## 1. Introduction

According to scholars in science, it is widely known that carbon monoxide contained in cigarette smoke decreases the oxygen that is supplied to the fetus. Steady supply of oxygen is vital for a baby's development, although the precise physiological effect of the lack of oxygen isn't yet scientifically understood. The birth weight of the newborn baby may be a critical component to determine whether there may be problems with the baby's health (decrease in oxygen, struggles in development,  high probability of death rate). With this in mind,  this study will focus on analyzing the statistical differences in birth weight between babies born to mothers who smoked during pregnancy and those who did not. This paper will discover through these differences by examining the numerical summary, graphical distribution and frequency of the birth weights.

## 2. Analysis

### 2.1 Numerical summary of the three distributions of birth weight

**Methods:**
First, we filtered out smokers and nonsmokers separately. This was done by using "setdiff" on the entire dataset and the smoker indices. This indicates that the nonsmoker dataset contains unknown, or irregular data (the value of column "smoke" is 9). To see whether this unknown data causes bias, we divided the nonsmoker dataset into "nonsmoker with irregular data" and "nonsmoker without irregular data". Then we compare between these three datasets by using "summary". We specifically looked at the "bwt" column statistics for analysis.

**Analysis:**
According to the summary, it was found that "nonsmoker with irregular" and "nonsmoker without irregular" datasets barely had any difference in all statistical aspects; they were almost the same, which means that the outliers don't have much statistical significance. The smoker dataset generally had lower "bwt" statistics, having a lower mean, 1st quartile, etc. Babies born with weight less than 5.5 pounds(88 oz) are considered small for their gestational age. The 1st quartile numbers for the three datasets were 102, 113 and 113. We found that the summary itself has a limitation in reflecting the exact proportion of small weight babies, since the 1st quartile values weren't really close to the threshold (88oz). And even though the mean of bwt was lower for the "smokers" dataset, the minimum value was a little higher than the "nonsmoker" datasets (58oz. > 55oz.). Thus, we figured that "summary" wasn't the perfect tool for statistical comparisons; a visual method would be better.

**Conclusions:**

While “summary” was quite helpful in that it displayed all the exact statistics and distributions for each datasets in one picture, it wasn’t possible for us to figure out the exact proportion of babies that weigh under 88 oz, which is a vital information needed for us to determine whether the difference between birth weights of babies born from mothers who smoke and mothers who don’t smoke.

## 2.2 Graphical distribution

**Methods:**

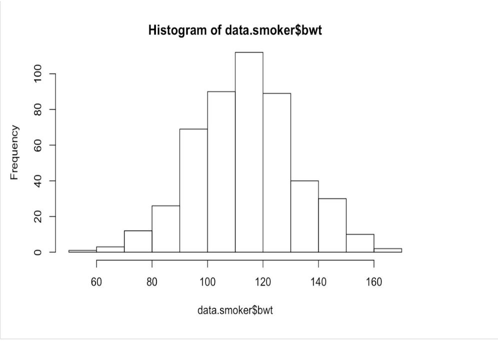
We used three different graphical distributions to compare the birth weight in ounces (bwt) of smokers and non-smokers. The different graphical distributions integrated were the histogram and boxplot. There were three datasets that were used to model the graphical distributions, and they were the following:

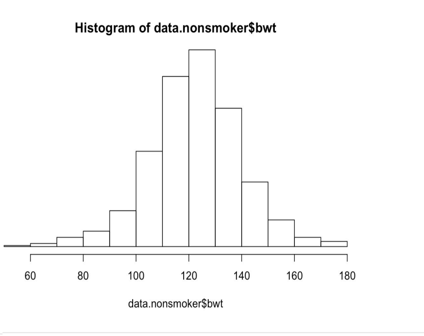
- Dataset that included only smokers (value = 1).
- Dataset that included non-smokers and unknown status (value = 0, value = 9).
- Data set that included only non-smokers (value = 0).

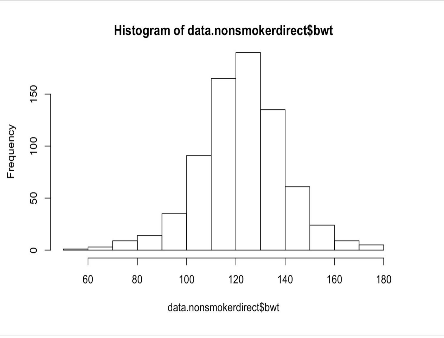
**Analysis:**

1) Three different histograms were modeled from the different datasets to view the density of the datasets. Histograms are important in data visualization, because they incorporate large amounts of data into frequency values. The higher the bar of each bin, the more relatively common that data is in the population. Histograms also directly tell us the shape of the distribution, which is useful for determining the modality of the shape of distribution.

