Maternal Smoking and Infant Health Development

Yong Liu, Yunlin Tang, Jian Jiao

MATH 189, Spring 2020, HW1

I. Introduction

Epidemiological study indicates that babies who have low birth weight and premature birth will be associated with lower survival rates, where the babies' maturity is measured by their birth weight and gestational age. However, the magazine New York Times on March 5, 1995, had stated: "Low Weights not solely to blame Surgeon General warning 'Smoking by pregnant women may result in fetal injury, premature birth, and low birth weight.'" In order to evaluate the reliability of the surgeon general's warning, the primary method is to compare the birth weights of babies born to smokers and to non-smokers. Specifically, the proposed question derived from this investigation is: "What is the difference in weight between babies born to mothers who smoked during pregnancy and those who did not? Is this difference important to the health and development of the baby (aka the death rate)?" The data used in this analysis was cited from the Child Health and Development Studies program, which collected the pregnancy records that occurred between 1960 and 1967 among women in the Kaiser Health Plan of the San Francisco region. It contains 1236 observations with 7 variables: birth weight in ounces, length of pregnancy in days(gestational age), parity, mother's age, mother's height, mother's prepregnancy weight, and smoke status of the mother. We found there is a tendency of decrement in babies' birth weight if mother smoked during gestation and babies born by mothers who smoked are likely to encounter early born and thus have a lower survival rate**.** In this report, we will introduce several methods (such as the numerical, graphical summary, and incidence statistics) and their analysis after performing the calculation in R in order to answer the investigating question.

II. Analysis

A. Numerical Summary

Method: We first divide the data into two groups: smoking and non-smoking. Then we analyze these two distributions of baby birth weight using statistics including average, standard deviation, quartiles, and kurtosis.

Analysis: After performing the respective calculation in R, we have obtained several meaningful numerical statistics from these two distributions. First, since average represents the center of data distribution. We use the mean of data to show that: the mean of smoking is 114.1, and the mean of non-smoking is 123.0. The difference in the mean is 8.9 ounces. Then we calculate the lower quartiles (25% of the distribution) for each group to show the lower side in the data. We obtain that LQ for the smoking group of 102 and LQ for the non-smoking group is 113; the difference between LQ from two groups is 11. After showing the lower quartiles, we calculate the variance to investigate the spread of two distributions since the variance shows how individual data varies from the center. The variance of the smoking group is 327.57 and the non-smoking group is 302.71. Finally, using the kurtosis measure can reveal how heavily the tails of distribution differ

from the tails of a normal distribution. The kurtosis of the smoking group is 3.0 and the non-smoking group is 4.0.

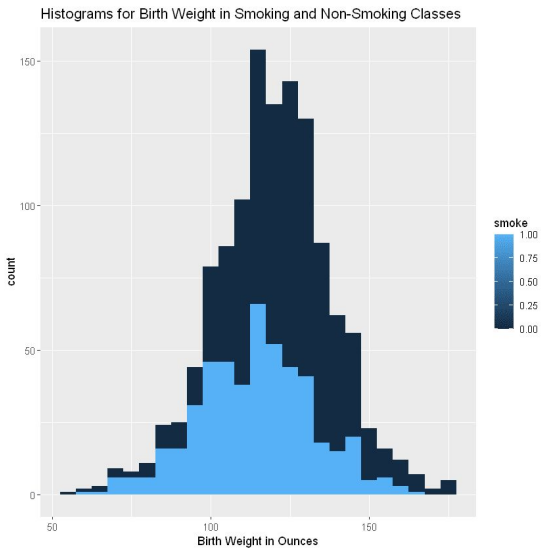
Conclusion: Our result shows that there is a difference of 8.9 between the mean birth weight of babies of the smoked and mothers who did not smoke. The Lower Quartile of these two categories shows there is an 11 ounces difference between the birth weight of babies of the smoked and mothers who did not smoke. These two differences show the tendency of decrement of babies' birth weight if the mother smoked. There is not much difference between the Variance of two distributions, which means that both are in a similar degree of distribution. The value of Kurtosis shows that Non-Smoked distribution is more centralized than Smoked, which means that smoked includes more tailed values (i.e. extreme values).

