Darren Chang
Shiyin Liang
Anthony Martinez

Homework 1

Introduction

Research Question: 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 of the baby?

For this assignment we were tasked with determining the difference in weight between babies born to mothers who smoked during pregnancy and those who did not. This is an important topic to research babies that are born underweight or overweight can have serious health problems. The data that was used for this assignment, (babies.txt)[6], was collected from all pregnancies that occurred between 1960 and 1967 among women in Kaiser Health Plan in Oakland, California.

We were able to determine that there is indeed a difference in birth weight of babies born to mothers who smoked versus those that did not. More specifically, it seems that babies born to mothers who did not smoke during pregnancy are born heavier than babies born to mothers who did smoke during pregnancy.

The rest of this paper will go over three different types of analysis--numerical, graphical and statistical testing--that allowed us to reach this conclusion. Similarly, there will be a discussion focused on the meaning of these results and future proposals for this project

Data

There are 7 different features in the dataset: birth weight of baby, gestational period, parity, age of mother, height of mother, weight of mother, and whether the mother was a smoker or not during the pregnancy. For the purposes of our research, we filtered out the rows where birth weight or smoker status was unknown and we only looked at the birth weight of the baby and smoker status of the mother. We also made note that smoker status is a binary categorical variable and birth weight is a continuous numerical variable when creating graphs and plots.

Methods

We used the programming language R for our data analysis. We used three methods for our analysis which were numerical, graphical, and incidence. Using these methods we were able to slowly build up to answering our research question.

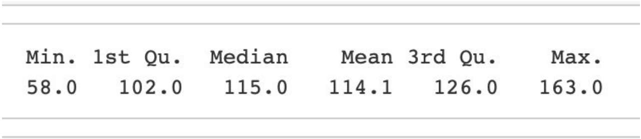
Analysis

Difference in Baby Weights Born to Smokers and Nonsmokers

Methods/Analysis

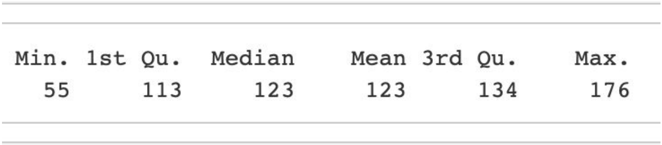
First, in our numerical analysis, we acquired the minimum, maximum, median, and mean statistics for baby weight of the two groups as shown in figure 1 and figure 2. Numerical analysis is important because it allows us to obtain numerical values that can describe the dataset at a quick glance. More specifically, the average allows us to have a singular value that generalizes each data group. Here, we see here that the average weight of a baby born to a mother who did not smoke is about 6 ounces higher than a baby born to a mother who did smoke and the values of the median, and maximum for the nonsmoker group are all higher than the values from the smoker group. This suggests that most non-smoking moms have heavier babies than smoking moms.

Figure 1: Summary Statistics of Smoker Group



```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 58.0   102.0   115.0   114.1   126.0   163.0 
```

Figure 2: Summary Statistics of Nonsmoker Group



```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  55     113     123     123     134     176 
```

Next, we started using graphical methods to compare the two distributions of birth weights. We can see from figure 3, 4, and 5, the peak frequency of birth weight differs between the two groups. The peak frequency in figure 3 shows that the most birth weights lie within the 110-120 ounce range. While the peak frequency in figure 4 shows that the most birth weights lie within the 120-130 ounce range. This shows that babies in our dataset born to mothers who did not smoke have a higher frequency of a heavier birth weight. Additionally, the histogram and density graph for non-smokers looks to be unimodal and suggests normality while the histogram for smokers looks more bimodal and is harder to determine normality.

