Anthony Ayon
Orafidiya Ooduah

Assignment 1

Introduction

The topic that we're researching the comparison between the birth weights of babies whose mothers smoked while pregnant vs those which mothers who have not. This study is based on the Surgeon General warning that advises pregnant women not to smoke during pregnancy as it may put their child at higher risk for fetal injury, premature birth and low birth weight. This data is a [art of the Child Health and Development Studies collected between 1960 and 1967 nad involved women in the Kaiser Health Plan in the San Francisco region. Our analysis was focused on the birth weights of babies whose mothers had smoked during the pregnancy vs the birthweight of babies whose mothers had not smoked during the pregnancy. We ultimately came to the conclusion that smoking did in fact have an effect on the birth weight of babies but was not very significant in regards of the babies birth weight.

In this report we focused on the possible correlation between child birth deficiencies due to smoking cigarettes while pregnant. We acquired a dataset featuring 23 variables mainly focusing on information from the newborn, mother and father. For this study the main variables we used where whether if the mother smoked or not and the baby's birth weight. Through the use of data modeling we were able to visualize any potential correlations between these factors as well as determine if this sample dataset would be representative of the entire population.

Structure

- **Smoker Data vs Nonsmoker Data Analysis**
- **Formal Checks for Normality**
- **Low Birth Weight Data - Smoker and Nonsmokers**
- **Theory**
- **Additional Analysis**
- **Results/Conclusion**

Smoker Data vs Nonsmoker Data Analysis

In this section we extracted and separated the dataset, "babies", into people who stated they smoked during their pregnancy and people who did not smoke. From there we individually took the mean and standard deviation in each dataset for the recorded birth weight of the baby. From *table 1,* we can see how the average baby birth weight

for smokers was approximately 9 oz lower than the nonsmoker data. On the other hand, both babies do not fall below 88 oz, which is categorized as small. Both birth weights fall within the range considered normal for a newborn baby.

Table 1

	Smoker Data	*Nonsmoker Data*
Sample Mean	**114.1 oz**	**123 oz**
Standard Deviation	**18.098 oz**	**17.398 oz**

In order to get a better visual of the data we decided to do some graphical comparison. As we can see from *fig1* , for both datasets there is clear normality. One factor to point out is the slight skewness to the left of the histogram of nonsmokers, which ties back to our table above, as it reflects the data's concentration of low 120 oz. It is also important to point out that there is 742 samples for nonsmokers while smokers only has 484, which can be visualized in *Fig 2.*

Fig 1

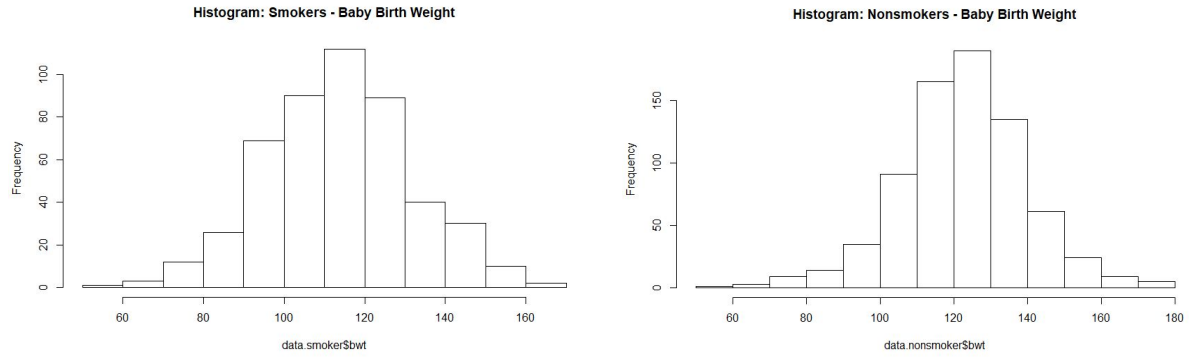
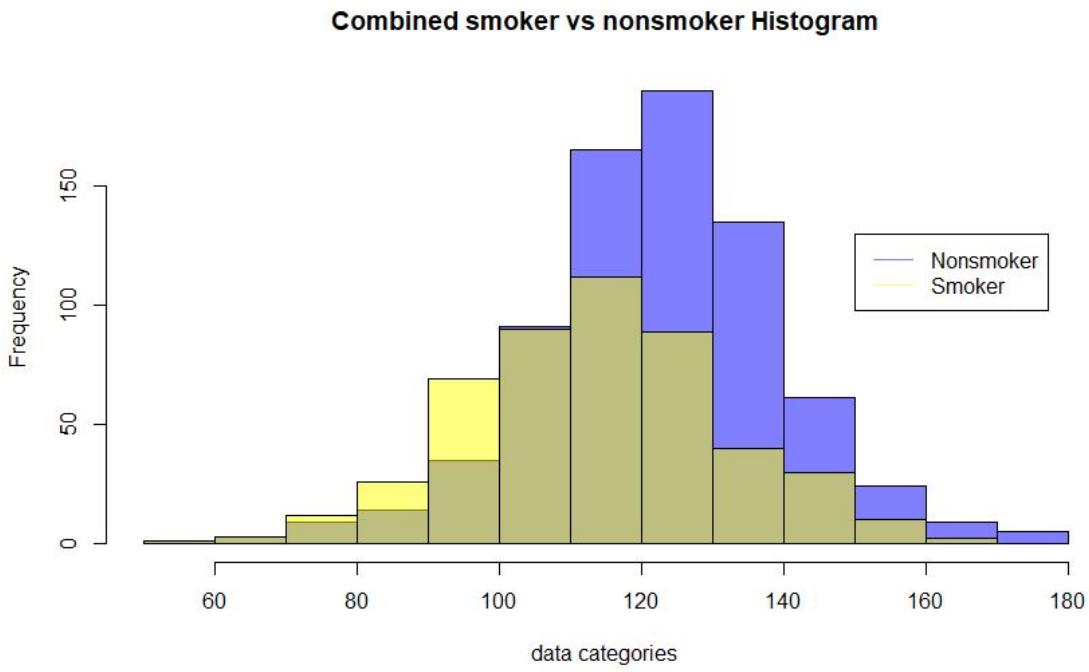


