Maternal Smoking and Infant Death Statistical Analysis

October 13, 2024

Contribution

Student 1 and Student 2 collaborated on the Introduction, analyses 1-5 in Section 2, and the Conclusion. Student 2 completed the Advanced Analysis section and wrote the code for the analysis and visualizations in the R-file, while Student 1 organized the data visualizations throughout the report. Both students worked together on formatting the PDF, and Student 2 completed the Appendix.

1 Introduction

This report investigates the potential differences in birth weight between babies born to mothers who smoked during pregnancy and those who did not. Epidemiological research has consistently linked smoking during pregnancy to low birth weights and preterm birth. Birth weight is an indicator of a newborn's health, with lower birth weights associated with an increased risk of infant mortality. This study seeks to quantify the effects of maternal smoking on infant health through statistical analysis using data from the Child Health and Development Studies (CHDS) dataset.

The CHDS dataset includes data on 1236 babies, all male, who were born between 1960 and 1967 to mothers enrolled in the Kaiser Health Plan in Oakland, California. Each baby was a single birth and survived for at least 28 days after birth. Variables in the dataset include birth weight, length of gestation, whether or not this was the mother's first pregnancy, mother's age, mother's height, mother's weight, and whether or not the mother was a smoker.

The primary objective is to answer the question: Is there a statistically significant difference in birth weights between babies born to mothers who smoked during pregnancy and those who did not? Specifically, we will explore the distribution of birth weights for both smokers and non-smokers, the incidence of low birth weight (under 100 ounces) in both groups, the potential variability in these comparisons based on different statistical methods. Using numerical, graphical, and incidence comparison approaches to answering this question, we found that maternal smoking significantly impacts birth weight. Babies born to mothers who smoked during pregnancy tended to have lower birth weights on average (~10 oz) compared to those born to non-smokers.

The report is structured as follows: Section 1 addresses each of the research questions outlined above, focusing on the numerical, graphical, and incidence comparisons between the smoking

and non-smoking groups; Section 2 covers the statistical analysis processes and results; Section 3 expands the analysis by examining other variables that may impact birth weight; and Section 4 discusses the conclusions and implications of the results. Overall, this report will determine whether the difference in weight between babies born to mothers who smoked during pregnancy and those who did not is significant for the health of the baby.

