Descriptive Statistics

Introduction

The exploration and analysis of data in question is initiated by performing some basic summary statistics. This facilitates some important insights that could be building blocks for the upcoming analysis.

```
> summary(Cancer)
       X                time          status         sex
 Min.   :  1   Min.   :  10   dead : 57   female:126
 1st Qu.: 52   1st Qu.:1525   alive:134   male  : 79
 Median :103   Median :2005   other: 14
 Mean   :103   Mean   :2153
 3rd Qu.:154   3rd Qu.:3042
 Max.   :205   Max.   :5565

      age              year      thickness           ulcer
 Min.   : 4.00   1972   :41   Min.   : 0.10   ulcer free:115
 1st Qu.:42.00   1973   :31   1st Qu.: 0.97   present    : 90
 Median :54.00   1971   :27   Median : 1.94
 Mean   :52.46   1968   :21   Mean   : 2.92
 3rd Qu.:65.00   1969   :21   3rd Qu.: 3.56
 Max.   :95.00   1967   :20   Max.   :17.42
                 (Other):44
```

Figure 1: Summary of all Variables

	Variable	Mean	Variance	StandardDeviation
time	time	2152.800000	1.259020e+06	1122.060667
age	age	52.463415	2.779460e+02	16.671711
thickness	thickness	2.919854	8.758242e+00	2.959433

Figure 2: Mean, Variance and Standard deviation of the continuous variables.

Notable Observations

Survival Time in days (time):

The survival time ranges from **10 to 5565** days for patients after the operation, and the high standard deviation of **1122.06** days points out its significant variability. Comparatively, median is smaller than mean which could mean it is a right-skewed distribution, this will be verified in further analysis.

Age in years(age):

The mean age of the patient's during operation is **52** years with a standard deviation of **16.67** years, this indicates moderate variance of age.

Thickness of tumour in mm (thickness):

2.92 mm is the mean thickness of tumour in patients, with a wide range from **0.10** mm to **17.42** mm.

Status:

A total of **134** people were still alive and **57** were declared dead up until the data was recorded. The ratio is tipped in the favor of people staying alive rather than dying post operation.

Gender of patient (sex):

The number of female patients (**126**) is higher than male patients (**79**) in this data.

Ulcer presence (ulcer):

There is not much of a disproportion in this variable, with **115** ulcer free patients and **90** of those with ulcer.

Year of operation (year):

Only 1972 and 1973 were the 2 years with more than 30 operations, **41** and **31** respectively.

Graphical Illustrations

To further analyze any emerging trends or aspects, a graphical summary of each variable is probed with. The left column of histograms comprises of only continuous variables under the scope whereas, the right column consists of all the categorical ones using GG-plot library (Wickham 2016). Noteworthy aspects of the data are mentioned below each graphic:

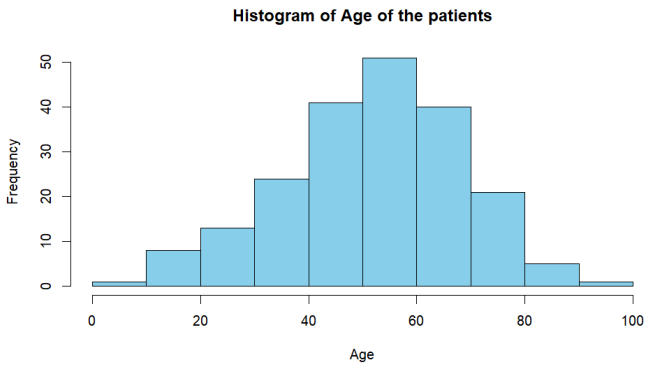

Figure 3: A bell-shaped curve in age distribution is observed.

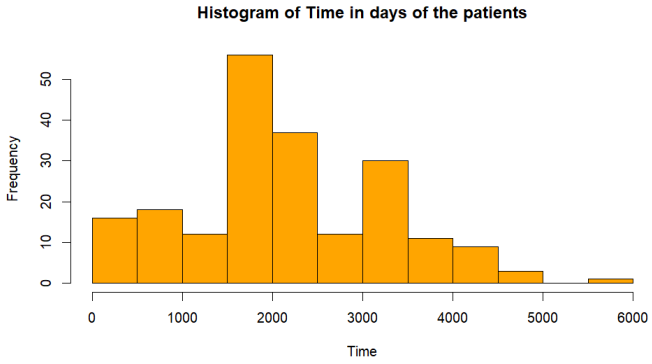

Figure 5: The no. of days is a bit right skewed but not entirely.