Fig 2:



In order to make direct comparison for both datasets we decided to use a boxplot and QQ-plot. These graphical representation will enable us to see some distinctive features of each dataset.

As we can see from *fig 3* the three boxes represent all smoker, nonsmoker and unknown data, respectively. For this case we will focus on smoker and nonsmoker data. For smokers, we can see the median ranges around 115 oz while for nonsmokers it's about 123 oz, fairly similar to their means. Just as our histogram illustrated, the core 50% of the weights stay relative to a certain range. In this case smokers range around 102 - 126 oz with a min of 58 and max of 163(which can be seen by the dark circles) while nonsmoker ranges around 113 - 134 oz with a min of 55 and max of 176. One distinction worth pointing out is all the dark circles present in the nonsmoker boxplot. This can be attributed the nonsmoker being a larger dataset hence there is more variability among the weights. Also by analyzing closely, with the exception of some circles, the lower quantiles of smokers (the bottom line) still fall under most nonsmoker circles.

Another visual representation similar to a box plot in QQplots. Similar to what we saw on boxplots, we can see a high concentration of points within the ranges specified above as well as the min and max values serving as outliers. As the QQplot shows, the

points do not follow the straight line path, which just demonstrates the disparity between normal distribution (also illustrated in *fig 2*).

Fig 3

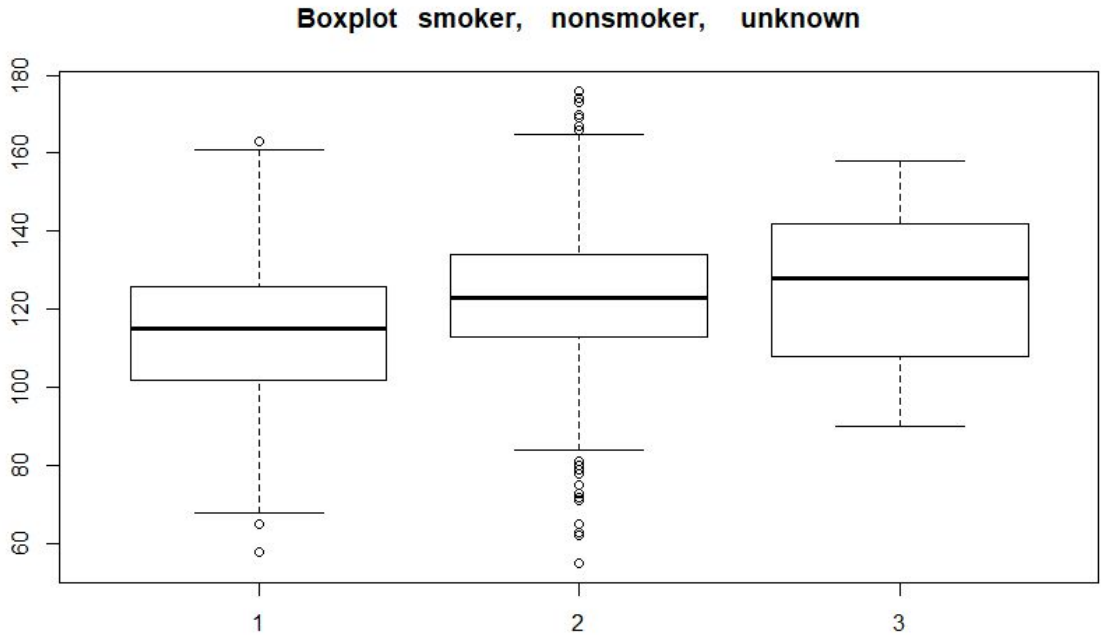
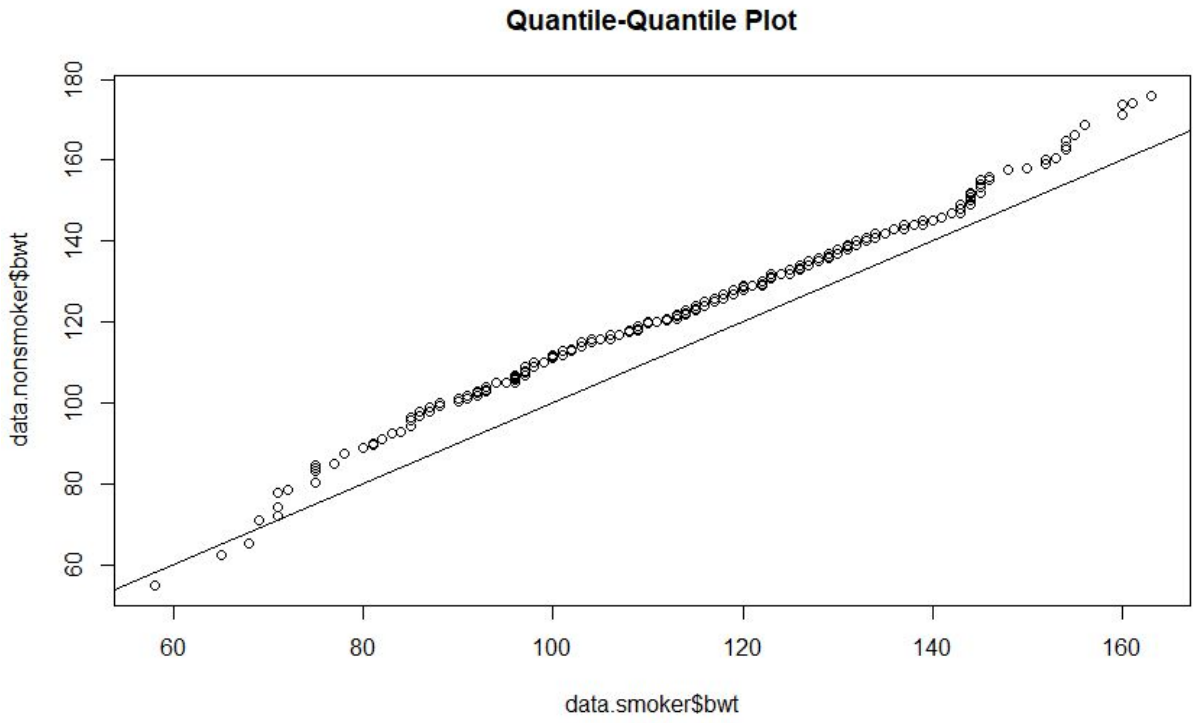


Fig 4



Formal Checks for Normality

After looking at some visual representation of the data and testing it on normality, in this section we will look at a more mathematical approach. Here we will use the Central Limit Theorem, kurtosis, and skewness to check for normality.

First off we took 100 random samples from each smoker and nonsmoker data and plotted it into a histogram. We used the Central Limit Theorem to make a histogram of those samples and just as *fig 5* demonstrates the data looks very similar to both's population histograms. As the samples increase the data becomes more normal.

We then looked into Kurtosis and Skewness, results are displayed in *table 2*. For Kurtosis if we run a series of repetitions for each length weight of both datasets, we get an average of 3. Meaning both smoker and nonsmoker Kurtosis results should approximate that value. For smokers is fairly close but not so much for nonsmokers which also justifies why the skewness is farther away from zero. Meaning nonsmoker data is somewhat skewed left. This was an observation pointed out in section one but now we have mathematical proof of it.

