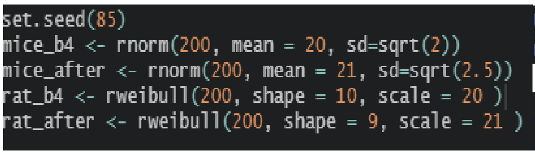
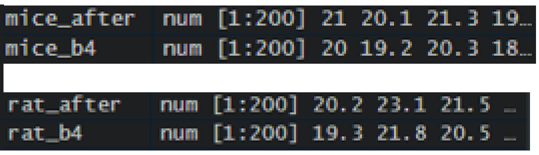
Hypothesis testing in R. Comparing means and fitting distributions

Data Generation. The data created for this coursework is artificially generated two data sets. The first data set contains 200 sets of values before treatment and 200 sets of values after treatment of mice weight records. The second data set contains 200 sets of values before treatment and 200 sets of values after treatment of rats' weight records. The first step to generate the artificial data set is to use “set.seed” function. This function will prevent the data that will be generated by the functions “rnorm” and “rweibull” from changing every time we run the program. The mice data sets (before, after treatment) were generated by using “rnorm” function. The “rnorm” function for generation of random deviates from the normal distribution that takes three parameters: n which is a number of samples to generate, mean and standard deviation. In the task description for mice data sets, we were given parameters of 200 sets of values, mean and variance whereas standard deviation is equal to the square root of the variance. The rats' data sets (before, after treatment) were generated by using “rweibull” function that generates random deviates for the Weibull distribution that takes three parameters: n which is a number of generated samples, shape and scale. All parameters were given in the task description. Please see the data generation code below:

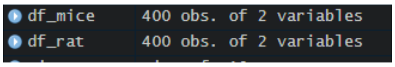


```
set.seed(85)
mice_b4 <- rnorm(200, mean = 20, sd=sqrt(2))
mice_after <- rnorm(200, mean = 21, sd=sqrt(2.5))
rat_b4 <- rweibull(200, shape = 10, scale = 20 )
rat_after <- rweibull(200, shape = 9, scale = 21 )
```



```
mice_after    num [1:200] 21 20.1 21.3 19…
mice_b4       num [1:200] 20 19.2 20.3 18…

rat_after     num [1:200] 20.2 23.1 21.5 …
rat_b4        num [1:200] 19.3 21.8 20.5 …
```

For some further experiments, I had to create data frame with accurate columns for weights and labels.



```
df_mice      400 obs. of 2 variables
df_rat       400 obs. of 2 variables
```

Density plots:

Figure1 represents the density plot for the rat variables. The blue curve describes rats before treatment whereas the red curve represents rats after treatment. Both examples are negatively skewered therefore the distributions have longer tails on the left side compared to the normal distribution. Two factors contribute to that statement. Firstly, long tails on the left side and short on the right side indicate negative skew. Secondly, the distances from the mean of each example to its mode. Both examples are unimodal. In addition, the blue curve smoothens at the peak whereas in comparison the red curve has one sharp peak therefore density is more focused in one weight range for the blue curve. The height of the red curve peak is a bit lower than the peak of the blue curve which indicates that the weight of the rats before treatment has a higher density in some weight range. Since the red curve is shifted more to the right, rats after treatment gained weight.

Figure2 is the density plot for the mice variables whereas blue is the mice before treatment and red after treatment. Both of the examples are unimodal. The distribution of the weight of mice before treatment is slightly skewed to the right which can be observed by looking at the tails and distance between mean and mode. The mice after treatment is skewered to the left with smaller differences from mean to mode distance and more even tails on both sides. Both of the examples are almost identical to the bell-shaped curve therefore close to normality. The blue curve has a higher peak than the red curve. This means that mice before treatment have higher density in some weight range. The red curve is more skewered to the left which indicates that mice after receiving the treatment gained weight. We know that data generated by ‘rnorm’ function is normally distributed although the sample size for our size is not large enough to prove it on the density graph.

