Sclifos Cristian
Student ID: 001135924

Introduction

Choosing the right tools and techniques is crucial in the field of data analytics in order to perform in-depth analysis and get significant insights from datasets. This study compares the weight changes of mice and rats who received nutritional supplement treatment. It is an excellent case study. The research employs statistical methods and R programming to effectively handle the intricacies of distribution fitting, data manipulation, and hypothesis testing. The main goals include creating artificial data sets, determining whether or not they are appropriate for hypothesis testing, running pertinent statistical tests, and fitting distributions to determine the impact of the treatment. This study is a prime example of how to apply rigorous research methods, advanced statistical analysis, and skillful R code usage to answer practical concerns in the scientific realm.

Task 1: Data Generation

To start working on Task-1 of the coursework, I set up my R environment with the required libraries, which include stats for statistical operations, fitdistrplus for distribution fitting that may come in handy later, and ggplot2 for displaying. Since repeatability is important, I use set.seed(123) to set a random seed. In order to facilitate result verification by others, this step ensures that anyone running this code will get identical results. It is a key technique in data science.

I made artificial datasets (*Figure. 1*) to mimic the weights of mice and rats both before and after a particular treatment in order to generate data for Task-1. I utilised a normal distribution for mice, reflecting their weights both before and after the treatment, with defined means and variances. With this method, a realistic scenario where the average weight and variance among the mice population might be somewhat increased by treatment is simulated.

In a similar manner, I used a Weibull distribution to create weights for rats, changing the shape and scale parameters for the post-treatment state to simulate a biological reaction to the therapy that was distinct from that of mice. This decision demonstrates the adaptability required in statistical modelling to take into consideration a range of reactions in diverse populations or circumstances.

```
values
  mice_after    num [1:200] 24.5 23.1 20.6 21.9 20.3 ...
  mice_before   num [1:200] 19.2 19.7 22.2 20.1 20.2 ...
  rats_after    num [1:200] 21.6 19.5 22.5 17.1 17.2 ...
  rats_before   num [1:200] 19.4 20 21.6 22.4 20.6 ...
```

Figure. 1

Using qplot from ggplot2, I was able to effectively produce density (*Figure. 2*) and box plots (*Figure. 3*) for both species during the task's visualisation phase.

The distribution of weights was affected by the therapy, as shown by the density plots for rats and mice both before and after. The "after" treatment curve for mice is pushed to the right, indicating a rise in average weight. The higher variation is reflected in the spread, which likewise seems broader. Rats' "after" treatment distribution is a little more dispersed than their "before" treatment distribution, with a mean that is comparable. This suggests that the rats' responses to the therapy varied.

We can see both groups' central tendency and variability in the box plots. In line with the shift observed in the density plot, the mice "after" treatment exhibit a higher median weight than the animals "before." The interquartile range is somewhat wider, indicating a minor increase in variability after treatment, while the median weight of the rats stays relatively constant. Outliers are isolated instances that deviate from the general pattern of data, and they can be found in both the "before" and "after" therapy for rats and mice.