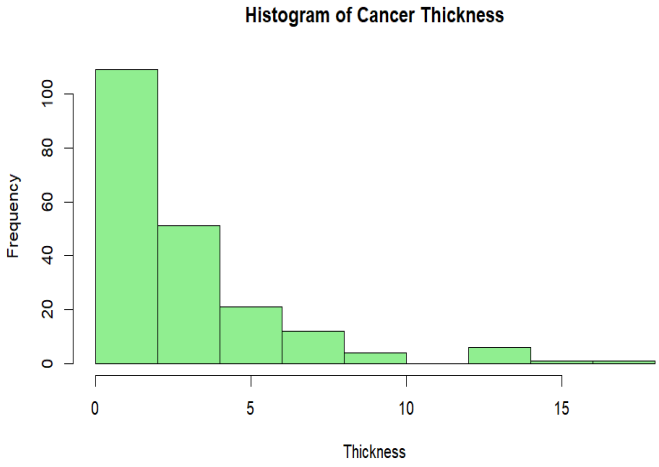

Figure 4: Thickness is completely right skewed.

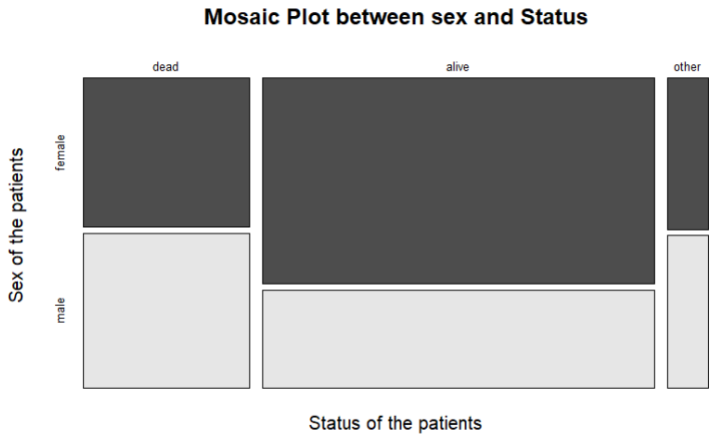

Figure 6: Status seems to be evenly distributed over gender apart from 'alive' patients where females dominate.

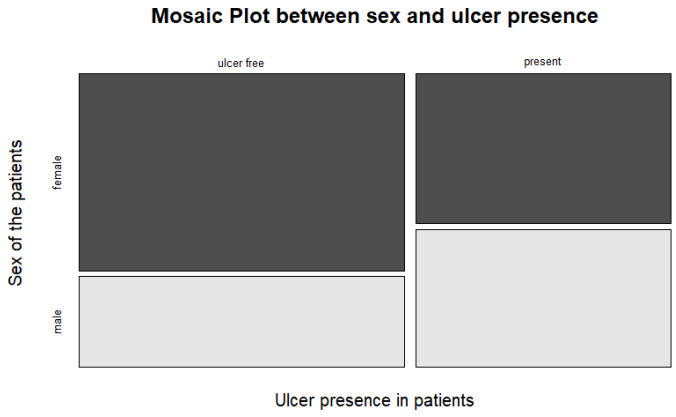

Figure 7: Ulcer free patients are females in most cases.

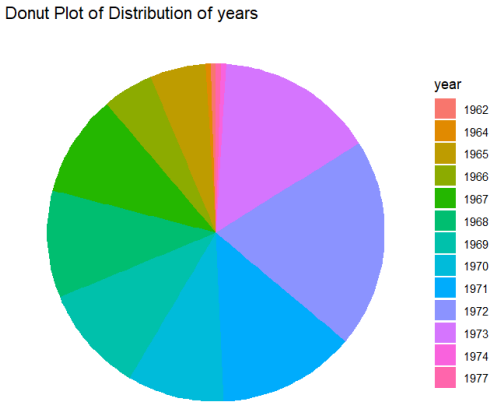

Figure 8: 1972 and 1973 are the years with most operation

Regression and Correlation Exploration

Introduction

After the individual analysis of variables, exploration between the continuous variables is started. Regression and Correlation Analysis is very useful in discovering any interdependencies between variables. For this phase, Performance Metrics library from R studio is used (Peterson and Carl 2020). The below image displays the correlation matrix between time, thickness, and age; it plots scatter plot with regression line between the three variables as well.

