Homework 1 Report

Antony Sikorski, Ayaan Saifi

4/5/2021

Analysis of the Effects of Smoking During Pregnancy on the Birthweight of Children

Author Contributions:

Ayaan Saifi: Data setup, Problem 2 histograms, Problem 3 83 and 88 cutoffs, Advanced analysis data cleaning and Gestation vs Birthweight, Advanced analysis scatter plot and t-test

Antony Sikorski: Problem 1, Problem 2 boxplot, Problem 3 93 cutoff and density plot, Advanced Analysis gestation vs Parity, advanced analysis bxoplot

The code was evenly split, and both authors also contributed equally to the write up of the report itself.

Introduction

In this study, we examine a data set that pertains to the birth weight of newborn babies based on multiple variables regarding their mothers. The data is sourced from Child Health and Development Studies (CHDS). The study was performed between 1960 to 1967, in which data was collected on 1236 babies. The babies were strictly males who all lived at least 28 days and were born from a single birth (no twins/triplets etc.). A quick demonstration of the format of the data set is presented below:

```
##   bwt gestation parity age height weight smoke
## 1 120       284      0  27     62    100     0
## 2 113       282      0  33     64    135     0
## 3 128       279      0  28     64    115     1
## 4 123       999      0  36     69    190     0
## 5 108       282      0  23     67    125     1
## 6 136       286      0  25     62     93     0
```

As seen above, the data measured multiple variables, which include the birth weight (oz), gestation period (days), parity (whether or not the mother has given birth before), age of mother (yrs), height of mother (in), weight of mother (lbs), and whether the mother smokes or not.

Our interest lies in studying which variables affect the birth weight of the child, primarily whether the mother smokes or not. Prior research shows that the birth weight and general health of a child can be negatively affected by a mother who smokes during pregnancy. The Surgeon General makes the claim that “Smoking by pregnant women may result in fetal injury, premature birth, and low birth weight”. In addition to this, epidemiological research has also shown that babies who are born early and born with a low birth weight have lower survival rates. We wish to examine this claim by comparing the birth weights of the children in both smoking and non-smoking mothers. We examine these claims by comparing the birth weights of babies based on whether their mothers smoke or not, and then do further analysis to determine whether the gestation period has an effect on the babies’ birth weight so that we can provide resources for future studies to predict which babies will potentially have lower survival rates.

Initial Analysis

Data Processing & Methods

To begin, we wished to study the effects of smoking strictly on birth weight, so we decided to clean the missing values from those categories. Since, birth weights did not have any missing values, we simply had to clean the missing values from the “smoke” category, which were labeled as the number 9 (it is a binary variable). To begin, we present a summary of our entire data frame:

```
##       bwt          gestation        parity            age       
##  Min.   : 55.0   Min.   :148.0   Min.   :0.0000   Min.   :15.00  
##  1st Qu.:109.0   1st Qu.:272.0   1st Qu.:0.0000   1st Qu.:23.00  
##  Median :120.0   Median :280.0   Median :0.0000   Median :26.00  
##  Mean   :119.5   Mean   :286.9   Mean   :0.2569   Mean   :27.35  
##  3rd Qu.:131.0   3rd Qu.:288.0   3rd Qu.:1.0000   3rd Qu.:31.00  
##  Max.   :176.0   Max.   :999.0   Max.   :1.0000   Max.   :99.00  
##      height          weight          smoke       
##  Min.   :53.00   Min.   : 87.0   Min.   :0.0000  
##  1st Qu.:62.00   1st Qu.:115.0   1st Qu.:0.0000  
##  Median :64.00   Median :126.0   Median :0.0000  
##  Mean   :64.67   Mean   :154.1   Mean   :0.3948  
##  3rd Qu.:66.00   3rd Qu.:140.0   3rd Qu.:1.0000  
##  Max.   :99.00   Max.   :999.0   Max.   :1.0000  
```

After this, we will split our cleaned data into two categories: the birth weights of babies of smokers and the birth weights of babies of non-smokers. We can then numerically summarize our data of these two categories by finding the mean, median, standard deviations, IQR, and general distribution of the data.