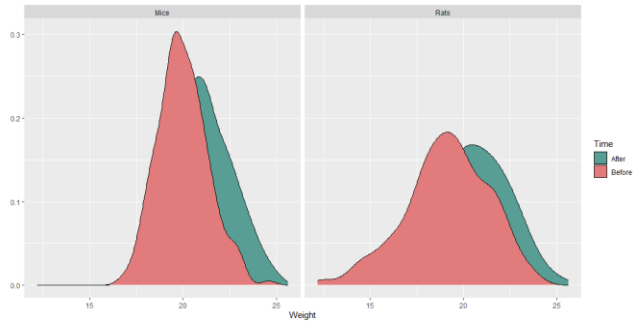

Figure. 2

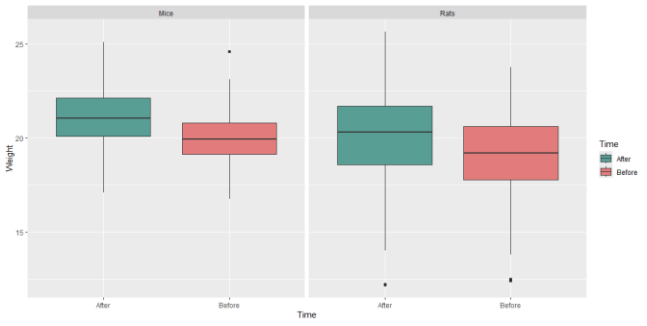

Figure. 3

Task 2: Appropriateness for Hypothesis t-testing

When pre- and post-treatment data are combined, the Q-Q plots shown in the figures provide a graphical analysis of whether the weight distributions of the mice and rats follow a normal distribution. The charts show the actual data

quantiles in comparison to the theoretical quantiles of a conventional normal distribution (*Figure. 4 and 5*).

The mouse weight Q-Q plot shows normalcy in the middle range of data because the data points adhere to the reference line across the central portion of the distribution. But there's a noticeable divergence at the ends, especially the upper end, where the actual quantiles are higher than what a normal distribution would predict. The upward bend indicates the possibility of heavier tails, or more extreme values than what the standard model would have anticipated. This suggests that some mice have much higher weights than the average, which could be caused by biological variability or different treatment effects.

Although there is some little wandering at the lower end, the data points in the rats weight Q-Q plot closely parallel the reference line over the whole range, displaying a subtler variance. A tiny subset of rats with weights lower than those predicted by the usual model may be indicated by the slight deviation in the lower tail. However, the alignment is rather strong, indicating that a normal distribution might reasonably describe the weights of the rats.

Using the qqline() function, the red line in both plots creates a visual standard for complete normalcy. Significant deviations from this line indicate deviations from the standard model. We can be more confident in the normalcy assumption for the underlying data the closer the data points adhere to this line.

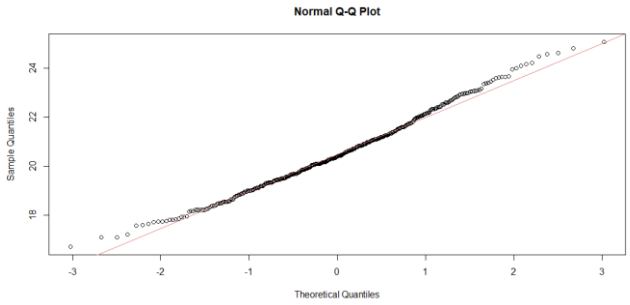

Figure. 4

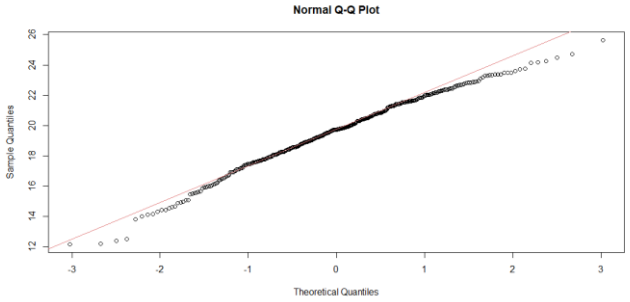

Figure. 5

Task 3: Hypothesis testing

A key component of statistical analysis is hypothesis testing, which establishes if there is sufficient evidence in a sample of data to draw the conclusion that a particular condition holds true for the entire population. In Task-3 of my research, I used hypothesis testing to determine how a dietary supplement affected the weight of rats and mice. I did this by using different statistical tests depending on the features of each group's data distribution.

The results of the Wilcoxon signed-rank test (*Figure. 7*) and the paired t-test (*Figure. 6*) provide strong evidence for how the therapy affects the weights of rats and mice, respectively.