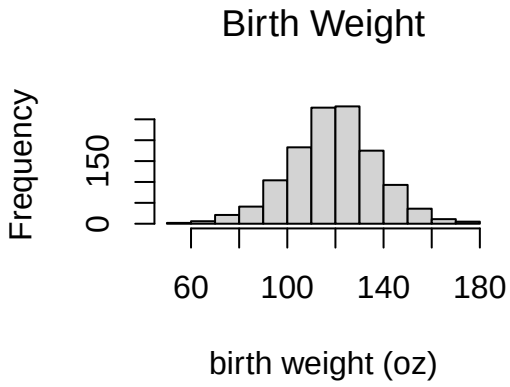
2 Analysis

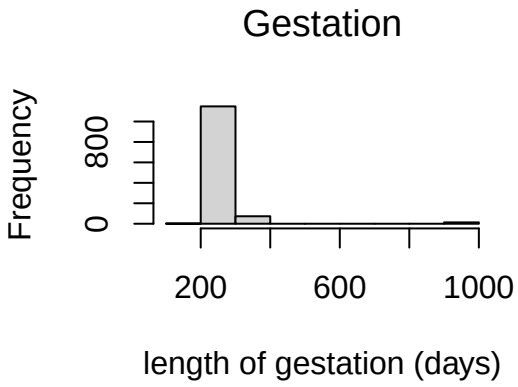
2.1. Data Processing

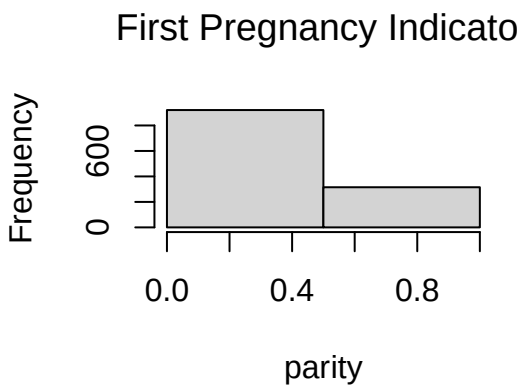
2.1.1. Methods

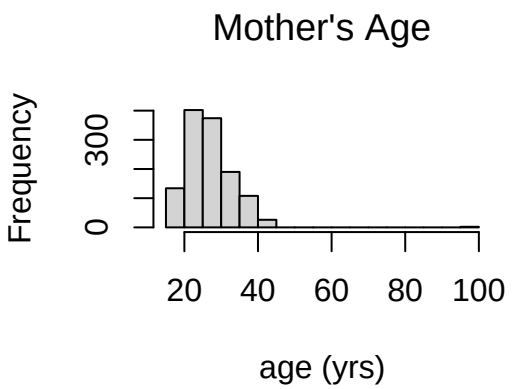
The first step of our analysis was to examine the dataset and understand the characteristics and distributions of each variable. We reviewed all variables and determined their type (numerical or categorical) and identified any outliers or irregularities. To help us understand the structure of the data, we looked at the distributions of each variable graphically using a histogram.

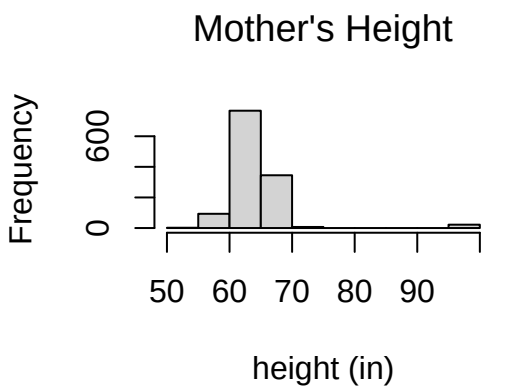
2.1.2. Analysis

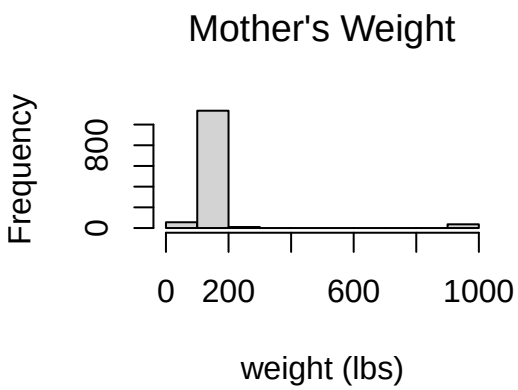


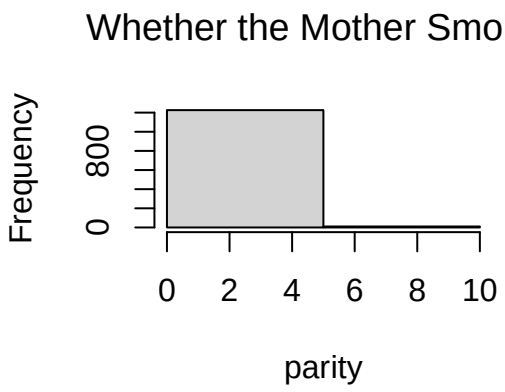












Several outliers and inconsistencies were identified in the data, which may distort the results of our analyses. The dataset includes a binary indicator for whether the pregnancy was the first for the mother. Since parity can either take on 0 (first pregnancy) or 1 (subsequent pregnancy), we removed any entries that were not 0 or 1. The current maximum height of a woman, based on the Guinness World Record, is 7 feet 0.7 inch (84.7 inches); we set the upper limit for height at 85 inches to account for potential data entry errors or unreasonable outliers. The longest pregnancy recorded lasted 375 days; gestation values exceeding this were excluded from the dataset as unrealistic outliers. We set an upper weight limit of 500 pounds, which is extremely high but not impossible. Women weighing over 500 lbs are not too rare, but are typically associated with severe health conditions, which would significantly skew the dataset. Excluding such outliers helps avoids distorting results. Similarly, we set the maximum reasonable age for a mother's age to be 60 years. Pregnancies above this age are exceptionally rare and could result from data recording errors.

After cleaning the data, we plotted the distributions of each variable to confirm that the cleaning process successfully addressed the outliers and inconsistencies. These graphs, presented in the Appendix, reflect the revised dataset and provide a more accurate representation of the variables.

Refer to the Appendix for the revised histograms of birth weight, gestation, parity, mother's age, height, and weight.