Fig 5

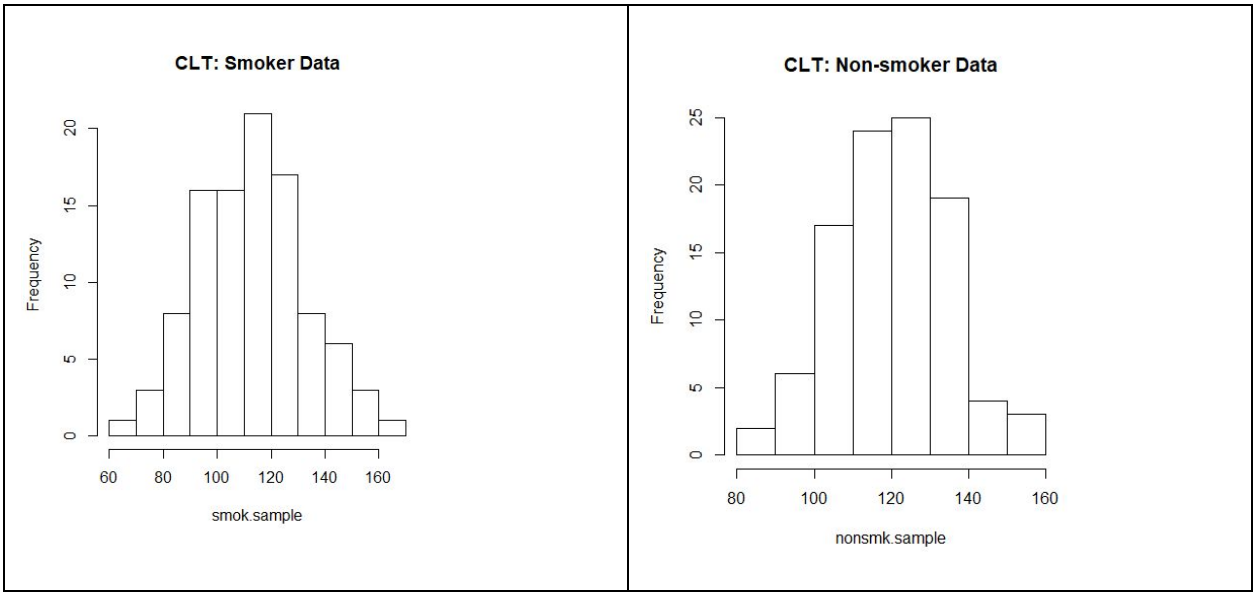


Table 2:

	Smokers	Nonsmokers
Kurtosis	2.988	4.037
Skewness	-0.0335	-0.186

Low Birth Weight Data - Smoker and Nonsmokers

For this section we extracted the babies categorized as low weight for both smokers and non-smokers. We then again acquired the specific means of each one and their standard deviation(see *table 3)*. We then plotted a histogram and QQplot for each individual dataset.