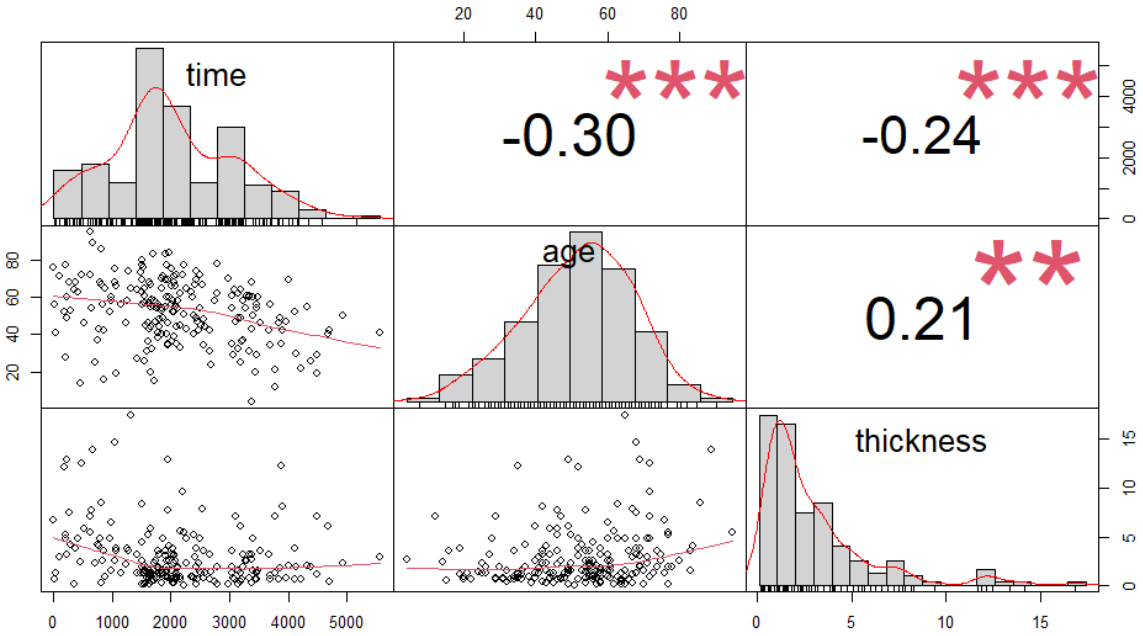

Figure 9: Correlation and Regression Matrix

Correlation Study

Correlation between time and thickness

The weak negative correlation between time and thickness can be implied by the coefficient value being **-0.24** as seen in figure 9. This means on average as survival time increases, tumour thickness tends to be smaller and smaller. (Nakagawa, Johnson 2017)

Correlation between time and age

Another case of weak negative correlation is observed here from figure 9 which indicates the correlation coefficient to be **-0.30**. This suggests that, on average, as survival time increases, patients tend to be slightly younger. (Nakagawa, Johnson 2017)

Correlation between thickness and age

The **0.21** value in figure 9 indicates a weak positive correlation here. This indicates that, on average, as tumour thickness increases, the patients tend to be of older age. (Nakagawa, Johnson 2017)

Regression Analysis and model evaluation

To analyze the regression model and its reliability, some of the important statistics are focused from the model evaluation.

Model evaluation of time ~ thickness

```
Call:
lm(formula = Cancer$time ~ Cancer$thickness)

Residuals:
    Min      1Q  Median      3Q     Max
-2325.4  -707.6  -210.6   744.9  3410.4

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       2413.41     107.39  22.473  < 2e-16 ***
Cancer$thickness   -89.25      25.86  -3.451 0.000679 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1093 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 10: Model summary of time ~ thickness.

The regression suggests a negative association, with larger values of `thickness` of tumour matching with lower `time` in days. Some of the important statistics from this evaluation were:

- **Intercept (2413.41):** It means that if the thickness of tumour is 0, the predicted survival time is **2413** days.
- **Thickness Coefficient (-89.25):** This model predicts that for each mm of thickness increasing, the time goes down by **89** days.
- **R-squared value (0.055):** The diminished R-squared value suggests a limited reliability in forecasting the potential lifespan of a patient solely relying on the thickness of their tumor. (Cheek, Bewick et al. 2003)

Model evaluation of time ~ age

```
Call:
lm(formula = Cancer$time ~ Cancer$age)

Residuals:
    Min      1Q  Median      3Q     Max
-2464.3  -646.2   -54.4   712.1  3179.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3217.448    247.879  12.980  < 2e-16 ***
Cancer$age   -20.293      4.504  -4.506 1.12e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1072 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 11: Model summary for time ~ age.