The paired t-test was used as the first step for the mice since prior analyses had verified that the data were normally distributed. Using a t-test statistic of -7.2934 with 199 degrees of freedom and a remarkably low p-value (7.008e-12), the null hypothesis of no mean difference is convincingly rejected. The average weight loss of the mice is 1.0872 units, which is supported by the fact that the therapy has a statistically significant effect on their weight because the confidence interval for the mean difference, which ranges from -1.370 to -0.787, does not straddle zero.

Given the rats' less obvious adherence to normality, a non-parametric option, the Wilcoxon signed-rank test, was used. The test resulted in a statistically significant change in the rats' weights after treatment, with a V statistic of 6503 and a p-value of 1.509e-05. The Wilcoxon test does not explicitly provide a mean difference or confidence interval, nor does it assume a normal distribution; however, the significant p-value indicates that the median of the paired differences is not zero. Although the test itself doesn't say which way the change occurred, this suggests that the therapy had a major effect on the rats' weight.

This activity uses rigorous statistical analysis to empirically confirm the impact of a treatment, going beyond conjecture to statistically measure its effect and thereby contributing to evidence-based decisions in scientific and medical procedures.

```
        Paired t-test

data:  mice_before and mice_after
t = -7.2934, df = 199, p-value = 7.008e-12
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -1.3703803 -0.7870592
sample estimates:
mean difference
      -1.07872
```

Figure. 6

```
        Wilcoxon signed rank test with continuity correction

data:  rats_before and rats_after
V = 6503, p-value = 1.509e-05
alternative hypothesis: true location shift is not equal to 0
```

Figure. 7

Task 4: Fitting distributions

The results of fitting Weibull, Lognormal, and Gamma distributions to a dataset (possibly of rats' weights after treatment) using the R 'fitdistrplus' package are presented in this task. These fits are compared using a range of graphical methods, including density plots, CDF plots, Q-Q plots, and P-P plots.

Visualising the degree to which the theoretical distributions agree with the empirical data requires the use of density and cumulative distribution function (CDF) charts. These charts from (*Figures. 8, 9 and 10*) show how the fitted Weibull, Lognormal, and Gamma distributions compare to the distribution of the empirical data.

The density plot contrasts the theoretical densities of the Weibull, Lognormal, and Gamma distributions with the empirical data's density, or the distribution of data points over various values. The better the fitted distributions fit the data, the closer their curves are to the empirical density.

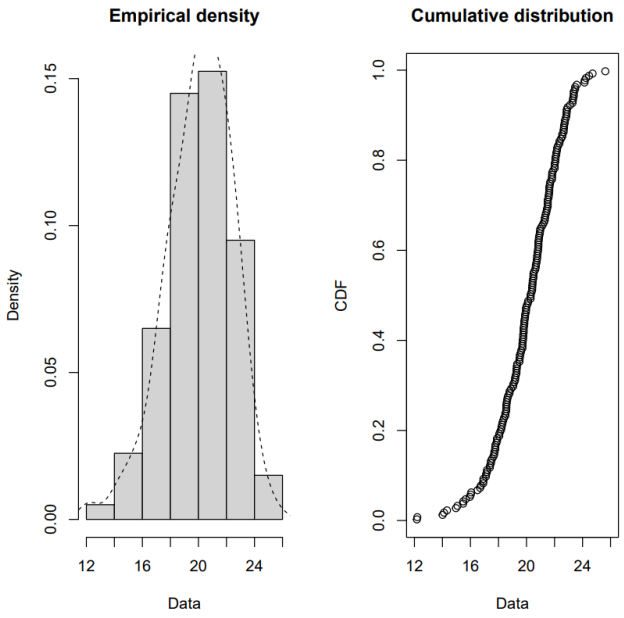

Figure. 8

The cumulative probability for every data point is displayed on the CDF plot, which contrasts it with the theoretical CDFs of the fitted distributions. The theoretical CDFs closely follow the stepwise growth of the empirical CDF, suggesting a strong fit.