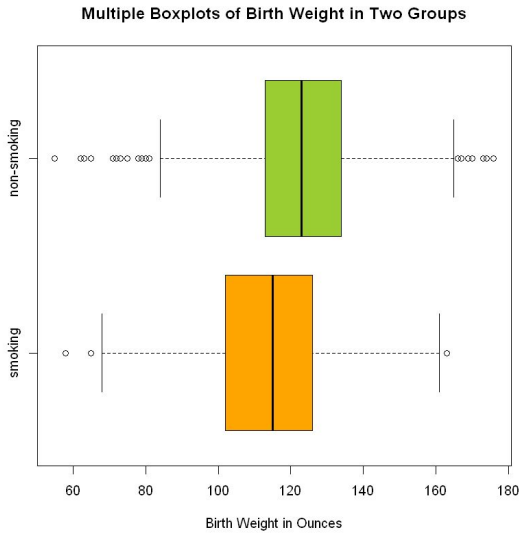
B. Graphical Summary

Method: In this step, we use several graphical methods to compare the two distributions of birth weight to represent the findings in the previous method. The data is also separated into two groups: the smoking group and the non-smoking group, which respectively represents the group of mothers who smoked during pregnancy and the group who did not. Three different types of graphs are generated: histogram, box plot, and quantile-quantile which enhance our understanding of the distributions of two groups and help us furthermore analyze the basic statistics of the data visually.

Analysis:

First, we produced the histogram of baby birth weight for each class. The histogram of birth weight in the smoking group (see Appendix 1) is unimodal which has a single prominent peak in the bin [110, 120). The shape of this histogram tends to be symmetric thus we suspect that the mean and median in this distribution will be roughly the same. Furthermore, this histogram also indicates that there are few outliers, such as light/heavy birth weight of babies. The histogram of baby birth weight in the non-smoking group (see Appendix 2) is also symmetric and unimodal. Based on the shape of this histogram graph, we suppose that the mean and median in this distribution are also roughly the same. However, the height of the 7th bin(left of the peak) is close to the peak and higher than the 9th bin(right of the peak). Thus we suspect that the mode would be lower than mean and median. In the combined histogram graph, the lighter blue histogram represents the distribution of baby birth weight in a smoked group (where 'smoke' variable equals to 1), and the darker blue represents the non-smoked group ('smoke' equals to 0). This combined graph shows that the distribution of the two classes is roughly similar. They are centered between 100 to 125 ounces. We can see that both histograms roughly have the symmetric shape and unimodal distribution without extreme long tails. Thus we suspect that the two groups will have a similar distribution and they are normally distributed.