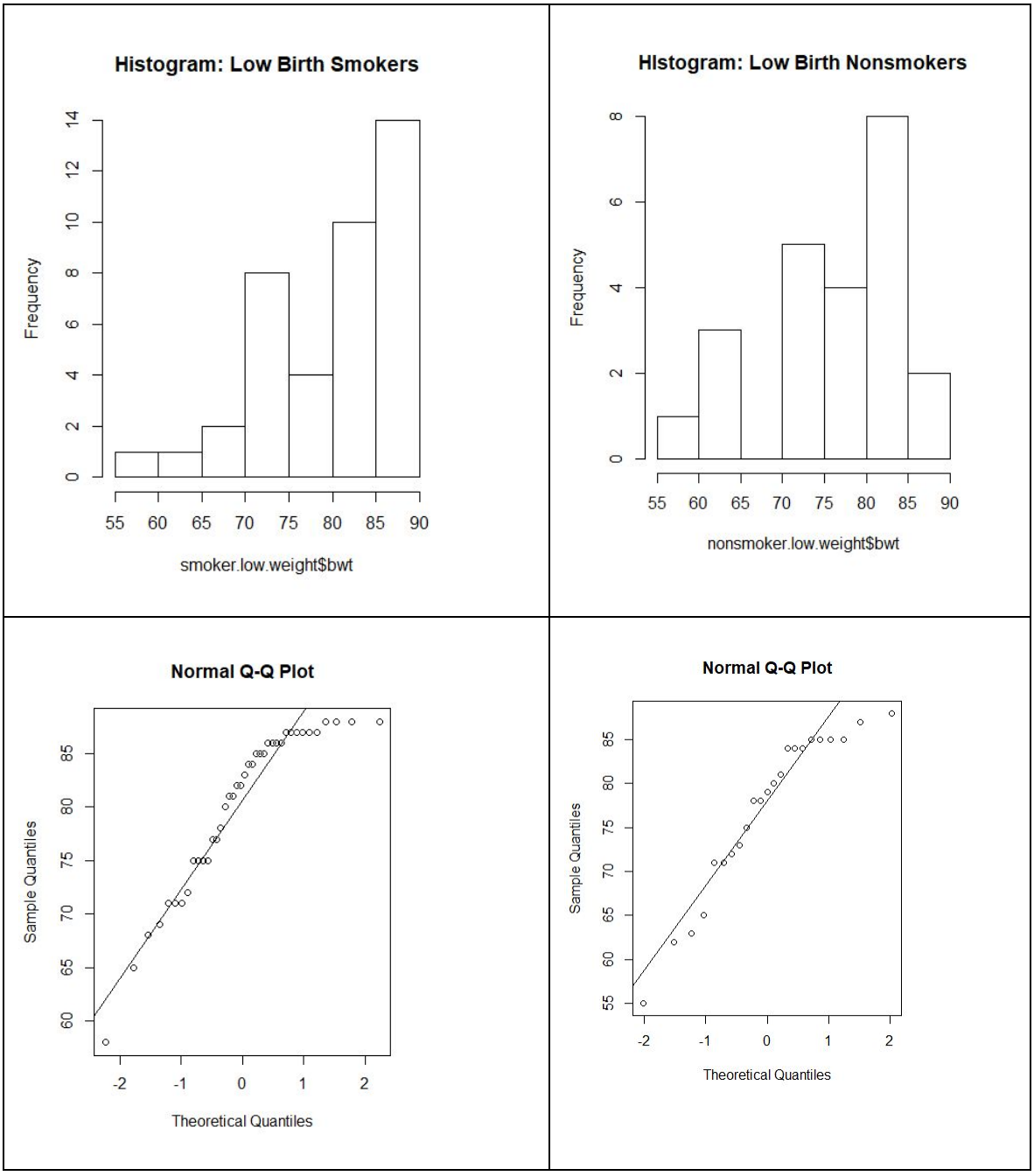
As illustrated in the histograms, the data seems to no longer be normally distributed and heavily skewed right. On one hand we do have more than double babies categorized as low weight for smokers with with a total 0f 40 weighing 88 oz or less. Nonetheless, the mean of nonsmokers is less than smokers, meaning we have more newborns at lower weights as the 50% quantile concentrated within the range of 70 -85oz.

When proportions for babies categorized as low weight are compared, one can see the biggest change among the smoker dataset, while the nonsmoker data stays relatively the same. As *table 3* shows, we have the biggest change in proportion present between 88 and 85 oz. Hence after these observations we can state that findings seem reliable considering the reflection on both population and low birth data. In the case of adding more or taking less data, it is highly unlikely there will be much change to what is currently present. Currently there is more data for nonsmokers which leads to set range among the most common weights. Thus, even if nonsmoking has more outliers, the difference in weight will not be affected by the addition of more data. Furthermore, the smoking data may also stay relatively the same considering most data weight values stayed above the acceptable birth weight. Meaning we can potentially see an increase in average if more smoking data is to be recorded.

Table 3

	Smoker	Nonsmokers
Mean	80.17 oz	76.96 oz
Standard Deviation	7.547 oz	9.097 oz
Proportion (<= 88 oz)	0.0826	0.031
Proportion (<= 85 oz)	0.0537	0.0283
Proportion (<= 90 oz)	0.0867	0.0363

Fig 6


Histogram: Low Birth Smokers
Frequency
0
2
4
6
8
10
12
14
55
60
65
70
75
80
85
90
smoker.low.weight$bwt
Histogram: Low Birth Nonsmokers
Frequency
0
2
4
6
8
55
60
65
70
75
80
85
90
nonsmoker.low.weight$bwt
Normal Q-Q Plot
Sample Quantiles
60
65
70
75
80
85
-2
-1
0
1
2
Theoretical Quantiles
Normal Q-Q Plot
Sample Quantiles
55
60
65
70
75
80
85
-2
-1
0
1
2
Theoretical Quantiles

Theory

Histograms

We utilized histograms to better understand the distribution of the data as it could be skewed left, right or follow a standard normal distribution. This can also visually tell us where the outliers in our data lie.