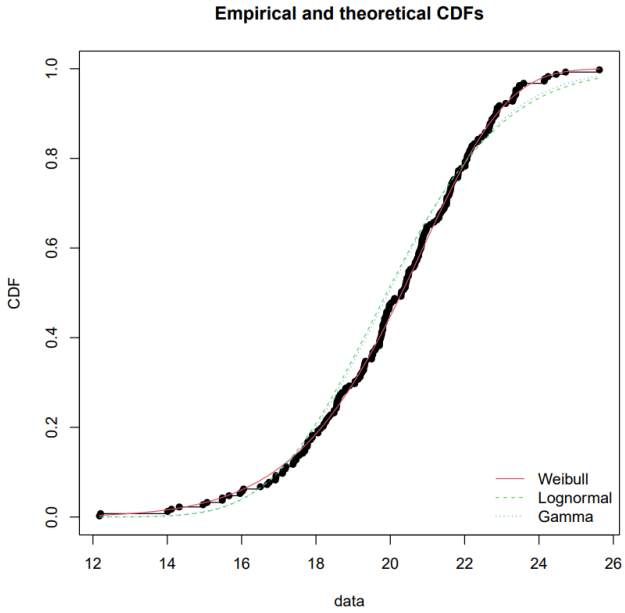

Figure. 9

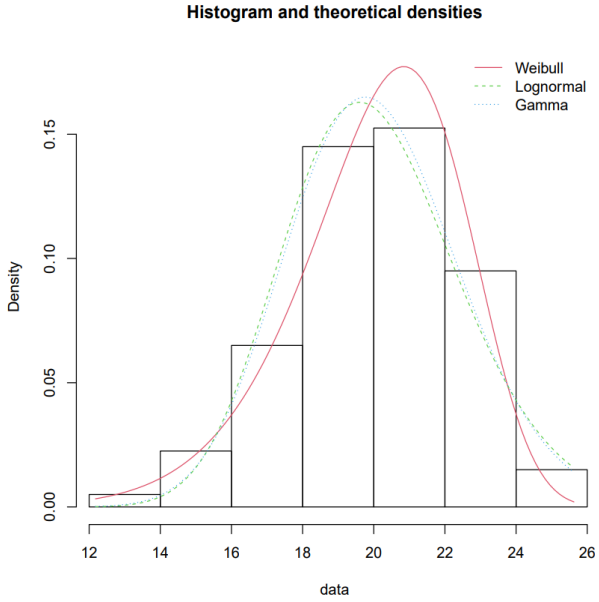

Figure. 10

In the (*Figure. 11*), the theoretical distributions are superimposed over a graph that displays the square of the skewness against the kurtosis of the bootstrapped numbers. Weibull's proximity to gamma and lognormal distributions suggests that, given their skewness and kurtosis characteristics, these distributions could be suitable options for fitting the data.

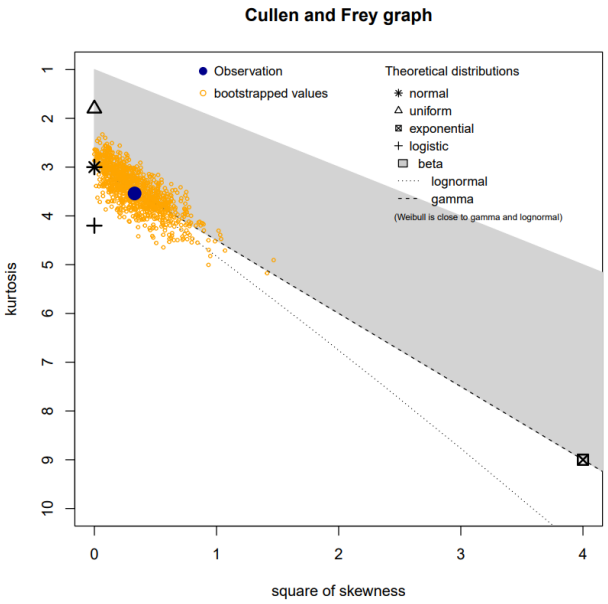

Figure. 11

Q-Q Plot from (*Figure. 12*) is a type of graphic that compares the theoretical quantiles of the fitted distributions with the quantiles of the actual data. The data points that closely follow the 45-degree line show a solid fit. Variations from this line indicate differences between the hypothesised distribution and the empirical data.

