MATH 189 HW1

Introduction

This study conducts an exploratory data analysis of the weight of babies born to mothers who smoked during pregnancy and those who did not. The aim of this analysis is to uncover any difference in weight between these two groups, and the role this difference plays in the health of the baby. This topic arises from a 1995 study conducted by the NYTimes, known for its unexpected finding that "ounce for ounce babies of smokers did not have a higher death rate than the babies of nonsmokers." Regardless, the article suggests that "smoking by pregnant women may result in fetal injury, premature birth and low birth weight." Furthermore, epidemiological studies have indicated that "smoking is responsible for 150g (5.3oz) of reduction in birth weight, with smoking mothers being twice as likely as non smoking mothers to have a low-birth-weight baby (under 2500g or 88oz)."

This study uses an expanded dataset from the Child Health and Development Studies (CHDS) which consists of all pregnancies that occurred between 1960 and 1967 among women enrolled in the Kaiser Health Plan and delivered their babies in Northern California. They also obtained prenatal care in the San Francisco area. Given the findings of previous studies and the expanded dataset, this study's analysis is hypothesized to reveal a trend signalling that non smoking mothers have low-birth-weight babies more often than smoking mothers on average.

However, there appear to be significant issues underpinning the data. Chiefly, the sample from which this data is collected is not representative of the entire population of mothers. The dataset only contains mothers from California who have higher than average income, education, healthcare, and prenatal care. It is likely that these variables act as confounders in the study which mask the effect of smoking on the health of the babies. To attempt to combat this effect, the study groups babies by their gestational age.

MATH 189 HW1

Weights of babies born from mothers who smoked during pregnancy will be contrasted with healthy weight ranges for babies born in their respective gestational age range.

The findings from this study conclude that birth weight is correlated with the mother's smoking status during pregnancy. This conclusion can be drawn for the following reasons:

- There exists a significant difference among the centrality (mean) between the two populations (smoking mothers and nonsmoking mothers).
- While the study does not conclude that babies belonging to mothers who smoke are consistently found to be underweight, the results indicate a correlation between mothers who smoke and low birth weight. This is supported by a 5% difference in the number of underweight between the two populations.

Body

Data

As mentioned above, this study uses an expanded dataset from the Child Health and Development Studies (CHDS) which consists of all pregnancies that occurred between 1960 and 1967 among women enrolled in the Kaiser Health Plan and delivered their babies in Northern California who also obtained prenatal care in the San Francisco area. This study is composed of 1236 single-birth male babies, all of whom lived at least 28 days. At birth, the measurements obtained are length, weight and head circumference of the baby. This analysis makes use of birth weight, gestational age, and smoking status. We classify birth weight as a continuous numerical variable and gestational age as a discrete numerical variable. Smoking status is labelled as a continuous variable for this study.

MATH 189 HW1

Methods

In order to make inferences about the data and give a well-informed response to the questions proposed in the write-up, several analyses were performed in this study.

1) To numerically understand the distribution of birth weight for babies born to women who smoked during their pregnancy and did not smoke during their pregnancy, the mean and standard deviation of each population were computed. These metrics allow a better understanding of the central tendency and spread of the data. An overlapping histogram of the two distributions was generated to visualize these metrics.
2) To graphically present the distribution of the weight of the two populations a box plot was created. The library *ggplot2* enhanced the quality of the graph with options to customize the color and width of the graph. To further understand the distribution of outliers seen in the box-plots, quantile-quantile plots were generated to visualize their location and concentration in the data.
3) Finally, in order to compare the frequency of low-birth-weight babies for the two populations this study attempts to identify underweight babies (under 88 oz) by their gestational age. The proportion of babies under this barrier was calculated for each population and contrasted with the proportion of babies in the healthy weight ranges to arrive at a conclusion.

Analysis

To begin exploring the data, summary statistics were calculated in Fig 1a. for the two populations consisting of babies born to mothers who smoked and did not smoke during pregnancy.