We plan to graphically represent our data through the use of histograms and box plots to gain a visual perspective on the differences in birth weights of children caused by smoking. We will perform a hypothesis test in order to determine whether the difference in means is significant enough to warrant the conclusion that smoking has some effect on the birth weights. Following this, we will perform analysis on the frequency of low weight babies from both smokers and non-smokers, along with a visual demonstration to better show the differences.

Analysis

First, we summarize the data numerically to gain a better understanding of the general distributions.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    58.0   102.0   115.0   114.1   126.0   163.0 
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##      55     113     123     123     134     176 
```

We also found the mean and standard deviation of both data sets, since they are not included in the above summary.

Mean (Smokers):

```
## [1] 114.1095
```

Standard Deviation (Smokers):

```
## [1] 18.09895
```

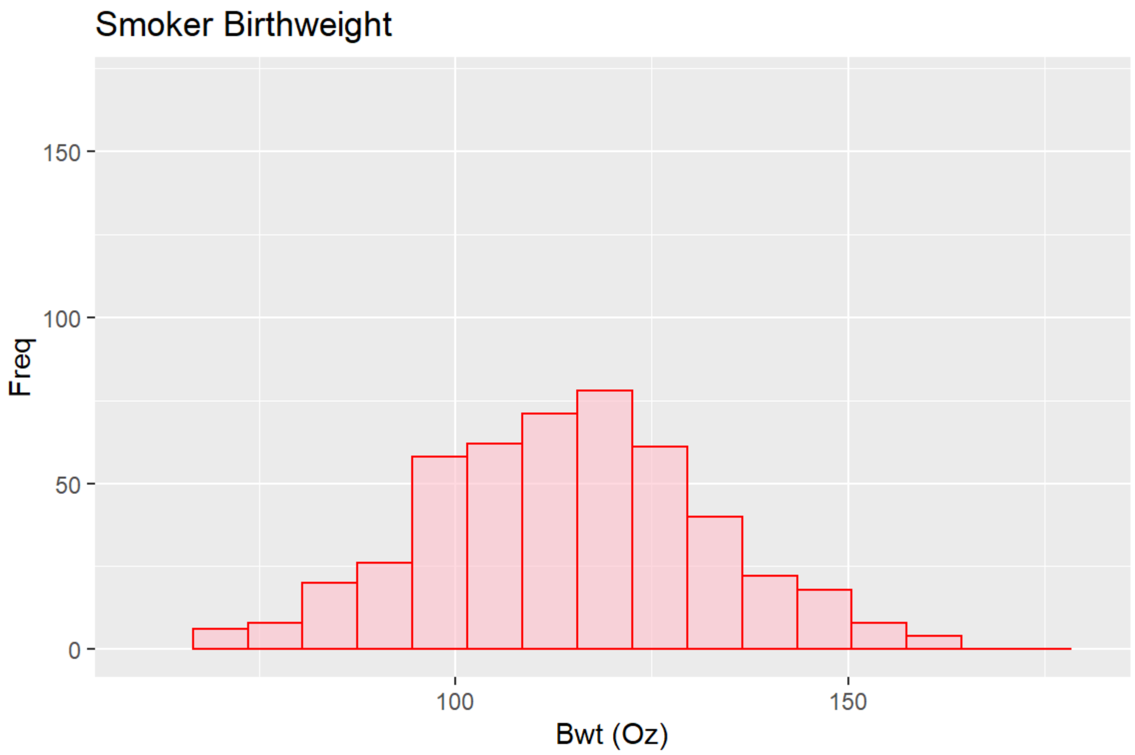
Mean (Non-Smokers):

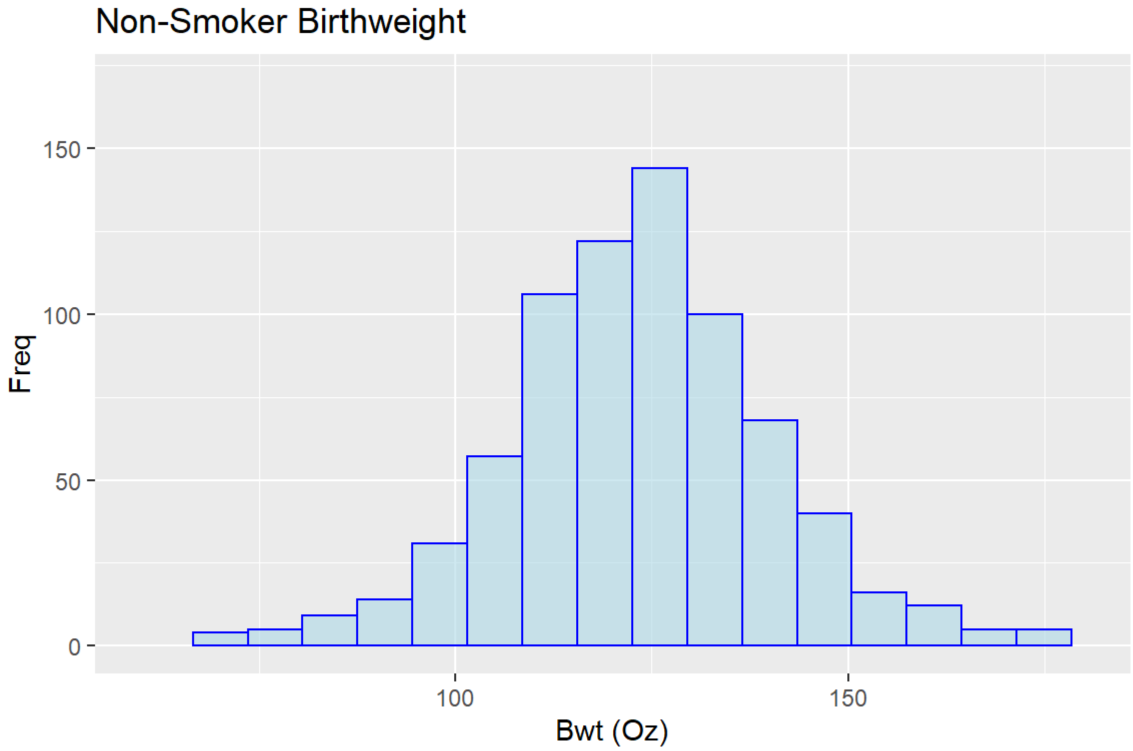
```
## [1] 123.0472
```

Standard Deviation (Non-Smokers)

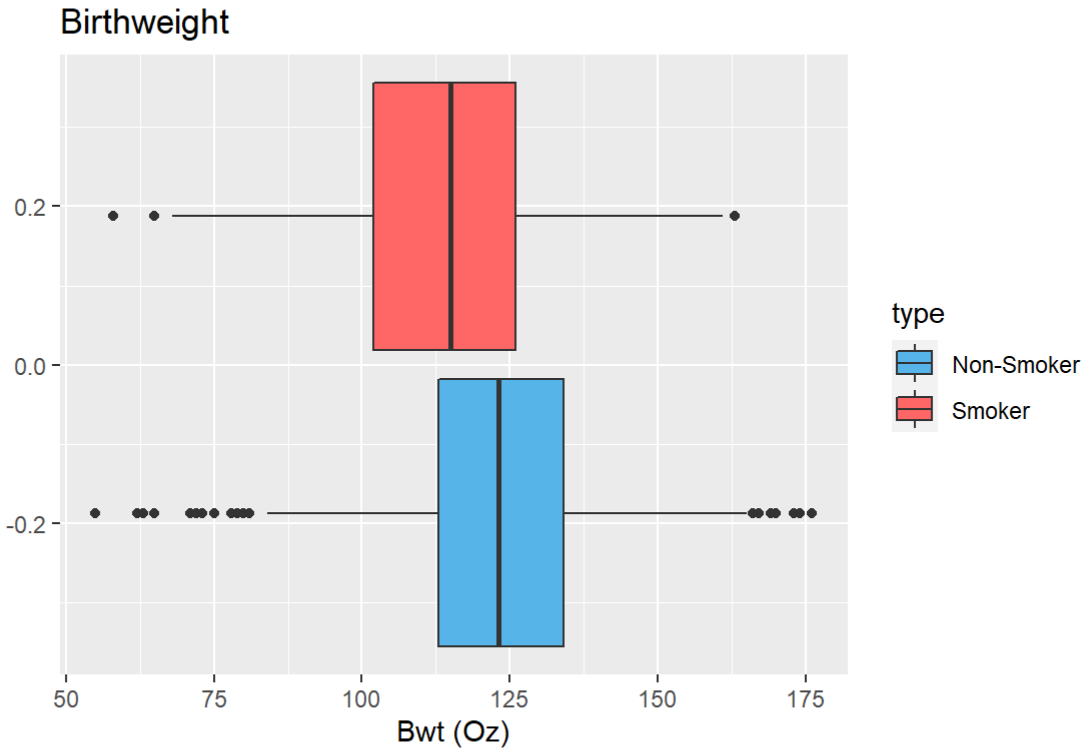
```
## [1] 17.39869
```

We then represent the two distribution separately using histograms to gain a sense of how the birth weights are distributed for both smokers and non-smokers.





We then plot both distributions as side-by-side box plots so we can see the differences between them.



Next, we perform a two sample Welch t-test in order to determine whether the difference in the means of the distribution is significant enough to prove that smoking affects the birth weight of the child.