The regression model predicts that as the patients get older the survival time tends to get smaller. Some of the important statistics from this evaluation were:

- **Intercept (3217.44):** It means that if the age is 0 years, the predicted survival time is **3217** days.
- **Age Coefficient (-20.29):** This model predicts that for each year a patient is older, the survival time goes down by **20** days.
- **R-squared value (0.090):** The diminished R-squared value suggests a limited reliability in predicting the potential number of days a patient might live based solely on their age using this model. (Cheek, Bewick et al. 2003)

Model evaluation of thickness ~ age

```
Call:
lm(formula = Cancer$thickness ~ Cancer$age)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6853 -1.7727 -0.9155  0.9558 14.0273

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.94105    0.67004   1.404  0.16170
Cancer$age   0.03772    0.01217   3.098  0.00222 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.899 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 12: Model Summary of thickness ~ age.

The model suggests that as the age increases, the predicted values of thickness of tumour also tend to increase. Some of the important statistics from this evaluation were:

- **Intercept (0.92):** It means that if the age is 0 years, the predicted thickness of the tumour will be **0.92** mm.
- **Age Coefficient (0.037):** This model predicts that for each year a patient is older, the thickness of the tumour increases **by 0.037** mm.
- **R-squared value (0.045):** The reduced R-squared value implies that the model lacks strong reliability in forecasting the thickness of a patient's tumor solely based on their age. (Cheek, Bewick et al. 2003)

Testing

Impact of Gender on different variables

To initiate the testing procedure, the data is grouped by the gender of patients. This helps in understanding the relationship better and provides some important insights into the exploration.

Impact of Gender on Survival time

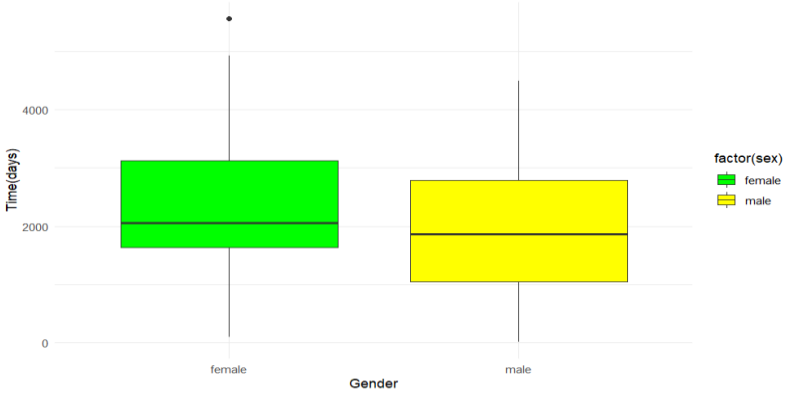

Figure 13: Box plot of survival time for different genders.

```
# A tibble: 2 × 5
  sex    num.obs mean_time sd_time se_time
  <fct>    <int>     <dbl>   <dbl>   <dbl>
1 female     126      2283    1090      97
2 male        79      1946    1148     129
>
```

Figure 14: Notable statistics of Survival time grouped over gender.

- The mean survival time difference between the 'female' and 'male' groups is evident, with females having, on average, a higher survival time.
- The standard deviation values are similar for both groups, indicating similar variability in survival time within each group.
- The standard error values provide an indication of the precision of the mean estimates. (Watson, Fitzallen 2017)

Impact of Gender on Age

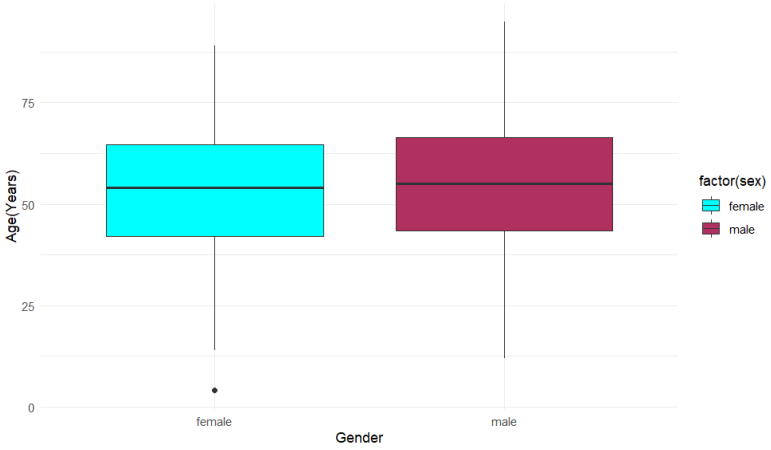

Figure 15: Box plot of Age distribution over different genders.