Figure1

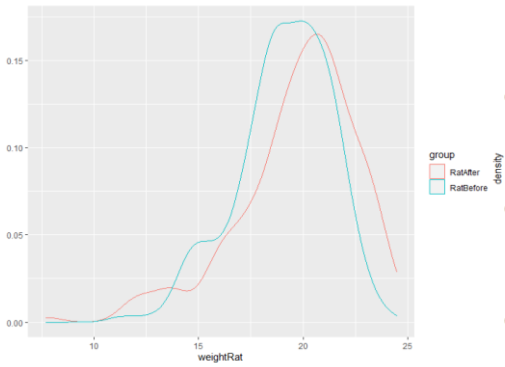
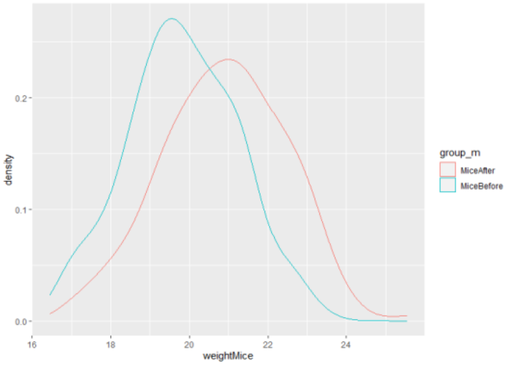


Figure2



Boxplots:

Figure3 is a boxplot representation for the mice dataset whereas red is mice after treatment and blue before treatment. The mice after the treatment weights range between around 17 to 24.1. There is only one outlier point at around 25.2 which means that this value point ranges away from the 25th or 75th percentiles (value is greater than 1.5 interquartile). This value represents a mouse, whose weight is quite different from the others in the data set. The line between the minimum and the lower quartile is longer than the line between the upper quartile and maximum. This means that the median is falling more into upper weights (closer to upper quartile) therefore it's left-skewed.

The mice before treatment weights range between around 16.5 to 21.2. There are no outliers points. The median line falls closer to the lower quartile which means data is skewed to the right. The line between the upper quartile and maximum is slightly longer which confirms positive skewness.

The range of the data that was reported for both examples shows that the weight of the mice after the treatment increases overall. In addition, one mouse after the treatment was significantly heavier than the rest of the reported weights.

Figure4 is a boxplot representation for the rats' dataset whereas red is rats after the treatment and blue before the treatment. The rats after the treatment weights range between around 13 to 24 with multiple outliers below the minimum value. The median is closer to the upper quartile which means that data is left-skewed. The peak of the weights can be observed between the median and upper quartile.

The rats' before treatment weight range between around 14 to 23. There are only a couple of outliers below the minimum value. The median is falling slightly closer into the upper quartile which means that data is left-skewed. As reported on the density plot Figure1, the peak of the rats before the treatment is smoothened at the top which can be observed on the boxplot as the median is located almost in the middle of the interquartile range (IQR).

The range of the data that was reported for both rats' examples showed that the weight of the rat after the treatment increased overall. Although rats gained weight after the treatment overall, they are multiple examples of rats that weight significantly decreases and is even lower than any recorded before the treatment. That can be observed by looking at the outliers. In addition, rats gained less weight after the treatment than mice which can be observed by comparison of the two ranges in which both data sets fall.

Figure3

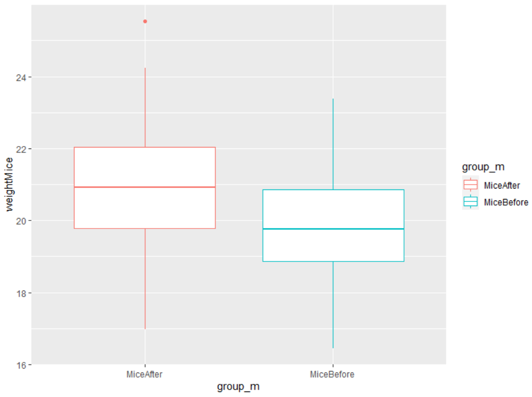
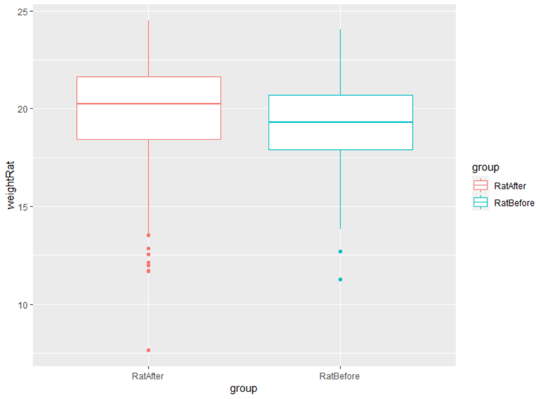