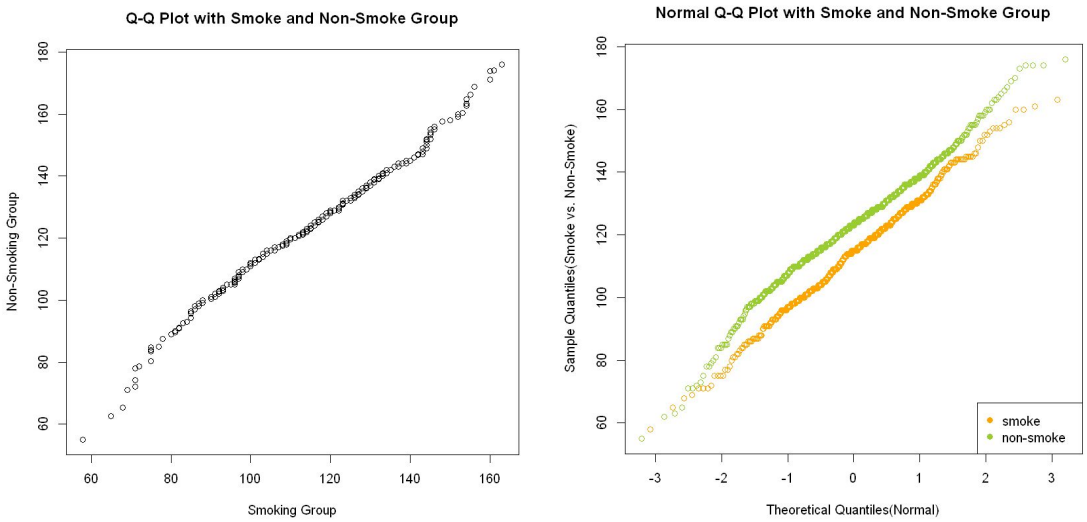




Second, we produce the boxplot for each group as well. Since the box plot can better visualize the distribution of the sample by indicating the "five-number summary" (lower whisker, Q1, median, Q3, and upper whisker), we choose also draw out the box plot for these two groups in order to further understand our data and the difference between the birth weight of two groups. On the bottom of the graph, we have the boxplot for the birth weight variable in the smoking group (orange color). In this group, we see that the lower whisker of birth weight will be around 70 ounces, and the upper whisker weight will be around 160 ounces. The Q1 and Q3 will roughly be 100 and 125 ounces respectively. Also, the median seems to be between 110 to 120, which can be confirmed by the calculation performed in step 1 by showing that the median is 115. On the top, we have the box plot for the birth weight in the non-smoking group. Based on this graph, we estimate the lower whisker will be around 80 and the upper whisker will be 162; also, the Q1, median and Q3 interval will be 110, 120, and 135 respectively. From this boxplot graph, we can compare the two distributions for their basic statistics. The shape of two boxplots for both groups have a similar pattern in which the interquartile ranges are evenly split by the upper and lower whisker. This observation indicates they possibly have the symmetric shapes of the distribution (we have also shown that both groups have symmetric distribution in the histogram step). However, the median of birth weight in the non-smoking group is 8 ounces(0.5 lbs) higher than the median of the smoking group. The interquartile range for the smoking group is larger than the non-smoking group, which indicates that the smoking group has greater variability and spread than the non-smoking group. There are more outliers in the non-smoking group compared to the smoking group. This difference in outliers may be caused by missingness of birth weight data and other potential confounders reasons.

Third, we also produce a quantile-quantile plot for each group. In this step, we will use a quantile-quantile plot to examine the similarity of birth weight distribution between the smoking and non-smoking group. By plotting the qqplot and qqnorm in R, we can visually inspect the properties of distributions in order to further develop our estimation and analysis to answer the research question. The following qqplot compares the data

distribution of birth weight variables between the smoking group and non-smoking group by pairing their respective sample quantiles. We see that the plot is fairly linear with some minor departures from a straight line. This observation indicates that the two distributions from 2 different groups will roughly have the same shape but different means or standard deviation, which confirms our analysis in the above histogram and box plot section. To further inspect whether these groups of data are normally distributed, we also create two normal quantile-quantile plots for each group. To create the combined normal quantile-quantile plot for both groups in R, we first separately draw out the normal q-q plot for each group, then overlay these two plots into one graph. In such a way, we can visually identify the normality of these distributions. The theoretical quantiles on the x-axis are the quantiles from the standard normal distribution with mean 0 and standard deviation 1. The y-axis includes the birth weight data from each group and plots them separately. From this graph, we can see that the points from two classes respectively form a straight line roughly. Therefore, it would be a fair assumption that the birth weight data from two groups (smoke and non-smoke) are normally distributed within their group.

