COMP1814 Coursework - 001084942

Task 1: Data Generation

The purpose of this investigation is to see whether 'Nutritional Supplement' had an impact on the weight of mice and rats. To examine this, the data first needed to be generated.

Two normal random variables were made for mice, 'mice_before', and 'mice_after'. This was done using the 'rnorm' function and consists of 3 parameters:

1. The number of datasets (x); in this case, we must produce 200 sets of values.
2. The mean is 20 for 'mice_before' and 21 for 'mice_after'.
3. The last parameter is the standard deviation, 2 and 2.5, respectively.

The primary assumption was that the 'mice_after' variable would have a broader range since it has a greater standard deviation and a higher mean, as that was what was assigned. Using the 'mean(x)' and 'IQR(x)' on the variables, the assumption was proven correct. As seen in the table below, 'mice_before' had a lower range (19.9) and a mean (2.93) than 'mice_after'. This suggests that the nutritional supplement may have positive effects on weight for mice.

Next, two additional values were created to simulate the rat's weights, named 'rats_before' and 'rats_after'. Both followed a Weibull distribution, so the 'rweibull' function was used. 'rats_before' had a shape of 10 and a scale of 20. 'rats_after' had a shape of 9 and a scale of 21. Using the inputs, it can be assumed that 'rats_after' would have a greater range as its scale was larger than 'rats_before'. Applying 'mean(x)' and 'IQR(x)' again, with respect to 'rats_before' and 'rats_after', gives a larger IQR and lower mean for 'rats_before'. This is because a higher shape value increases its scale so that the value has a larger range. Overall, this suggests that the nutritional supplement increases rats' weight, although further investigation is required.

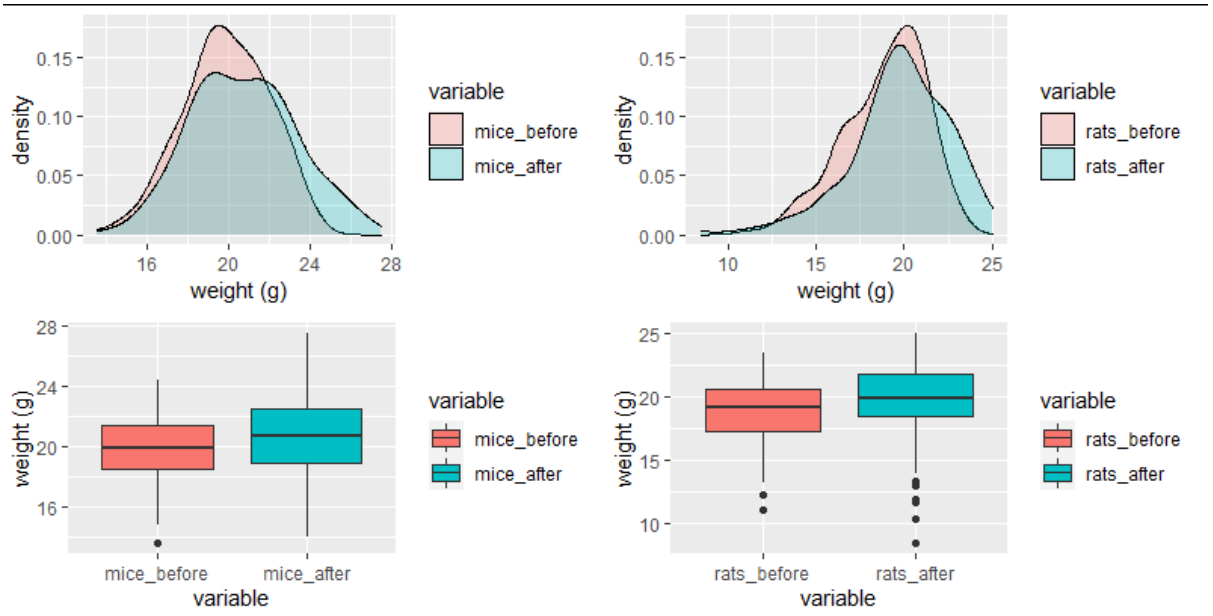
Test	mice_before	mice_after	rats_before	rats_after
Mean	19.87869	20.73398	18.86941	19.74168
IQR	2.928404	3.649012	3.31902	3.296799

