Data Analysis on the Effects of the Birth Weight of a Baby from a Mother's Smoking Habits.

Introduction

Our study will focus on the difference in baby birth weight from mothers who smoked during pregnancy and those who didn't. The dataset for this analysis was an observational study that took place from 1960 - 1967 and observed all pregnancies during that time that occured in the Kaiser Health Plan within San Francisco. The dataset consists of 1236 observations, in which 484 of those observations were smokers and 742 of those observations were non-smokers. 10 observations did not answer or were not tested and so are omitted from our analysis. Our main analysis will look into the birth weight of the baby in ounces and a binary value stating whether the mother smoked or not.

Some framing issues with this dataset that is necessary to state is how representative the sample is for the rest of the population. 39.4% of the dataset were mothers who smoked, which contradicts with the 15% stated in the 1996 study. This demonstrates an issue of the time setting of the study and the changes in smoking habits and ingredients within cigarettes. Furthermore, the study only took place in San Francisco, the smoking habits and consumption of second hand smoke could be different to the rest of the population and show bias within the results. Moreover, there is no variable that quantifies the smoking patterns between individuals, although it makes the analysis accessible, it is difficult to categorise all smokers into one group and makes it difficult to analyse whether this is harming the health of the baby. Lastly, another confounding issue was eating habits, smokers tend to eat less and as a result, this will affect the weight of the babies. The confounding issues makes it difficult to analyze the causal relation between smoking and the baby's weight.

Our data analysis will gain insight on the characteristics in spread and average location of the birth weight of babies between smokers and non smokers. Looking deeper, we will analyze the subclass of babies who are low birth weights and compare between the two groups. This will answer whether all babies are on average lower weight when mothers smoke and/or there is a larger group of babies that classify as low birth weight. Our paper will conduct analysis and will provide a numerical summary of the distribution between two groups and also check the normality between distributions. From there, we will check proportional values between low birth weights. Lastly, we will look into the age of mother to look at an example of a confounding issue and check the relation between one another and issues that come upon that.

Analysis

Spread and Locational Distribution Between Smoker and Non-Smokers.

	Birth Weight in Ounces (Smoker)	**Birth Weight in Ounces (Non-Smoker)**

Mean	**114.1**	**123.0**
Median	**115.0**	**123.0**
Min	**58.0**	**55.0**
Max	**163.0**	**176.0**
Standard Deviation	**18.10**	**17.4**
IQR	**24.0**	**21.0**

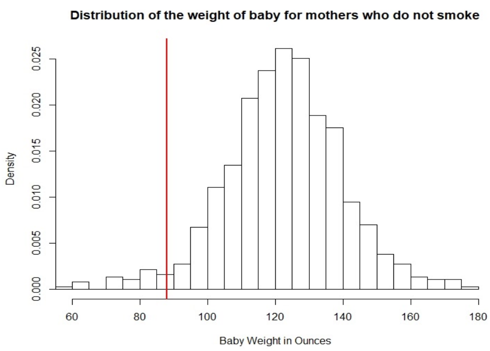
Table demonstrating a numerical summary of characteristics shown in birth weight of babies from both groups. (Appendix 1.1)

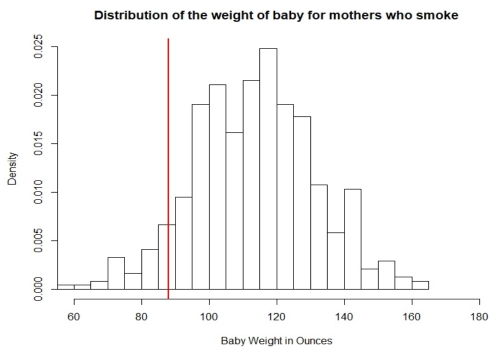
First, we summarized the characteristics the distributions showed for the two groups in order to gain insight on spread and location of data. Our data shows that there is a difference of 8.9 for the mean weight between the two groups, where the non-smoker group is showing a larger weight. This is shown graphically with the peak of the smoker group being shifted to the left. (Appendix 1.2) This appears to be a potentially significant difference considering the proportion of this change compared to the actual weights of the babies.

On the other hand, the spread and distribution of weights show similar traits with both histograms being unimodal and resemble gaussian functions.(Appendix 1.2) Both show their graphical distributions being skewed slightly to the left, as the median is slightly above the mean for both datasets. (Appendix 1.2) Moreover, the standard deviation and interquartile range show a small difference between one another.