```
# A tibble: 2 × 5
  sex    num.obs mean_age sd_age se_age
  <fct>    <int>    <dbl>  <dbl>  <dbl>
1 female     126       52     16      1
2 male        79       54     18      2
>
```

Figure 16: Notable Statistics of Age grouped over gender.

- The mean age difference between the 'female' and 'male' groups is small, and its statistical significance would depend on further testing.
- The larger standard deviation in the 'male' group suggests a more diverse age distribution within this group.
- The higher standard error in the 'male' group indicates more variability in the sample means, possibly reflecting the smaller sample size. (Watson, Fitzallen 2017)

Impact of Gender on Thickness

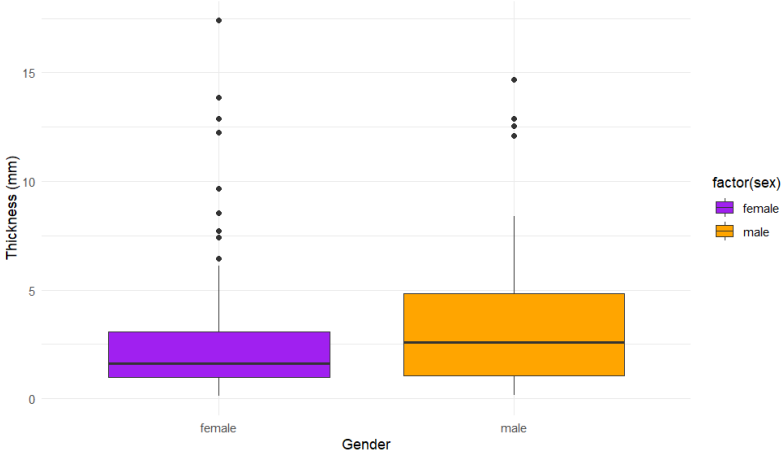

Figure 17: Box plot of Thickness over different genders.

```
# A tibble: 2 × 5
  sex    num.obs mean_thickness sd_thickness se_thickness
  <fct>    <int>          <dbl>        <dbl>        <dbl>
1 female     126              2            3            0
2 male        79              4            3            0
>
```

Figure 18: Notable Statistics of Thickness grouped over gender.

- The observed mean thickness difference between the 'female' and 'male' groups is apparent, indicating that, on average, males have a higher thickness.
- The standard deviation values are the same for both groups, indicating similar variability in thickness within each group.
- The standard error of 0 suggests no variability in the sample means, which may be due to rounding or reporting practices.

Two Sample t-test

The two-sample t-test is used to compare the differences in statistics of different groups. Common knowledge is that if $n < 30$ then one should use the t distribution for the test, in fact, more is true. Technically, if σ is unknown (as it almost always is) one should use the t distribution for $n \geq 30$. For that reason, there is no standard z-test function built into R.

Welch Two Sample t-test of age grouped by sex

```
        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 female and group male is not equal to 0
95 percent confidence interval:
 -7.162764  2.492280
sample estimates:
mean in group female   mean in group male
            51.56349             53.89873
```

Figure 19: t-test of age by sex.

- The calculated t-statistic for the test is -0.955, showing how much the sample means differ from each other using standard error as a measure. The degrees of freedom for the t-distribution are approximated at 154.42, considering the difference in variance between the two groups.
- In the hypothesis test, the null hypothesis assumes that the real difference in means between the female and male groups is 0, while the alternative hypothesis proposes that the true difference is not equal to 0.
- The 95% confidence interval facilitates a range of predicted values for the actual difference in means. Since the interval includes 0, it suggests that we don't have enough evidence to reject the null hypothesis. (Posten, 2014)

Welch Two Sample t-test of thickness grouped by sex