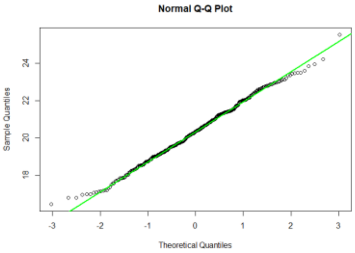
Figure4

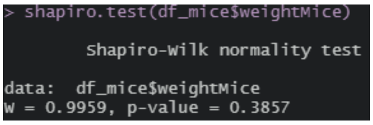


Appropriateness for Hypothesis t-testing

Mice combined before and after

The Quantile-Quantile plot for mice datasets implies normal distribution as the scattered data points are formed along the reference line. The Shapiro Wilk test p-value is above 0.05 which accepts the null hypothesis therefore normality can be assumed. Further, the analysis hypothesis of the mice dataset can be performed by a parametric test such as paired t-test. The t-test assumes that the data is normally distributed. It makes assumptions about the parameters of the population distribution from which the sample is taken.

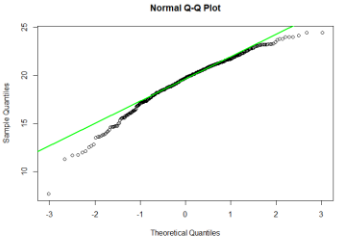


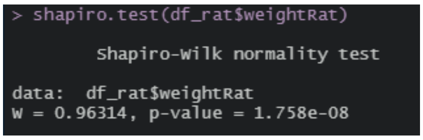


```
> shapiro.test(df_mice$weightMice)

        Shapiro-Wilk normality test

data:  df_mice$weightMice
W = 0.9959, p-value = 0.3857
```

Rats combined before and after

The Quantile-Quantile plot for rats datasets where data is not fully falling along the reference line. The data falls outside the reference line on both tails. This implies negative skewness. The Shapiro Wilk test rejects the null hypothesis as the p-value is less than 0.05. The data significantly deviate from the normal distribution. Further analysis of the hypothesis can be performed by using a non-parametric test. This test does not require a distribution to meet the assumptions to be analyzed.





```
> shapiro.test(df_rat$weightRat)

        Shapiro-Wilk normality test

data:  df_rat$weightRat
W = 0.96314, p-value = 1.758e-08
```

Hypothesis testing Definitions below explained based on (Hayes, 2022).

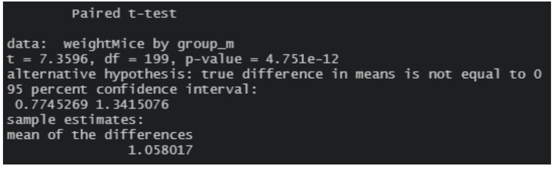
Paired t-test. T-test static “t” value which is a ratio of the difference between the mean of the two samples sets and the variation in these sample sets. A large value of T indicates a stronger argument to reject the null hypothesis.

Degrees of freedom is the maximum number of logically independent values, that may be different, in the data sample.

P-value which is the probability that sample results occurred by chance. The p-value for the mice data set is less than the significance level and the null hypothesis can be rejected. Therefore, the average weight of the mice before treatment is significantly different from the average weight of the mice after treatment. The alternative hypothesis is correct.

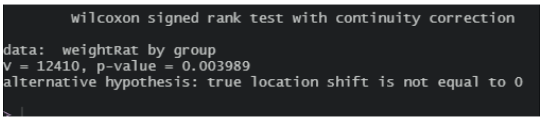
Confidence Interval is a probability that a population parameter will fall between a pair of values around the mean. This experiment was conducted using a confidence level of 95%.

Sample estimates are the difference between two sample means in the dataset.



