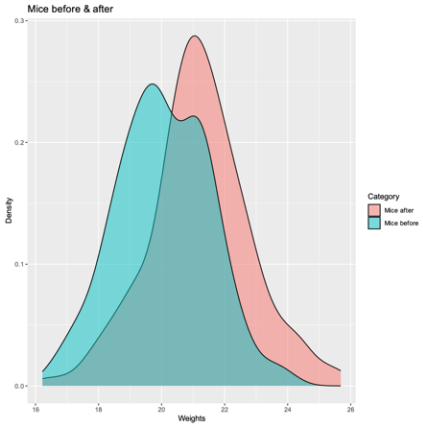
Matthew Vallance (001225832) – COMP1814: Hypothesis testing in R

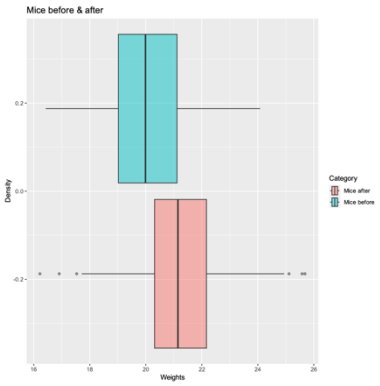
Hypotheses testing in R

Task 1 – Data generation

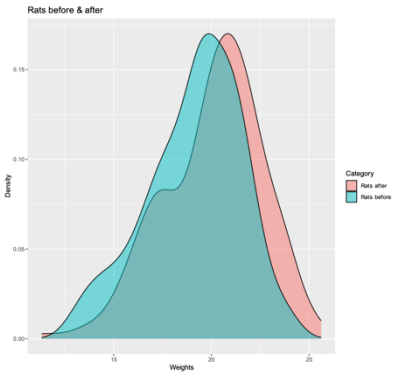
In task 1, we generate synthetic data for mice and rats. The aim is to create 200 datapoints of their weights before and after receiving treatment. We then analyze and compare this data using a density plot and a boxplot.



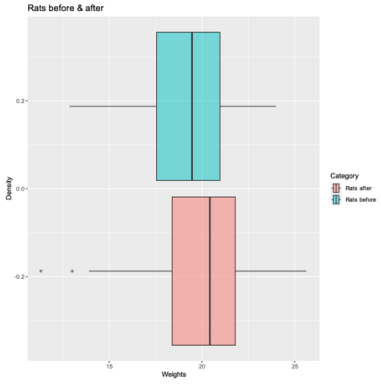
First, we look at the mice data we generated. The density chart (left) displays the distribution of the mice weights before and after treatment. The chart tells us that the weights of the mice before treatment have a bimodal distribution, this could tell us that there is a clear difference in the weights of male and female mice for example, though we would need a more in-depth dataset to confirm this hypothesis. The weights before treatment have a peak at around 19.8 and a peak of about 21.1 after treatment. This suggests that the treatment increases the weight the mice by about 1.3 on average. The chart also tells us that the variance of the weights is slightly larger after treatment than it is before as the weights are a little more dispersed.



The boxplot (right) tells us that the interquartile range increased, indicating more variability in the weights after treatment. There are some outliers in the weights after treatment, which suggests there are anomalies, which is expected as the data was generated with a higher mean and variance. The density plot and boxplot both show distribution of the weights, but they display some different features. The density plot shows the shape and spread, whereas the boxplot shows the median, quartiles and outliers.



Next, we will look at the rats dataset we generated, we used a Weibull distribution here so we expect a different set of outcomes.
This density chart (left) shows the distribution of the rats' weights before and after treatment. The weights of the rats before treatment have a right-skewed distribution. The weights after treatment have a more symmetrical distribution. This means that before, more rats had a lower weight than after. The chart also tells us that the mean weight of the rats has increased after treatment, as the top of the curve shifts slightly further to the right. The variance of the weights also increases after treatment, we can see this because the curve is generally a bit wider and flatter.



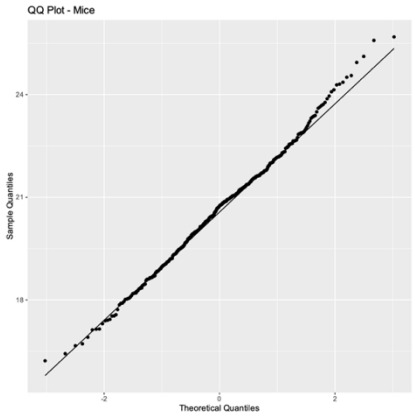
The boxplot (right) tells us the median and the interquartile range are higher after treatment, showing a clear increase in the mean and variance of the weights. There are two outliers in the after group, significantly less than the mice dataset, so there aren't as many anomalies in this data.

Both boxplots and density charts come to the same conclusion which is that the treatment clearly affect the weight of mice and rats respectively.

