Hadil Bati
Student ID: 001090152

HYPOTHESIS TESTING IN R
Comparing means and fitting distributions

INTRODUCTION:

200 mice were given a treatment "Nutritional Supplement" within 6 months, following an investigation on how the treatment has an impact on the weight of mice. For an extensive research within this investigation, an addition 200 Rats were given the same treatment however results were found to differ. The way in which data was generated within the investigation, was by measuring the weight of the mice before and after treatment, as well as measuring the weight of rats before and after treatment. This gives us 200 sets of values before and after treatment for the two groups.

The aim of the study which I am following, is to create the two datasets for each set of mice and rats from artificial data and follow through with an investigation on hypothesis testing as well as normality testing within the variation of the two datasets. The study's expected results include the various aspects that; the Mice dataset would come from a normal distribution whilst rats come from a Weibull, as the values would have a distribution close to normal. Further comparisons would be associated with the Shapiro-Wilk test, and that Mice will most likely pass the normality test, therefore rejecting the null hypothesis, and going on to test the assumptions through a paired t-test, whilst the Rats will most likely not pass the normality test, following through with a non-parametric test and therefore accepting the null hypothesis.

TASK 1- DATA GENERATION:

To begin the study and investigation on the effect of the treatment "Nutritional Supplement" that was given to the 200 mice and rats across a 6-month period, the generation of artificial data was the necessary starting step. Within this task, I created the two separate datasets for Mice and Rats by creating functions that I named **'generateMiceData'** and **'generateRatsData'**. The Mice data is required to come from a normal distribution which is defined by the following probability density function where the μ is the mean and the σ^2 is the variance.

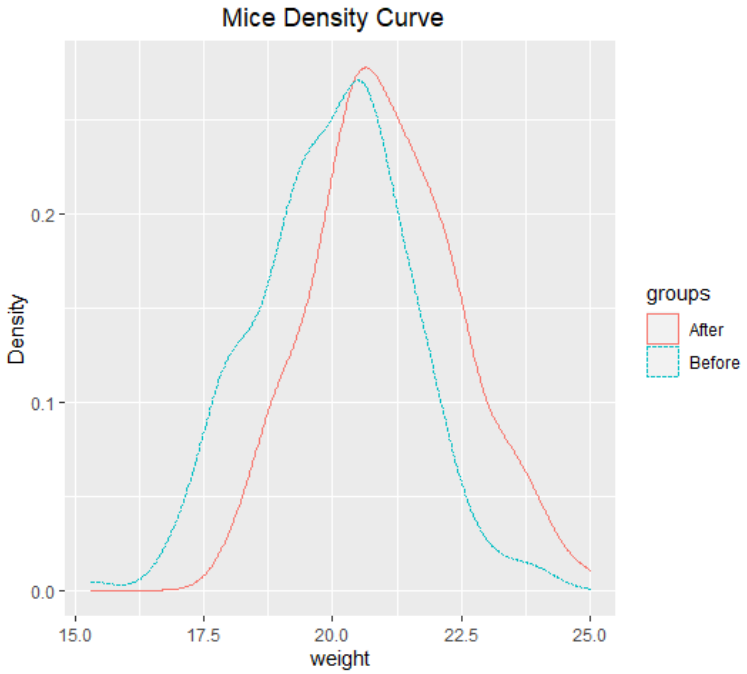
Within the Mice, I created variables for the **weight_before**, in which I used the command: rnorm(200, mean = 20, sqrt(2)) and the **weight_after** using the command: rnorm(200, mean = 21, sqrt(2.5)). Rnorm is used for generating the random variates and for that, I generated the "before" ~ $N\ (\mu^1\ ,\ \sigma^2)$ mice data coming from a normal distribution with the mean = 20 and the variance = 2, therefore making use of its square root. Likewise, with the "after" ~ $N\ (\mu^1\ ,\ \sigma^2)$ Mice data, I generated 200 more random variates with rnorm, coming from a normal distribution of mean = 21 and the variance = 2.5. Lastly, creating a data.frame() with the weight, including the different groups.