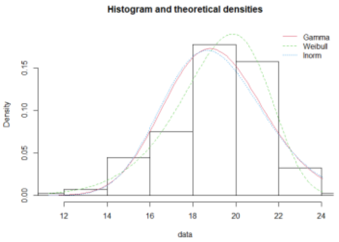
```
        Paired t-test

data:  weightMice by group_m
t = 7.3596, df = 199, p-value = 4.751e-12
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.7745269 1.3415076
sample estimates:
mean of the differences
               1.058017
```

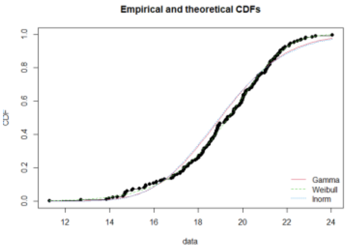
The non-parametric test “wilcox”. The p-value of this test is less than the significance level of $\alpha = 0.05$. Therefore, the median value weights of rats before the treatment is significantly different from the median value weight of the rats after the treatment. The null hypothesis is rejected and the alternative hypothesis is correct. . This is a paired sample Wilcox test with the following hypothesis: H0: rats before and after the treatment are equal, H1: rats before and after are different.

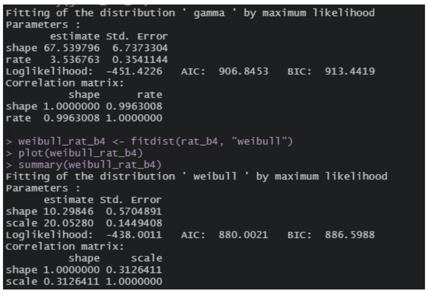


```
        Wilcoxon signed rank test with continuity correction

data:  weightRat by group
V = 12410, p-value = 0.003989
alternative hypothesis: true location shift is not equal to 0
```

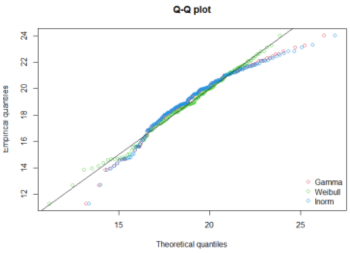
Fitting distributions Rat before:

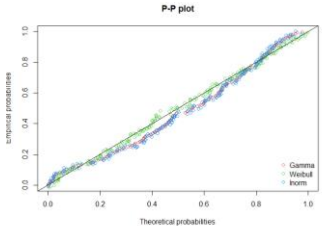


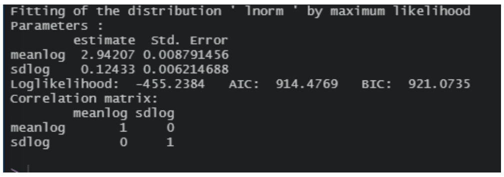




```
Fitting of the distribution ' gamma ' by maximum likelihood
Parameters :
       estimate Std. Error
shape 67.539796  6.7373304
rate   3.536763  0.3541144
Loglikelihood:  -451.4226   AIC:  906.8453   BIC:  913.4419
Correlation matrix:
          shape      rate
shape 1.0000000 0.9963008
rate  0.9963008 1.0000000

> weibull_rat_b4 <- fitdist(rat_b4, "weibull")
> plot(weibull_rat_b4)
> summary(weibull_rat_b4)
Fitting of the distribution ' weibull ' by maximum likelihood
Parameters :
      estimate Std. Error
shape 10.29846  0.5704891
scale 20.05280  0.1449408
Loglikelihood:  -438.0011   AIC:  880.0021   BIC:  886.5988
Correlation matrix:
          shape     scale
shape 1.0000000 0.3126411
scale 0.3126411 1.0000000
```

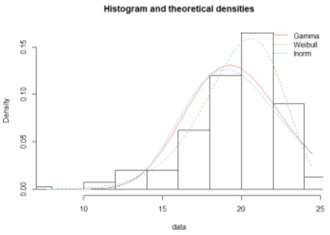


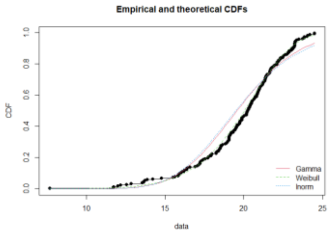


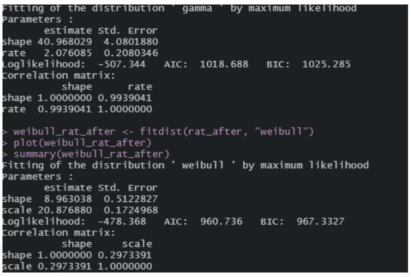


```
Fitting of the distribution ' lnorm ' by maximum likelihood
Parameters :
        estimate  Std. Error
meanlog  2.94207 0.008791456
sdlog    0.12433 0.006214688
Loglikelihood:  -455.2384   AIC:  914.4769   BIC:  921.0735
Correlation matrix:
        meanlog sdlog
meanlog       1     0
sdlog         0     1

>
```