2.1.3. Conclusion

We concluded that the dataset does not qualify as a simple random sample because of the method in which it was collected. In order to qualify as a simple random sample, every pregnant women would need to have had an equal chance of being selected. However, in this dataset the pregnant women selected were limited to a certain time frame (1960-1967) and were limited to a specific health plan in a single region. Therefore, this dataset does not qualify as a simple random sample. This means that the dataset is not representative of the entire population and the results of the analysis cannot be generalized.

2.2. Numerical Analysis of Birth Weight Distributions

2.2.1. Methods

Now that we had examined the dataset and each variable as a whole, we looked at the two distributions of birth weights for babies born to women who smoked during their pregnancy and for babies born to women who did not smoke during their pregnancy. Summarizing these distributions numerically, we calculated the minimum, maximum, mean, median, standard deviation, and quartile values. This will allow us to understand the basic characteristics of the data before we perform any visualizations.

2.2.2. Analysis

We found that the numerical summaries were as follows:

Table 1: Numerical Differences in Distributions of Smoking and Non-Smoking Mothers

	Smokers	Non_Smokers
Minimum	58.0000	55.0000
Maximum	163.0000	176.0000
Mean	113.8192	123.0853
Median	115.0000	123.0000
Standard Deviation	18.2950	17.4237
Q1	101.0000	113.0000
Q3	126.0000	134.0000

Most of these numerical summaries show that the birth weights of babies of mothers who are non-smokers are larger than the birth weights of babies of mothers who are smokers. To determine whether or not those differences are significant would require further analysis.

2.2.3. Conclusion

We found that the mean and median birth weights of the smokers were 114.1095 and 115 respectively while the mean and median birth weights of the non-smokers were 123.0472 and 123 respectively. The lower mean and median birth weights for smokers suggest that smoking may contribute lower birth weights. Since the means and medians of the two distributions were very similar, we can infer that the shapes of the two distributions are almost symmetrical.

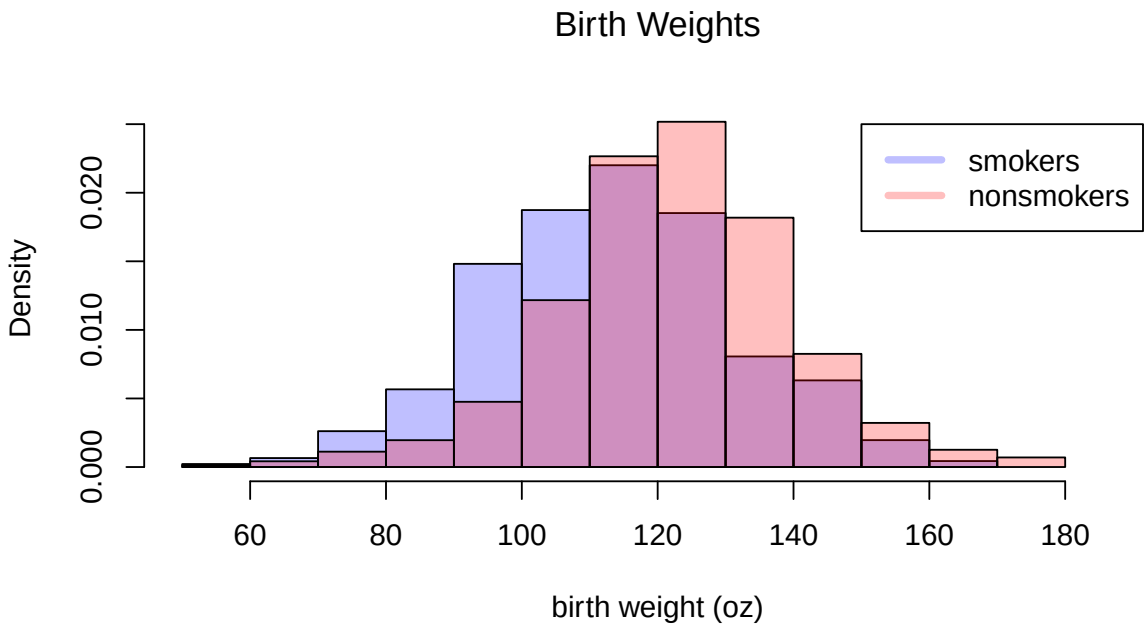
2.3. Graphical Analysis of Birth Weight Distributions

2.3.1. Methods

Now that we've summarized the two distributions numerically, we will visualize the distribution of birth weights for both smokers and non-smokers to explore the data more comprehensively.

