SUMMARY STATISTICS

Mean

Time	Age	Year	Thickness
2153	52	1970	2.91

Average survival time of the patient is 2153 days and their average age is 52. The operation was held in 1970 in average. The thickness of the tumour among the patient is 2.91mm in average.

Median

Time	Age	Year	Thickness
2005	54	1970	1.94

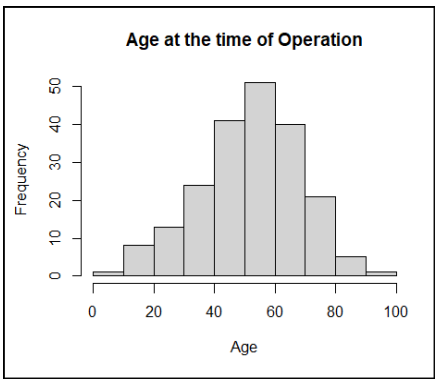
The median survival time tends to be 2005 and the median age is likely to by 54.

Mode

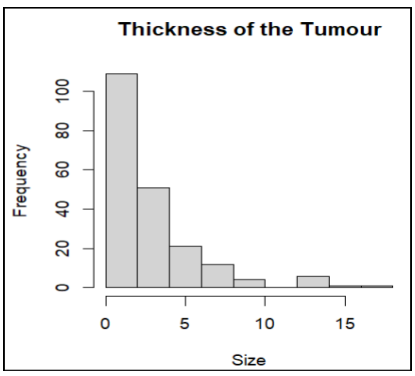
Status	Sex	Ulcer
Died: 57 Alive: 134 Death from other causes: 14	Male: 79 Female: 126	Present: 79 Absent: 126

According to the finding we can identify that the majority of them were alive during the end of the study. Also, it can be identified that females were the one who were affected mostly be melanoma. Majority did not have the ulcer present in their tumour.

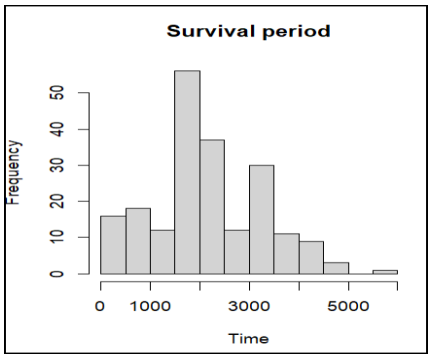
GRAPHICAL SUMMARIES



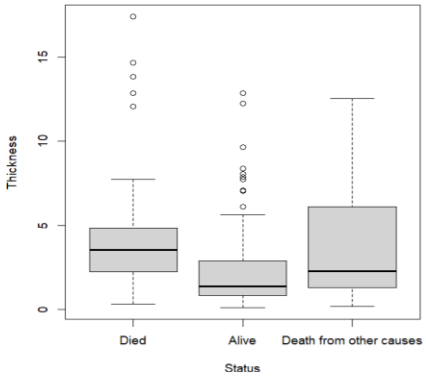
As per the above histogram of age, we can identify that the age distribution is roughly symmetric. The centre of the age distribution is around 50 years at a frequency of 50.



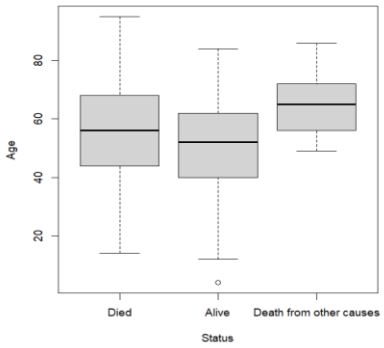
The above image is a right skewed histogram of Thickness. The high frequency lies around 1-2mm with a frequency of 110.



The above histogram depicts the frequency of survival days. It is a continuous distribution where there is a high frequency around 1500-2000 and another peak at 3000-3500days. The most common frequency of survival days is of 1500-2000 days at the frequency of 55.