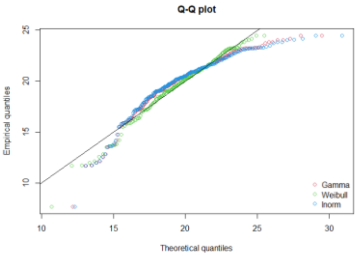
Rat After:

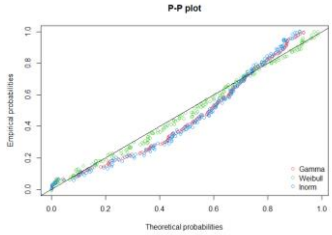


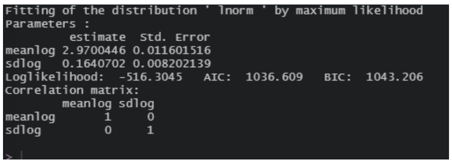




```
Fitting of the distribution ' gamma ' by maximum likelihood
Parameters :
       estimate Std. Error
shape 40.968029  4.0801880
rate   2.076085  0.2080346
Loglikelihood:  -507.344   AIC:  1018.688   BIC:  1025.285
Correlation matrix:
          shape      rate
shape 1.0000000 0.9939041
rate  0.9939041 1.0000000

> weibull_rat_after <- fitdist(rat_after, "weibull")
> plot(weibull_rat_after)
> summary(weibull_rat_after)
Fitting of the distribution ' weibull ' by maximum likelihood
Parameters :
       estimate Std. Error
shape  8.963038  0.5122827
scale 20.876880  0.1724968
Loglikelihood:  -478.368   AIC:  960.736   BIC:  967.3327
Correlation matrix:
          shape     scale
shape 1.0000000 0.2973391
scale 0.2973391 1.0000000
```







```
Fitting of the distribution ' lnorm ' by maximum likelihood
Parameters :
         estimate  Std. Error
meanlog 2.9700446 0.011601516
sdlog   0.1640702 0.008202139
Loglikelihood:  -516.3045   AIC:  1036.609   BIC:  1043.206
Correlation matrix:
        meanlog sdlog
meanlog       1     0
sdlog         0     1

>
```

In the experiment above, three different distributions perform testing on rats datasets: Weibull, Lognormal and Gamma. For both examples, there are various plots and summaries on the right side to examine which of the distributions fits our datasets best. In both examples, density plots show that Weibull distribution is the one that is skewed to the leftmost and it fits our description. The Q-Q and P-P plots show that Weibull distribution falls the most on the reference line which is another evidence that our data is Weibull distribution. The CDF function plots show that the empirical cumulative distribution for Weibull distribution data matches perfectly whereas the other two distributions are falling outside the reference line.

In the summaries on the right side, we can distinguish values such as Loglikelihood, Akaike information criterion (AIC) and Bayesian Information Criterion (BIC). These values will prove which of the distribution is the best fit for our data set.

The ideal model for our case should have the highest Loglikelihood and lowest values of AIC and BIC. The highest Loglikelihood can be observed in Weibull distribution for rats after the treatment and it equals -478.368. The lowest values of AIC and BIC can be observed in the Weibull distribution as well. The value of AIC is equal to 960.736 and BIC equal to 967.3327. In the two other distribution examples, the values of Loglikelihood were above -500 and the values of AIC and BIC above 1000. The same results with the highest Longlikelihood and lowest AIC and BIC were reported in rats before the treatment data set for Weibull distribution.

To conclude, the tests run above showed that Weibull distribution is the best fit for our rats' datasets.

References

Hayes, A., 2022. *Investopedia.* [Online]
Available at: https://www.investopedia.com/terms/t/t-test.asp
[Accessed 21 03 2022].