To compare the values, 'qplots' can provide the framework for various plots, such as histograms or scatter plots. In this case, 'density' was selected. Qplots require data frames, so 'mice_before' and 'mice_after' were merged into one dataframe named 'mice', and 'rats_before' and 'rats_after' were joined into a separate dataframe called 'rats'. Four variables were created, each corresponding to a different dataset within the two dataframes, and 'density' was selected. The 'gridExtra' function allowed all four graphs to be visualised side by side (as seen below). Using the prior information, the mice dataset's density will most likely be higher in 'mice_before'as the standard deviation is lower than 'mice_after'. In the rat's dataset, it can also be assumed that 'rats_after' will have a higher density than 'rats_before' due to the higher mean and lower IQR, as seen in the table.

When the data was plotted, 'mice_after' demonstrated a lower density than 'mice_before'. Also, the density positioning was higher, with a more significant number of mice weighing up to 28g. 'mice_before' had a peak (mode) at approximately 19g. In contrast, after the supplement, there are now two peaks between 19g and 22g. This suggests that the mice have gained weight after treatment, such that most are concentrated around 19-22 grams. In addition, the 'rats_before' dataset had a mode of 21g, which decreased to approximately 20g in 'rats_after' as the density shifted towards the upper end of the graph. This suggests that the rats were gaining weight.

A second investigation was carried out using qplots, this time with a boxplot. The dataframes were already created, so they were reused. From the density plot and the data gathered in the table above, we can assume that; 'mice_after' will have a larger range and higher mean than 'mice_before', and that 'rats_before' and 'rats_after' will have negative skews. 'rats_after' will also have a higher upper quartile than 'rats_before', as seen in the density plot.

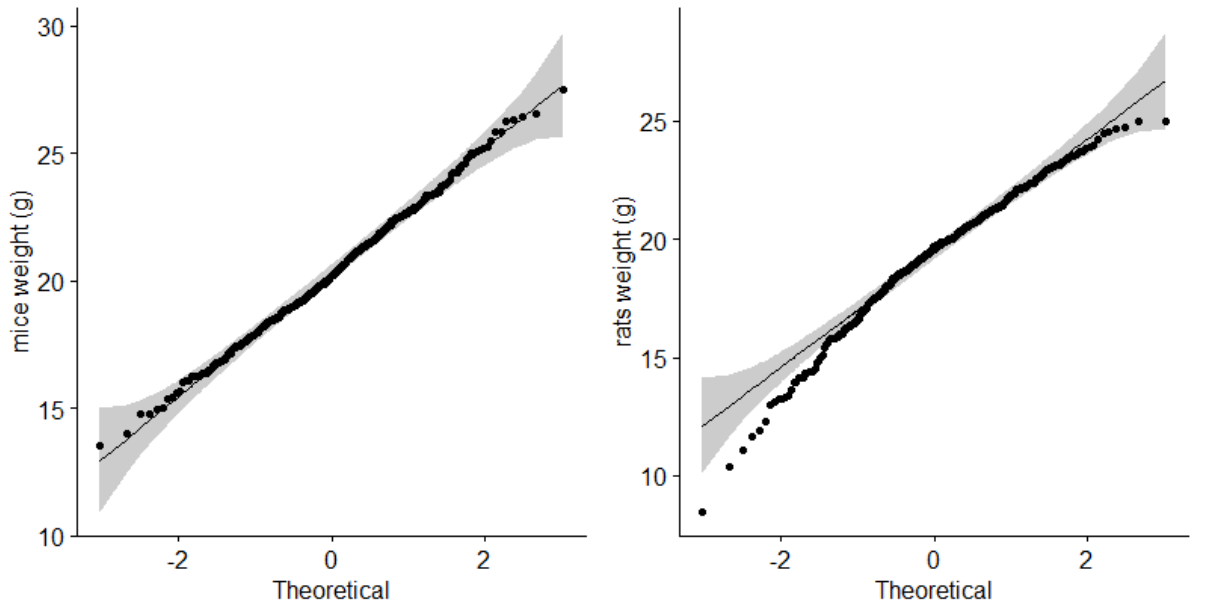
By interpreting the plot, the mice's mean weight has increased from approximately 20g to 21g. For rats, the mean weight has increased from approximately 19g to 20g. Mice maintained a normal distribution with no skew. However, rats moved from a negative skew to a slight positive skew, suggesting a significant increase in the distribution of rats in the upper quartile and proving the assumptions made earlier correct. There is evidence in both boxplots that suggests the nutritional supplement has caused a weight increase.