As per the above box plot above, we can identify that the range of thickness is approximately same for the three status. However, those who were alive during the end of the study had a lower thickness size and those that died has a high thickness. The variation in the size of the tumour is great for those who died from other causes and least for the one who are alive.



As per the above box plot above, we can identify that the range of age is quite same for the three status. However, those who were dead and alive are in around same age. The variation in the size of the tumour happens to be quite similar for those who are dead and alive but little lower for those who died from other causes.

REGRESSION ANALYSIS AND CORRELATION COMPUTATIONS

Time vs Thickness

```
> cor(thickness,time, method="pearson")
[1] -0.2354087
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.2565053  0.4365428   9.750  < 2e-16 ***
Dataset$time -0.0006209  0.0001799  -3.451 0.000679 ***
```

As per the coefficient estimate, the intercept of the variables is 4.26. This indicates that in average, a 4.26mm thickness of tumour results to the survival of the patients. The slope value is -0.00062, this indicates that a 1mm increase in the thickness will reduce the survival days by 0.00062. P value tests whether the estimates of the intercept and slope are equal to 0 or not. If they are equal to 0 then they don't have much use in the model. In general, the intercept does not provide much value to the model, hence its p-value is not considered. However, the p-value of the slope is expected to be < 0.05 in order for it to be statistically strong. If it is less than 0.05, then there is less likelihood of this happening due to chance. A significant p-value for thickness will give us a reliable number of survival days. Three asterisks in the model represent a highly significant p-value. In this scenario the p-value is 0.000679 which is less than 0.05 hence we can assume that the thickness size gives us a reliable survival day. However, as the value is less than 1, it is likely that the relationship between both the variables is not very much strong.

Age vs Time

```
> cor(age, time, method="pearson")
[1] -0.3015179
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.1079361  2.4125775  25.743  < 2e-16 ***
Dataset$time -0.0044800  0.0009943  -4.506 1.12e-05 ***
```

A negative figure of -0.30 again indicates that there is a negative correlation between age and survival time. For instance, as age of the patient increase the survival time is likely to reduce and vice versa. According to the derived results we can identify that the intercept is -0.004 which means that the increase in the age will reduce the survival time by 0.004 days. The p-value in the given is significantly lower than the default value of 0.05 which indicates that the age provides us with reliable survival time.

Age vs Thickness

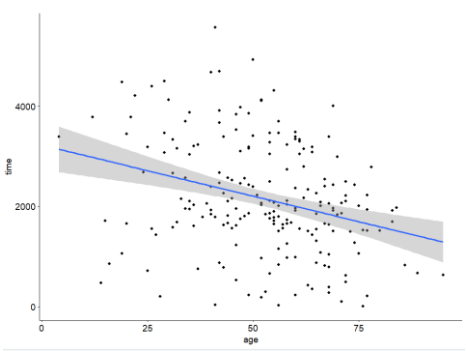
```
> cor(age, thickness, method="pearson")
[1] 0.2124798
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        48.9684     1.6043  30.524  < 2e-16 ***
Dataset$thickness   1.1970     0.3864   3.098  0.00222 **
```

A positive figure of 0.21 indicates that Age and Thickness has a positive correlation. As the age of the patient increase, the thickness of the tumour is also likely to increase. As the figure is very much less than 1, we can come assume that the there is no strong relationship between both the variables. However, the p-value of 0.0022 is again lower than of 0.05 so we can assume that age provides a reliable measure of the thickness.