2.3.2. Analysis

We plotted two graphs, one with the birth weight distribution of women who are non-smokers and another with the birth weight distribution of women who are smokers. Overlaying the two distributions allows us to observe any visual differences between the groups.



2.3.3. Conclusion

Based on the layering of the two distributions, the distribution of the baby weights of the mothers who are smokers is shifted to the left of the the distribution of the baby weights of the mothers who are non-smokers. This shows that overall, the baby weights of mothers who are smokers is less than the baby weights of mothers who are non-smokers. To determine whether or not this amount is significant, further analysis would need to be done.

2.4. Incidence Analysis of Low Birth Weights

2.4.1. Methods

Now that we've done both numerical and graphical analysis, we want to use one final comparison approach: incidence comparison. Using 100 ounces as the threshold of a low-weight baby, we found the percentages of babies that weigh under that threshold for both smoking and non-smoking mothers. We then adjusted the threshold to observe any changes in the incidences of low-weight births.

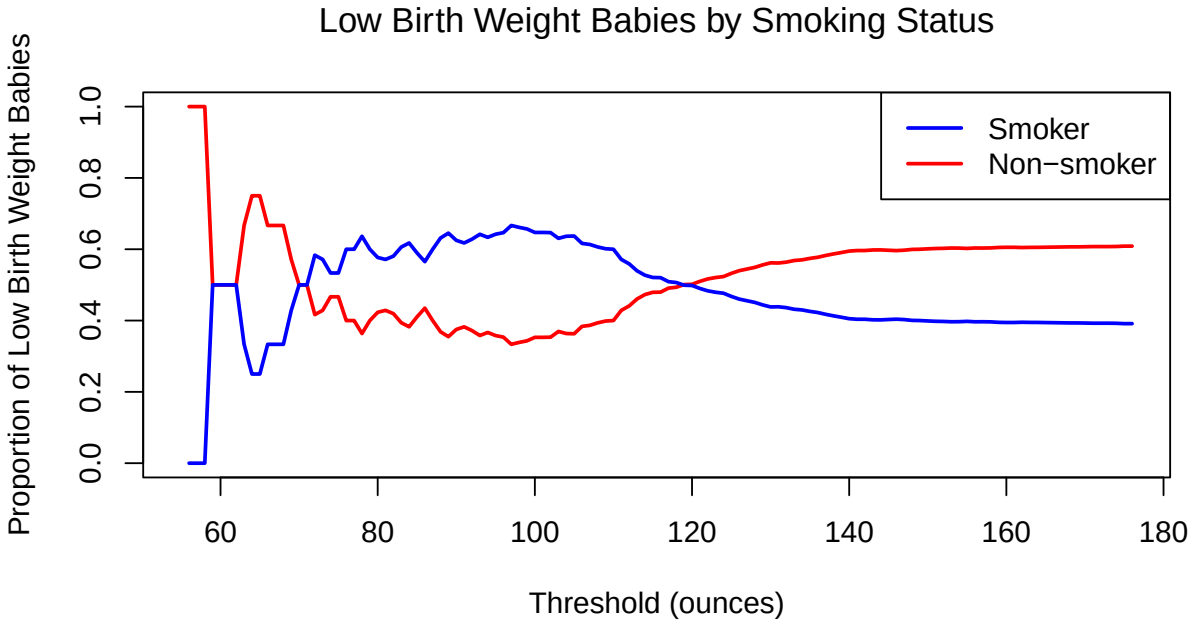
2.4.2. Analysis

We found that the percentages of low-weight babies are as follows:

Percentage of low-weight babies in smoking mothers: 64.70588%

Percentage of low-weight babies in non-smoking mothers: 35.29412%

Then, we wanted to visualize the changes in the incidences of low birth weights as the threshold changes. So, we used every integer between the minimum and maximum birth weights as a threshold and recorded how the percentages of low-weight babies in smoking vs. non-smoking mothers changed. The results of this are shown below.



2.4.3. Conclusion

As shown in the graph above, the incidences of low-weight babies in smoking mothers is greater than the incidences of low-weight babies in non-smoking mothers when the threshold

for low-weight babies is between 70 and 120. Outside of that range, the results are flipped.