Task 2 - Appropriateness for Hypothesis t-testing

To continue with the investigation, hypothesis t-testing was used to check for normal distribution within the datasets. QQ plots are a qualitative test that provides a graph of the samples against each quantile. If the data forms a straight diagonal line, the data is normally distributed. There may be some deviation at the upper and lower bounds of the line, but the

line's center is the most significant part. By plotting the mice dataset, it was assumed that they would follow a normal distribution as the data was generated using the 'rnorm' function to create random variates. By analysing the QQplot, it is clear that the plots form a diagonal line with some deviation at the ends of the line. This suggests there is most likely a normal distribution for the weight of mice. The distribution plot showed a skew for rats, so they would not likely follow a normal distribution. By analyzing the output for rats, the assumption was correct. The plots stray from the diagonal line more than mice, suggesting the data is not normally distributed.



Another way to test for normal distribution is a quantitative Shapiro-Wilk test. The Shapiro-Wilk test takes the dataset and returns a 'p-value'. If the value is less than 0.05, the null hypothesis is accepted. The null hypothesis states that there is no significant difference between the sets of data. In this case, the null hypothesis (H0) would mean a normal distribution between the mice's weights (or rat's). Looking at the QQ plot for the mice dataset, it would be predicted that there is a normal distribution. The p-value returned is 0.8057, which exceeds the 0.05 limit, so the null hypothesis is accepted. It can be assumed that there is a significant difference - the mice have a normal distribution.

Shapiro-Wilk normality test	W	P-value
Mice	0.99747	0.8057
Rats	0.9765	0.00000444

Using the Shapiro-Wilks test to investigate this further, the p-values returned for rats was 0.00000444. As the p-value was less than 0.05, mice are significantly different from a normal distribution. This infers that quantitative and qualitative tests support the hypothesis that rat's weight does not follow a normal distribution.

In summary, mice passed both quantitative and qualitative tests for normality; therefore, the dataset likely follows a normal distribution. However, rats failed both tests, so it can be confirmed that the data does not follow a normal distribution. To further investigate the

distribution, more tests will need to be carried out. Other parametric or non-parametric tests should be carried out to investigate further this hypothesis, like Spearman's or Pearson's tests.

Now that mice have been proven to follow a normal distribution, the most appropriate way to test that the nutritional supplement has impacted the mice's weight is a paired t-test. This will be covered in the next section.

Task 3 - Hypothesis Testing

A paired t-test is a parametric test used to deduce whether there is statistical evidence that the mean weight difference between mice before and after is significantly different from zero.

To do this, the mice dataframe was divided into two parts, ‘X’ and ‘Y’. ‘X’ represented ‘mice_before’ and ‘Y’ represented ‘mice_after’. Both were mapped to one another when passed into the ‘t.test()’ function. The function does three things:

- Calculates the difference between each pair, i.e. before and after
- Calculates the mean and standard deviation of the differences
- Compares the average difference to 0, if there is a significant difference greater than the confidence level (0.95) the null hypothesis is rejected.

For this context, if there is a statistical difference between the weights of the mice before and after, then we can assume that the mice’s weight has changed.

Using the previous data gathered, it can be inferred that mice's weight before and after will be significantly different. Therefore the null hypothesis is that there is no significant difference between mice's weight before and after consuming the supplement. The alternative hypothesis is that there is a significant difference between weights.