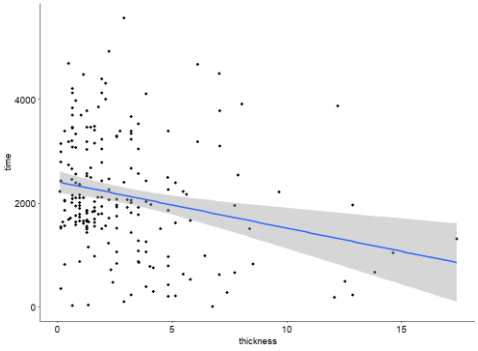
RELATIONSHIP BETWEEN THICKNESS, TIME AND AGE VARIABLES

Thickness vs Time



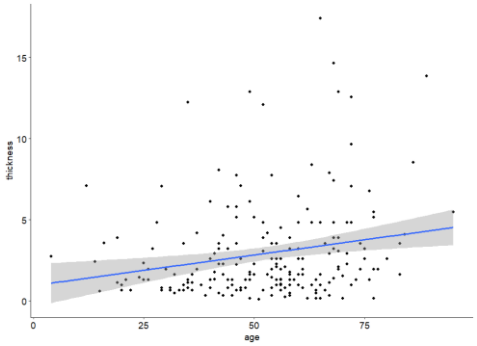
According to the above scatter plot. We can identify that both thickness and time have an inverse relationship. An increase in the thickness is likely to reduce the survival time of the patient

Age vs Time



Again, age and time also tends to have an inverse relationship. According to the diagram we can identify that as the age of the patient increase the survival time tends to go down.

Age vs Thickness

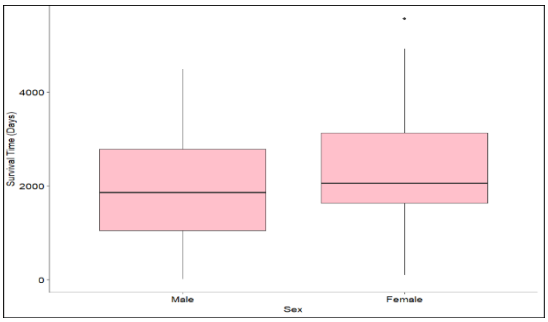


According to the above diagram we can note that age and thickness have a positive relationship. As the age increases the thickness of the tumour is likely to be low.

TWO SAMPLE SIGNIFICANCE TEST

In this task we will use the T Test because the 'n' is less than 30. We will not be able to use Z test as we do not know the population standard deviation. There for we will use the T Test as it is one of the best significance testing tools in statistics.

Time



According the given box plot the mean survival day of the Female group is higher compared to Male. The Q1 and Q3 of the Female is higher compared to Males.

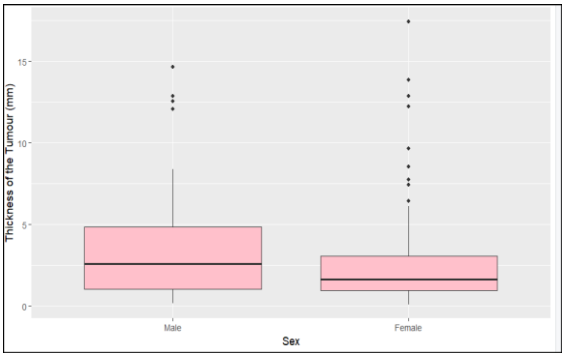
In this particular scenario the null hypothesis would be that there is no difference between the mean survival time among Male and Female. The alternative hypothesis is that there is a difference between the mean survival time. We set up a null hypothesis with the idea of rejecting it. We normally assume that the null hypothesis is the one that is not true.

```
data:  time by sex
t = -2.0848, df = 159.27, p-value = 0.03868
```

```
sample estimates:
  mean in group Male mean in group Female
            1945.709             2282.643
```

The p value is 0.03868. The default level of significance of p is 0.05. Since the p value which we have got is quite smaller than 0.05, it is reasonable to reject the Null Hypothesis and accept the alternate hypothesis. Therefore, we can derive to a conclusion that sex of the people do affect the survival time.

Thickness



In this particular scenario the null hypothesis would be that there is no difference between the mean thickness among Male and Female. The alternative hypothesis is that there is a difference between the mean thickness.

