Applied Analytics Assignment 2

An investigation on the impact of smoking on heart rate and it's association with gender

Pratham Radhakrishna S3997064 and Timothy Rohith Bharath S4042762

Last updated: 02 June, 2024

Introduction

-Smoking is a prevalent public health issue associated with numerous diseases, including cardiovascular diseases, respiratory illnesses, and various forms of cancer.
-Understanding the physiological impact of smoking is crucial for developing strategies to mitigate its adverse effects.
-The dataset from Kaggle, titled "Body Signal of Smoking," provides an opportunity to explore the relationship between smoking and various body signals.
-The dataset contains information on several physiological variables, including age, gender, height, weight, and various body signal measurements.
-This rich dataset allows for multiple statistical analyses to understand the impact of smoking on the human body.

Problem Statement

In this analysis, we aim to explore the physiological differences between smokers and non-smokers and identify key physiological predictors associated with smoking.

Specifically, we will address the following questions:

1. Does smoking have a significant effect on heart rate in a sample population?

2. Is there an association between smoking status and gender in the same population?

Data

- The data 'Body signals of smoking' was collected from the Kaggle website. https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking/data
- It contains important health-related information, including demographic, physical, and biochemical measurements data of both smoking and non-smoking groups.
- It consists of 55692 observations and 27 variables including some key variables like age of the individual, gender, height, weight, blood pressure, cholestrol and smoking status.

1. Age
   Type: Numeric
   Age is a crucial variable that affects various health outcomes. It is important for adjusting other health measurements, as risk factors and normal values can change with age.

2. Height
   Type: Numeric
   Height is used to calculate body mass index (BMI) when combined with weight. BMI is an important metric for assessing overweight and obesity status.

Data contd

3. Systolic Blood Pressure
   Type: Numeric
   Systolic blood pressure is a critical measure of cardiovascular health. High systolic blood pressure (hypertension) is a risk factor for heart disease and stroke.

4. Cholesterol Type: Numeric
   Cholesterol levels are important indicators of cardiovascular health. High cholesterol levels can lead to atherosclerosis and increase the risk of heart disease.

5. Smoking Status
   Type: Categorical
   Levels:
   Smoker (1): Indicates individuals who currently smoke. Non-smoker (0): Indicates individuals who do not currently smoke. Smoking status is a key variable in assessing the impact of smoking on health outcomes. It is associated with various adverse health effects, including increased heart rate, hypertension, and elevated cholesterol levels.

Descriptive Statistics and Visualization

In this section, we will summarize the important variables related to smoking status and use visualization to highlight interesting features of the data. We'll also explain how data issues such as missing values and outliers were handled.

Summarizing Key Variables

We'll focus on variables like **age**, **`systolic blood pressure`**, and **`smoking status`**.

```
# Display summary statistics
summary_stats <- dataset %>%
  select(age, `height(cm)`, `weight(kg)`, systolic, Cholesterol, smoking) %>%
  summary()

print(knitr::kable(summary_stats))
```

```
##
##
##  |   |     age       |  height(cm)  |  weight(kg)   |   systolic   | Cholesterol  |      smoking     |
## |:--|:--------------|:-------------|:--------------|:-------------|:-------------|:-----------------|
## |   |Min.   :20.00 |Min.   :130.0 |Min.   : 30.00 |Min.   : 71.0 |Min.   : 55.0 |Non-smoker:35237 |
## |   |1st Qu.:40.00 |1st Qu.:160.0 |1st Qu.: 55.00 |1st Qu.:112.0 |1st Qu.:172.0 |Smoker    :20455 |
## |   |Median :40.00 |Median :165.0 |Median : 65.00 |Median :120.0 |Median :195.0 |NA               |
## |   |Mean   :44.18 |Mean   :164.6 |Mean   : 65.86 |Mean   :121.5 |Mean   :196.9 |NA               |
## |   |3rd Qu.:55.00 |3rd Qu.:170.0 |3rd Qu.: 75.00 |3rd Qu.:130.0 |3rd Qu.:220.0 |NA               |
## |   |Max.   :85.00 |Max.   :190.0 |Max.   :135.00 |Max.   :240.0 |Max.   :445.0 |NA               |
```