```
        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 female and group male is not equal to 0
95 percent confidence interval:
 -1.9775560 -0.2718653
sample estimates:
mean in group female   mean in group male
            2.486429             3.611139
```

Figure 20: t-test of thickness by sex.

- The calculated t-statistic for the test is -2.6059, indicating the extent to which the sample means differ in terms of standard errors. The degrees of freedom for the t-distribution are determined as 149.09, accounting for the difference in variance between the two groups.
- In the hypothesis test, the null hypothesis assumes that the actual difference in means between the female and male groups is equal to 0, while the alternative hypothesis suggests that the true difference is not equal to 0.
- The 95% confidence interval facilitates a range of predicted values for the actual difference in means. Since the interval includes 0, it suggests that we don't have enough evidence to reject the null hypothesis. (Posten, 2014)

Welch Two sample t-test of time grouped by sex

```
        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 female and group male is not equal to 0
95 percent confidence interval:
  17.74767 656.12032
sample estimates:
mean in group female   mean in group male
             2282.643             1945.709
```

Figure 21: t-test of time by sex.

- The calculated t-statistic for the test is -0.955, showing how much the sample means differ from each other using standard error as a measure. The degrees of freedom for the t-distribution are calculated as 159.27, in accordance with the difference in variance between the two groups.
- In the hypothesis test, the null hypothesis assumes that the actual difference in means between the female and male groups is equal to 0, while the alternative hypothesis suggests that the true difference is not equal to 0.
- The 95% confidence interval facilitates a range of predicted values for the actual difference in means. Since the interval includes 0, it suggests that we don't have enough evidence to reject the null hypothesis. (Posten, 2014)

Analysis of Data Distribution

A deeper analysis of data distribution will be conducted using QQ-plots. This is important because the t-test is only reliable if the data is normally distributed. QQ-plots help in determining if the data is distributed normally or exponentially. QQ-plots graphically display the deviations between the theoretical and Sample values. (Kim, Park 2019)

Distribution of age grouped by gender

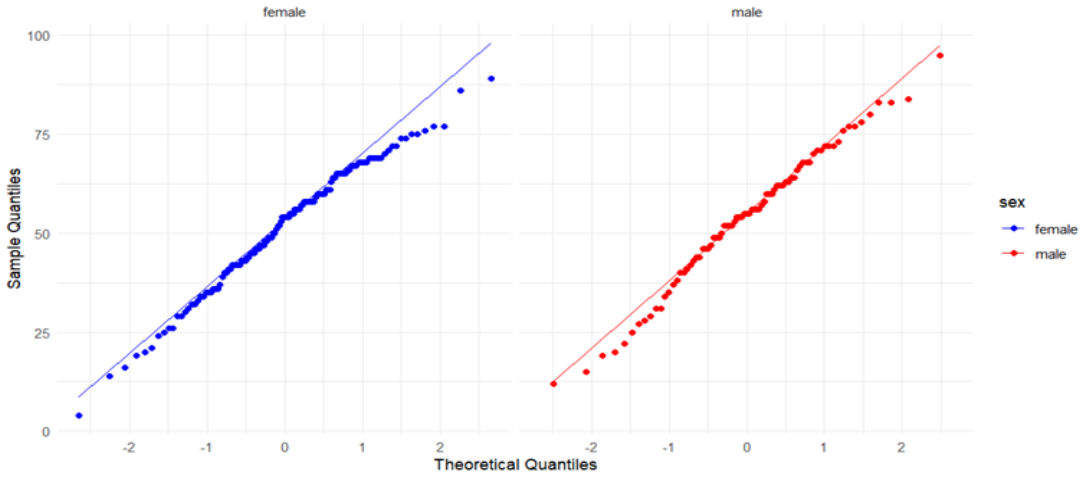

Figure 22: Normality test of age by gender.

The data points in figure 22 are close to the line which implies that both the genders indicate a normal distribution. This validates our methodology of using two sample t-test. (Limpert, Stahel 2011)

Distribution of thickness grouped by gender

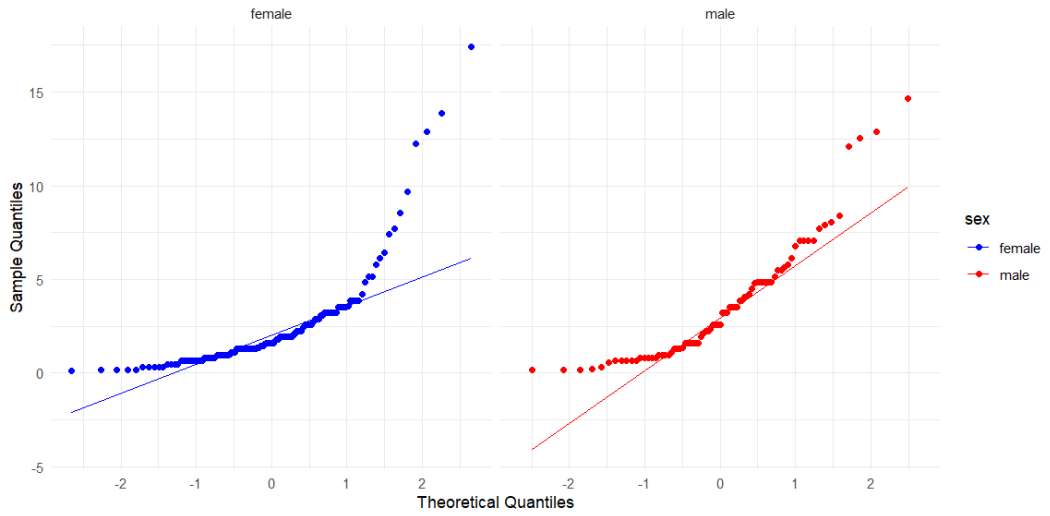

Figure 23: Normality test of age by gender.