Contrasting to the way in which the Mice data was generated, the Rats data followed a use of rweibull() with the "before" treatment generating 200 random deviates with a shape of 10 and scale of 20, coming from the Weibull distribution, as well as the "after" treatment, generating 200 more deviates with a shape of 9 and

scale of 21. Creating the Rats data.frame() by similarly creating a column for the weight which includes the groups of the 'before' and 'after'.

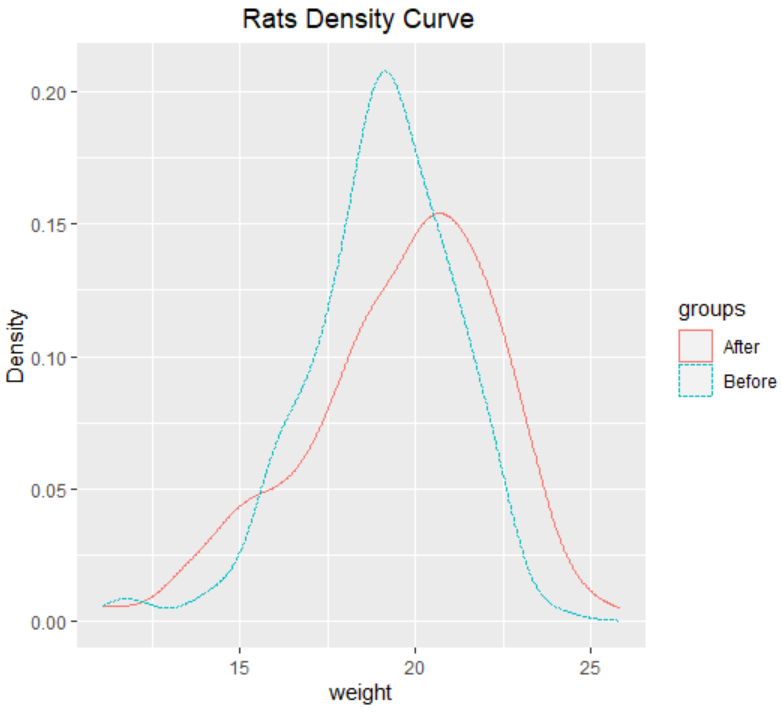
Following the generation of the data for both Mice and Rats datasets, I used the function 'Qplot' with 'geom – density' to create visual plot of comparison for the Mice(before, after) as well as the Rats(before, after). Figure 1 follows the display of comparison of the groups within the Mice dataset.

Figure 1: 'Mice Density Curve.'



Using the Qplot to create visual representation of the "before" and "after" groups within the Mice dataset. With the sampling data of 200 rnorm weight values, both groups are not highly differentiated in their normal distribution, as they both have a symmetric, single peaked and bell-shaped curve. Despite the "Before" group having a slightly asymmetric gradual increase of the Density whilst the "After" group having a slightly asymmetric decrease of the Density curve. With the "Before" group having the smaller mean and variance unit than the slightly bigger mean and variance within the normal distribution of the "After" group, both display a similar stance within the visual Qplot.

Figure 2: 'Rats Density Curve.'



Using the Qplot to create the visual representation of data within the Rats dataset, we see a drastic difference in density between the two groups. Whilst the "Before" has the smoother distribution of weight with a normal curve being symmetric, single peaked and bell-shaped, the "After" group is found to have a smaller and slightly more asymmetric density curve. The drastic difference in the densities of the two groups theoretically could be due to the Weibull distribution, which creates variables that are not distributed normally, resulting in density peaks with a skewed and more asymmetric form of a density curve.

Hadil Bati
Student ID: 001090152

Figure 3: 'Mice Boxplot.'