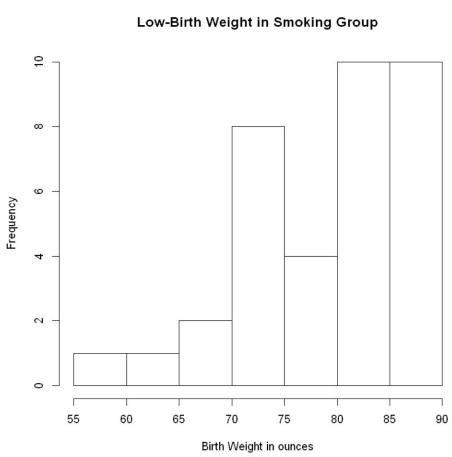


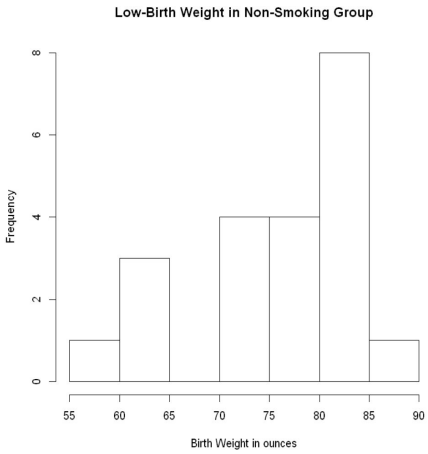
Conclusion: After plotting the histograms, box plots, and quantile-quantile plots of birth weight data for both groups, we have enhanced our knowledge toward the dataset. In conclusion, we found that the distributions of birth weight in two groups both have symmetric and unimodal shape, and these two groups of data which were extracted from the same dataset would both potentially have a normal distribution. By comparing their plots visually, we also found that the quartiles of birth weights in the non-smoking group are roughly higher than the smoking group, and the data from the latter group is more spread than the former.

C. Incidence Comparison

Method: In this step , we will explore the difference of low-weight-baby frequency between smoked group and non-smoked group; The baby who weighs under 5.5 pounds which is 88 ounce is considered as low-weight.

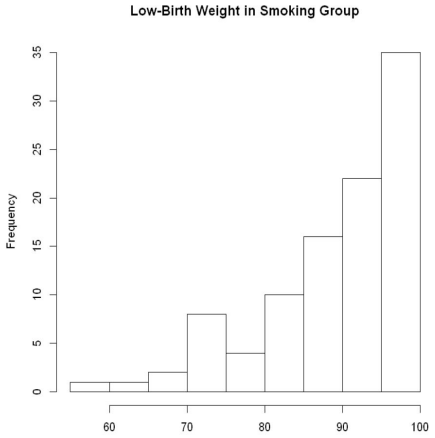
Analysis:

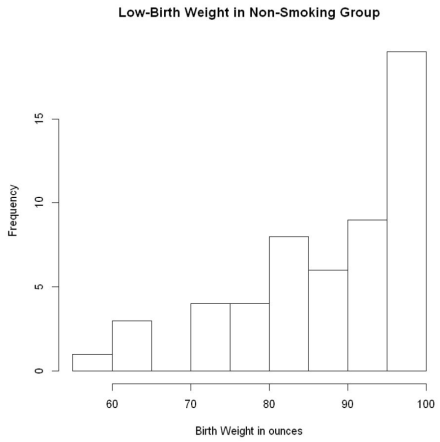




Based on the histograms, we can see that the Low-Birth Weight in Smoking Group is roughly A Left skewed distribution, however the Non-smoking group tends not to be a left skewed distribution. We estimate that the reason behind this is because the Smoking group has more low-birth-weight babies, which means more data to show in the graph. Additionally, the sum of bar areas under Smoking-graph is larger than those of Non-smoking. To test whether our estimate is reliable we will change the upper boundary of low-birth-weight.

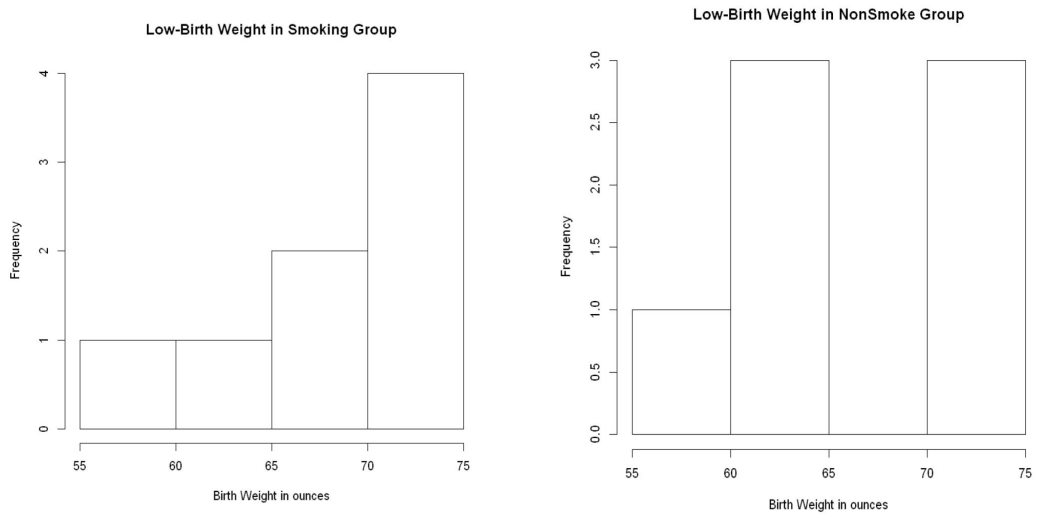
First, we try increase the boundary of low-birth weight to 100 ounces, which means there will be more low-weight-baby