The data points in this case for both genders seem to be far from the line, indicating that this is not a normal distribution. For this instance, we use the large sample size to validate our use of two sample t-test. (Limpert, Stahel 2011)

Distribution of time grouped by gender

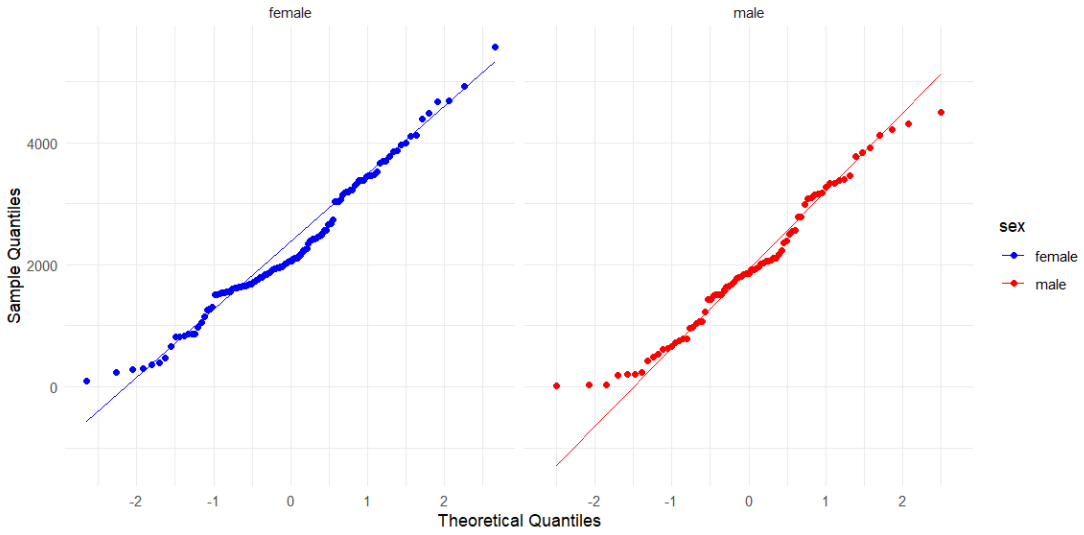

Figure 24: Normality test of time by gender.

The proximity of data points in both genders to the line suggests a distribution that approximates normality. This reaffirms the appropriateness of our methodology employing the two-sample t-test. (Limpert, Stahel 2011)

Insights and Suggestions

Insights Generated from the Data

1. Thickness Differences:

Statistical Significance: The analysis yielded a statistically significant distinction in mean thickness between females and males (p-value = 0.01009). The negative test statistic suggests that, on average, females exhibit significantly lower tumor thickness than males.

Clinical Significance: The 95% confidence interval for the true difference in means, which is (-1.98, -0.27), aligns with the rejection of the null hypothesis, indicating statistical significance. However, it's important to note that this difference, while statistically significant, may not be clinically significant due to its relatively small magnitude. The practical implications of the observed difference should be carefully considered in a clinical context, taking into account the potential impact on patient outcomes or treatment strategies.

2. Survival Time Differences:

Statistical Significance: The analysis also found a statistically significant difference in mean survival time between females and males (p-value = 0.03868). The positive values within the 95% confidence interval (17.75, 656.12) suggest that, on average, females have a higher survival time compared to males.

Clinical Implications: The practical significance of this finding should be carefully interpreted. While statistically significant, the wide confidence interval indicates substantial uncertainty in the true difference, and clinical context is crucial for interpretation.

3. Age Differences:

Statistical Insignificance: The analysis revealed a statistically significant difference in mean survival time between females and males (p-value = 0.03868). The positive values within the 95% confidence interval (17.75, 656.12) indicate that, on average, females exhibit a higher survival time compared to males.

Clinical Consideration: The negative test statistic implies a slightly lower average age for females, but this difference is not considered significant at the 0.05 significance level. It might be relevant to explore potential clinical implications or reasons for any observed trends.

Suggestions for Future

1. Exploring Other Factors:

Integrating additional factors into the analysis, such as tumor type or specific clinical characteristics. This inclusion would contribute to a more comprehensive understanding of the observed differences in tumor thickness and survival time. The incorporation of these factors could lead to a more nuanced interpretation of the data and provide valuable insights into the potential influencing variables. (Watson, Fitzallen 2017)