2.5 Comparative Analysis of Various Analysis Types

Numerical Analysis

The first type of analysis used was numerical, where we compared the minimum, maximum, mean, standard deviation, median, and quartile values of the two distributions (non-smoking women and smoking women). Numerical analysis provide concrete quantifications that allowed us to highlight the key information in the data to directly compare the two distributions. However, using numerical analysis dismisses information about the shape of the data that may be important for analysis. For example, if the data were skewed, the means and standard deviations would be less reliable due to their susceptibility to outliers.

Graphical Analysis

The second type of analysis used was graphical, where we graphed the two distributions to visualize their general shapes. While graphs make it easier to visualize shape, trends, patterns, and outliers, they don't provide specific numerical values or statistical evidence. So while graphical analysis allows us to visualize any trends, numerical is analysis is still necessary for proving the significance of those observations. In other words, numerical and graphical analysis complement each other and can be used hand-in-hand for statistical analysis.

Incidence Analysis

The last type of analysis used was incidence analysis, where we looked at rates of low-births in each distribution at different thresholds.

Conclusion

The results of each analysis strongly suggest that smoking has a significant effect on birth weight and increases the risk of preterm births. The summary statistics reveal that mothers who smoke have babies with lower birth weights on average, the graph of the distribution of smokers was shifted to the left of the graph of the distribution of non-smokers, and the proportion of low birth weights among smokers (64.71%) is substantially higher than that among non-smokers (35.29%).

3 Advanced Analysis

We will explore the relationship between maternal smoking and gestation length, focusing on preterm births (defined as gestation periods less than 259 days). Smoking is known to

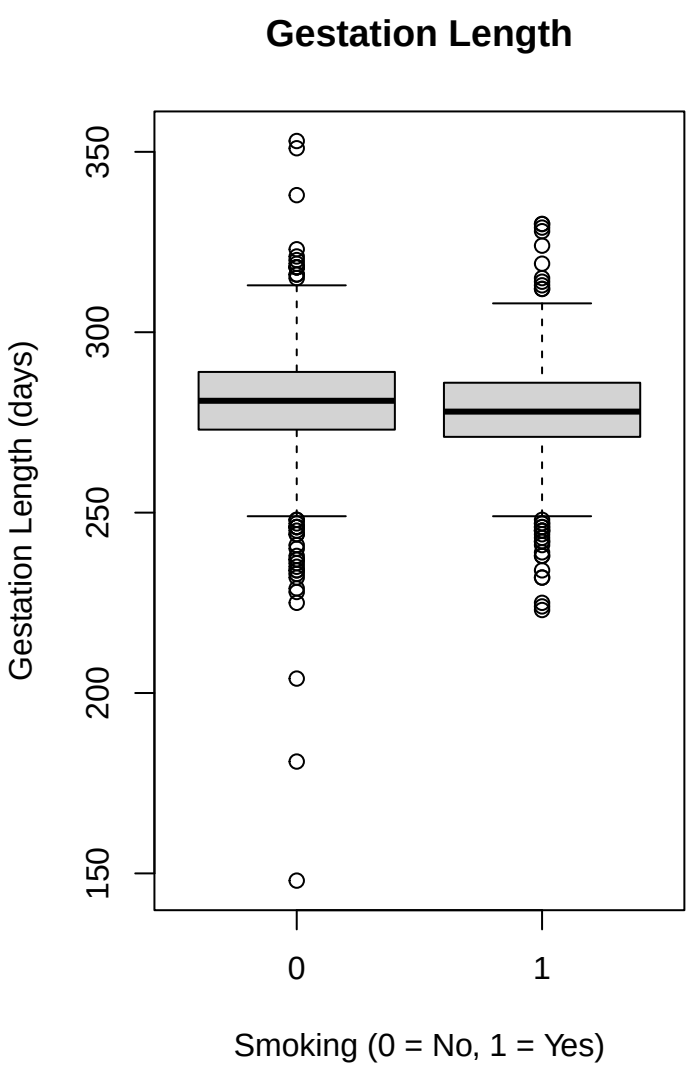
be associated with shortened gestation and increased rates of preterm birth, and advertedly lower birth weights.