Standard Normal Distribution

The standard normal distribution has 3 properties; its graph is bell shaped, its mean is 0 and its standard deviation is 1. This distribution follows the 68-95-99.7 rule which details that 68% of the area under the curve is within 1 unit, or standard deviation, of the center, 95% of the area under the curve is within 2 units of its center. We can use this rule in order to check the normality of a distribution of data. This provides us a way to summarize data.

Quantile Plots

We use quantile plot in order to compare the distribution of our data against a theoretical distribution such as the normal or exponential distributions. Any deviation from the line indicate differences between the data and the targeted distribution. If the distributions are identical the plot should have an intercept of 0 with a slope of one. A nonzero intercept indicates a shift distributions and a non unit slope indicates a scale change.

Boxplots

The box plot displays the distribution of the data. It is separtaed into first quartile, median, third quartile, and maximum with each approximately being 25%.

Central Limit Theorem

Let X1, X2, . . . , Xn be an i.i.d. random sample from a distribution with mean μ and standard deviation σ.

Kurtosis

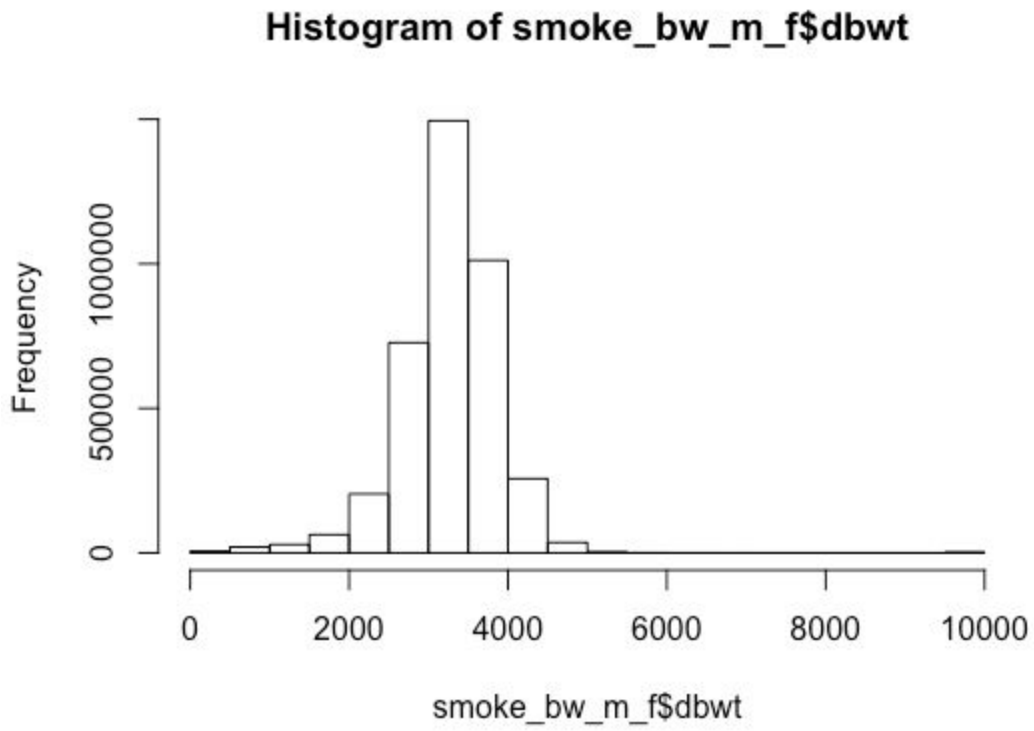
Kurtosis is the average of the fourth power of the standardized data. It is calculated by the formula,