Figure 3: Histogram for Smokers

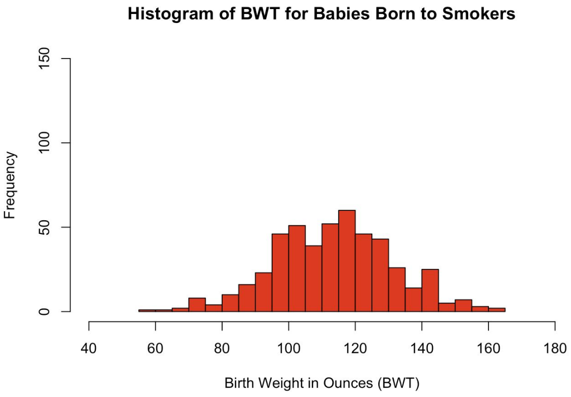


Figure 4: Histogram for Non-smokers

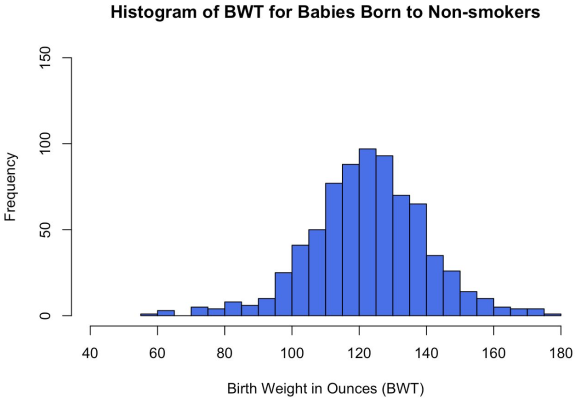
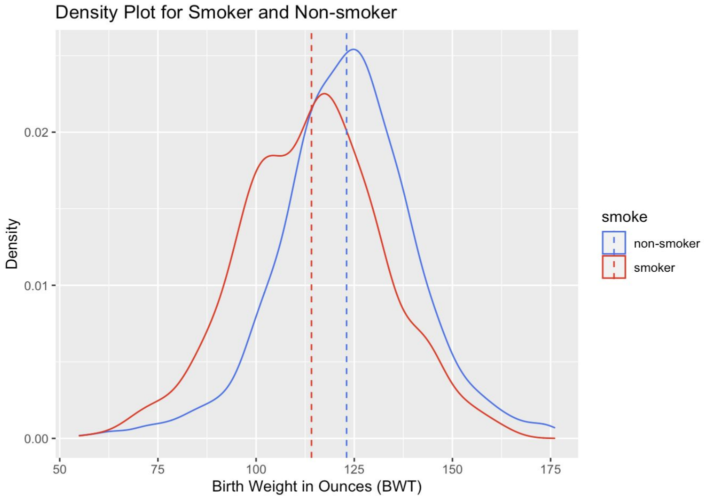
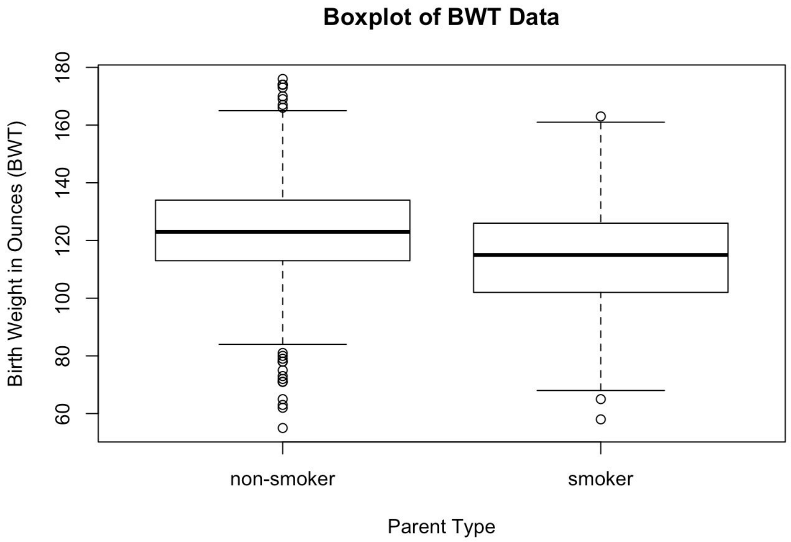


Figure 5: Density Plot for Smokers and Non-smokers



We used box plots to compare the distributions and identify potential outliers of the two groups. By placing both box plots on the same graph we were able to easily identify the differences in the two distributions. The box plot for mothers who are smokers looks similar to its counterpart, but is shifted down. Meaning everything including its median, minimum, max, quartiles, and outliers are all shifted down as well. We can also see that there are a lot more outliers in the non-smoker group than the smoker group which suggests more variance and are potential sources of problems in the statistical analysis we will later do.

Figure 6: Box Plot for Smokers and Non-Smokers



The final type of graph we implemented were the Q-Q (Quantile-Quantile) plots. Unlike our observations from the histograms, it shows us that the smoking group actually has a normal distribution while non-smokers do not seem to have a normal distribution.