- Histogram of bwt for smoker babies (smoker$bwt)
- Histogram of bwt for non-smoker babies (nonsmokerdirect$bwt)
- Histogram of bwt for non-smoker (including unknown) babies. (nonsmoker$bwt)



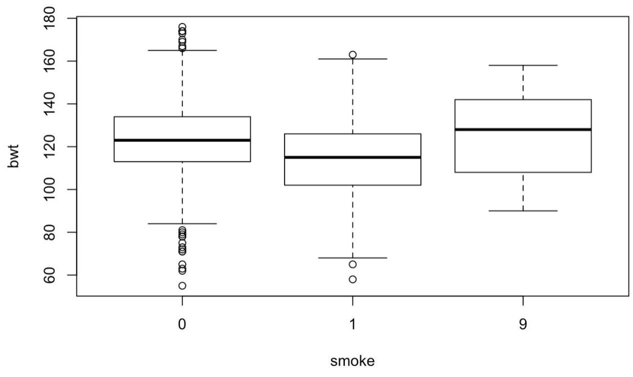




- From the three histograms plotted above, we can see that the average bwt of non-smoker babies would be higher based on the distribution of height of bins in the nonsmoker and nonsmoker direct dataset compared to the smoker dataset. We can also see that there were babies that weighed more than 170 ounces in the non-smoker dataset, while there were no babies in the smoker dataset that weighed more than 170 ounces. In general, we could see that the average weight of babies in the smoker dataset is lower, leading to the possibility that frequency of low birth weight babies are higher in smoker dataset. We can infer that the distribution of the three histograms are unimodal, because the histograms have a single prominent peak. We can also observe that the distributions are not skewed, but are all symmetric because there exists no long tails for the distributions. Also, inferring from the shape and  summary of the three datasets, we can see that there are no unusual outliers that exist in the datasets.

- Shown in Appendix A, we plotted the histogram of the entire dataset that includes smokers, non-smokers, and unknown values in the "smoke" column. From this entire histogram, we can see that the frequency of bwt was the highest around the weight of 120 ounces, which is understandable, because the mean statistic of bwt is 119.6 ounces. The histogram of the entire dataset is also unimodal and symmetric, and no unusual outliers exist.

2)



- A box plot was modeled based on the three different values in the "smoke" column of the entire dataset. The y-axis represents the frequency of bwt in ounces, and the y-axis shows the different values of the smoke column (0, 1, 9). The smoker (x-axis = 0) has many suspected outliers above its upper whisker and lower whiskers. Identifying outliers are important using the box plot because they help us identify extreme skew in the data distribution, identify data collection and entry errors. From the numerical analysis, we can assume that there are many outliers below the lower whisker, because some babies might have been born prematurely earlier than the average gestational period of 40 weeks. The outliers above the upper whisker might have occurred, because some babies can stay in the utero longer than the average gestational period. We can also observe from the box plot that the non-smoker dataset has a lot more outliers than the unknown or smoker dataset. I believe that this is the case, because the number of rows for non-smokers is 752 entries, whilst smokers has 484 entries, and unknown only has 10 entries. The box plot also shows the median, first quartile, and third quartile. The line below the median would be the Q1, line above the median would be the Q3, and the height of the box itself would be the interquartile range (IQR). From the box plot, we could definitely see that the birth weight of smoker babies was smaller than the non-smoker dataset.

- **Non-smoker analysis of Q1, Q3, median:**

- 25% of babies weigh more than 135 ounces by evaluating the Q3 (75% below third quartile)
- 25% of babies weigh less than 115 ounces by evaluating the Q1 (75% below first quartile)
- 50% of babies weigh below or above 122 ounces.

- **Smoker analysis of Q1, Q3, median:**

- 25% of babies weigh more than 125 ounces by evaluating the Q3 (75% below third quartile)
- 25% of babies weigh less than 102 ounces by evaluating the Q1 (75% below first quartile)
- 50% of babies weigh below or above 115 ounces.

**Conclusions:**

- The histogram and box plots definitively show that the birth weight of smoker dataset babies are smaller than the non-smoker dataset. From the box plot, we can see that the distance from the third quartile and lower whisker is 102 oz ~ 70 oz, indicating that the smoker dataset may produce a higher proportion of babies that are low weight compared to the non-smoker dataset. Although we cannot directly tell the difference of low birth weight proportions from the histogram of box plot, the graphs both show that the bwt of the smoker dataset is lower than the non-smoker dataset, indicating the possibility of low birthweight babies for the smoker dataset. Directly comparing the frequency would better help us determine if there are differences between the smoker and non-smoker datasets.