Output	Value
T-test statistic	-3.4107
Degrees of freedom	199
P-value	0.0007845
Confidence interval	0.95 (-1.3497852, -0.3607937)
Sample estimates	-0.8552895

The table above is formed by extracting the key values from the paired t-test output. The first value is the t-test statistic, a number that illustrates the standardised difference in the variation. The larger the absolute value, the more significant the reason to reject the null hypothesis. In this case, the t-test statistic was -3.4107. This is not enough to reject the null hypothesis, so the other values will also be examined. The second statistic is the degrees of freedom. This is calculated by subtracting one from the number of samples (200-1), leaving 199. This value represents the estimated number of independent samples that were used in the calculation. 200 was used instead of 400 as the t-test was paired; the mice in the 'X' are

the same as the mice in 'Y'. This value is used to determine the t-distribution, which is used to calculate p-values and t-values. The next value is the p-value, 0.0007845. The p-value is used to measure the probability that the difference observed was by chance and thus reject the null hypothesis. A value greater than 0.05 would otherwise suggest that the null hypothesis is true. Since the value is less than 0.05, the mice's weight did change significantly, and the null hypothesis is rejected. We can then assume that nutritional supplements did have an impact on their weight. For this investigation, the confidence interval was 0.95 or 95%. The value is the range at which the values are most likely to lie. In this case, 95% of the mice lie between the upper and lower bounds. The upper and lower bounds given were -1.3497852 and -0.3607937, respectively. This is the range that 95% of the mice lie with respect to the weight difference before and after the treatment, excluding the 5% discrepancy. The final value from the paired t-test is the sample estimates. This represents the mean difference in weight between mice before and after. The value given is -0.8552895, which indicates that the mean weight increased by a value of 0.8552895.

To conclude, the evidence from the paired t-test comparing the weight of 'mice_before' and 'mice_after' heavily suggests a change in the mice's weight from using the nutritional supplement seen by the sample estimates. However, as the p-value does not exceed the threshold we cannot assume that the supplements changed the weight significantly.

To test the rat's dataset, a non-parametric test on the rat's dataset was required. Non-parametric tests differ from parametric tests by not making assumptions about the distribution of the data. This means that the rat's weights are assumed not to have a normal distribution, which is most likely true. The non-parametric test used on the rats dataset is the Wilcoxon signed-rank test with continuity correction. Like the t-test, this test investigates the significance of two or more data pairs (rats) and returns a value to determine said significance. A null hypothesis is also needed, H0, stating that the mean difference in weight for rats before and after is zero. Alongside the alternative hypothesis, H1, where the mean difference is not zero.

The test works by:

1. Calculating the difference between the rats before and after weight
2. Ordering the pairs from smallest to largest absolute values
3. Ranking the pairs from smallest to the largest difference
4. Calculating W
5. Calculating Z
6. Calculating P

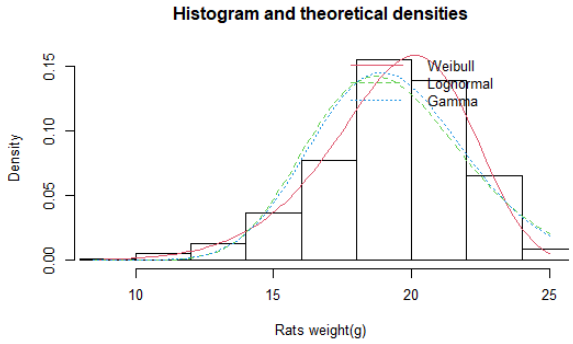
Output	Value
V	7055
P-value	0.0002584
Confidence interval	95 (-1.5724236, -0.4246354)
Sample estimates	-0.8992499

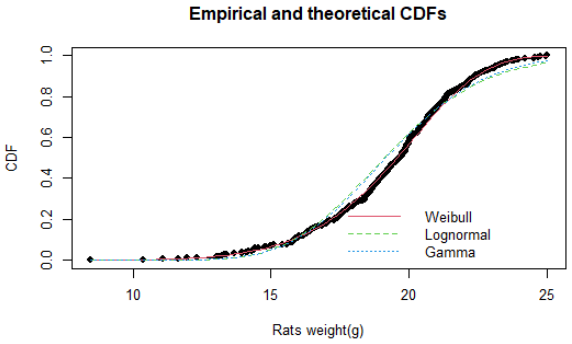
Of the values returned in the test, the most significant is the p-value and sample estimates. Like the t-test, the confidence interval contains 95% of the mean weight difference between the rat population before and after. The sample estimate is -0.8992499, meaning that the average increase in weight was 0.8992499 grams after the supplement. The p-value obtained 0.0002584, which is less than the 0.05 critical value and is statistically significant. The null hypothesis (H0) can be rejected from this alone, and the alternative hypothesis accepted. Therefore there is a statistical difference between the weight of rats before and after the nutritional supplement.

