qqline, we can see that the qqline is quite a bit off visually from how to data points are plotted. This would suggest that the data distribution is not normal.

*Shapiro-Wilk Test 4.3*

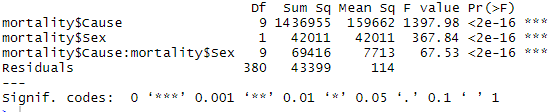
Our value received for P is less than our chosen alpha which would suggest that a Kruskal-Wallis test would be more appropriate. However, we will still go ahead and use the two way Anova test to test the means.

T**est**

*Box-plot of Mortality Rates against Gender 4.4* 

In this boxplot, we can see that one of our independent variables being gender, has varying impacts on the number of deaths. This suggests that there is a difference in the means of this variable. However, a two-way anova test will give us greater insights into the difference.

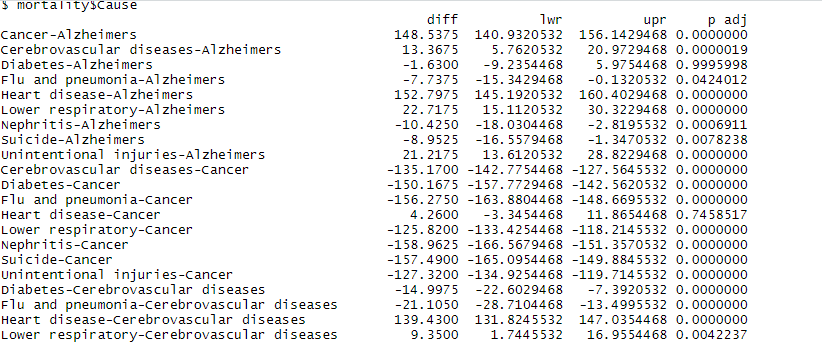
*Two-way Anova Testing 4.5*



From applying the two-way anova test, we can see that the P-value of the variables on its impact on the mortality rates are very low, meaning that there is a significant difference between the means of the variables. From the test, we can see that the probability of ariving at these results by chance once again will be very slim denoting a great variance in the impact size of the different vraibales, causes of death and sex. The null hypothesis can be rejected.

**Post Hoc Testing**

Let’s now conduct a post hoc test to understand the relationship of individual sub factors with themselves. Its not so necessary to conduct a post hoc test on gender as there are only 2 sub-factors. We will conduct a post hoc Tukey test on the causes of deaths.

*Post-hoc Tukey Testing 4.6*

From the post hoc Tukey test, we can see that some causes of deaths have very similar means such as cancer and heart disease whereas deaths caused by lower respiratory and cancer have significant difference in means. We can be 99% sure that the mean mortality rates caused by Suicide and Nephritis and Unintentional injuries and Lower respiratory will have equal means.

**Interpretation**

Causes of deaths and gender both have a significant impact on the number of deaths. We can see that there is a link between alzheimers and diabetes, they have equal means. This may suggest that a person who dies from either one may possibly have the other as well. It is reported that the two are highly correlated with each other (Robertson, 2019).

From this study, we can see which specific causes of deaths seem to have similar means. This may suggest that these specific causes of deaths are highly related to each other. The recommendation would be to further understand these pairs of causes of deaths to derive further insights. This may help insurance companies offer better medical insurance packages understanding that certain causes of deaths are highly related to other causes of death as well.

Bibliography

Animals.mom.me. (2019). *Similar Characteristics in Rats & Humans*. [online] Available at: https://animals.mom.me/similar-characteristics-in-rats-humans-7683023.html [Accessed 7 Jul. 2019].

Bradburn, S. (2019). *What Is And How To Calculate Cohen's d? - Top Tip Bio*. [online] Top Tip Bio. Available at: https://toptipbio.com/cohens-d/ [Accessed 7 Jul. 2019].

Consultancy.uk. (2019). *Sick staff cost British firms £77 billion annually in lost productivity*. [online] Available at: https://www.consultancy.uk/news/15551/sick-staff-cost-british-firms-77-billion-annually-in-lost-productivity [Accessed 7 Jul. 2019].

Kalish, M. (2019). *Similar Characteristics in Rats & Humans*. [online] Animals.mom.me. Available at: https://animals.mom.me/similar-characteristics-in-rats-humans-7683023.html [Accessed 7 Jul. 2019].

Robertson, C. (2019). *Diabetes and Alzheimer's: What's the link?*. [online] Medical News Today. Available at: https://www.medicalnewstoday.com/articles/324458.php [Accessed 7 Jul. 2019].

Seman, D. (n.d.). [online] Pdfs.semanticscholar.org. Available at: https://pdfs.semanticscholar.org/b624/d7669390c74133daea658eede5a80a3c1044.pdf [Accessed 7 Jul. 2019].

Statistics Solutions. (2019). *Statistical Power Analysis - Statistics Solutions*. [online] Available at: https://www.statisticssolutions.com/statistical-power-analysis/ [Accessed 7 Jul. 2019].

Statistics.laerd.com. (2019). *Kruskal-Wallis H Test in SPSS Statistics | Procedure, output and interpretation of the output using a relevant example.*. [online] Available at: https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php [Accessed 7 Jul. 2019].

Statistics.laerd.com. (2019). *Wilcoxon Signed Rank Test in SPSS Statistics - procedure, output and interpretation of output using a relevant example.*. [online] Available at: https://statistics.laerd.com/spss-tutorials/wilcoxon-signed-rank-test-using-spss-statistics.php [Accessed 7 Jul. 2019].