We begin by summarizing the gestation lengths for both smokers and non-smokers:

Table 2: Numerical Differences in Distributions of Smoking and Non-Smoking Mothers

	Smokers	Non_Smokers
Min	223.00000	148.00000
Max	330.00000	353.00000
Mean	277.89760	279.87413
Median	278.00000	281.00000
Standard Deviation	15.20143	16.47282
Q1	271.00000	273.00000
Q3	286.00000	289.00000

The following boxplot shows the distribution of gestation lengths for smokers and non-smokers:



To determine if the observed difference in gestation lengths between smokers and non-smokers is statistically significant, we will perform a t-test, providing the t-value, p-value, and 95% Confidence Interval. A p-value less than 0.05 would indicate a statistically significant difference in gestation lengths between smokers and non-smokers.

From the summary statistics, we observe that the mean, median, first and third quartile gestation length is shorter for smokers compared to non-smokers. The boxplot supports this difference, showing a slight leftward shift in the distribution of gestation lengths for smokers. The t-test confirms whether this difference is statistically significant, with a p-value of 0.03895697, refer to Appendix: T-Test for Gestation Length. Smoking appears to be associated with shorter gestation periods, which can increase the risk of preterm birth.

4 Discussion

The goal of this analysis was to find out if there was a statistically significant difference between the birth weights of babies that were born from smoking mothers and non-smoking mothers. Using numerical, graphical, and incidence comparisons, we found that all three types of analysis suggest that there is a difference. Specifically, babies that were born from smoking mothers tend to be lighter than babies that were born from non-smoking mothers. We even looked into other variables, like gestation period, further showing how smoking mothers tend to give birth to babies of lower weights.

However, our results cannot be generalized to the entire population. In other words, we cannot say that all smoking mothers tend to give birth to babies of lower weights. This is due to the nature of our sample, which wasn't representative of the entire population that we wanted to study (pregnant mothers).

Other limitations of this dataset may also affect the applicability of our results.

- The dataset only provides data about women in 1960 to 1967, meaning the results from this analysis may no longer be relevant to women in the present time.
- The absence of data on female babies and multiple-birth babies also limits the applicability of our findings.
- Recording only one occurrence of the mother's weight, even though how it changes throughout the course of the pregnancy may also be an important factor to consider in the analysis.
- The aforementioned limitations with specific locations and health plans.

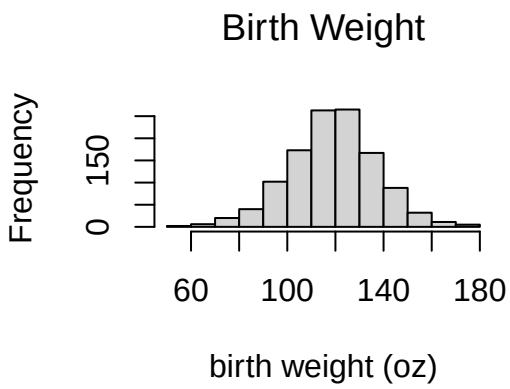
We acknowledge the presence of these limitations and understand their affects on the generalizability of our results.

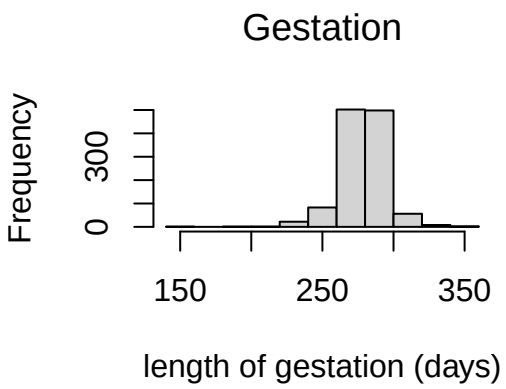
Other observations made during the analysis process...

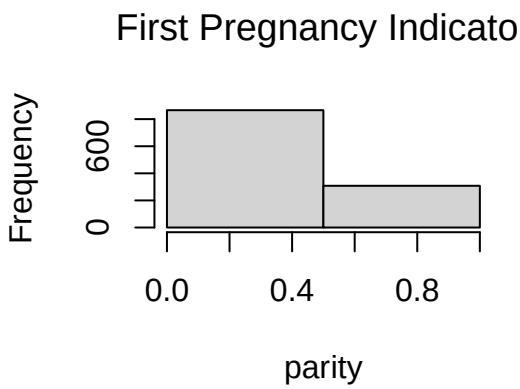
Future considerations include looking at...