Skewness

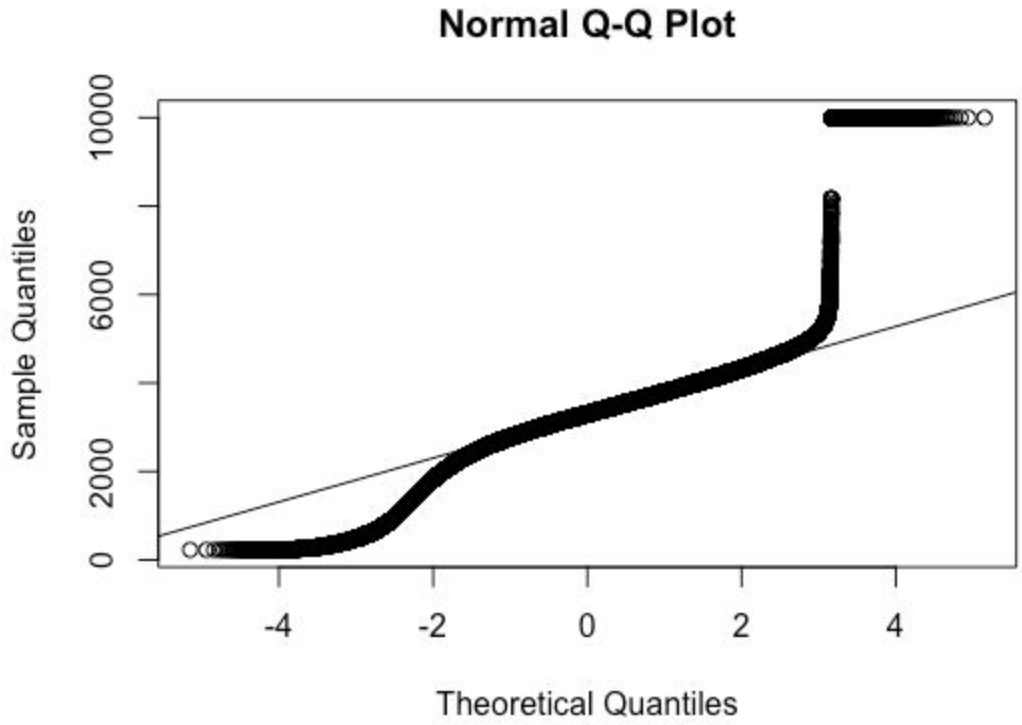
Skewness is the average of the third power of the standardized data. Skewness is a way to check for normality

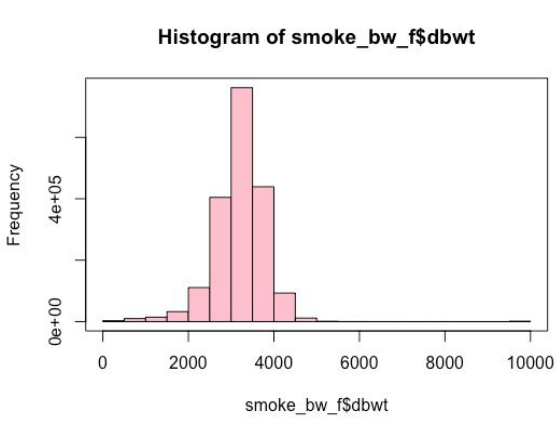
Additional Analysis

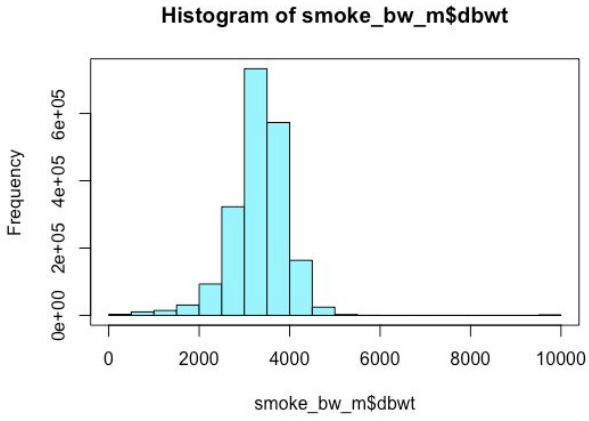
For our extra analysis we used the NCHS Vital Statistics Natality Birth data from 2017. Our goal was to figure out if smoking mothers babies were affected differently based on the gender of the baby.