	Mean weight (oz)	Standard deviation (oz)
Non-smokers	123.09	17.42
Smokers	113.82	18.29

Fig 1a. Summary Statistics

It appears that smokers who have babies are 10oz lighter than babies born to non-smokers. Both distributions appear to have a similar spread. However, these statistics alone are not enough to understand the distribution of data. An overlapping histogram was generated in Fig 1b. to visualize the two distributions.

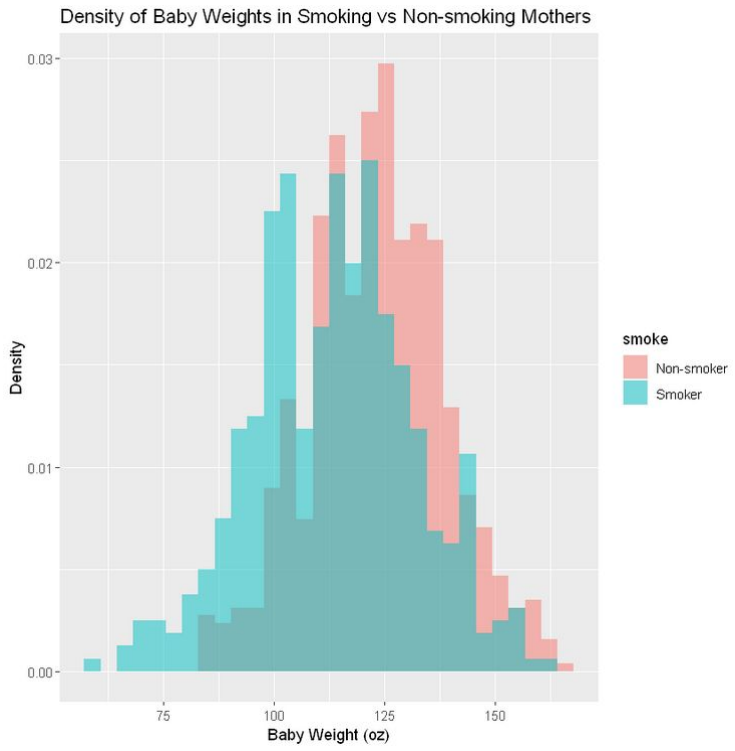

Figure 1b. Overlapping Histogram

Both distributions appear to be roughly normal, with non-smokers having a slight right skew and smokers having a slight left skew. This histogram helps visualize the difference in mean weight of the two populations. More babies appear to be in the unhealthy range (< 88oz) in the distribution of smokers.

To ensure only valid data is used, observations with an unknown smoking status were filtered from the dataset. Across each continuous and numerical variable from the dataset, outliers were identified and removed. The resulting box-plots in Fig 2a, 2b, and 2c. visualize the wrangled data.

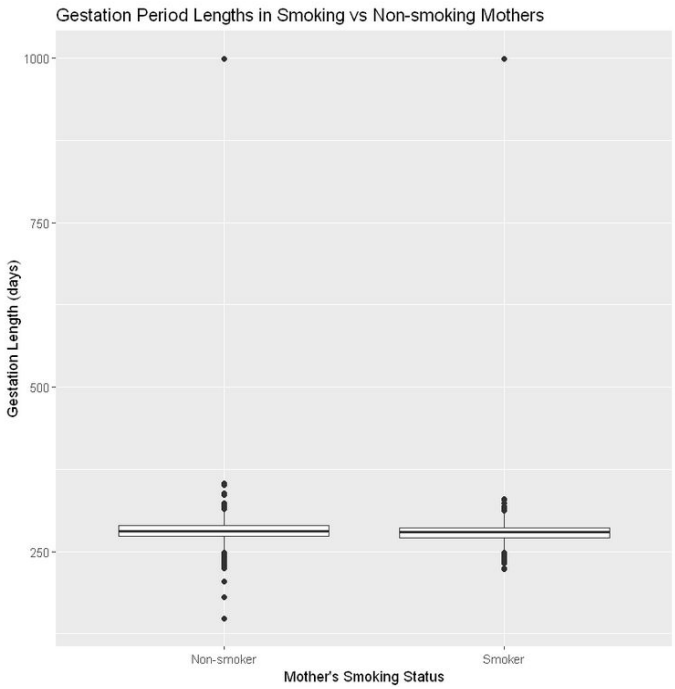

Fig 2a. Extreme outliers seen with gestation lengths above 300 days.