```
##
##  Welch Two Sample t-test
##
## data:  data_smoker$bwt and data_nonsmoker$bwt
## t = -8.5813, df = 1003.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.98148  -6.89385
## sample estimates:
## mean of x mean of y
##  114.1095  123.0472
```

Finally, we analyze the frequency of low birth weights in both distributions. The general cutoff for a low weight baby is roughly less than or equal to 2500 grams, which converts to approximately 88 ounces (oz). We chose to perform analysis on the frequency of low weight babies based on multiple definitions, so we set two other cutoffs at 83 ounces and 93 ounces. This was done to avoid inconsistencies based on the source that defines “low birth weight”, and to analyze how the frequencies change with different cutoffs.

Frequency of low weight births for smokers (cutoff at 88 oz):

```
## [1] 0.08264463
```

Frequency of low weight births for non-smokers (cutoff at 88 oz):

```
## [1] 0.0309973
```

Frequency of low weight births for smokers (cutoff at 83 oz):

```
## [1] 0.04338843
```

Frequency of low weight births for non-smokers (cutoff at 83 oz):

```
## [1] 0.01886792
```

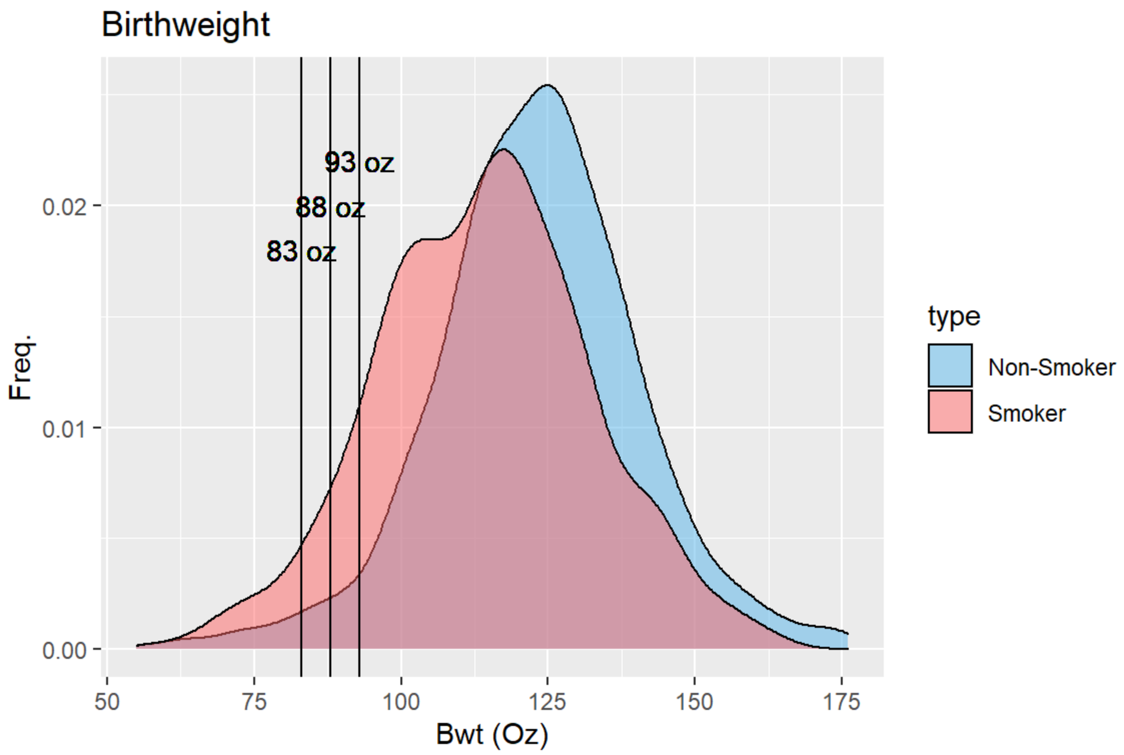
Frequency of low weight births for smokers (cutoff at 93 oz):