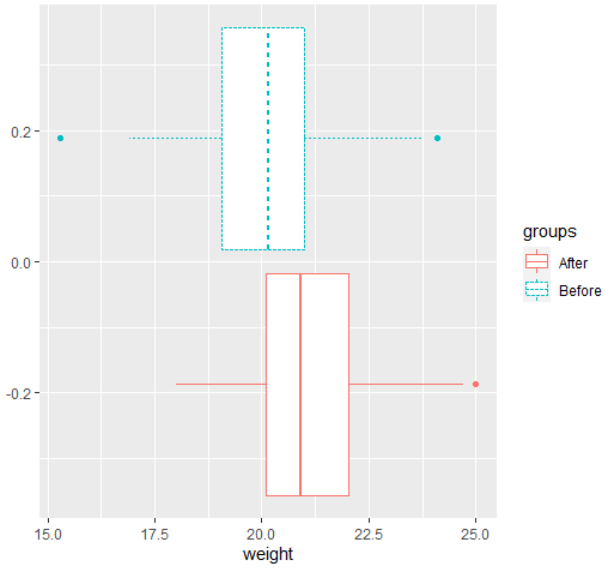
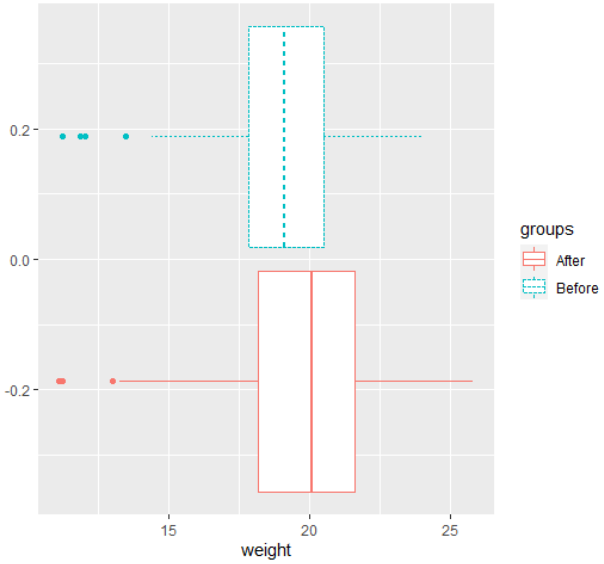


Figure 4: 'Rats Boxplot.'



Here we have plotted the weight and colour by group. Performing the same operation of the Qplot, using geom = 'boxplots. Boxplots are a way of displaying the distribution of the data based on the 5-number summary which is the min, the first Quartile, the median, third Quartile and the max. In the Mice boxplot, we can see that the 'after' group has a higher max as opposed to the 'before'. There are also various outliers plotted.

Within the Rats data boxplot, there are many more outliers past the minimum of the groups. It is also found that the 'after' group has a larger Inter-quartile Range (IQR), finding it has a higher third Quartile and first Quartile, as opposed to the 'before' group.

Figure 5: 'Paired Data for Mice.'

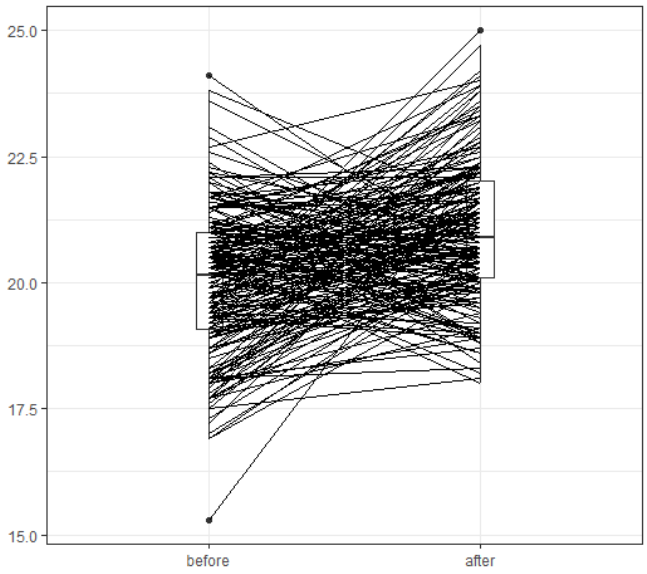
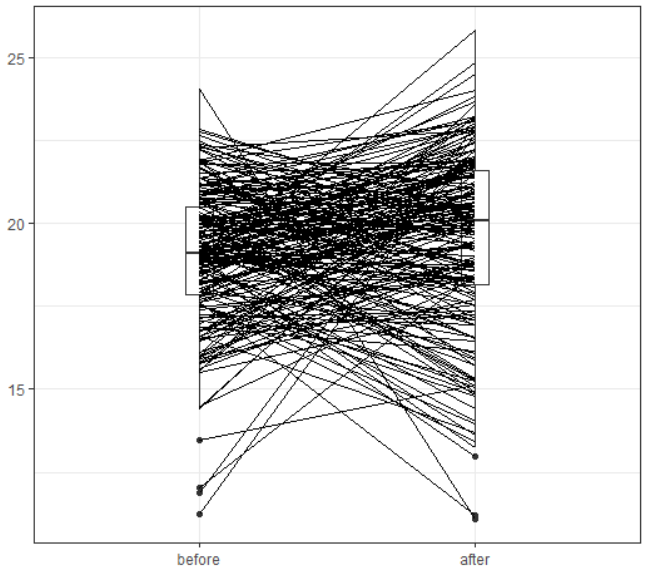