Task 2 - Appropriateness for hypothesis testing

In task 2, we discuss the appropriateness of whether our generated datasets pass the normality qualitatively (using QQ plots) and normality quantitatively (Shapiro-Wilk test). QQ plots (quantile-quantile plots), compare the quantiles of a dataset with the quantiles of a theoretical distribution, in this case the normal distribution. If the points on the QQ plot are close to the straight line, it shows that the dataset passes the test.

The Shapiro-Wilk test also checks to see whether the data is normally distributed. It checks against the null hypothesis, which is that the data is normally distributed, that is to say it assumes the data is normally distributed and then provides a p-value. If the p-value is less than 0.05, we can assume that the data is not normally distributed, and if it's more than 0.05 we can assume it's normally distributed, though we cannot prove that it is just from this measure. It also gives us a W number which is the difference between the observed distribution and the expected normal distribution (Statistics Kingdom, n.d.). The closer this number to 1, the more likely it is the data is normally distributed.

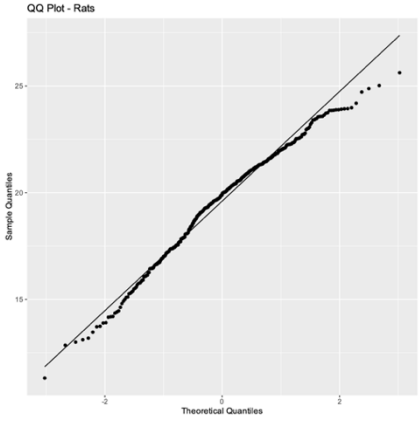


First, we will analyze the mice dataset generated in task 1. The QQ plot (left) shows how the quantiles of the mice dataset compare to the quantiles of the normal distribution. The points are generally close to the normal distribution line, therefor the mice dataset follows the normal distribution relatively well. There are some outliers at the start and end of the normal distribution line, but they are not very extreme, which further confirms that the data is normally distributed.

We then perform the Shapiro-Wilk test on this data.
The test (using R command: "`shapiro.test(mice$values)`") gave me a p-value of 0.5069. The p-value is greater than 0.05, so we can assume it's normally distributed. The test also gave me a W number of 0.99639, which also gives us a decent confidence the data is normally distributed. This supports what we look at above from the QQ plot. So, we can say that the mice dataset we generated in task 1 has a normal distribution both by looking at the plot and using both outputs of the Shapiro-Wilk test. This data passes both the normality qualitatively and normality quantitatively tests.

Next, we analyze the rats dataset. By looking at the QQ plot (below), we can clearly the points deviate from the distribution line, therefor, we can determine that it is not normally distributed.

We then perform the Shapiro-Wilk test on the rats data.
The test (using R command: "`shapiro.test(rats$values)`") gave me a p-value of 0.0001508. The p-value is significantly less than 0.05, so we cannot assume it's normally distributed in this case. The test also gave us a W number of 0.98341, which also concludes that the data is not normally distributed. This confirms what we see from the QQ plot, which tells us that the data points deviate from the normal distribution line. So, we can say that the rats dataset does not have a normal distribution both by looking at the plot and using both outputs of the Shapiro-Wilk test. This data does not pass either the normality qualitatively or the normality quantitatively test.



Task 3 - Hypothesis testing

In task 3a, we perform a paired t-test on the mice dataset which compares the means of two samples and determines whether they are statistically different from each other (Frost, n.d.). The null hypothesis of this test is that the mean difference is zero and the alternative is that they are not zero. The table shows the results of the t-test, using the mice weights before and after as the samples.

Output	**Value (rounded 4dp)**	
T-test statistics	-7.2251	
Degrees of freedom	199	
P-value	1.044249e-11 (not rounded)	
Confidence interval	-1.4721	-0.8409
Sample estimates	-1.1565	

Firstly, the T-test statistics here show that there is a significant difference between the mice weights before and after the treatment. The negative value indicates that the mice before treatment have a lower mean than after, which tells us that the mice generally have a lower weight before treatment.

Next, degrees of freedom represent the total number of values that can vary in the data, this is calculated by subtracting 1 from the total number of items in our data, therefor 199 tells us that the t-test is considering all 200 points of the mice dataset in our results.
The p-value is below 0.05 and is very near to zero (1.044249e-11 or 0.00000000001044249), this tells us that there is a high-chance we can reject the null hypothesis with this data, thus the mean difference is not zero.

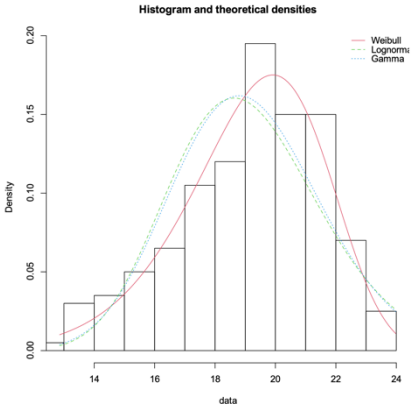
The confidence interval is given in a range which is -1.4721 to -0.8409. The default confidence interval is set to 95% in R; therefore, this range is the true difference in means between the two groups. Since it doesn't include zero, it suggests there is a significant difference between pairs.