```
## [1] 0.1198347
```

Frequency of low weight births for non-smokers (cutoff at 93 oz):

```
## [1] 0.04716981
```

Visually, we can demonstrate the frequencies at each cutoff using a labeled density plot, which gives a good perspective of how changes in the definition of “low birth weight” affect frequency in the respective distributions.



Conclusions of Initial Analysis

We can now begin to draw conclusions surrounding the relationship between smoking and birth weight from our analysis and visual representations above. To begin, the numerical analysis shows that the smoker and non-smoker mean birth weights are roughly 123 ounces and 114 ounces, respectively. The medians are 123 and 115 ounces. The distributions are quite symmetric in nature, and look approximately normal because of the similarities in the mean, median, and spread. The standard deviations are 18 and 17, meaning that the distributions themselves are also quite similar, with the principal difference being their means.

The histogram of the non-smokers has larger frequencies, but this is simply due to the sample size of the non-smokers being higher than that of the smokers. The two distributions are better compared together on a side-by-side box plot, where their similarities are seen, but there appears to be a general trend of non-smoker babies having a larger birth weight. The difference in the means is reinforced by a two sample Welch t-test. The null hypothesis is that the two means are the same, with the alternative hypothesis claiming that they are different. The test was done with a significance value of alpha = 0.05. The p-value that is found is extremely low (2.2e-16), meaning that there is a significant difference in the means of the two distributions.