Figure 6: 'Paired Data for Rats.'



Using the library PairedData, I had also created the Paired Data boxplots which are visual insight on the paired t-test for the Mice and Rats data to be done in hypothesis testing. There is a connection of the 'before' and 'after' data within the datasets.

TASK 2- APPROPRIATENESS FOR HYPOTHESIS T-TESTING:

Within this task, I tested the appropriateness of the Mice and Rats datasets for the hypothesis t-testing, through examining whether the data passes normality qualitatively and quantitatively using the QQ plot and the Shapiro-Wilk test following a series of assumptions assigned.

Figure 7 below displays the Normal qualitative test of the Q-Q Plot for the Mice Dataset.

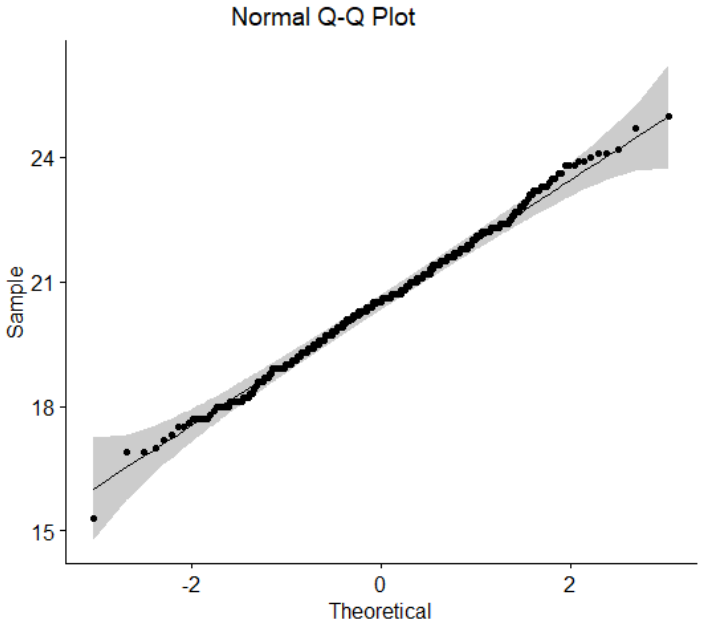


Figure 7: 'Normal Q-Q Plot' for Mice.

I examined whether the Mice data passes normality qualitatively, through the usage of the ggqqplot function from the ggpubr library. Similarly, to the Qplot density curve of the Mice groups within Figure 1, with Figure 7 we can see that the Mice dataset passes the normality qualitatively as the data is Normally distributed by being evenly aligned with the standard normal variate. We can therefore assume the normality.