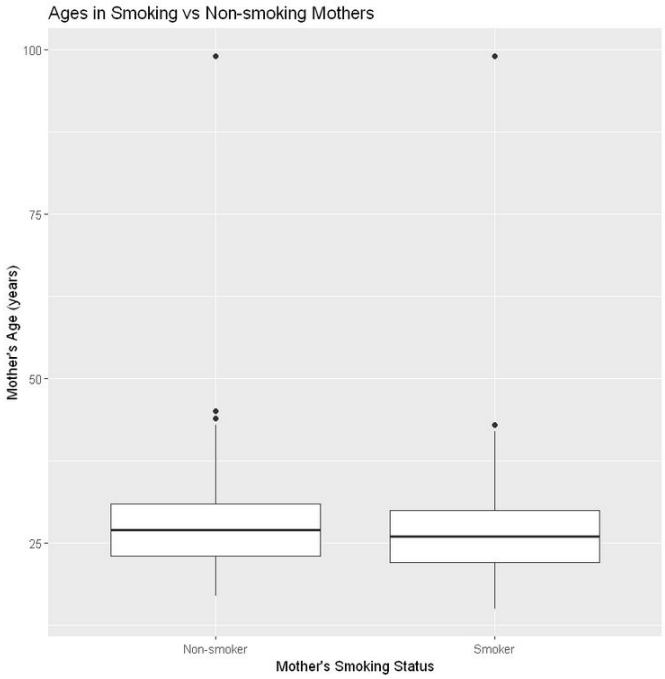

Fig 2b. Extreme outliers are observed with mothers above 45 years old.

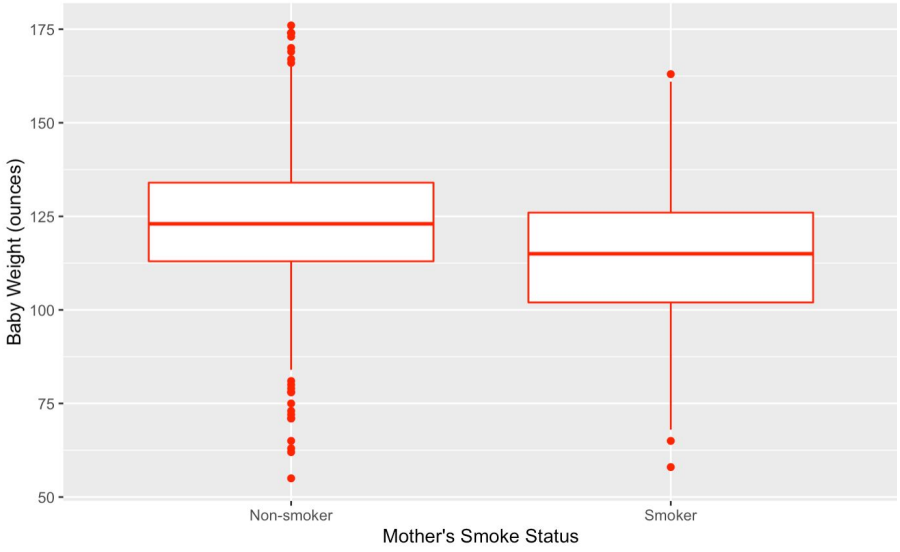

Fig 2c. Resultant boxplot with extreme outliers removed

MATH 189 HW1

From the lecture slides, it is noted that the weight of healthy babies ranges from 88oz to 140oz. Taking 88oz as the lowest possible weight of a healthy baby, the table below shows the count of the smoker and nonsmoker populations.