Figure 7: QQ Plot for Smokers

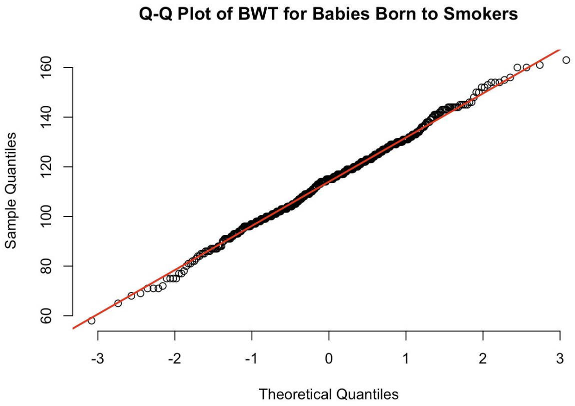
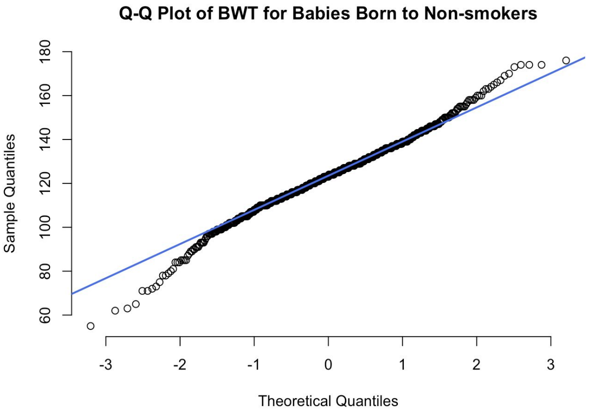


Figure 8: QQ Plot for Non-smokers



Conclusion

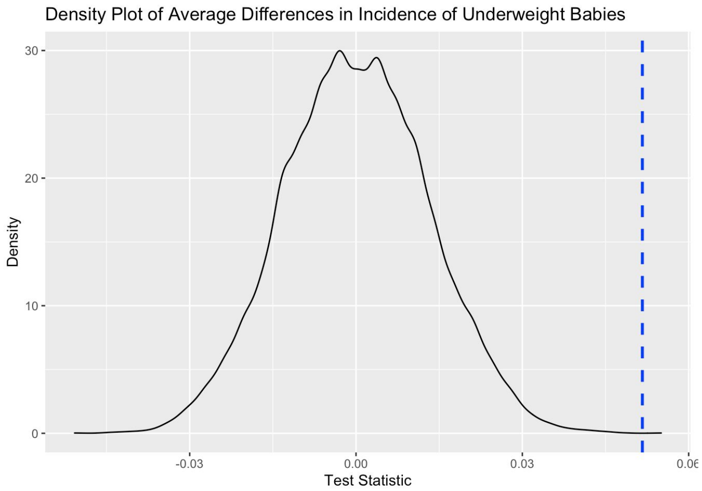
From the numerical and graphical data analysis methods, we were able to visually see that there is a difference in birth weight between babies born from mothers who smoked during pregnancy and those that did not. It’s clear from these analyses that the average of birth weight is higher for babies whose mother did not smoke during pregnancy in our dataset and similarly, they have a higher frequency of heavier babies. The smoking group has a normal distribution while the nonsmoking group does not.

Significance of Difference

Methods & Analysis

To see if variability in sampling would give more or less cases of low birth weight for each group, we ran a simulation where we shuffled birth weights and recorded the difference in proportion of underweight babies for each group 10,000 times. Our null hypothesis is that the proportion of underweight babies born to smoking mothers is not different from the proportion of underweight babies born to non-smoking mothers. In other words, the two groups come from the same distribution and any difference in proportions we observed is due to random chance. Our alternate hypothesis is that smoking mothers have more underweight babies. What we learned in our simulation is that, in 10,000 trials, we never see a difference in birth weights as big as 0.05 ounces. Here our p-value is 0 which means under our null hypothesis that the two distributions are the same, the probability of seeing a difference as extreme as our observed difference is exceedingly small. What this tells us is that smoking mothers and non-smoking mothers don't come from the same distribution after all and that there is something causing smoking mothers to have more underweight babies than non-smoking mothers.

Figure 9: Density Plot for Average Difference in Incidence of Underweight Babies



Since the Q-Q plot for the we also conducted a hypothesis test to see if the difference in proportions between the two groups is statistically significant. Our null hypothesis and alternative hypothesis are the same as our simulation. We found the Z statistic with:

$$\widehat{P}_{all} = \frac{40+23}{484+742}$$

$$n_n = samples\ of\ nonsmokers\ = 742$$

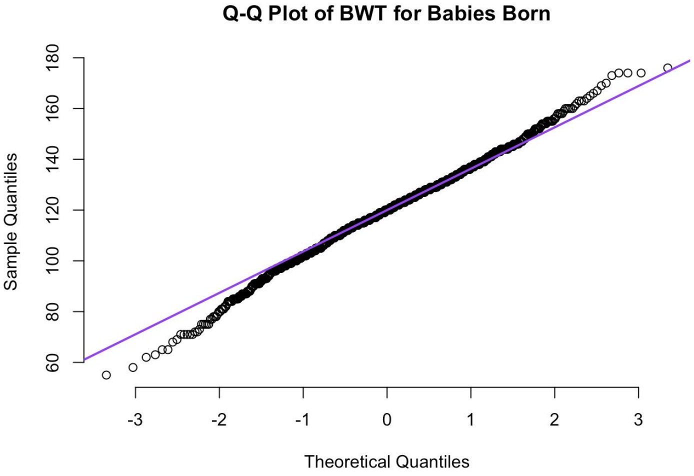
$n_s = samples\ of\ smokers\ = 484$

$$\sigma_{\widehat{P}_{smoker} - \widehat{P}_{nonsmoker}} = \sqrt{\frac{\widehat{P}_{all}(1 - \widehat{P}_{all})}{n_s} + \frac{\widehat{P}_{all}(1 - \widehat{P}_{all})}{n_n}}$$