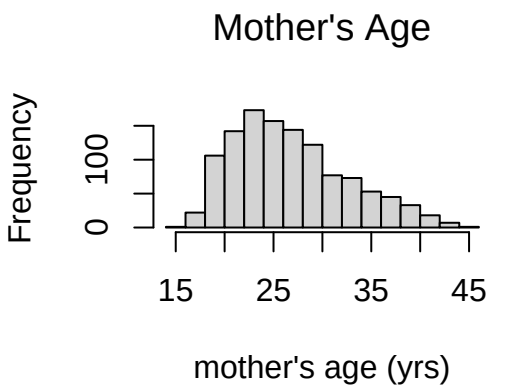
5 Appendix

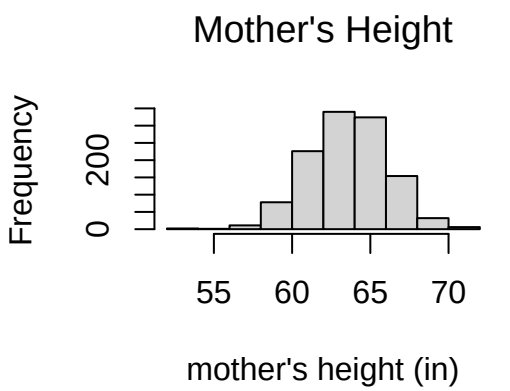
Distribution of the Cleaned Data

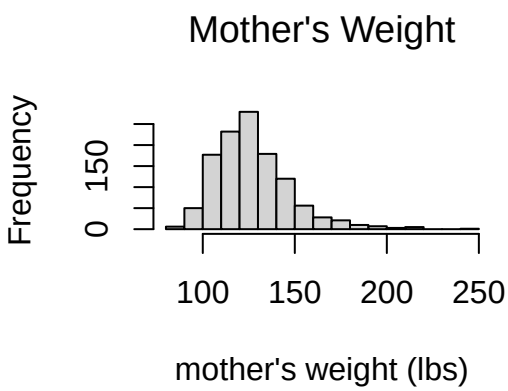


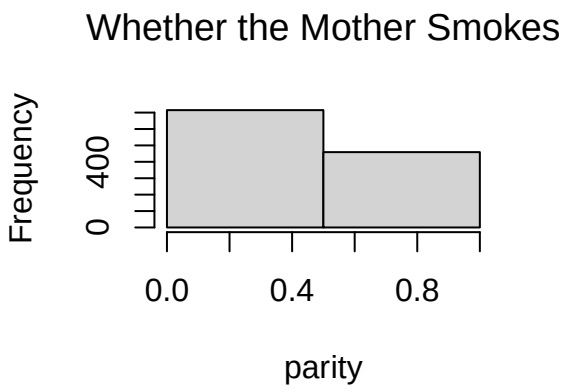












T-Test for Gestation Length

This test is used to determine if there is a statistically significant difference in the means of two independent groups.

We assume:

1. **Independence:** The data from the two groups (smokers and non-smokers) are independent. Since each row in our dataset represents a unique pregnancy, the assumption of independence is satisfied.
2. **Normality:** The distribution of gestation lengths in each group is approximately normal. While the histograms suggest some deviations from normality, given the large sample sizes, the Central Limit Theorem justifies use of the T-test.
3. **Equal Variance:** The t-test assumes that the variances in gestation length between the two groups are roughly equal.

Hypotheses:

- **Null Hypothesis (H0):** The mean gestation length for mothers who smoke is equal to the mean gestation length for mothers who do not smoke.
- **Alternative Hypothesis (H1):** The mean gestation length for mothers who smoke is different from the mean gestation length for mothers who do not smoke.

Test Statistic:

Using the `t.test()` function, we computed the test statistic and p-value with a **95% confidence level**.

- **t-statistic**: The computed t-statistic was **-2.0670**.
- **Degrees of Freedom (df)**: 1172.
- **p-value**: The p-value for the t-test was **0.03895697**, which is far below the 0.05 significance level. This allows us to reject the null hypothesis.

Conclusion:

Since the p-value is less than 0.05, we reject the null hypothesis and conclude there is a statistically significant difference in the mean gestation length between smokers and non-smokers. Smokers tend to have shorter gestation periods, which aligns with our expectations based on prior epidemiological studies. This is further supported by the increased proportion of under-weight births observed in the smoking group.