Handling Missing Data and Outliers

Missing Data: We can check for missing data and handle it appropriately by either removing the rows with missing values or imputing them.

```
# Check for missing data
missing_data <- colSums(is.na(dataset))
print(knitr::kable(missing_data))
```

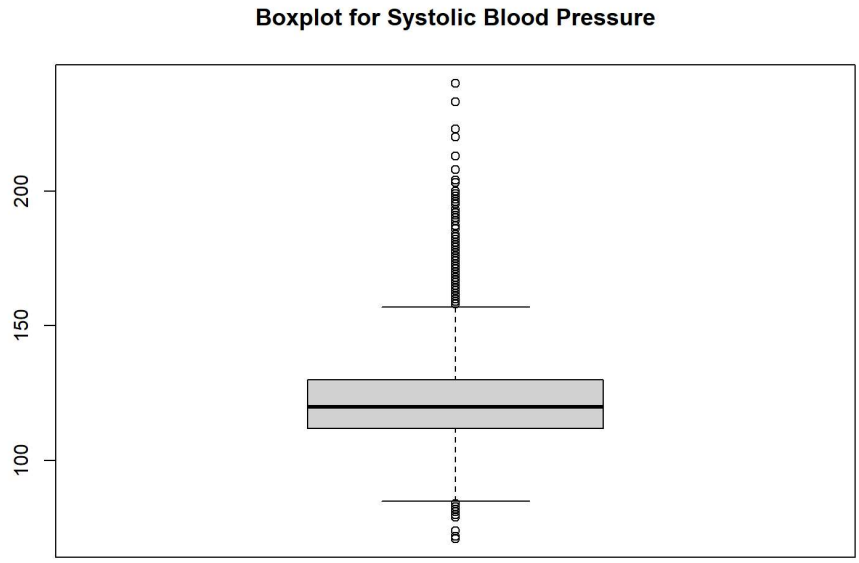
```
##
##
## |                   |  x|
## |:------------------|--:|
## |ID                 |  0|
## |gender             |  0|
## |age                |  0|
## |height(cm)         |  0|
## |weight(kg)         |  0|
## |waist(cm)          |  0|
## |eyesight(left)     |  0|
## |eyesight(right)    |  0|
## |hearing(left)      |  0|
## |hearing(right)     |  0|
## |systolic           |  0|
## |relaxation         |  0|
## |fasting blood sugar |  0|
## |Cholesterol        |  0|
## |triglyceride       |  0|
## |HDL                |  0|
## |LDL                |  0|
## |hemoglobin         |  0|
## |Urine protein      |  0|
## |serum creatinine   |  0|
## |AST                |  0|
## |ALT                |  0|
## |Gtp                |  0|
## |oral               |  0|
## |dental caries      |  0|
## |tartar             |  0|
## |smoking            |  0|
```

The dataset did not contain any missing values.

Outliers:

We can detect and handle outliers using interquartile range (IQR) or other methods. For simplicity, we might visualize and remove extreme outliers.

```
# Visualize outliers using boxplots
boxplot(dataset$systolic, main="Boxplot for Systolic Blood Pressure")
```




```
# Remove extreme outliers if necessary (example: removing systolic BP > 180)
dataset <- dataset %>% filter(systolic <= 180)

# Re-check the summary statistics after cleaning
cleaned_summary <- dataset %>%
  select(age,`height(cm)`, `weight(kg)`, systolic, Cholesterol, smoking) %>%
  summary()

print(knitr::kable(cleaned_summary))
```