### 2.3 Frequency

**Methods:** We compared between smoker and nonsmoker without irregular to get precise analysis. Firstly,we set up the standard birth weight (stand.weight) as ounce, which is 5.5*16 = 88. If the birth weight is less than the weight, we consider it as low birth weight. Through the step, we found the indices of the low birth weight on each column and counted them. As a result, we got counts of two variables, lbwt(low birth weight) and rbwt(regular birth weight), in each category. In order to compare two categories, we calculated the proportion of the low birth weight babies. Also, it could have noise in the data such as parent's health, and affect our results. Therefore, we subtract 3 from the smoker's low birth weight, and add 3 from the nonsmoker's low birth weight. Also we add the two side t-test to check if our data is reliable. The results are as the following:

- Smoker's low birth weight babies: 36
- NonSmoker's low birth weight babies: 22
- Smoker's low birth weight proportion: 7.4%
- Nonsmoker's low birth weight proportion: 3%
- Manipulated proportion of smoker's low birth weight: 6.8%
- Manipulated proportion of smoker's low birth weight: 3.3%
- H0(Null): Birth weight mean of nonsmoker babies are same with smoker babies
- H1(Alternative): Birth weight mean of nonsmoker babies are not same with smoker babies
- Significance level for t-test: 0.05, 95% confidence interval
- P-value : 2.2e-16

**Analysis:** When we compare the two proportions, the proportion of smoker's low birth weights babies are 4.4% higher than 3. Under our results, we can assume that the smoker has more proportion of low weight babies since there are significant differences between two proportions. When we look at the results of manipulating the few data to avoid bias, there are only 3.5%. There are still significant differences between two proportions. We can consider that our estimates are quite reliable data. If we want to make sure the data is reliable, we can use the hypothesis test by using t-test. To reject our value, there will be no difference between smoker and nonsmoker data. When we use a two-sided t-test, p-value is less than our significance value, 0.05. Thus, we can say two data's distributions are different. Therefore, our results can be occured. Even though we got these results , we cannot conclude that nonsmokers' babies are healthier than smokers' babies, since we can not know the overall distribution by only observing the frequency of low birth weight.

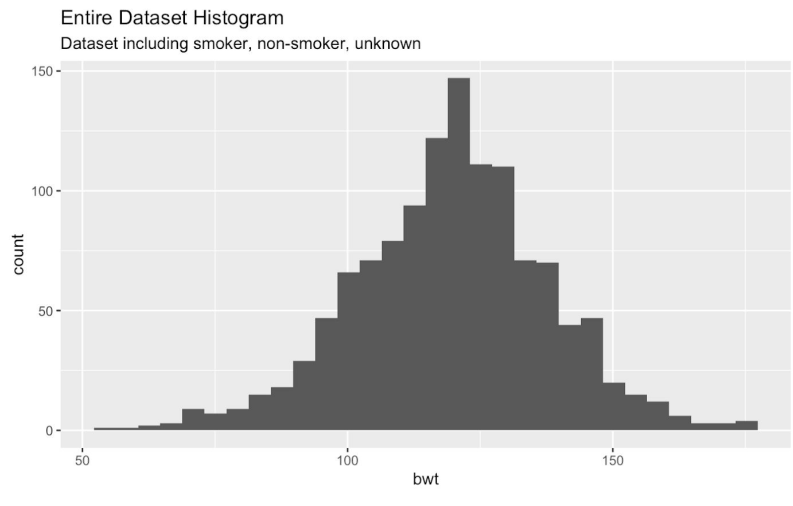
**Conclusions:** In conclusion, the nonsmokers' babies have a lower proportion of the low birth weight than smokers' babies. Thus, we can say smokers' babies have more probability to get lower birth weight.  Based on these results, when we have a specific range, the frequency is a useful statistical method to compare specific values like our experiment. However, this way has some limitations when we want to compare overall data since we cannot know the distribution of our observed data.

## 3. Conclusion(s)/Discussion

- The numerical summary displays the exact statistics and distributions of the birth weight data, but did not allow us to find the exact proportions of low birth weight babies (<88oz.) Visualizations of the data allowed us to easily see the distribution, shape, and helped us infer statistics. Through visualization, we were able to see that the birth weight of the smoker dataset was smaller than the non-smoker dataset, but visualization did not allow us to directly find the exact frequencies of low birth weight babies. On the other hand, calculating the frequencies directly using the smoker and non-smoker dataset helped us compute directly that the dataset of smoker babies had a higher frequency of low birth weight babies. We used hypothesis testing using the t-test and found that our directly computed result is reliable. Although there were limitations on using the dataset, we were able to incorporate different methods to conclude that the dataset of smoker babies had a higher frequency of low birth weight babies.

## 4. Appendix

- Appendix A:



- Histogram plotted using entire babies.txt dataset that includes smokers, non-smokers, and unknown in the "smoke" column. Plotted using ggplot2 in R.