In this situation, both graphs have a left-skewed graph. Non-smoking graphs also have a left-skew because as we increase the boundary of low-weights , there is more data collected. Overall, the smoking group still has a larger sum of bar area.

What if we decrease the boundary of low-birth weight to 75 ounces, which means there will be less low-weight-babies



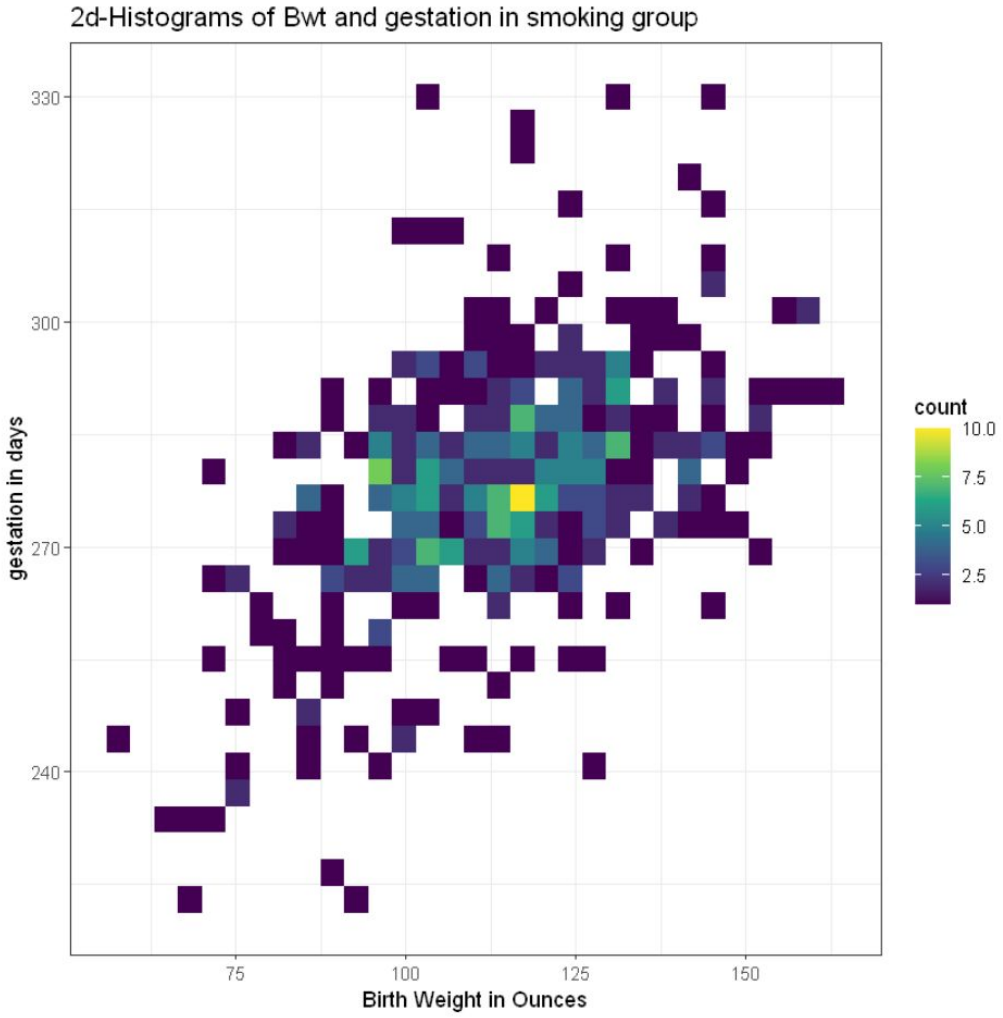
In this situation, as you can see from the graph, the smoking group has a better left skewed distribution than those of in Non-smoking group. Since lower boundary means less data collected, smoking groups still have more babies who weigh under 75 ounces. Therefore, Smoking groups tends to have a left-skewed distribution.Overall, the smoking group still has a larger sum of bar area even

