Table of Contents

1.0 Background

1.1 Introduction

This research aims to investigate the correlation between many clinical and demographic variables, such as age, gender, smoking history, tumor size, tumor location, cancer stage, therapy type, and survival months, in lung cancer patients. The main objective is to examine the correlation between these characteristics and assess their impact on the survival rates of lung cancer patients. Various statistical methods will be employed to estimate the associations between these variables.

1.2 Methodology

The data set utilized in this study is secondary data obtained from an online source, comprising a population of 100 lung cancer patients. The selected data should include variables such as patient age, gender, smoking history, tumor size, tumor location, cancer stage, treatment type, and survival months. The analysis will employ hypothesis testing, correlation, regression, chi-square test for independence, and ANOVA to test the sample data. The data will be analyzed using RStudio to generate graphical presentations and perform basic calculations, leading to the conclusions drawn from the findings.

2.0 Data Analysis and Results

2.1 Hypothesis

In this Analysis, we will use two variables smoking histories and Tumor size, where we wish to determine if there is any difference in mean tumor size between smokers(current smoker, former smoker) and non-smokers(never smoked) at a 95% confidence level, the variance unknown. The two samples above are independent since the sample values from one population are not related to or somehow paired or matched with the sample values selected from the other population. From the sample data, we calculate the frequency(n), mean($\overline{x}$), and standard deviation(s).

```
> clean_lung_cancer <- clean_lung_cancer %>%
+     mutate(Smoking_Status = case_when(
+         Smoking_History %in% c("Current Smoker", "Former Smoker") ~ "Smoker",
+         Smoking_History == "Never Smoked" ~ "Non-Smoker"
+     ))
>
> # Calculate summary statistics
> summary_stats <- clean_lung_cancer %>%
+     group_by(Smoking_Status) %>%
+     summarise(
+         n = n(),
+         mean= mean(Tumor_Size_mm),
+         sd = sd(Tumor_Size_mm)
+     )
>
> print(summary_stats)
# A tibble: 2 x 4
  Smoking_Status     n  mean    sd
  <chr>          <int> <dbl> <dbl>
1 Non-Smoker        43  54.5  24.3
2 Smoker            58  49.4  28.0
```

Figure 2.1.1: *Mean, frequency, and standard deviation calculation*

The results from the above code are grouped:

$n_1 = 43$	$n_2 = 58$
$\overline{x_1} = 54.5$	$\overline{x_2} = 49.4$
$s_1 = 24.3$	$s_2 = 28.0$

Here,

Group1 is the non-smoker's tumor data

Group2 is the smoker's tumor data

1. Hypothesis statement:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Where,

μ_1 = Mean of tumor size in non-smoker

μ_2 = Mean of tumor size in the smoker

2. Test statistics, critical value, and degree of freedom:

Given a confidence level of 95%, $\alpha = 0.05$. The mathematical equation for test statistics for two samples Independent hypothesis testing where the variances are unknown and unequal is,

$$T_0^* = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n1} + \frac{S_2^2}{n2}}}$$

The degree of freedom can be obtained using the formula,

$$v = \frac{\left(\frac{S_1^2}{n1} + \frac{S_2^2}{n2}\right)^2}{\frac{\left(S_1^2/n1\right)^2}{n1-1} + \frac{\left(S_2^2/n2\right)^2}{n2-1}}.$$

Using R studio we calculate the values of test statistics, degree of freedom and critical value,

```
> xbar1 = 54.5
> xbar2 = 49.4
>
> s1 = 24.3
> s2 = 28.0
>
> n1 = 43 # Non-Smoker
> n2 = 58 # Smoker
>
> t0 = (xbar1-xbar2-0)/(sqrt((s1^2/n1)+(s2^2/n2)))
>
> v = ((s1^2/n1)+(s2^2/n2))^2/((((s1^2/n1)^2)/(n1-1))+(((s2^2/n2)^2)/(n2-1)))
>
> alpha = 0.05
> t.alpha = qt(alpha/2, floor(v))
>
> cat("Critical value: ",t.alpha)
Critical value:  -1.984984
> cat("Degree of freedom: ",v)
Degree of freedom:  96.49047
> cat("Test Statistic (t0): ", t0)
Test Statistic (t0):  0.9769906
>
```

Test statistic, $t_0 = 0.977$

Degrees of freedom, v = 96.4905

Therefore using $\alpha = 0.05$, the critical value is,

$$t_0 > t_{0.025,96.2} = 1.9850 \text{ or } t_0 < t_{0.025,96.2} = -1.985$$

3. Decision:

   Since, $t_0 = 0.977 < t_{0.025,96.2} = 1.985$, we reject the null hypothesis. There is evidence to conclude that the mean tumor size in smokers is different from the mean tumor size in non-smokers.