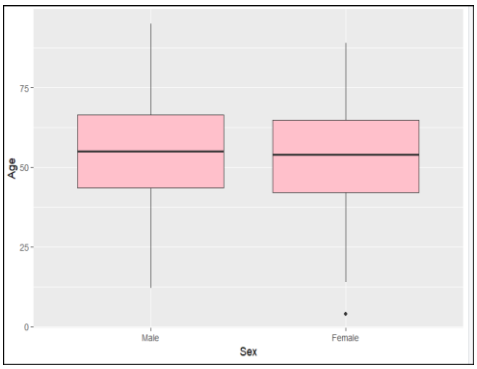
According to the given box pox, we can observe that that the mean thickness in the males is higher than that of the female and hence we can reject the null hypothesis. Let's perform the T-Test in order to further identify the accuracy.

```
data:  thickness by sex
t = 2.6059, df = 149.09, p-value = 0.01009
```

```
sample estimates:
  mean in group Male mean in group Female
            3.611139             2.486429
```

As per the t-test, the p value is 0.01 which is lower than the significance of 0.05 hence it is safe to reject the null hypothesis and assume that there is difference between the both mean values. The sample estimate of the mean in male group is 3.61 and female group of 2.49 hence we can agree with the output of the box plot given above.

Age



In this case the null hypothesis would be that there is no difference between the mean age among Male and Female.

The alternative hypothesis is that there is a difference between the mean age.

The box plot illustrates that the mean thickness in the males is somewhat close to the mean thickness in the female group. Let's perform the T-Test in order to further identify the accuracy.

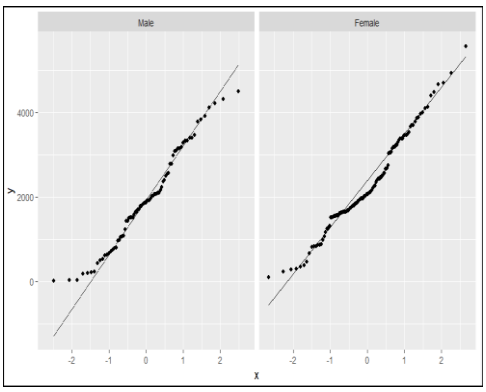
```
sample estimates:
  mean in group Male mean in group Female
            53.89873             51.56349
```

```
data:  age by sex
t = 0.95559, df = 154.42, p-value = 0.3408
```

As per the above t-test output, the p value is 0.34 which is significantly higher than the default value of 0.05. Hence in this scenario we will need to accept the null hypothesis and reject the alternate hypothesis. The mean estimates also provides a value which is very close between the two group therefore we can conclude that that the mean of both groups does not have any difference.

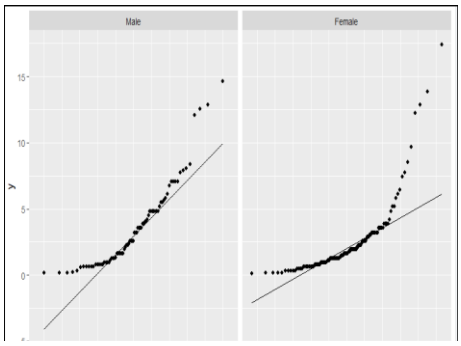
QQ PLOTS

Time grouped by Gender



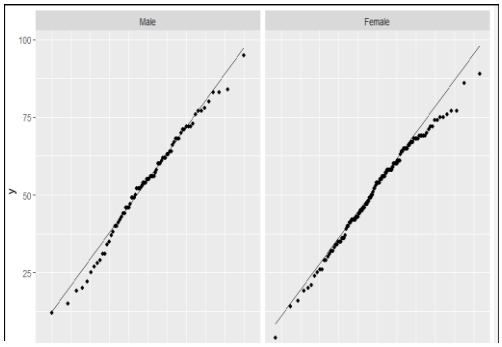
As per the above Q-Q Plot we are able to observe that some of the data points near the tails does not lie within the straight line, however for most part of the sample data appears to be normally distributed hence we can conclude that the dataset follows a normal distribution.

Thickness grouped by Gender



As per the above Q-Q plot it is visible that the datapoints are departed away from the straight line hence we can conclude that the dataset does not appear to follow a normal distribution. When compared between the both, female data points are departed far away than those of male data points near the tail.