Since it is unreliable to simply conclude this via a t-test, which primarily considers means, we also examined the frequency of low weight births within the two distributions. A low weight birth is traditionally defined as less than or equal to 88 ounces, but we chose to perform the same calculations using cutoffs of 83 and 93 ounces to get a more holistic view. Although it is quite obviously represented on our density plot, all of the calculations concluded that the frequency of low birth weights was much higher for smokers than for non-smokers. In fact, the frequencies for 88 ounces being the cut off were roughly: 8% low birth weights for smokers, and 3% low birth weights for non-smokers.

Further Analysis

Data Processing and Methods

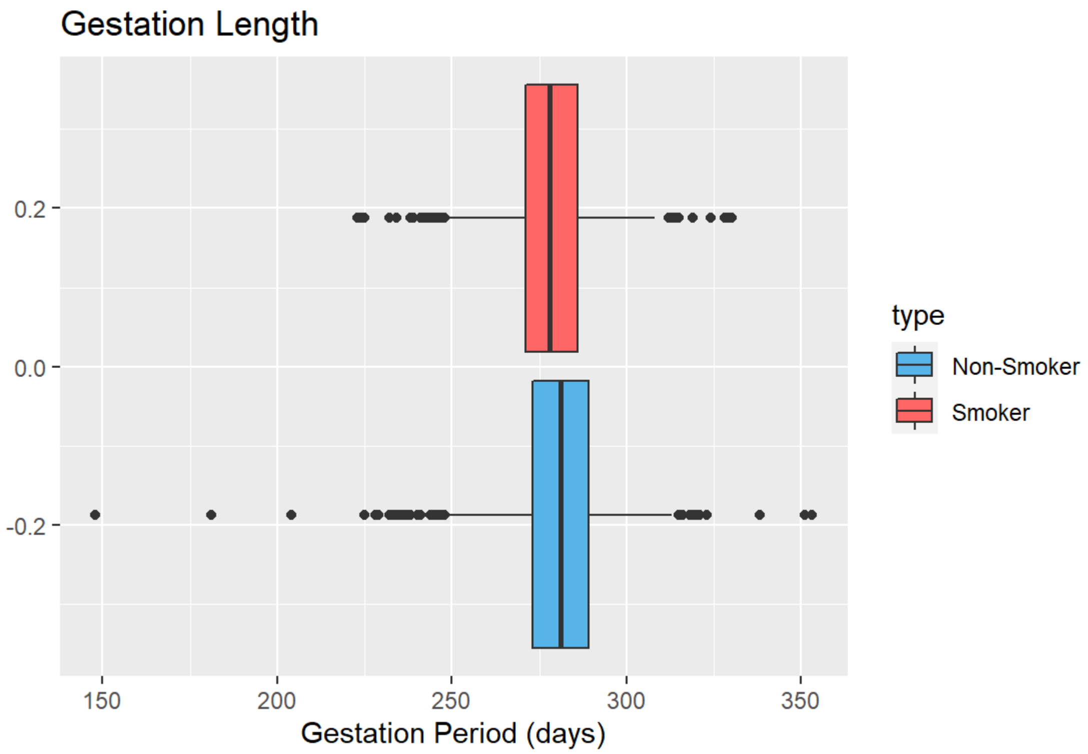
To further analyze the data, we chose to factor in the gestation period of the pregnancy. We want to see whether or not smoking has a significant effect on the gestation period, and whether the gestation period has an effect on the birth weight of the child. In order to do this, we first chose to completely clean all of the missing values from the data rather than just focusing on cleaning the smoking and birth weight variables. After this, we can create separate distributions for smokers and non-smokers.

After that, we will visually show the effects that smoking has on the gestation period, and see if the differences in the means are significant enough to conclude a difference that is not a result of a statistical or sampling anomaly. We will do this by yet again utilizing the two sample Welch t-test.

After this, we will visually demonstrate the correlation between the gestation period and the birth weight in an attempt to form a loose conclusion about other ways that smoking may affect the birth weight of a child.

Analysis

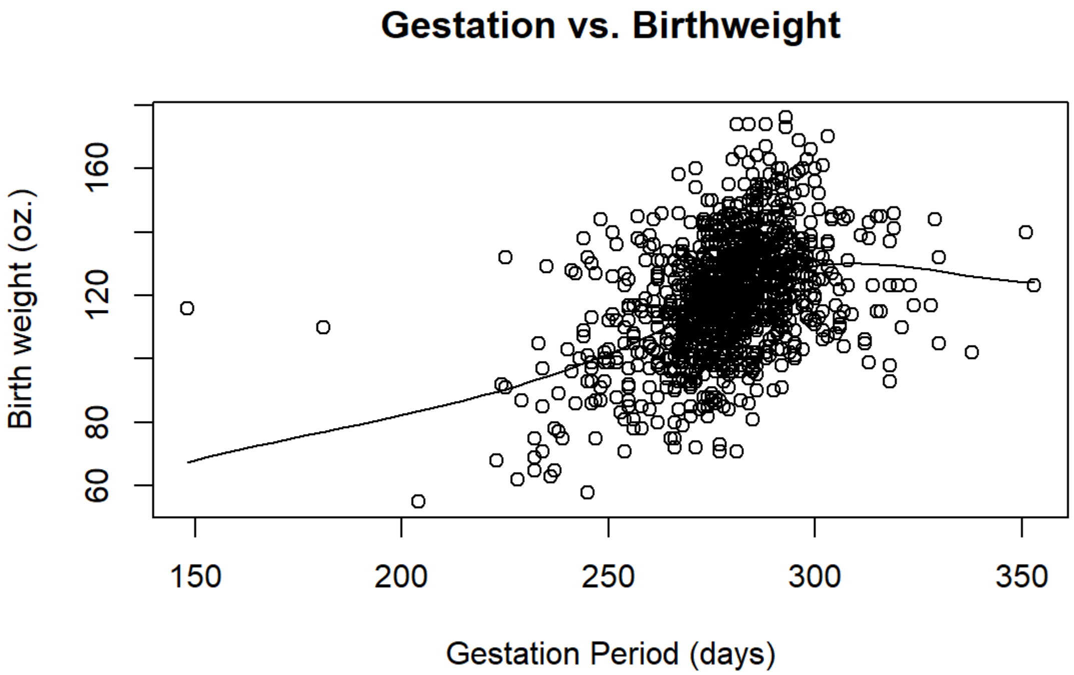
First, we plot a combined box plot which shows the length of the gestation periods for both smokers and non-smokers.



Since it is clear that there is a difference in the distributions, especially the means, we perform a two sample Welch t-test to determine how significant that difference is.