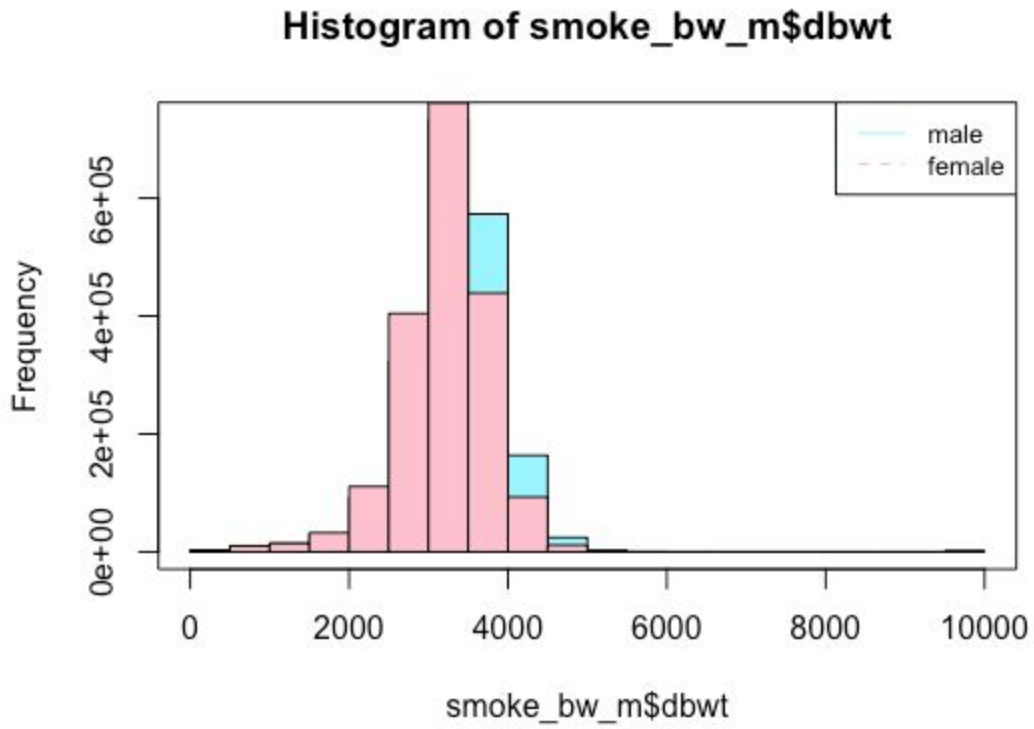


We first plotted of babies with smoking mothers in terms of grams. We found that the average weight for both male and female babies with smoking mothers was approximately 3267 with a standard deviation of 621.


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
0
2000
6000
10000
-4
-2
0
2
4


Histogram of smoke_bw_f$dbwt
Frequency
0e+00
4e+05
0
2000
4000
6000
8000
10000
smoke_bw_f$dbwt


Histogram of smoke_bw_m$dbwt
Frequency
0e+00
2e+05
4e+05
6e+05
0
2000
4000
6000
8000
10000
smoke_bw_m$dbwt



We then compared the histograms for the weight of male and female babies born to smoking mothers. Overall there wasn't a significant difference as male babies on average weighed about 3323 grams with a standard deviation of 633 and female babies on average weighed about 3208 grams with a standard deviation of 603. In conclusion it is recommended that women do not smoke during pregnancy but if they end up doing so, it won't have a significant effect dependant on the gender of their baby.

Results/Conclusion

We broke our analysis down into 3 types of observations ; numerical, graphical and incidence. From our numerical methods we found that the averages were approximately 9 ounces apart were the babies with smoking mothers weight averaged out to 122.7776 ounce while babies with nonsmokers birth weight averaged at 114.1095.

When running our graphical analysis we used found the averages reflected in the difference between each histogram. The weight histogram for non-smokers seems slightly skewed left which indicates that it is less likely for the baby to be low weight. We used box plots as well which was the most clear in displaying the difference between the effects between babies with smoking mothers and babies with non smoking mothers. The Normal QQ plots showed that these two samples of weight from smokers and nonsmokers are fairly normally distributed, hence, validating any comparisons among the two.

Another aspect we took into account was applying the central limit theorem into both of our weight datasets. As we ran each approximation we plotted each result from smoker ("CLT: smoker data") and nonsmokers ("CLT: Nonsmoker Data") into a histogram. The results were surprising as the histograms seemed relatively skewed left with smokers having a closer bell shaped curve.

In all after taking a close look at the data and modeling results, we have concluded that regardless if more samples where added to the overall dataset, the outcome would still be the same. This in part, because the rate at which newborns with data recorded as smoked had a substantial and clear difference over non smokers. As portrayed in the boxplot model, there is a clear distinction especially in newborns under the smokers that tend to overall average less weight than nonsmokers. On the other hand it was surprising to see that the data for low birth values portrayed a higher weight average for smokers than nonsmokers. This may be attributed to the nonsmoker data recordings being substantially larger than smoking data. Hence, newborn's low weights are more vulnerable to be affected by external factors unrelated to smoking. In all we cannot reject the hypothesis that smoking affects baby birth weights.