2. Clinical Relevance:

Evaluating the clinical relevance of statistically significant differences could be helpful. This can be done by assessing whether the observed variations in tumor thickness and survival time have practical implications for patient outcomes or treatment strategies.

3. Outlier Identification:

Investigation of outliers in the dataset, especially in the thickness and survival time variables. Outliers could potentially influence the statistical results and warrant further exploration or exclusion if appropriately validated.

4. Longitudinal Analysis:

Conducting a longitudinal analysis to explore how these variables change over time. This may provide insights into the dynamics of tumor development, survival, and aging. (Limpert, Stahel 2011)

5. Validation and Robustness:

Validating the findings with an independent dataset to ensure the robustness of the observed differences. Sensitivity analysis or alternative statistical approaches could also be employed to assess the stability of the results.

6. Collaboration with Clinical Experts:

Collaborating with clinical experts to gain domain-specific insights. Their expertise can provide valuable context for understanding the clinical significance of the observed differences and guide further investigations. (Watson, Fitzallen 2017)

Conclusion

In conclusion, while statistical significance is important, a comprehensive understanding of the clinical context and potential confounding factors is crucial for drawing meaningful conclusions from the data. Further exploration and collaboration with domain experts will enhance the depth and applicability of the insights derived from the analysis.

References

- CHEEK, L., BEWICK, V. and BALL, J., 2003. Statistics review 7: Correlation and regression | Critical Care.
- KIM, T.K. and PARK, J.H., 2019. More about the basic assumptions of t-test: normality and sample size. *Korean Journal of Anesthesiology*, **72**(4), pp. 331-335.
- LIMPERT, E. and STAHEL, W.A., 2011. Problems with Using the Normal Distribution – and Ways to Improve Quality and Efficiency of Data Analysis. *PLOS ONE*, **6**(7), pp. e21403.
- NAKAGAWA, S. and JOHNSON, P.C.D., 2017. The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded | Journal of The Royal Society Interface.
- POSTEN, H.O., Robustness of the Two-Sample T-Test | SpringerLink. *Department of Statistics, University of Connecticut Storrs, Connecticut, 06268, USA*, .
- TWISK, J.W., 2004. Longitudinal Data Analysis. A Comparison Between Generalized Estimating Equations and Random Coefficient Analysis | European Journal of Epidemiology.
- WATSON, J. and FITZALLEN, N., 2017. The Practice of Statistics | SpringerLink.
- Peterson, Brian G., and Peter Carl. 2020. “PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis.” https://CRAN.R-project.org/package=PerformanceAnalytics.
- Wickham, Hadley. 2016. “Ggplot2: Elegant Graphics for Data Analysis.” https://ggplot2.tidyverse.org.