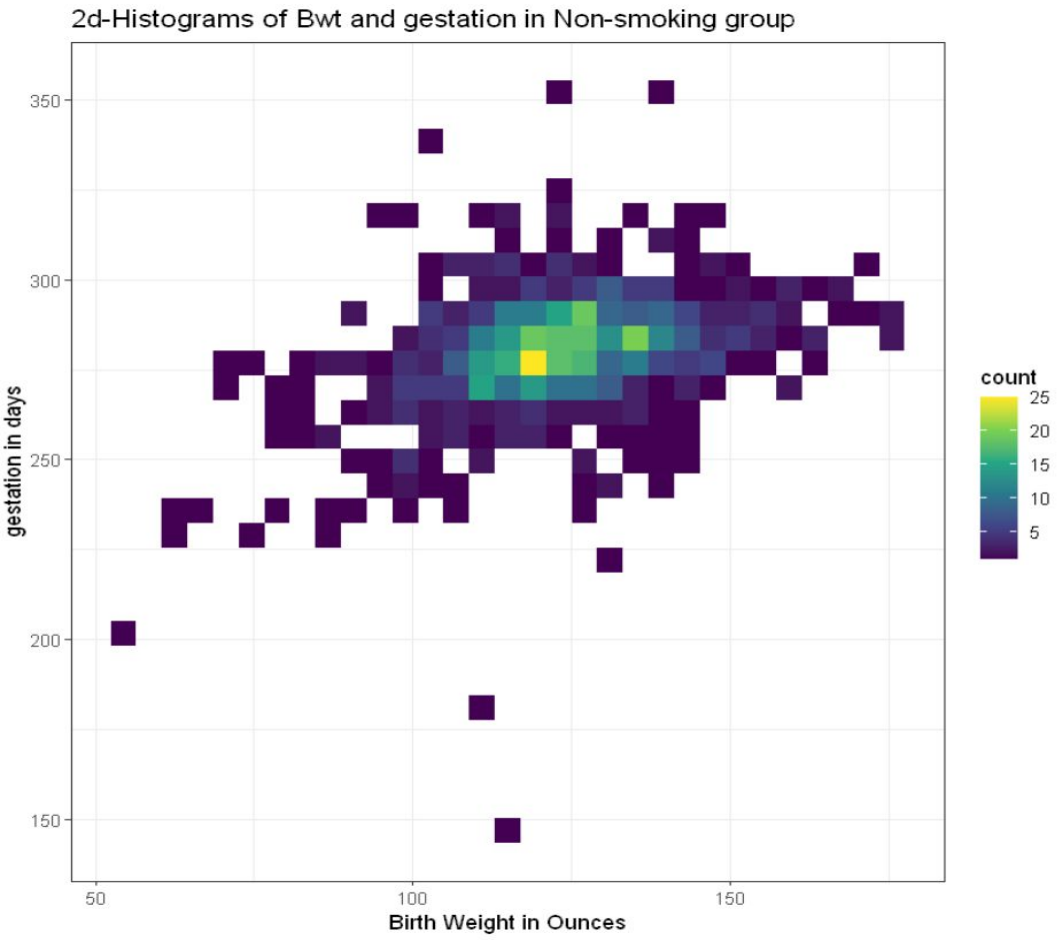
Conclusion:

Low-weight-babies in the smoking group have a better left skewed distribution than those in Non-smoking group because the smoking group has more data to be collected if we set a low-weight boundary. Therefore, the smoking group contains more low-weight babies than the Non-smoking group.