Class (count) / Weight (ounces)	< 88	>= 88
Smoker	40	444
Non-Smoker	23	719

Fig 3a.

Proportion of babies in each population under the barrier of 88 oz:

Smoker	40 / (40 + 444) * 100 = **8.2 %**
Non-Smoker	23 / (23 + 719) * 100 = **3.1 %**

Fig 3b.

We observe a significant difference of more than 5% between the smoker and non-smoker populations in Fig 3b. Furthermore, from the boxplot in Fig 2c, there appear to be a large number of outliers under 88 ounces within the non-smoker population. The presence of these outliers shows that underweight babies are not at all common in this distribution, meaning that non-smoker mothers tend to produce babies in the healthy range of weight. On the other hand, the box plot from the smoker population barely reveals any outliers, although the distribution does contain underweight babies. This indicates that babies who weigh less than 88 ounces are quite common within the population of babies born to smokers. Therefore, this study provides sufficient evidence to suggest that smoking is negatively correlated with the birth weight

of babies. However, a hypothesis test would have to be conducted to determine whether or not such a difference is reliable.

Advanced Analysis

To further understand the concentration and location of outliers in the two distributions, quantile-quantile plots were generated according to the mother's smoking status in Figure 4a and 4b.

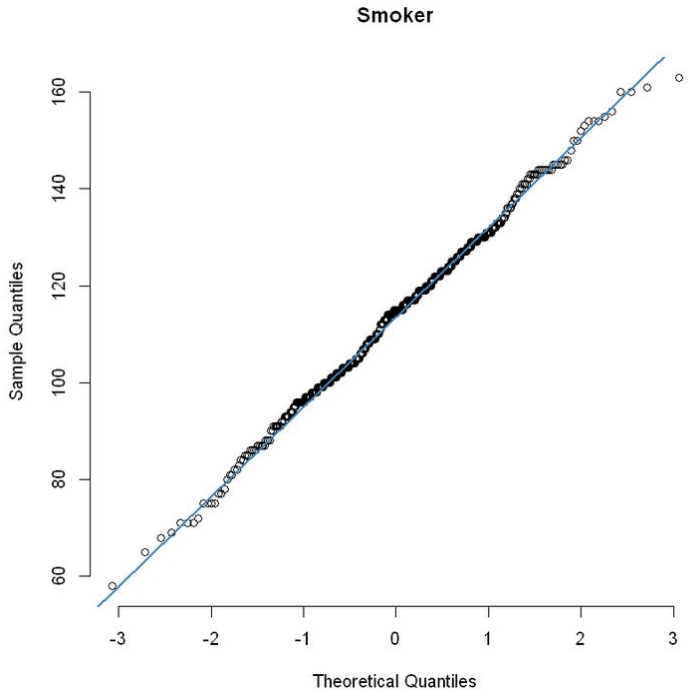

Figure 4a.