The sample estimate is the mean of the differences in the data as a whole. The value of -1.1565 tells us that the mice before have a lower mean weight than mice after. We would expect this as throughout our tests we've found that the treatment does affect the weight of the mice.

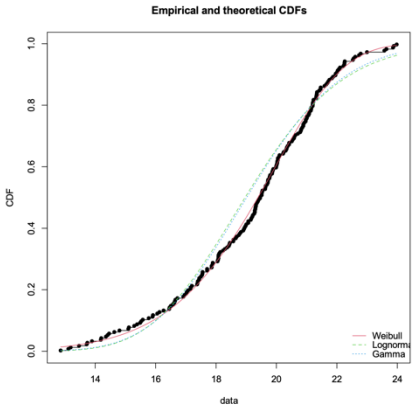
In task 3b, we perform a non-parametric test on the rats' data, for this we use the Wilcoxon test. The Wilcoxon rank sum test compares the ranks of our two independent samples (rats before and after treatment), it tests the null hypothesis that the two samples come from the same population, this is tested against the alternative that it is not. The output for the rats' dataset is a W statistic of 15454 and a p-value of 8.438e-05 or 0.00008438. The W statistic is simply a sum of all the datapoints in both samples. As above, the p-value is below 0.05, this tells us that there is strong evidence to suggest that we should reject the null hypothesis with this data. From performing this test, we can safely say that this data doesn't come from the same population.

From these results, if this were real data, you would have some cause for concern as this data was from two different sets of rats, or that the results of the weights of the rats could have been altered after the trial had completed.

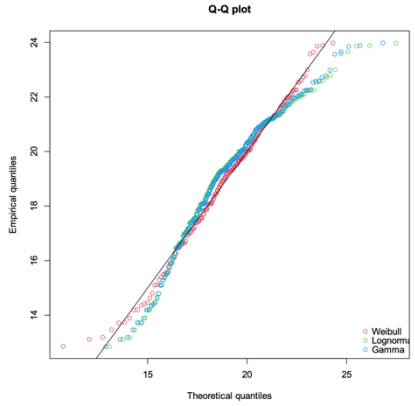
Task 4 - Fitting distributions



In task 4, we are required to examine the best-fit distribution for the rats dataset, I decided to use the original data before treatment here. The first chart is a Histogram (left) and theoretical densities chart. In this, we have fitted three distributions to the dataset: Weibull, Lognormal and Gamma. These curves illustrate how well the data fits these distributions. We can see that the data follows the Weibull distribution, as the top of each bar is roughly in line with the Weibull line, however it's not that clear from this chart alone. The data is not normally distributed as there isn't a clear bell-shaped curve here. We can also see in this chart that the data doesn't as closely fit into either of the other distributions.

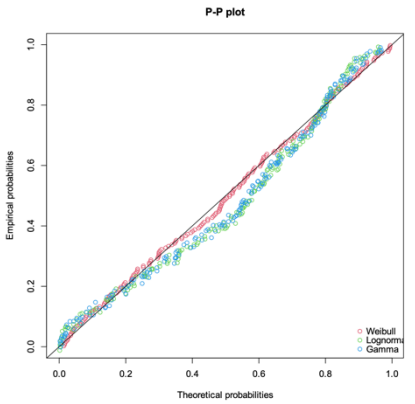


From this CDF (right), we are comparing the weibull, lognormal and gamma distribution to the empirical data, from the plot we can also see that the data aligns closely with the Weibull distribution, which further proves that the weibull distribution fits the theoretical data, some discrepensis could be explained by outliers in the datasets, we could further imrpove the accuracy with a larger sample size. This gives us a much better idea of which distribution this data follows, you can clearly tell from this chart when compared to the densities above.



The QQ plot (left) is comparing the theoretical quantiles (colored dots) to empirical quantiles (straight black line). All of the distributions follow relatively close here, however the Weibull distribution is closest overall to the emperical quantiles, this further proves that the data follows the Weibull distribution.

The PP plot (below) is vizually similar to the QQ plot, however, it is comparing the theoretical probabilities (colored dots) to empirical probabilities (straight black line). Again, from this chart we can see that the Weibull distribution is the closest fit. At point , both the Lognormal and Gamma distributions drift away from the emprical probabilities, we can clearly see in this plot that the data follows the Weibull distribution.



In conclusion, from the four plots we have generated in this task, three of them (CDF, QQ and PP) show that the data closely follows the Weibull distribution. The first chart doesn't show this as clearly however it still points to the same conclusion.

References

Frost, J., n.d. *Statistics by Jim.* [Online]
Available at: https://statisticsbyjim.com/hypothesis-testing/t-test/
[Accessed 10 March 2024].
Statistics Kingdom, n.d. *Statistics Kingdom.* [Online]
Available at: https://www.statskingdom.com/doc_shapiro_wilk.html
[Accessed 8 March 2024].