Age grouped by Gender



As per the above Q-Q plot the data points lie within the range of the straight line. Some datapoints of the female are departed slightly near the tail yet we can consider the data to be normally distributed.

OVERALL DISCUSSION AND RECCOMMENDATION FOR IMPROVEMENT

Using descriptive statistic, we identified that malignant melanoma is more likely to affect females and those who are aged around 52 and many more useful information, however using descriptive statistic will allow the data to be clearer, to identify patterns however it will not allow us to conclude as it does confirm any assumptions we made.

By performing regression analysis, we were able to identify that there is a relationship between variables like time and age, thickness and time. It allows to determine the areas that need more focus to get the desired outcome. However, there are also some limitations in the model that needs to be considered while making decision based on this model. While interpreting the model we mainly considered the p-value and the slope value whereas other factors are ignored. Therefore, for a better model accuracy it is best if we can take other factors such as R^2 into consideration as well.

Correlation does not equal causation. For example, in the earlier section we discussed that age affects the thickness of the tumour. We cannot 100% confirm that age factor influenced the thickness and not the other way around. We will need to use this as a starting point to further analyse the relationship between the variables. In the same time, we need to ensure that we use the right variables to analyse the relationship.

The boxplot for thickness grouped by sex provides us an understanding that the tumour size in female is lower comparative to the tumours size of the males. The plot also provides outliers, although it seems like a noise in the data model, it needs to be positively approached. Outliers can be present due to various reason that is why it is important for us to identify and analyse them further. Because, outliers tend to communicate something unique on the situation. When they are further analysed it assist in solving a problem in an efficient way.

When performing t-test we identified that for Age, the p value was 0.34 which greater than the default value of 0.05 hence we accepted the null hypothesis. However, the output of the mean estimates shows a different mean value for male and female.

There could be a possibility for the p-value to be so higher just by a chance. In this case we can apply the p-hacking by drawing a number of samples and selecting the best fit for our model.

References

1. www.youtube.com. (n.d.). *Regression Analysis in Rstudio: The Poor Mice:(*. [online] Available at: https://www.youtube.com/watch?v=aMilLhMylqY [Accessed 18 Jan. 2023].
2. Github.io. (2015). *Quick Guide: Interpreting Simple Linear Model Output in R*. [online] Available at: https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R.
3. Qualtrics. (n.d.). *Regression Analysis: The Ultimate Guide*. [online] Available at: https://www.qualtrics.com/uk/experience-management/research/regression-analysis/.
4. www.linkedin.com. (n.d.). *Importance of Outliers*. [online] Available at: https://www.linkedin.com/pulse/importance-outliers-abhirami-a/ [Accessed 19 Jan. 2023].
5. Github.io. (2015). *Quick Guide: Interpreting Simple Linear Model Output in R*. [online] Available at: https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R.
6. statistics.laerd.com. (n.d.). *Independent-samples t-test using R, Excel and RStudio (page 4) | Interpreting and reporting the results for an independent-samples t-test*. [online] Available at: https://statistics.laerd.com/r-tutorials/independent-samples-t-test-using-r-excel-and-rstudio-4.php [Accessed 19 Jan. 2023].
7. desktop.arcgis.com. (n.d.). *Normal QQ plot and general QQ plot—ArcMap | Documentation*. [online] Available at: https://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/normal-qq-plot-and-general-qq-plot.htm#:~:text=Points%20on%20the%20Normal%20QQ.