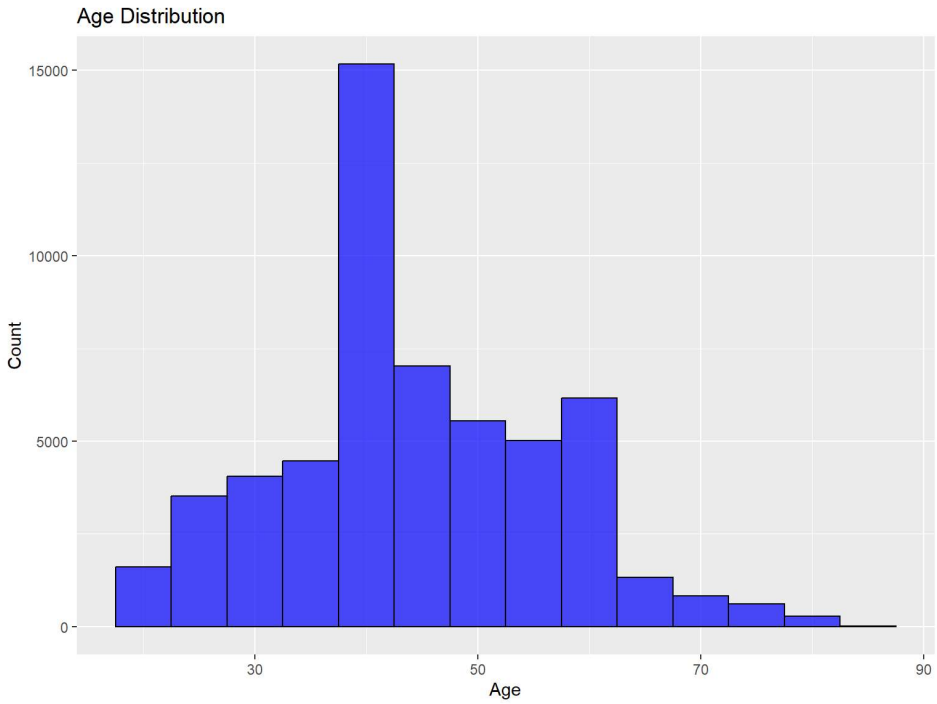
```
##
##
## |   |     age       |  height(cm)  |  weight(kg)   |   systolic   | Cholesterol  |      smoking     |
## |:--|:--------------|:-------------|:--------------|:-------------|:-------------|:-----------------|
## |   |Min.   :20.00 |Min.   :130.0 |Min.   : 30.00 |Min.   : 71.0 |Min.   : 55.0 |Non-smoker:35203 |
## |   |1st Qu.:40.00 |1st Qu.:160.0 |1st Qu.: 55.00 |1st Qu.:112.0 |1st Qu.:172.0 |Smoker    :20429 |
## |   |Median :40.00 |Median :165.0 |Median : 65.00 |Median :120.0 |Median :195.0 |NA               |
## |   |Mean   :44.18 |Mean   :164.7 |Mean   : 65.86 |Mean   :121.4 |Mean   :196.9 |NA               |
## |   |3rd Qu.:55.00 |3rd Qu.:170.0 |3rd Qu.: 75.00 |3rd Qu.:130.0 |3rd Qu.:220.0 |NA               |
## |   |Max.   :85.00 |Max.   :190.0 |Max.   :135.00 |Max.   :180.0 |Max.   :445.0 |NA               |
```

Visualizing the Distribution of Important Variables

We'll create visualizations for **age**, **height**, **weight**, **systolic blood and pressure**, and their relationship with **smoking status**.

1. Age distribution

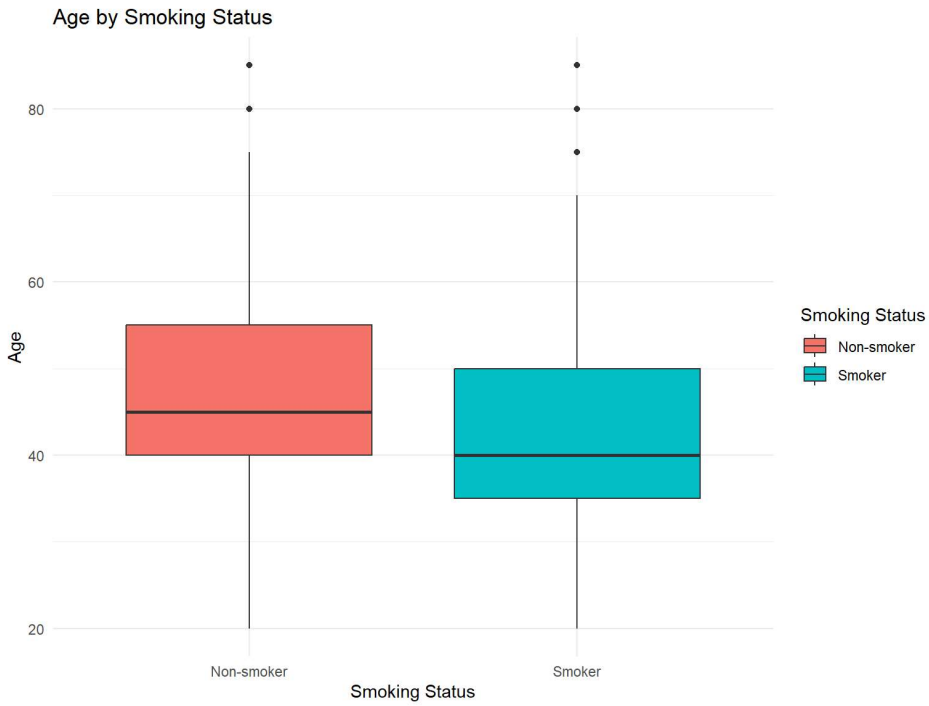
```
# Histogram for age
p1 <- ggplot(dataset, aes(x=age)) +
  geom_histogram(binwidth=5, fill="blue", color="black", alpha=0.7) +
  labs(title="Age Distribution", x="Age", y="Count")
p1
```