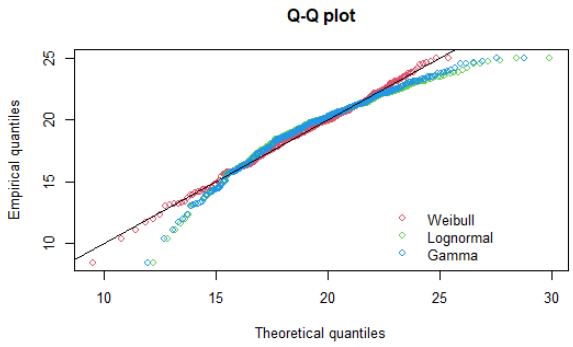
Task 4 - Fitting Distributions

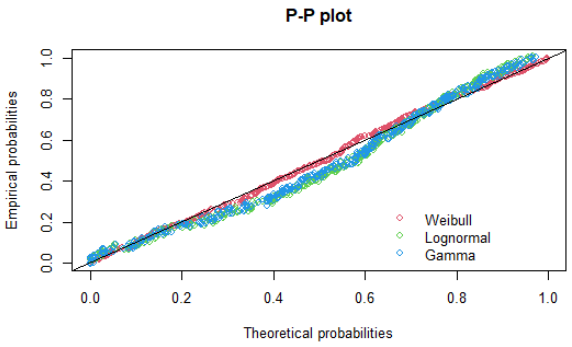
For the final investigation, the 'fitdist' function will examine the rat's dataset for the best-fit distribution (Weibull, Lognormal, or Gamma). Four plots will be used; Density, CDF, QQ, and PP. Of all the distributions, Weibull will be the best distribution for the rat's dataset as it was used to create the data at the start of the investigation. However, for the sake of investigating the similarities between distributions, Lognormal and Gamma will also be compared.

The density diagram displays the density of the rat's weights at 1 gram intervals. It can be seen that there is a negative skew with a mode of 20-21 grams. Of the three distributions, the Weibull accurately represents the density as its mode and skew are the most alike. Both Lognormal and gamma have similar distributions; however, they do not accurately demonstrate the rats' distribution. By looking at the CDF, there is a slight ascent with a sharp plateau at the end, suggesting a mode at 20-22 - which coincides with the density plot. Of the three densities plotted against the graph, Weibull deviates the least from the CDF. Lognormal and gamma follow a sharper path that exceeds the values of the plots. There is a similar output in the QQ plot. While lognormal and gamma densities rarely follow the diagonal line by either going above or below, Weibull follows the line almost precisely, with minimal deviation. For the final graph, the PP plot, there is a more significant variation from the diagonal line from Lognormal and Gamma distributions. Weibull follows the plot with a minor divergence from the diagonal line, with the lowest deviation. By visually assessing the four graphs' outputs, it can be assumed that the Weibull distribution is the best fit. This is because the Weibull distribution is the closest fit to each of the graphs' actual line. Therefore provides the best distribution for the rat's dataset.


Histogram and theoretical densities
Density
0.00
0.05
0.10
0.15
10
15
20
25
Rats weight(g)
Weibull
Lognormal
Gamma


Empirical and theoretical CDFs
CDF
0.0
0.2
0.4
0.6
0.8
1.0
10
15
20
25
Rats weight(g)
Weibull
Lognormal
Gamma


Q-Q plot
Empirical quantiles
10
15
20
25
10
15
20
25
30
Theoretical quantiles
Weibull
Lognormal
Gamma


P-P plot
Empirical probabilities
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Theoretical probabilities
Weibull
Lognormal
Gamma