2.2 Correlation

The correlation test assesses the association between the age of lung cancer onset and the month of survival in a sample of 100 patients with lung cancer. We employ Pearson's method to get a correlation coefficient as both datasets are of the ratio type.

Sample of correlation coefficient:

$$r = \frac{\sum xy - \left(\sum x \sum y\right)/n}{\sqrt{\left[\left(\sum x^2\right) - \left(\sum x\right)^2/n\right]\left[\left(\sum y^2\right) - \left(\sum y\right)^2/n\right]}}$$

Whereby,

r = Sample correlation coefficient

n = Sample size

x = Value of the independent variable

y = Value of the dependent variable

We use n=100, x =Age of Patients with Lung Cancer, and y = Survival Month of Patients with Lung Cancer to calculate the correlation coefficient by using RStudio.

S

Figure 2.2.1: *Scatter plot of Survival Month of patient with lung cancer and Age of patient with lung cancer*

```
> library(readxl)
> library(ggplot2)
> library(gganimate)
>
> data <- read_excel("C:/Users/User/Downloads/clean_lung_cancer.xlsx")
>
> data <- na.omit(data[, c("Age", "Survival_Months")])
>
> x <- data$Age
> y <- data$Survival_Months
>
> correlation_test <- cor.test(x, y)
> print(correlation_test)

        Pearson's product-moment correlation

data:  x and y
t = 0.79065, df = 98, p-value = 0.4311
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1186594  0.2717824
sample estimates:
       cor
0.07961431

>
> p <- ggplot(data, aes(x = Age, y = Survival_Months)) +
+   geom_point(color = "blue", size = 3) +
+   geom_smooth(method = "lm", formula = y ~ x, col = "red") +
+   labs(title = "Relationship between Age and Survival Months",
+        x = "Age",
+        y = "Survival Months") +
+   theme_minimal()
>
> print(p)
>
> anim <- p +
+   transition_reveal(Age) +
+   labs(title = "Age vs. Survival Months: {frame_along}",
+        x = "Age",
+        y = "Survival Months")
>
> animate(anim, renderer = gifski_renderer("animated_plot.gif"))
```

Figure 2.2.2: *The result of correlation, r, and test statistic, t by using RStudio*

Based on Figure 2.2.2, the correlation coefficient, $r = 0.07961431$. A scatter plot and correlation coefficient, r, indicate that there is a very weak positive relationship between age and survival months of lung cancer patients. This means that there is little to no association between age and the number of months patients survive. The scatter plot shows that the points are widely dispersed, and the slight positive trend indicated by the regression line suggests that as age increases, the survival months may slightly increase.

Significance Test for correlation

We decided to test whether there is any evidence of a linear relationship between the age of lung cancer patients and their survival months at the 0.05 level of significance.

Null hypothesis, H0: $\rho = 0$ (no linear correlation)

Alternative hypothesis, H1: $\rho \neq 0$ (linear correlation exists)

$\alpha = 0.05$, degrees of freedom,

df = 100 − 2 = 98.

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

$$t = \frac{0.0796}{\sqrt{\frac{1-(0.0796)^2}{100-2}}}$$

$$t = 0.79065$$

Based on **Figure 2.2.2**, the test statistic, *t*, is approximately 0.79065. The p-value is 0.4311. The 95% confidence interval for the correlation coefficient is (-0.1186594, 0.271824). Since the p-value of 0.4311 is greater than the significance level of 0.05, we fail to reject the null hypothesis. There is insufficient evidence of a linear relationship between age and survival months of lung cancer patients at the 0.05 significance level.

2.3 Regression

In regression analysis, we will use variables such as survival months and patients' ages. We will be testing whether the survival months depend on the age of the patients. Here the Independent variable(x) is age and the dependent variable(y) is Survival months. This model is a simple linear regression since the changes in the months of survival are assumed to be caused by the age of the patient.