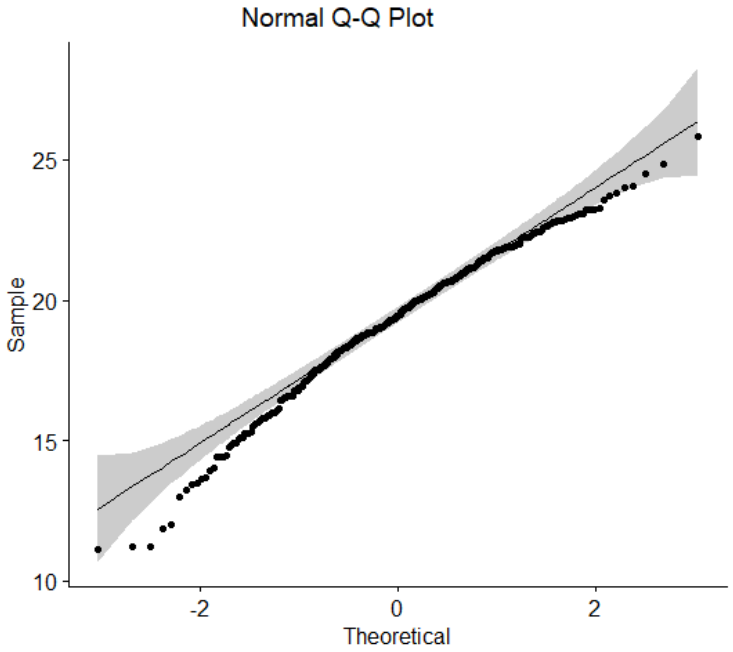


Figure 8: 'Normal Q-Q Plot' for Rats.

Within the Q-Q Plot of the Rats dataset, we can conclusively examine that the data does not pass normality qualitatively, as the Q-Q plot is skewed. With the theoretical quantiles on the x-axis and the sample quantiles on the y-axis we can see a peculiar shape on the Q-Q plot as a measure of skewness. We can distinctly see this through the bottom and top ends of the plot deviating from the straight line.

We previously had a similar insight to the way in which the Rats Q-Q Plot is to display as in Figure 2, the Rats groups had a density curve which was skewed and likewise asymmetric. Conclusively, we cannot assume the normality.

The Shapiro-Wilk test a quantitative normality test which is used in order to ascertain whether data shows a serious deviation from normality or not. It is a further and more significant inspection into the data as opposed to the visual inspection within plots. The Shapiro-Wilk's test is based on the correlations which are found between the

data and the corresponding normal scores given. It is a preliminary test which also further checks the paired t-test assumptions.

As a parametric and normality test, it examines the null and alternative hypotheses much like the paired t-test. In statistics, we can define the null hypothesis (H_0) as:

1. $H_0 : m = 0$ - is the mean difference (m) equal to 0?
2. $H_0 : m \leq 0$ - is the mean difference (m) less than 0?
1. $H_0 : m \geq 0$ - is the mean difference (m) greater than 0?

To accept our null hypothesis, the p-value needs to be ≥ α.
We can also define the alternative hypotheses (H_a) as:

1. $H_a : m \neq 0$ - (different)
2. $H_a : m > 0$ - (greater)
3. $H_a : m < 0$ - (less)

Some of these assumptions regarding the data of which we have about the Mice and Rats, includes:

1) Are the two samples paired?
   - Yes, the samples are paired as the data has been generated from measuring the weight of the same mice twice, similarly to the rats.
2) Is the data a large sample?
   - Yes, as the n > 30. The sample size is larger than 30, therefore we are not required to check if the differences of the groups follow a normal distribution.

Following the usage of the Shapiro.test() function for the Mice weight "before" and "after", we get various outputs which are displayed in Figure 9.

```
        Shapiro-Wilk normality test

data:  mst
W = 0.99499, p-value = 0.7491
```

Figure 9: 'Shapiro-Wilk test' for Mice.

From the output given, we can examine that the p-value is >0.05. This essentially means that the p-value of the data is greater than the significant level 0.05, therefore implying that the Mice data is not significantly different from normal distribution, in result we can assume the normality. Conclusively the paired t-test would be an appropriate test to test the hypothesis.

To further investigate the differentiation between the Mice and Rats datasets, we commence in the Shapiro-Wilk test on the Rats dataset to see if it passes normality quantitatively or not. The Figure below displays the output results.

```
        Shapiro-Wilk normality test

data:  rst
W = 0.98302, p-value = 0.01624
```

Figure 10: 'Shapiro-Wilk test' for Rats.

When I did the preliminary Shapiro-Wilk test on the Rats dataset, to check paired t-test assumptions, I got the p-value as <0.05 which means that it is greater than the significant

level 0.05, therefore being significantly different from normal distribution, and you cannot assume the normality.