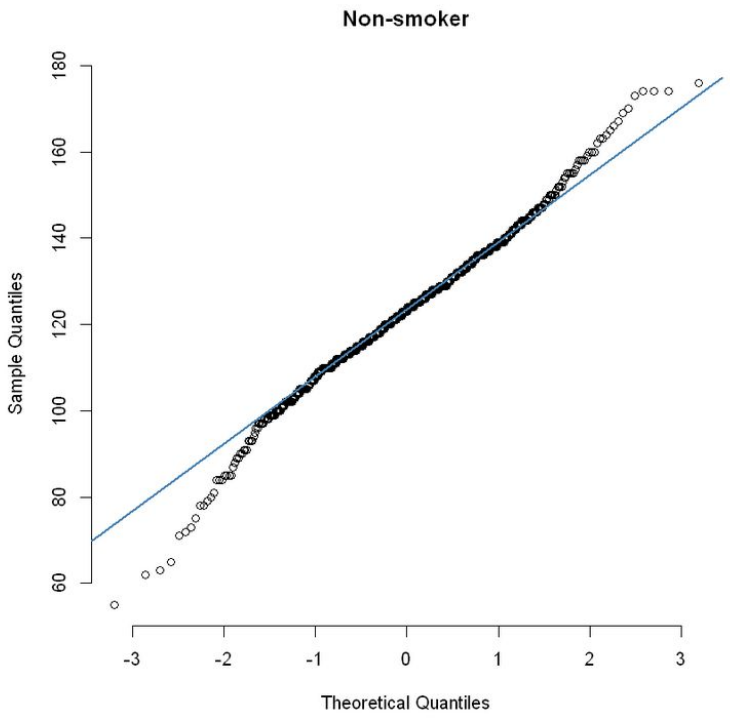

Figure 4b.

Both the smokers and nonsmokers weight distributions are normal. However, the non-smokers tails are fatter, per the deviation from the normal line at the ends of its distribution.

Conclusion

The purpose of this study was to explore the difference in weight between babies born to mothers who smoked during pregnancy and those who did not. The analysis presented throughout this paper attempts to understand whether such a difference is important for the health of the baby. Although it seems apparent from Fig 1a and 1b that the average weight of the babies from smokers and non-smokers differ significantly, it was only once the data distributions were investigated that a correlation between these variables could be uncovered. The study goes on to explore the distribution of each population by

introducing box plots (Fig 2a, 2b, 2c) and quantile-quantile plots (Figure 4a, 4b). These reveal the existence of outliers and anomalies in the data.

At this point, the study finally attempts to answer the ultimate question, whether the smoking status of the pregnant mother plays a role in the babies birth weight, and consequently, their health. This was done by checking if such babies had birth weights outside of the desired weight range, 88 oz to 144 oz. Tabular data (Figure 3a, 3b) is introduced to point out a 5% difference in the number of babies falling below this range between the smoker population and nonsmoker population. To track the birth weight of babies through their growth, Figure 5. study groups babies by their gestational age and contrasts birth weights among the two populations. It is apparent that there is a positive correlation between gestational age and birth weight, but also that smokers consistently have lower birth weight babies at most stages of gestation. While such a difference appears alarming, it is not evidence that smoking status causes a lower birth weight in babies. Instead, this study concludes that there is a negative correlation between birth weights of babies and smoking during pregnancy.

Appendix

To further the analysis and arrive at a conclusion, baby weights were binned by gestation age and it is observed that baby weights are positively correlated with gestation period.

Smokers		
Gestational Age (weeks)	Mean (g)	SD (g)
32	83.67	13.58
33	76.00	12.17
34	84.85	18.06
35	88.25	17.36
36	91.49	17.44
37	95.95	17.62
38	101.47	16.81
39	107.63	17.53
40	110.47	17.72
41	112.43	18.19
42	112.98	18.10
43	113.29	18.17
44	113.36	18.16
45	113.50	18.23
46+	113.53	18.19

Non-smokers		
Gestational Age (weeks)	Mean (g)	SD (g)
32	111.25	18.64
33	107.625	17.66
34	109.69	17.88
35	111.14	18.04
36	111.94	17.32
37	113.62	16.52
38	115.60	15.29
39	118.26	14.85
40	121.37	15.11
41	122.78	15.13
42	123.46	15.26
43	123.40	15.22
44	123.47	15.20
45	123.43	15.23
46+	123.42	15.22

Figure 5. Summary statistics by gestational periods of smokers vs non-smokers

MATH 189 HW1

Binning the means and standard deviations of the non-smoking mother's babies and the smoking mother's babies by gestational periods, we see that in general, birth weights increase with gestational age, and that the smoking mother's babies are much lower weight than the non-smoking mother's babies at earlier gestational periods.

Contributions

Arjun Sawhney (arsawhne@ucsd.edu) - Report Structure, Introduction, Body (Data, Analysis)

Noah Inada (ninada@ucsd.edu) - R code, Body (Analysis, Advanced Analysis), Appendix

Raimundo Castro (rac045@ucsd.edu) - Body (Analysis, Methods), Conclusion