Appendix

```
> Dataset$sex <- factor(Dataset$sex,
+                       levels = c(1,0),
+                       labels = c("Male", "Female"))
>
> Dataset$ulcer <- factor(Dataset$ulcer,
+                         levels = c(1,0),
+                         labels = c("Present", "Absent"))
>
> Dataset$status <- factor(Dataset$status,
+                          levels = c(1,2,3),
+                          labels = c("Died", "Alive", "Death from other causes"))
> View(Dataset)
> summary(Dataset)
      ...1          time                          status        sex          age
 Min.   :  1   Min.   :  10   Died                  : 57   Male  : 79   Min.   : 4.00
 1st Qu.: 52   1st Qu.:1525   Alive                 :134   Female:126   1st Qu.:42.00
 Median :103   Median :2005   Death from other causes: 14                Median :54.00
 Mean   :103   Mean   :2153                                             Mean   :52.46
 3rd Qu.:154   3rd Qu.:3042                                             3rd Qu.:65.00
 Max.   :205   Max.   :5565                                             Max.   :95.00
      year        thickness        ulcer
 Min.   :1962   Min.   : 0.10   Present: 90
 1st Qu.:1968   1st Qu.: 0.97   Absent :115
 Median :1970   Median : 1.94
 Mean   :1970   Mean   : 2.92
 3rd Qu.:1972   3rd Qu.: 3.56
 Max.   :1977   Max.   :17.42
```

Figure 1

```
> hist(Dataset$age,
+      main="Age at the time of Operation",
+      xlab="Age",
+      ylab="Frequency")
```

Figure 2

```
> hist(Dataset$thickness,
+      main="Thickness of the Tumour",
+      xlab="Size",
+      ylab="Frequency")
```

Figure 3

```
> hist(Dataset$time,
+      main="Survival period",
+      xlab="Time",
+      ylab="Frequency")
```

Figure 4

```
> boxplot(Dataset$thickness ~ Dataset$status,
+         xlab="Status",
+         ylab="Thickness")
```

Figure 5

```
> boxplot(Dataset$age ~ Dataset$status,
+         xlab="Status",
+         ylab="Age")
```

Figure 6

```
> library(ggplot2)
> View(Dataset)
> par(mfrow=c(1,2))
> Melanoma_lmmodel1 = lm(formula = Dataset$thickness ~ Dataset$time)
> summary(Melanoma_lmmodel1)

Call:
lm(formula = Dataset$thickness ~ Dataset$time)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8761 -1.8576 -0.8658  0.8727 13.9781

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.2565053  0.4365428   9.750  < 2e-16 ***
Dataset$time -0.0006209  0.0001799  -3.451 0.000679 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.883 on 203 degrees of freedom
Multiple R-squared:  0.05542,   Adjusted R-squared:  0.05076
F-statistic: 11.91 on 1 and 203 DF,  p-value: 0.0006793
```

Figure 7

```
> View(Dataset)
> Melanoma_lmModel2 = lm(formula = Dataset$age ~ Dataset$time)
> summary(Melanoma_lmModel2)

Call:
lm(formula = Dataset$age ~ Dataset$time)

Residuals:
   Min     1Q Median     3Q    Max
-46.01 -10.64   1.40  12.20  35.71

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.1079361  2.4125775  25.743  < 2e-16 ***
Dataset$time -0.0044800  0.0009943  -4.506 1.12e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.93 on 203 degrees of freedom
Multiple R-squared:  0.09091,   Adjusted R-squared:  0.08643
F-statistic:  20.3 on 1 and 203 DF,  p-value: 1.116e-05
```

Figure 8

```
> View(Dataset)
> Melanoma_lmModel3 = lm(formula = Dataset$age ~ Dataset$thickness)
> summary(Melanoma_lmModel3)

Call:
lm(formula = Dataset$age ~ Dataset$thickness)

Residuals:
    Min      1Q  Median      3Q     Max
-48.248 -10.823   2.254  12.794  39.472

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        48.9684     1.6043  30.524  < 2e-16 ***
Dataset$thickness   1.1970     0.3864   3.098  0.00222 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.33 on 203 degrees of freedom
Multiple R-squared:  0.04515,   Adjusted R-squared:  0.04044
F-statistic: 9.598 on 1 and 203 DF,  p-value: 0.002223
```