The histogram shows a moderately bell-shaped distribution, suggesting that the ages are approximately normally distributed.
The peak of the histogram is around 40 years, indicating that most of the people in the dataset are in their 40s.

Visualizing the Distribution of Important Variables contd

```
# Boxplot for age by smoking status
ggplot(data = dataset, aes(x = smoking, y = age,
                           fill = factor(smoking))) +
  geom_boxplot() +
  theme_minimal() +
  labs(fill = "Smoking Status", title="Age by Smoking Status", x="Smoking Status", y="Age")
```



From the boxplot, it is evident that the median for the non smoking group is larger than that of the smoking group. This suggests that older individuals are less likely to smoke compared to the younger groups.

Visualizing the Distribution of Important Variables contd

2. Gender

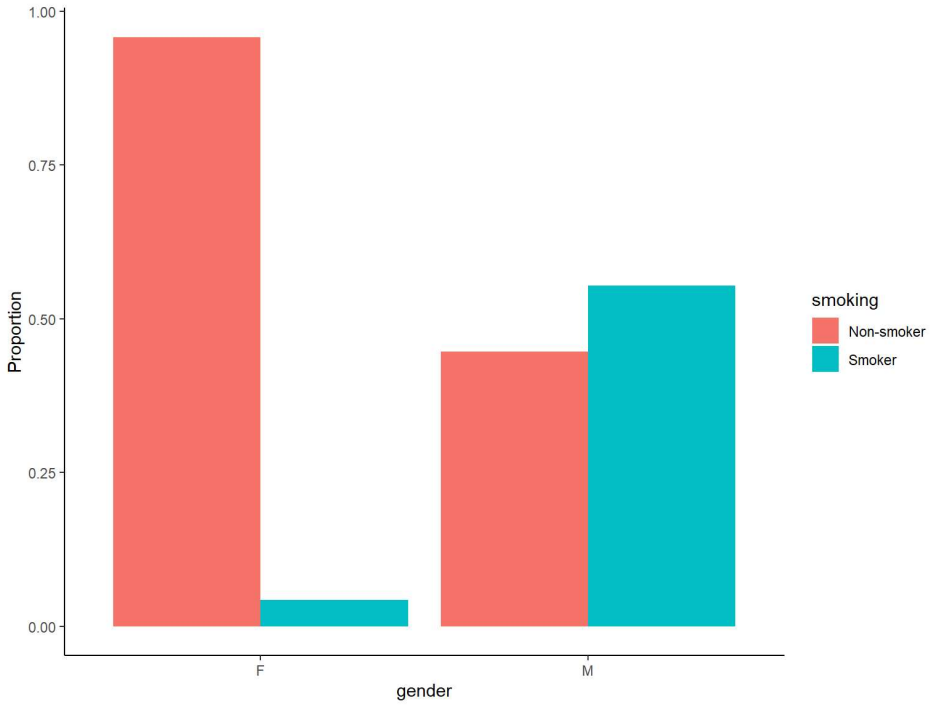
```
#To calculate proportions of smokers in male and female categories
gender_match_prop <- dataset %>% group_by(gender, smoking) %>%
  summarise(no = n()) %>%
  mutate(Proportion = no / sum(no))
```

```
## `summarise()` has grouped output by 'gender'. You can override using the
## `.groups` argument.
```

```
knitr::kable(gender_match_prop)
```