$$z = \frac{\widehat{P}_{smoker} - \widehat{P}_{nonsmoker}}{\sigma_{\widehat{P}_{smoker} - \widehat{P}_{nonsmoker}}} = \frac{0.08264463 - 0.0309973}{0.0129} = 4$$

The Z statistic resulted in 4. This tells us that under the assumption the null hypothesis is true, the observed difference between our sample proportions of smokers and non-smokers is 4 standard deviations above the mean in this sampling distribution. Using the z-score table, we found that the p-value is 0.00003. This means that in a world where the proportion of underweight babies born to smoking mothers and non-smoking mothers is the same, there is only 0.003% of a chance that we would observe a difference in proportion of underweight babies as extreme as we did in our data. Since our p-value is less than our significant level of 1%, our null hypothesis doesn't look likely.

Figure 10: QQ Plot for Smokers and Non-smokers



Conclusion

We can not conclude that smoking causes underweight babies since we are not conducting an experiment. However, both our simulation test and hypothesis test show that the observed difference in incidence of underweight babies between smoking and non-smoking mothers is statistically significant since our p-values are under our significance level of 0.01. When we add or take away underweight babies for each class, we see differences between the proportion of incidences for each class but rarely as big as the difference we observed.

Conclusion

In this report, we set out to find the difference between baby weights born to mothers who smoked during pregnancy and mothers who did not smoke during pregnancy. From our numerical analysis and graphical analysis, we saw that, in our dataset at least, babies born to smoking mothers weigh less than babies born to non smoking mothers. From our statistical analysis, we can conclude that the differences in weights between the two groups observed in our numerical and graphical analysis is not due to random chance. There is indeed a variable causing smoking mothers to have higher incidences of underweight babies but whether that variable is smoking, we can not confirm. Our findings in this report are supported by other studies as well such as the one done by the Brighton Collaboration Low Birth Weight Working Group[2].

A pitfall of our project is that we cannot fully answer the question “Is this difference important to the health of the baby?” So far, our analyses show evidence that smoking mothers and nonsmoking mothers do not have the same incidence of underweight babies and suggests--not prove--that smoking mothers in the San Francisco area have a higher incidence of underweight babies. Literature provided by sources like Urban Child Institute[3], American Pregnancy Association[4], and BMC Pregnancy and Childbirth Group[5] all warns that being clinically underweight can affect brain development and even lead to infant fatality. Thus, our report can weakly answer that the difference in incidence of underweight babies between smokers and nonsmokers is indeed detrimental to the baby’s health. However, our report can not fully conclude that smoking while pregnant is detrimental to the health of the baby since we did not conduct an experiment. Furthermore, our findings in this report can only be generalized to women covered under the Kaiser Health Plan, in the San Francisco area and for the years 1960 and 1967 because that is where our data derives[6].

In the future, if we were to expand on this report, we would focus on trying to answer whether the difference between weights of babies born to smoking and nonsmoking mothers is important to the health of babies. We would utilize more data from our dataset such as gestational age and weight of mothers to check for confounding variables that may also affect a baby’s weight besides the mothers smoking habits during pregnancy. We would also have to get data from more diverse sources so that we can generalize our new findings to women in general.

Appendix

1. http://acsweb.ucsd.edu/~djc035/Assignment1.html
2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5710991/

3. http://www.urbanchildinstitute.org/articles/policy-briefs/prematurity-and-low-birth-weight
4. https://americanpregnancy.org/pregnancy-health/smoking-during-pregnancy/
5. https://bmcpregnancychildbirth.biomedcentral.com/articles/10.1186/s12884-018-1694-4
6. babies.txt

Contributions:

Darren: R Code, Wrote analysis for histogram, box plots, density plots, and QQ plots. Proofread report.
Sueanne: Wrote Conclusion, wrote section under Significance of Difference, edited section under Difference in Baby Weights Born to Smokers and Nonsmokers, researched and calculated technique for advanced analysis
Anthony: Wrote Introduction, Edited Conclusion, Wrote part of Baby Weights Born to Smokers and Nonsmokers section, wrote analysis for histogram, researched other papers