**Hypothesis testing in R.**

**Comparing means and fitting distributions.**

**Task 1**

To create the required dataset of mice, the “rnorm(n=200, mean=20, sd=sqrt(2))” function was used to generate 200 random observations following a normal distribution with specified parameters for “before treatment”. It is worth noting that the function requires standard deviation which is a square root of variance. The same function with different parameters was used to create “After treatment” records.

Similarly, “rweibull(n=200, shape=10, scale = 20)” function was used for “Before treatment” and “rweibull(n=200, shape=9, scale = 21)” for “After treatment” was used to create a dataset for rats.

Next, the density quick plot was used to compare the “Before” and “After” observations for both mice and rats datasets. Figure 1 demonstrates that most mice’s weights were centered around 20 grams before the treatment, but then dispersed with the new mean of 21.

A graph of a weight scale

Description automatically generated with medium confidence A graph of a normal distribution

Description automatically generated

Figure 1. Density plot for the “Mice” dataset. Figure 2. Density plot for the “Rats” dataset.

Figure 2 shows the distribution of the “Rats” dataset. One can immediately conclude that this is not a normal distribution, because neither of the graphs in Figure 2 are symmetrical. This plot also demonstrates that the rats’ weight also became more dispersed after treatment with the new scale of 21 compared to the original distribution, which was mostly centred around 20 grams.

A graph with a row of rectangular objects

Description automatically generated with medium confidence A graph with a couple of rectangular objects

Description automatically generated with medium confidence

Figure 3. Boxplot for “Mice” dataset. Figure 4. Boxplot for "Rats” dataset.

Both Figures 3 and 4 demonstrate that the weight of mice and rats increased after treatment. Moreover, boxplots show the outliers in both experiments as dots. Both boxes seem to be sliced exactly in the middle, and their tails look symmetrical, suggesting that the records in Mice dataset follow a normal distribution. This is not the case for the Rats dataset as both boxes are not sliced in the middle.

**Task 2**

A graph with a line

Description automatically generated A graph with a line

Description automatically generated

Figure 5. Q-Q plot for the Mice dataset. Figure 6. Q-Q plot for the Rats dataset.

Q-Q plot demonstrates whether data is normally distributed. Since the values closely lay on the line in Figure 5, the weights of mice are most likely normally distributed. However, because in Figure 6, the smaller values lay quite far from the line, there is a high chance that the data is not normally distributed.

A screenshot of a computer code

Description automatically generatedShapiro-Wilk test is a quantitative test to check if the data is normally distributed. If the value of p is significantly more than α = 5% (or 0.05) then, the data is normally distributed which is the case for mice. If it is less, then, the data does not follow a normal distribution, which is the case for rats. It is vital to know the distribution of the population as that helps to choose the right tests for the hypothesis. If the populations is normally distributed, a paired t-test can be used, otherwise – a different test has to be performed

Figure 7. Shapiro-Wilk Test results.

**Task 3**

A computer screen with white text

Description automatically generatedA paired t-test can compare two sets of data (i.e. “Before” and “After” treatment), and check if there is a difference between them. In this case, it will prove or disprove the hypothesis that the data has / has not changed. To begin with, the null hypothesis (H0) is: there is no change in the weight of mice. Mean of weight\_before = mean of weight\_after. The alternative hypothesis (H1) is: there is a change and mean of weight\_before =/= mean of weight\_after. The function “t.test(mice$weight ~ mice$weight\_group, data = mice, paired = TRUE)” provides the following results (Figure 8).

Figure 8. Results of paired t-test for the Mice dataset.

The critical value (t) represents the significance of the results. Df stands for degrees of freedom, and shows the number of variable observations. Because there are 200 observations with fixed mean values, just one observation has to be fixed with the degree of freedom = 199. Similar to Shapiro-Wilk test, if P-value is smaller than α = 0.05, the null hypothesis can be rejected. Since p is about 0.0000000002, there is enough evidence to reject H0. Therefore, H1 hypothesis is true, and there is a change in the weight after the treatment. Additionally, that also means that the difference between two means (“Before” and “After”) is not equal to 0. There is also 95% confidence that the true difference between the two means is in the interval from 0.6861657 and 1.2618343. The “sample estimates” provide the mean difference of 0.974, which indeed lay in the given range. To conclude, the weight of mice after the treatment has changed.

A screenshot of a computer program

Description automatically generatedA similar test is performed on the “Rats” dataset. However, a different test has to be performed since the data does not follow normal distribution. One of the approaches would be to use Wilcoxon rank sum test. The outputs are similar to t-test above, with p-value also being significantly less than α (0.000006 < 0.05). Therefore, the H0 (there is no change in weight after treatment) can be rejected, and alternative hypothesis (mean weight\_before =/= mean weight\_after, implying a change in weight after treatment). This time, there is a 95% confidence that the difference in location is in the range from 0.6300304 to 1.5500756, with the real difference in location being equal to 1.099935. To summarise, there was a change in rats’ weight after the treatment.

Figure 9. Results of Wilcoxon rank sum test for the “Rats” dataset.

**Task 4**

A graph of a normal distribution

Description automatically generated A graph with numbers and lines

Description automatically generated

Figure 10. Density and CDF function for the “Rats” dataset. Figure 11. Cullen and Frey graph for the “Rats” dataset.

Function “plotdist(rats$weight, histo = TRUE, demp = TRUE)” shows what a well-fitted distribution would look like (Figure 10) by providing density and cumulative distribution (CDF) functions of the data. Whereas, “descdist(rats$weight, discrete=FALSE, boot=1000)” can provide a hint on which distributions are most appropriate for the data. In this case, logarithmic, gamma, and Weibull can be suggested (Figure 11).

After fitting all three distributions (Figure 12), the summary is then requested in Figure 13. The model with the largest Loglikelihood, and the lowest Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) (Figure 13). After careful consideration, Weibul distribution would be the best fit for this data. Next, Figure 14 demonstrates how each of the A black background with white text

Description automatically generateddistributions would fit the data.

Figure 12. Fitting Weibul, lnorm and gamma distributions to the “Rats” dataset.

A screenshot of a computer program

Description automatically generated A group of graphs showing different types of data

Description automatically generated

A graph of a number of data

Description automatically generatedFigure 13. Summary of all distribution fittings. Figure 14. Density, CDF, QQ and PP lots of all three distributions.

After plotting Weibull distribution and the weights of the rats, it is clear, that the data follows the distribution closely.

Figure 15. Plotting Weibull against the data.