Overall, the distributions share similar traits in spread, but the mean is shifted to the left for the smoker group. This could indicate that we could observe a relation between smoking and birth weights, as it appears to show it having an effect on the whole group.

Frequency of Low Birth Weights



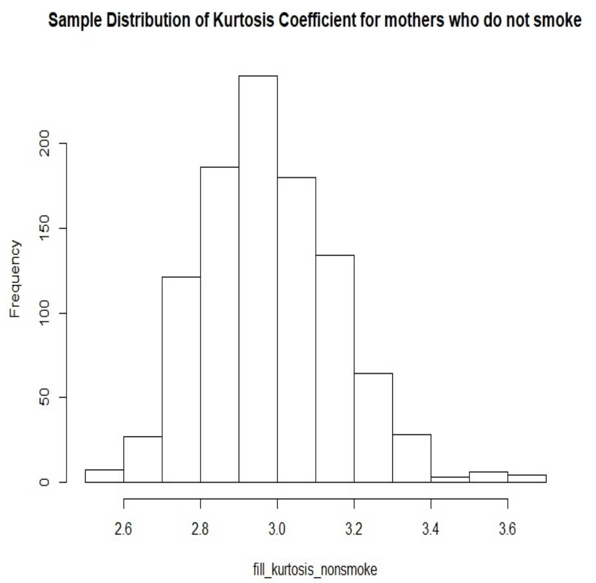


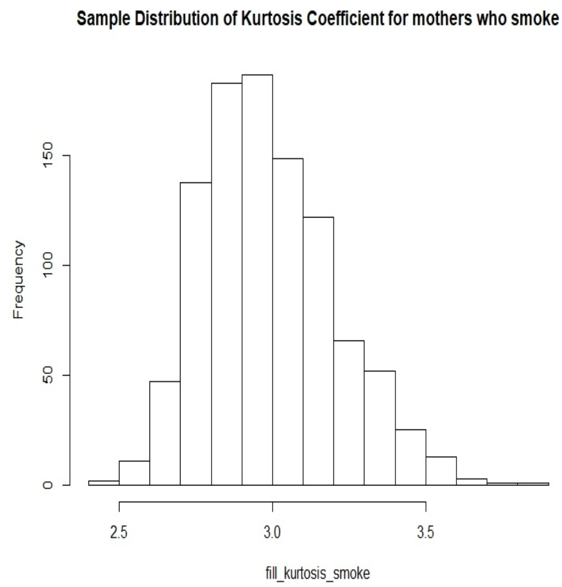
We compared the proportions of the two groups (smokers and non-smokers) that classified as low birth weight within their groups. Proportions were used as our frequency value due to sample size of groups being unequal. This was conducted by summing up the area left to the histogram from 88 ounces.

Our analysis shows that 7.4% of babies from the smoker group were classified as low weight, compared to a 2.9% from the non-smoker group. Changing the classified low weight limit by 5 ounces (83 ounces and 93 ounces) showed little difference in the proportion difference of percentages between low birth weight babies in non smoker and smoker groups. (Appendix 1.3) This demonstrates that our estimates from the sample are somewhat reliable to the dataset, and are robust.

The results appear to show a strong difference in the number of babies that are classified as low birth weight between the smoker group and non-smoker group.

Normality of Distributions





To check whether the distributions of mothers who smoked and did not smoke are normal distribution, we decided to use the Kurtosis coefficient for comparison.(Appendix 1.4) The observation statistic of Kurtosis coefficients are 2.975698 for mothers who smoked and 4.026186 for mothers who did not smoke. After taking 1000 samples with a population size of 1226, we produced the histogram above.

After comparing the observation statistic of Kurtosis coefficient to the histogram of samples, the Kurtosis coefficient of smoking mothers is close to the mean of its histogram of samples whereas the Kurtosis coefficient of non-smoking mothers is considered as an outlier.

Having a value greater than 3 indicates heavier tails. This demonstrates that from the smoking group there's a distribution that indicates more extreme values, and a larger difference in weights. This could indicate that smoking groups lead to more values that are within the tails range.

Effect of Mother's Age (Additional Analysis)

	Proportion of Low Birth Weight (%)	
Grouping of Mothers	Old	Young
Smokers	7.78	7.08
Non-Smokers	3.76	2.17

Another study completed by the Okinawa Child Health Study Group looked at the effects of age of the mother's smoking and found that maternal smoking is even more harmful given the mother's age. We decided to conduct the proportion of low birth weight analysis again. This time, however, we will be grouping by age of younger mothers and older mothers. This was done by calculating the median of the age of mothers and splitting the group based on the median. Moreover, we had to remove more values due to missing values in the age column.