gender	smoking	no	Proportion
F	Non-smoker	19415	0.9576777
F	Smoker	858	0.0423223
M	Non-smoker	15788	0.4465058
M	Smoker	19571	0.5534942

```
ggplot(data = gender_match_prop,
       aes(x = gender, y = Proportion, fill=smoking)) +
  geom_bar(position="dodge", stat="identity") +
  theme_classic()
```

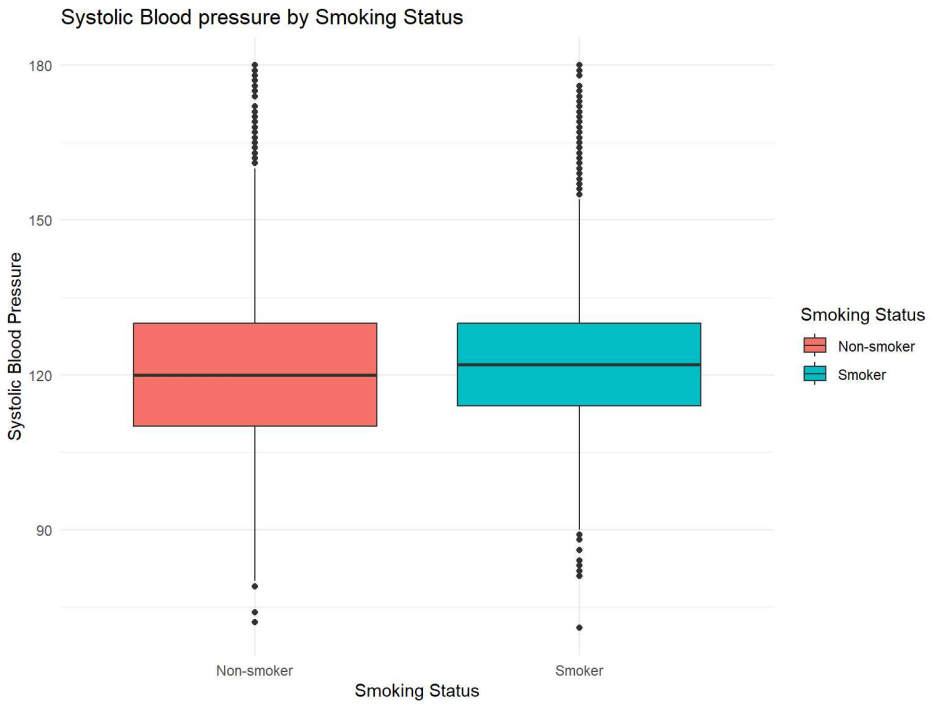


The visualisation shows that proportion of females in the dataset who are smokers are very low, at about 4.23%, and majority of them are non-smokers. However, it is quite different for males, with majority of them being smokers at 55.34%.