D. 2D-advanced Analysis:

After we analyze the difference in bwt between smoking and non-smoking groups, we are still not able to answer the question that how does this difference relate to health of the baby. Survival rates are a good feature to explore here. Since, the babies who are both low-weighted and early-born tend to have lower survival rate, we will explore how the relationship of baby weights and gestation differ in Smoking group and Non-smoke.





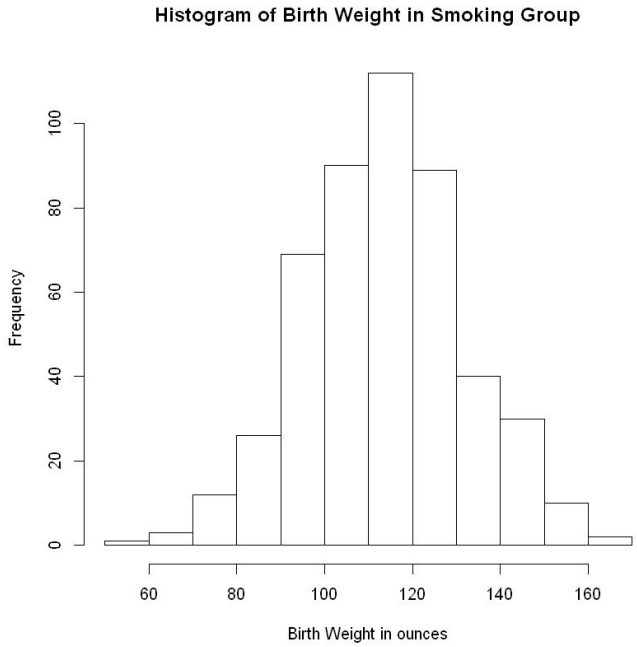
As we can see from the 2d histogram graph, it is obvious that the smoking group contains more babies that are both low-weight and early born. Since babies that are both low-weight and early born have lower survival rates, the babies born by smoking mothers are more likely to have lower survival rates than those of non-smoking.

III. Importance and Conclusion

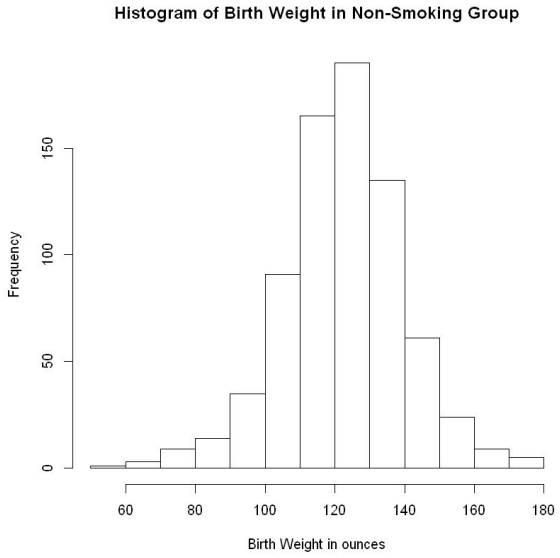
We answered the research question of finding differences in weight between babies born to mothers who smoked during pregnancy and those who did not in three ways (numerical, graphical, incidence). We have found that there is a tendency of decrement in babies' birth weight if mother smoked during gestation. Besides that, we have also researched on birth weight difference importance to the health and development of the baby (aka the death rate) by plotting data in 2-d histograms and we found that babies borned by smoking mother are more likely to have borned early, have lower birth weight and thus could have lower survival rates than those of non-smoking.

IV. Appendix

1. Histogram of Birth Weight in Smoking Group



2. Histogram of Birth Weight in Non-Smoking Group



V. Contribution Statement

Jian worked on Numerical Summary, Conclusion, setting the composition and proofreading.

Yunlin worked on Introduction, Graphical Summary, integrating analysis and code parts into consistent structure and format.

Yong offered the framework of this project and worked on incidence analysis(section C) and advanced analysis (section D)