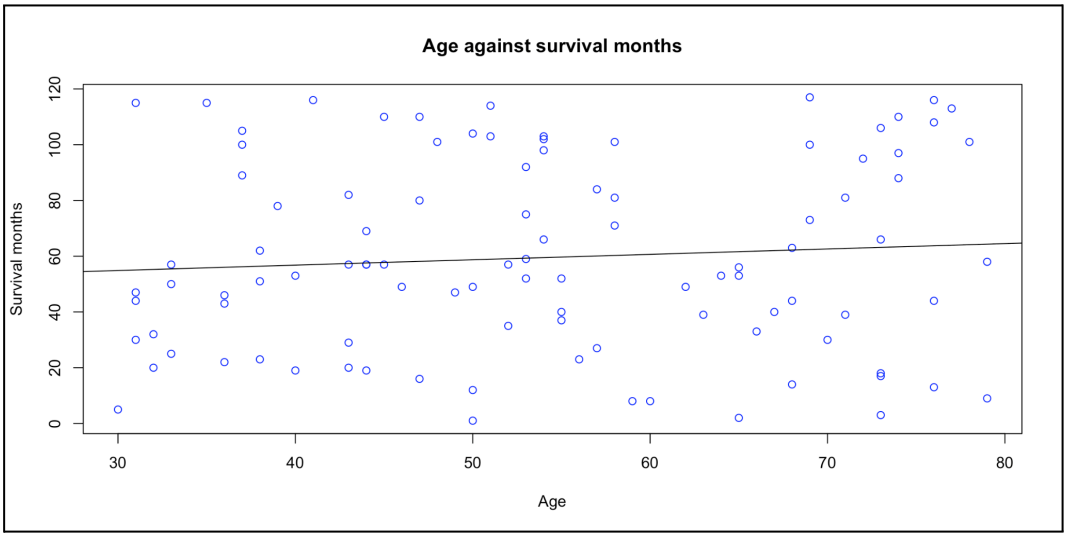
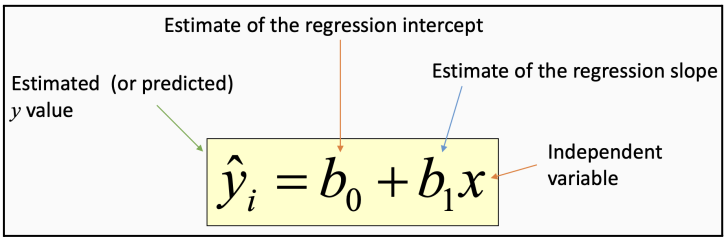


Figure 2.3.1: *Age against Survival months*

1. Estimated Regression Model:



From the equation stated above,

b_0 is the estimated average value of y (survival months) when the value of x (age) is 0.

b_1 is the estimated change in the average value of y (survival months) due to a one-unit change in x (age).

The formulas for b_0 and b_1 are,

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

Implementing the formulas in R studio,

```
> mean(x)
[1] 53.91089
> mean(y)
[1] 59.49505
```

```
> #Regression
> x <- data$Age  #Independent variable
> y <-data$Survival_Months  #dependent variable
> n <- 100
> sum(x)
[1] 5445
> sum(y)
[1] 6009
> sum(x^2)
[1] 314465
> sum(x*y)
[1] 327998
> b1 <-(sum(x*y)-(sum(x)*sum(y)/n))/(sum(x^2)-((sum(x)^2)/n))
> print(b1)
[1] 0.04492417
```

By using RStudio, we get $b_1 = 0.0449$, $b_0 = 57.0732$

Substitute the values of b_0 and b_1 into the regression model equation:

$$\hat{y}_i = 57.0732 + 0.0449x$$

Interpretation:

1. The survival months are not zero for any age groups, so $b_0 = 57.0732$ indicates that the survival months are within the range of months observed, 57.0732 months is the period that is not explained by the age factor.
2. Here, $b_1 = 0.0449$ tells us that the average Age increases by 0.0449 years on average for each additional month of survival.

2. Coefficient of Determination:

$$R^2 = \frac{SSR}{SST}$$

From the equation stated above,

SSR is the sum of squares explained by regression.

SST is the total sum of squares.

The formulas for SSR and SST are,

$$SSR = \sum (\hat{y} - \bar{y})^2 \qquad SST = \sum (y - \bar{y})^2$$

By Implementing the formulas in R studio,

```
> yhat <- b0+(b1*x)
> ssr <- sum((yhat-mean(y))^2)
> sst <- sum((y-mean(y))^2)
> r2 <- ssr/sst
>
> print(r2)
[1] 0.0003613267
```

We get R^2 approximately equal to 0.0004

Interpretation:

The value of $R^2 = 0.0004$ indicates that only 0.04% of the variance in patient survival months is explained by their age. This means that the relationship between age and survival months is very weak.

2.4 Chi-Square Test of Independence

The Chi-Square Test of Independence tests the relationship between two nominal variables. The test involves organizing data into a contingency table with each row corresponding to a category of one variable and each column corresponding to a category of another variable. In this part of the data set analysis, we aim to determine whether there is a relationship between the gender of lung cancer patients and the location of their tumors.