Adding onto this, we applied a linear regression model on both variables, birth weight and smoking, over the age of the mother and found a positive slope between both variables. (Appendix 1.5) This indicates that there is an association and could indicate a possible positive bias within the data.

This demonstrates that the older group did experience having a larger proportion of low birth weight in both groups. This was by a small margin and so we could conclude that although there's a relation between the age of mother, the smoking habits appear to be more important based on this analysis.

Conclusion

Our studies focused on analyzing the difference in birth weights between mothers who smoke and do not smoke. We produced analysis that looked into the data as a whole and data into a subclass of birthweight. Both results demonstrated the same results that there's a decrease in average birth weight if the individual's mother smoked. However, we cannot confirm this entirely, as not enough analysis was conducted. For our next analysis, we could look into bootstrapping this sample and confirm if the difference is significant.

Overall, the frequency indicator of the low birth weight was the most impactful to our conclusion. This is because it analyzed differences in subclass weight, where the weight was small enough to test the hypothesis of it affecting the health of the baby.

Another aspect of the analysis was to answer whether this difference is important to the health of the baby. Studies have stated that babies born smaller have lower survival rates based on epidemiological studies. However, this does not indicate a causal relation between a baby's weight and smoking habits. We inferred this by analyzing another variable, a mother's age, to demonstrate that there are other confounders that will have an effect on the birth weight. For further analysis, we could take these other variables into account and use instrument variables in order to calculate a regression that shows a more accurate portrayal of causality.

4. Appendix/Appendices

1.1 Output of Code that ran the summary of both smoker and non-smoker groups.

Summary of 'bwt' for mothers who smoked:

Min. 1st Qu. Median Mean 3rd Qu. Max.

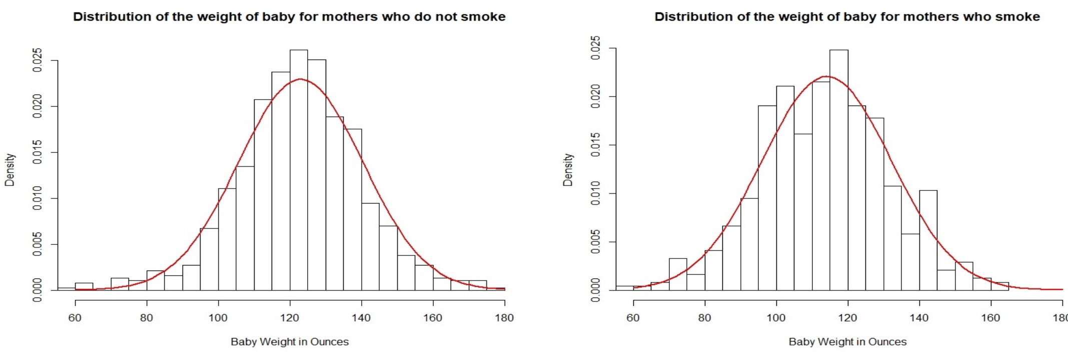
58.0 102.0 115.0 114.1 126.0 163.0

Summary of 'bwt' for mothers who did not smoked:

Min. 1st Qu. Median Mean 3rd Qu. Max.

55.0 113.0 123.0 123.0 134.0 176.0

1.2 Histogram distribution



1.3 Proportional Values between different Low Weight Classifiers.

Baby's Weight in Ounces	Proportion less than baby's weight	Proportion more than baby's weight
88 (Smokers)	0.0296	0.9704
88 (Non-Smokers)	0.0744	0.9256
93 (Smokers)	0.107	0.893

93 (Non-Smokers)	0.0417	0.958
85 (Smokers)	0.0475	0.952
85 (Non-Smokers)	0.0229	0.977

1.4 Kurtosis Value - Measures the size of the distribution's tails and is often used as a measure of the similarity of one distribution to a normal distribution. Value of 3 is often associated with a normal distribution.

1.5 Age of Mother
Linear Regression Model called between two variables

Coefficients:
(Intercept) bwt
25.98067 0.01066

Coefficients:
(Intercept) smoke
27.254484 0.001686

Contribution Statement
Introduction: Haley Lai
Analysis: Everyone
Conclusion: Rick Boelen
Rmd File: Hwang Min Yu