Conclusively, since the data is not normally distributed, a non-parametric test such as the paired Wilcoxon test would be more applicable for the Rats dataset, as opposed to the paired t-test.

TASK 3- HYPOTHESIS T-TESTING:

The preliminary Shapiro-Wilk tests were conducted on the Mice and Rats datasets to check the paired t-test assumptions, furthermore, see which follow up hypothesis test is appropriate for each based on their results in passing or not passing normality.

With understanding that the Mice dataset has passed the normality test, the paired t-test would be required for examining the hypothesis testing. Using the t.test() function within R, Figure 11 displays the output results of the Mice Dataset.

```
        Paired t-test

data:  weight by groups
t = 7.1221, df = 199, p-value = 1.899e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.7589151 1.3400849
sample estimates:
mean of the differences
                 1.0495
```

Figure 11: 'Paired t-test results' on Mice.

With investigating the results of the paired t-test on the Mice dataset, we can come to various conclusions and understandings of the hypothesis as well as analyse the different parts of it. Essentially the t-test statistic value = 7.1221, whilst the degrees of freedom(df) = 199. The degrees of freedom refer to the number of independent values that the t-test analysis can estimate. Therefore, there is a df = 199, which are the number of values that are free to vary. Following the df, you have the given P-value which in this result is = 1.899^{-11}. The p-value is very essential to the hypothesis test as it gives the insight on whether to accept or reject the hypothesis.

In this instance, the p-value is less than the significant level α = 0.05. This means we can reject the null hypothesis and therefore conclude that the average weight of the mice before the treatment is significantly different from the average weight of the mice after the treatment with the given p-value = 1.899^{-11}. The alternative hypothesis given within the results is $\boldsymbol{H_a : m \neq 0}$.

Another aspect of the paired t-test result is the given confidence interval (conf.int) of the mean differences at 95% is = [0.7589151, 1.3400849]. Furthermore, the sample estimates are the mean differences between the pairs of the Mice data, giving us the mean = 1.0495

As opposed to the Mice data which we tested the hypothesis of through the paired t-test, the Rats data previously did not pass normality and was not normally

distributed, therefore following the non-parametric Wilcoxon test. This can be done through the R function wilcox.test().

```
        Wilcoxon signed rank test with continuity correction

data:  weight by groups
V = 12003, p-value = 0.0172
alternative hypothesis: true location shift is not equal to 0
```

When performing the Wilcoxon-test on the Rats dataset, we find the p-value of the test is 0.0172, which is less than the significance level alpha(α)= 0.05. We can therefore conclude that the median weight of the Rats before the treatment is significantly different from the median weight of the Rats after treatment with a p-value = 0.0172.

TASK 4- FITTING DISTRIBUTIONS:

Within Task 4, I used the function 'fitdist' from the 'fitdistrplus' library in R, to examine the best-fit distribution of the Rats dataset. In this, I fit a Weibull, lognormal and a Gamma distribution within the Density, CDF, Normal QQ and PP plots. These are shown in the figures below.

These are the fitting distributions of Weibull, Gamma and Lnorm. Essentially, with fitting distributions the best fit model is found to follow several criterions, one of which is the Loglikelihood. In accordance with that criteria, the Higher the Likelihood the more fitted the model is, regarding the data. Furthermore, there is also the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) which are required to be taken in consideration when understanding the best fitted model of distribution. The AIC and BIC tend to penalise the models based on the likelihood value; therefore, it is found that the lower the AIC/BIC, the better the model as a candidate.