```
##
##  Welch Two Sample t-test
##
## data:  data_smokerXX$gestation and data_nonsmokerXX$gestation
## t = -2.1034, df = 1032.4, p-value = 0.03567
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.8203886 -0.1326562
## sample estimates:
## mean of x mean of y
##  277.8976  279.8741
```

Finally we plot a scatter plot that shows the correlation between the gestation period and the birth weight of the baby, and create a trend line to demonstrate the general trend of the distribution.



Conclusions of Further Analysis

After interpreting the box plot, we can see that the two distributions are clearly not equal. Not factoring in the outliers, the two have a similar spread, yet there exists a general trend where babies of smokers have a shorter gestation period. In order to support this conclusion, we perform the t-test, in which the null hypothesis is that the true difference in the means is 0, while the alternative hypothesis states that the true means are not equal to each other. The test was done with an alpha = 0.05 significance level, and our calculated p-value is 0.036, which means that there is a significant difference in the means.

Since we have determined that smoking affects the gestation period, we yet again relate it to the birth weight of the child. The general trend that is shown on the scatter plot is that the longer the gestation period, the higher the birth weight of the child. Since our data points to the fact that smokers generally have shorter gestation periods, we can reinforce the fact that smoking has a negative effect on the birth weight of the child.

Conclusion and Discussion

The goal of this study was to determine the effects of smoking on the birth weight of newborn children. Various studies have found that smoking has a negative impact on the health of newborns, and as it is a common goal of society to increase infant survival rate, we chose to investigate this claim. To begin, we found a clear difference in the distributions of birth weights for smokers and non-smokers. We used numerical and visual analysis, along with hypothesis tests, to prove our suspicions. Previous studies have predicted an average drop of approximately 150 grams (roughly 5 ounces) of weight between smoking and non-smoking babies, yet our study found that on average the non-smoker babies were 255 grams (9 ounces) heavier. To prove that the difference in the means was significant, we found that the p-value of the t-test of the two means was 2.2e-16, which is significantly smaller than our alpha = 0.05. This means that there is a statistically significant difference in the means of the two distributions. In addition to this, we analyzed the frequency of low weight babies with cutoffs of 83, 88, and 93 ounces, and found that the frequency of low weight babies was much higher when the mother smoked. All of these measures point to the conclusion that smoking reduces the birth weight of newborn children.

To reinforce our previous conclusion, we chose to examine the effects of smoking on the gestation period of the baby. We found that smoking reduces the average gestation period by a statistically significant amount, since our t-test between the means of gestation period concluded with a p-value of 0.036, which is smaller than our alpha = 0.05. We also found that a lower gestation period generally means a lower birth weight, which routes back to smoking generally lowering the birth weight of a newborn. Since there is strong evidence from various research sources, along with the Surgeon General himself, that low birth weight and a short gestation period generally result in poor infant health and a lower survival rate, we can conclude that smoking has negative effects on the overall health of the baby, and advise against doing so during pregnancy.

Despite seemingly strong evidence, there is much improvement that can be done to our study. To begin, the study was done with only male babies, which entirely leaves out any data from female babies. In addition to this, our study did not have any twin or triplet births, which could be interesting to examine since those are more dangerous than the birth of a single child. Our study also did not have even sample sizes for smokers and non smokers, with the non smoker population being about 1.5 time larger than the smoker population. Finally, the data for our analysis is sourced from a study that was conducted from 1960 to 1967, which is long enough ago that it would be wise to revisit this topic in a new study. Further research could examine both genders, and have a smaller difference in the smoker and non smoker populations. In addition to this, we make the assumption that the “smoking” discussed in this study is tobacco, since that was what was primarily popular during those years. It would be interesting to examine the effects of not only tobacco, but also of e-cigarettes and marijuana, which are extremely prevalent in today’s society. These further studies will hopefully reinforce the idea that smoking has negative effects on the health of newborns, and discourage mothers from doing so during pregnancy.