Visualizing the Distribution of Important Variables contd

4. Blood pressure

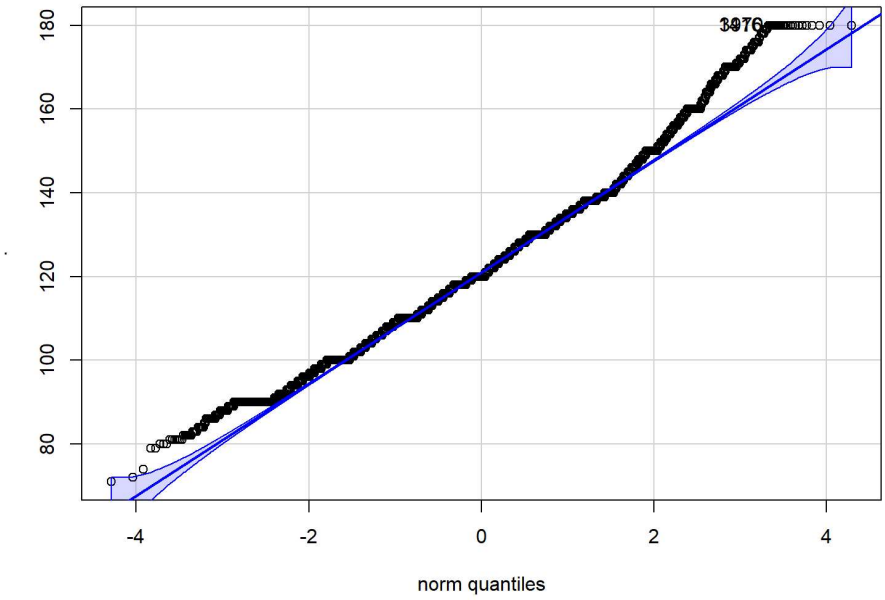
```
# Boxplot for systolic blood pressure by smoking status
ggplot(data = dataset, aes(x = smoking, y = systolic,
                       fill = factor(smoking))) +
  geom_boxplot() +
  theme_minimal() +
  labs(fill = "Smoking Status", title="Systolic Blood pressure by Smoking Status", x="Smoking Status", y="Systolic Blood Pressure")
```



The boxplot shows that the systolic blood pressure of the smoking group is slightly higher than that of the non-smoking group.

Visualizing the Distribution of Important Variables contd

```
dataset$systolic %>% qqPlot(dist="norm")
```




```
## [1] 1416 3970
```

From the QQ plot, as majority of the data points lie close to the diagonal reference line, it suggests that the systolic blood pressure values are approximately normally distributed.

Hypothesis Testing and Confidence Interval

1. Formulate Hypotheses:

- **Null Hypothesis (**H0H0): There is no difference in the mean systolic blood pressure between smokers and non-smokers.

$$H_0 : \mu_{\text{smokers}} = \mu_{\text{non-smokers}}$$

- **Alternative Hypothesis (**H1H1): There is a difference in the mean systolic blood pressure between smokers and non-smokers.

$$H_0 : \mu_{\text{smokers}} \neq \mu_{\text{non-smokers}}$$

Hypothesis Testing and Confidence Interval Contd

2. Check Assumptions:

- The data is approximately normally distributed within each group.
- The variances of the two groups are not assumed to be equal (hence using Welch's t-test).

```
# Split data into smokers and non-smokers
smokers <- filter(dataset, smoking == 1)
non_smokers <- filter(dataset, smoking == 0)
```

```
# Calculate summary statistics
mean_smokers <- mean(smokers$systolic, na.rm = TRUE)
mean_non_smokers <- mean(non_smokers$systolic, na.rm = TRUE)
var_smokers <- var(smokers$systolic, na.rm = TRUE)
var_non_smokers <- var(non_smokers$systolic, na.rm = TRUE)
n_smokers <- sum(!is.na(smokers$systolic))
n_non_smokers <- sum(!is.na(non_smokers$systolic))
```

```
#Perform Welch's t-test
t_test_result <- t.test(systolic ~ smoking, data = dataset, var.equal = FALSE)

# Calculate 95% confidence interval for the difference in means
conf_interval <- t_test_result$conf.int

# Print results
t_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  systolic by smoking
## t = -17.708, df = 45132, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Non-smoker and group Smoker is not equal to 0
## 95 percent confidence interval:
##  -2.279534 -1.825190
## sample estimates:
## mean in group Non-smoker     mean in group Smoker
##                 120.6609                 122.7133
```

```
conf_interval
```

```
## [1] -2.279534 -1.825190
## attr(,"conf.level")
## [1] 0.95
```

Hypothesis Testing and Confidence Interval Contd

Output Summary:

- **Test Statistic (**tt): -17.708
- **Degrees of Freedom (**dfdf): 45132
- **p-value**: < 2.2e-16
- **95% Confidence Interval for the Difference in Means**: [-2.279534, -1.825190]
- **Sample Estimates**:
  - *Mean systolic blood pressure in group 0 (non-smokers): 120.6609*
  - *Mean systolic blood pressure in group 1 (smokers): 122.7133*

Hypothesis Testing and Confidence Interval Contd

Interpretation:

1. **Test Statistic and p-value**:
    - *The test statistic ($t=-17.708$t=−17.708) is highly significant.*
    - *The p-value (< 2.2e-16) is far less than the common significance level of 0.05, indicating that we reject the null hypothesis.*
2. **Confidence Interval**:
    - *The 95% confidence interval for the difference in means is between -2.279534 and -1.825190. This interval does not include 0, which further supports the rejection of the null hypothesis.*
3. **Mean Estimates**:
    - *The mean systolic blood pressure for non-smokers is 120.6609.*
    - *The mean systolic blood pressure for smokers is 122.7133.*
    - *The difference in means is approximately 2.05 units (smokers have higher systolic blood pressure on average).*

Categorical association

To investigate the association between categorical variables (e.g., smoking status and gender), we can use a chi-squared test of independence. This test helps determine if there is a significant association between two categorical variables.

Step-by-Step Procedure:

1. Formulate Hypotheses:

- **Null Hypothesis ($H0$):** There is no association between smoking status and gender.
  $H0$: Smoking status is independent of gender
- **Alternative Hypothesis ($H1$):** There is an association between smoking status and gender.
  $H1$: Smoking status is not independent of gender.

2. Check Assumptions:

- The data should be in the form of counts or frequencies.
- The expected frequency in each cell of the contingency table should be at least 5.

```r
# Create a contingency table for gender and smoking status
contingency_table <- table(dataset$gender, dataset$smoking)

# Perform the chi-squared test of independence
chi_squared_result <- chisq.test(contingency_table)

# Print the contingency table and the chi-squared test results
contingency_table
```

```
##
##     Non-smoker Smoker
##   F      19415    858
##   M      15788  19571
```

```r
chi_squared_result
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contingency_table
## X-squared = 14487, df = 1, p-value < 2.2e-16
```

Categorical association contd

Interpretation:

1. **Contingency Table**:
    - *The table shows the counts of non-smokers (0) and smokers (1) for each gender (F and M).*
    - *There are significantly more male smokers compared to female smokers.*
2. **Chi-squared Test Statistic and p-value**:
    - *The test statistic (X2X2) is 14,487.*
    - *The p-value is less than 2.2e-16, which is extremely small.*
3. **Degrees of Freedom**:
    - *The degrees of freedom (df) for the test is 1.*

Results and Discussion

The investigation into the relationship between smoking status and heart rate and gender yielded two key findings:

- From the Hypothesis test, the analysis clearly shows that there is a significant difference in systolic blood pressure between smokers and non-smokers. Smokers tend to have a higher systolic blood pressure than non-smokers by about 1.83 to 2.28 units on average. This difference is statistically significant, indicating that smoking is associated with higher systolic blood pressure.
- From the Categorical association conducted for gender and smoking status, the p-value (< 2.2e-16) wass much smaller than the typical significance level (0.05), leading us to reject the null hypothesis. This result indicates that there is a highly significant association between gender and smoking status. Specifically, the data suggests that males are much more likely to be smokers compared to females.
- Future research should aim to include more diverse samples to enhance the generalizability of the findings, and consider additional factors to deepen our understanding of these relationships.

References

- Dataset: https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking/data
- Biswal, A. (2024) What is hypothesis testing in statistics? types and examples, Simplilearn.com. Available at: https://www.simplilearn.com/tutorials/statistics-tutorial/hypothesis-testing-in-statistics#:~:text=Hypothesis%20Testing%20is%20a%20type,relationship%20between%202%20statistical%20variables
- GeeksforGeeks (2023) Create boxplot with respect to two factors using GGPLOT2 in R, GeeksforGeeks. Available at: https://www.geeksforgeeks.org/create-boxplot-with-respect-to-two-factors-using-ggplot2-in-r/
- Zach BobbittHey there. My name is Zach Bobbitt. I have a Masters of Science degree in Applied Statistics and I've worked on machine learning algorithms for professional businesses in both healthcare and retail. I'm passionate about statistics (2022) How to perform Welch's T-test in R, Statology. Available at: https://www.statology.org/welch-t-test-in-r/
- Team, D. (2021) Chi-square test in R: Explore the examples and essential concepts!, DataFlair. Available at: https://data-flair.training/blogs/chi-square-test-in-r/