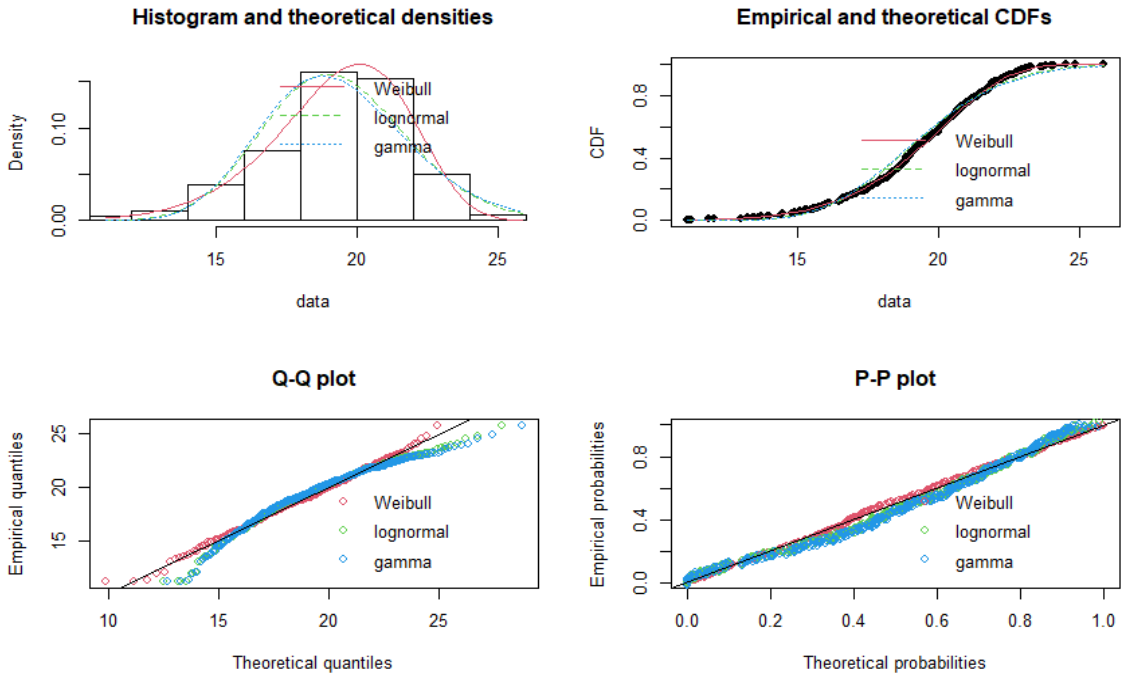
By investigating the Rats Dataset results, we can see that the Weibull distribution is found to be the best fitted model, as it follows the criterion- having the highest Likelihood at -915.6517 and the lowest AIC/BIC at AIC = 1835.303 and BIC = 1843.286

```
Fitting of the distribution ' weibull ' by maximum likelihood
Parameters :
       estimate Std. Error
shape  9.306927  0.3579303
scale 20.329965  0.1150649
Loglikelihood:  -915.6517   AIC:  1835.303   BIC:  1843.286
Correlation matrix:
          shape     scale
shape 1.0000000 0.3147656
scale 0.3147656 1.0000000


Fitting of the distribution ' gamma ' by maximum likelihood
Parameters :
       estimate Std. Error
shape 57.580692  4.0598033
rate   2.984721  0.2113587
Loglikelihood:  -938.5168   AIC:  1881.034   BIC:  1889.017
Correlation matrix:
          shape      rate
shape 1.0000000 0.9956615
rate  0.9956615 1.0000000
```

Hadil Bati
Student ID: 001090152

```
Fitting of the distribution ' lnorm ' by maximum likelihood
Parameters :
         estimate  Std. Error
meanlog 2.950974 0.006752002
sdlog   0.135040 0.004773208
Loglikelihood:  -947.0914   AIC:  1898.183   BIC:  1906.166
Correlation matrix:
         meanlog sdlog
meanlog       1     0
sdlog         0     1
```



These diagrams display all three distributions over a series of plots and through examining all the different plots, we can get visual representation that the Weibull distribution from the Rats data is the best fit model as it has the least skewed, symmetric, and single-peaked bell shape curve within the Histogram and theoretical densities model. It also is the most aligned with the standard variant line within the Q-Q plot, P-P plot as well as the Empirical and theoretical CDFs. Moreover, after commencing the study by using the Weibull distribution to generate the Rats Data and continuing to investigate the different normality tests and further hypothesis testing, we can understand that in these fitting distributions curves, the distribution which is best fit to the Data is the one that it was generated from- being the Weibull distribution.