H_0: The tumor location is independent of the gender of the patient.
H_1: The tumor location is dependent on the gender of the patient.

The formula of the test statistic is:

$$X^2 = \sum \frac{(Oi-Ei)^2}{Ei}$$

```
> library (readxl)
>
> setwd ("~/Desktop")
>
> MyData <- read_xlsx("clean_lung_cancer.xlsx")

> MyData2 <- table (MyData$Gender, MyData$Tumor_Location)
> MyData2

        Lower Lobe Middle Lobe Upper Lobe
  Female         14          14         13
  Male           22          20         18
> chisq.test(MyData2)

        Pearson's Chi-squared test

data:  MyData2
X-squared = 0.071319, df = 2, p-value =
0.965
```

Figure 2.4.1: *Calculation and contingency table drawn using RStudio*

After conducting the chi-square test of independence using R programming the chi-square statistic is 0.071319 with 2 degrees of freedom and a p-value of 0.965.

Since the p-value (0.965) is significantly higher than the commonly used significance level of 0.05 we do not reject the null hypothesis. This means based on the sample data provided there is no statistically significant association between the gender of the patient and the location of the tumor. Therefore any observed differences in tumor location distributions between males and females in this sample are likely due to random chance rather than a true underlying relationship.

2.5 Anova

In this ANOVA test, we evaluate if there are significant differences between the means of patient survival months among three categories based on their smoking history (Current smoker, Former smoker, Never smoker) at a 0.05 significance level.

Hypothesis:

H_0: $\mu_1 = \mu_2 = \mu_3$ (All means are equal)

H_1:(At least one of the means is different)

Where:

$\mu_1 \equiv$ survival months mean for current smokers

$\mu_2 \equiv$ survival months mean for former smokers

$\mu_3 \equiv$ survival months mean for never-smokers

Formula:

$$F = \frac{variance\ between\ samples}{variance\ within\ samples}$$

```
> library(tidyverse)
> library(readxl)
>
> lung_cancer_data <- read_excel("lung_cancer_data.xlsx")
>
>
> model <- aov(Survival_Months ~ Smoking_History, data = lung_cancer_data)
> summary(model)
                Df Sum Sq Mean Sq F value Pr(>F)
Smoking_History  2   1841   920.7   0.785  0.459
Residuals       98 115008  1173.5
```

Figure 2.5.1: *The ANOVA test was done using RStudio*

We can see from the results in **Figure 2.5.1** that the p-value (0.459) is greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis which means that there is not enough evidence to conclude that there are significant differences in the means of survival months among the three smoking history categories.

3.0 Discussion and Conclusion

From the hypothesis test, we can conclude that there is a significant difference in tumor size between smokers and non-smokers. This suggests that smoking history may be associated with tumor size in lung cancer patients. The correlation analysis revealed a very weak positive relationship ($r = 0.0796$) between the age of lung cancer onset and survival months. However, this correlation was not statistically significant ($p = 0.4311$) indicating that there is insufficient evidence to conclude a linear relationship between age and survival duration in lung cancer patients. The regression analysis further supported this finding with only 0.04% ($R^2 = 0.0004$) of the variance in patient survival months explained by age. The regression equation ($y = 57.0732 + 0.0449x$) suggests that for each year increase in age, the survival months increase by only 0.0449 months on average. However, given the very low R^2 value, this model has very little predictive power. The Chi-Square test of independence showed no statistically significant association between patient gender and tumor location ($p = 0.965$). This suggests that the location of lung tumors is not dependent on the patient's gender in our sample. The ANOVA test comparing survival months among different smoking history categories (current smoker, former smoker, never smoker) did not find significant differences ($p > 0.05$). This indicates that in our sample smoking history categories do not significantly affect the duration of survival for lung cancer patients.

In conclusion, while our analysis found a significant difference in tumor size between smokers and non-smokers most of our tests did not reveal strong relationships between the variables studied. Age showed little correlation with survival time, gender was not associated with tumor location, and smoking history categories did not significantly affect survival duration. These findings highlight the complex nature of lung cancer and suggest that other factors not included in this study may play important roles in determining patient outcomes. Further research with larger sample sizes and additional variables may be necessary to uncover more significant relationships and predictors of lung cancer progression and survival.

4.0 References

Lung Cancer Prediction. (2024, May 29). Kaggle. https://www.kaggle.com/datasets/rashadrmammadov/lung-cancer-prediction?resource=download