Figure 9

```
> qplot(x = sex, y = time,
+       geom = "boxplot", data = Dataset,
+       xlab = "Sex",
+       ylab = "Survival Time (Days)",
+       fill = I("pink"))
>
> Dataset %>%
+   group_by(sex) %>%
+   summarize(num.obs = n(),
+             mean_time = round(mean(time), 0),
+             sd_time = round(sd(time), 0),
+             se_time = round(sd(time) / sqrt(num.obs), 0))
# A tibble: 2 × 5
  sex    num.obs mean_time sd_time se_time
  <fct>    <int>     <dbl>   <dbl>   <dbl>
1 Male        79      1946    1148     129
2 Female     126      2283    1090      97
>
> time_t_test <- t.test(time ~ sex, data = Dataset)
> time_t_test

        Welch Two Sample t-test

data:  time by sex
t = -2.0848, df = 159.27, p-value = 0.03868
alternative hypothesis: true difference in means between group Male and group Femal
e is not equal to 0
95 percent confidence interval:
 -656.12032  -17.74767
sample estimates:
  mean in group Male mean in group Female
            1945.709             2282.643
```

Figure 10

```
> library(ggplot2)
> qplot(x = sex, y = thickness,
+       geom = "boxplot", data = Dataset,
+       xlab = "Sex",
+       ylab = "Thickness of the Tumour (mm)",
+       fill = I("pink"))
>
>
> Dataset %>%
+   group_by(sex) %>%
+   summarize(num.obs = n(),
+             mean_time = round(mean(thickness), 0),
+             sd_thickness = round(sd(thickness), 0),
+             se_thickness = round(sd(thickness) / sqrt(num.obs), 0))
# A tibble: 2 × 5
  sex    num.obs mean_time sd_thickness se_thickness
  <fct>    <int>     <dbl>        <dbl>        <dbl>
1 Male        79         4            3            0
2 Female     126         2            3            0
>
> thickness_t_test <- t.test(thickness ~ sex, data = Dataset)
> thickness_t_test

        Welch Two Sample t-test

data:  thickness by sex
t = 2.6059, df = 149.09, p-value = 0.01009
alternative hypothesis: true difference in means between group Male and group Femal
e is not equal to 0
95 percent confidence interval:
 0.2718653 1.9775560
sample estimates:
  mean in group Male mean in group Female
            3.611139             2.486429
```

Figure 11

```
> library(ggplot2)
> qplot(x = sex, y = age,
+       geom = "boxplot", data = Dataset,
+       xlab = "Sex",
+       ylab = "Age",
+       fill = I("pink"))
>
>
> Dataset %>%
+   group_by(sex) %>%
+   summarize(num.obs = n(),
+             mean_time = round(mean(age), 0),
+             sd_age = round(sd(age), 0),
+             se_age = round(sd(age) / sqrt(num.obs), 0))
# A tibble: 2 × 5
  sex    num.obs mean_time sd_age se_age
  <fct>    <int>     <dbl>  <dbl>  <dbl>
1 Male        79        54     18      2
2 Female     126        52     16      1
>
> age_t_test <- t.test(age ~ sex, data = Dataset)
> age_t_test

        Welch Two Sample t-test

data:  age by sex
t = 0.95559, df = 154.42, p-value = 0.3408
alternative hypothesis: true difference in means between group Male and group Femal
e is not equal to 0
95 percent confidence interval:
 -2.492280  7.162764
sample estimates:
  mean in group Male mean in group Female
            53.89873             51.56349
```

Figure 12

```
> p_time <- ggplot(data = Dataset, aes(sample = time))
> p_time + stat_qq() + stat_qq_line() + facet_grid(. ~ sex)
>
>
> #end
> p_thickness <- ggplot(data = Dataset, aes(sample = thickness))
> p_thickness + stat_qq() + stat_qq_line() + facet_grid(. ~ sex)
>
> #end
> p_age <- ggplot(data = Dataset, aes(sample = age))
> p_age + stat_qq() + stat_qq_line() + facet_grid(. ~ sex)
>
```

Figure 13