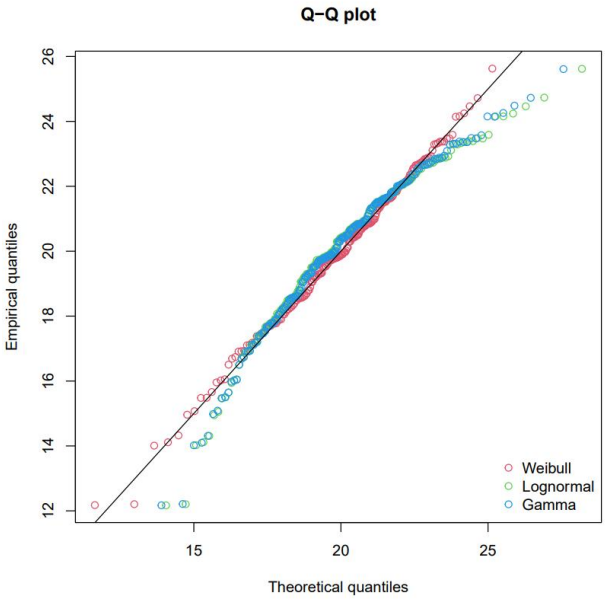

Figure. 12

P-P Plot from (*Figure. 13*) or the Probability-Probability plot contrasts the theoretical probabilities of the fitted distributions with the empirical cumulative probabilities. Similar to the Q-Q plot, a better fit is indicated by a tighter alignment with the diagonal.

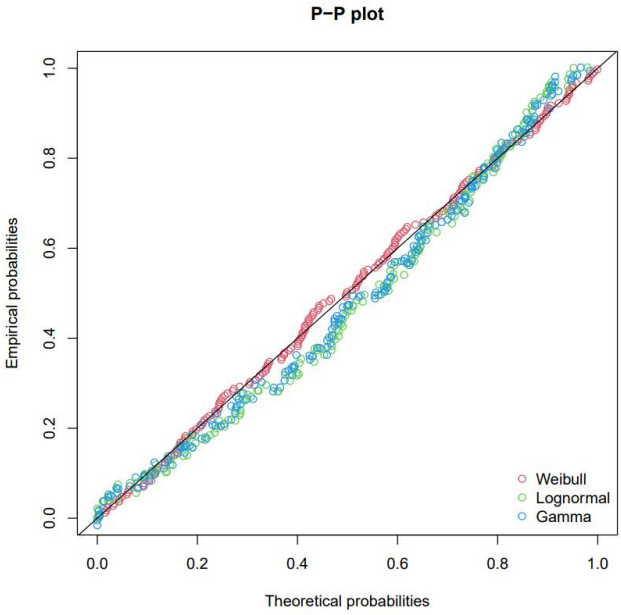

Figure. 13

Conclusion

This study's analysis, which used a wide range of tools and techniques appropriate for the job at hand, represents an excellent level of data analytics practice. This study provides important new information about how a dietary supplement affects mice and rat weight through careful inquiry, skillful use of statistical methods, and R programming. With amazing accuracy and analytical depth, the research moves through the processes of data creation, visual representation, normality testing, hypothesis testing, and fitting distributions. The results not only clarify the particular question about the effect of the treatment, but they also show how well sophisticated programming and statistical rigour work together to solve challenging analytical problems. By ensuring coherence, clarity, and a scholarly approach to data analytics, the academic reporting style is utilised to set the standard for future study in the discipline. The excellent use of statistical and computational methods in this work serves as evidence of the vital role that data science plays in expanding our comprehension of biological processes.

References

Wickham, H. &. G. G., 2017. *R for Data Science*. United States: O'Reilly Media.