Effects of Smoking Mothers on Newborn Birthweight

Ben Pfeffer

February 2021

Abstract

Smoking has been linked to many health issues in recent years, mostly affecting the smoker. However, ideas have developed about a possible link between a pregnant woman smoking and low birthweight of the mother’s newborn. Using a subset of data on maternal smoking from the Child Health and Development Studies in R, descriptive statistics and visualizations are created to get a better understanding of the relationship between maternal smoking and the newborn’s birthweight. Results show that there seems to be a clear distinction that newborns from smoking mothers tend to have a lower, more random birthweight than newborns from non-smoking mothers. If the data were truly random, a hypothesis could be supported with a correlation between a mother’s smoking habits and newborn birthweight, keeping in mind that correlation does not equal causation. However, there are issues with the data that result in bias due to the convenience sample taken.

Introduction & Background

Concern has arisen in recent decades regarding the effects that mothers smoking during pregnancy has on their unborn children. These effects on their children, that result after mothers smoke during pregnancy, are known to be hazardous and have been known about by scientists and the general population for decades. These hazardous effects range many aspects of the child’s health. Low birthweight, sudden infant death syndrome, birth defects, reduced birth circumference, altered brainstem development, altered lung structure, cerebral palsy, future obesity, and future smoking habits are all hazardous effects that maternal smoking has on the child. In addition, the U.S. Public Health Service recently reported that if all pregnant women in the United States stopped smoking, there would be an estimated 11% reduction in stillbirths and a 5% reduction in newborn deaths [1]. One of the most common issues is low birthweight of babies born to smoking mothers. A low birthweight has major risks when it comes to the survival of the child. The infant’s body is much weaker than normal, and the baby might have a harder time eating, gaining weight, and fighting infection. Since the infant has such low body fat, the baby will also have difficulty staying warm. The low birthweight issue often comes with a premature birth which leads to more problems when these issues are combined. These issues include low oxygen levels at birth, infant respiratory distress syndrome, neurological problems and gastrointestinal problems. [2] Many of the overarching problems surrounding maternal smoking stem from the issue of low birthweight. However, since there is a lack of research in the connection between low birthweight and maternal smoking, it is essential to develop a stronger understanding of this theorized connection between smoking and low newborn birthweight. In order to look at this theorized connection, data from the Child Health and Development Studies database can be used to compare the smoking habits of pregnant women to the birthweight of their children [4]. Through this idea, a conclusion will be made concerning the possible correlation between a pregnant woman’s smoking habits and their children’s birthweights.

Methods

The data used is a subset of the Child Health and Development Studies and was pulled from Berkeley’s Statlab [3] under the name “Birth weight I”. This data was imported to RStudio using R for analysis. The data was then separated into two subsets: one containing non-smoking mothers and one containing smoking
mothers. This allows for a distinction between the two groups whose distinction between one another is the purpose of the study. After this separation, one histogram was plotted for non-smoking mothers and another was created for smoking mothers. Histograms show the frequency in “bins” of values on the x-axis. These histograms were analyzed individually and were then overlaid onto one another. To get a better view of this data, density plots (smooth histograms) were created and overlaid using the “ggplot2” package in R. Then, the differences in the smooth histograms were analyzed. After seeing the data visually due to the histograms, descriptive statistics were calculated for each dataset (smoking and non-smoking mothers). These descriptive statistics included: mean, median, standard deviation, skewness, kurtosis, and the 5-point summary containing the minimum, quartile 1, median, quartile 2, and maximum was constructed. This data was then analyzed to better understand the data shape. The skewness was analyzed to understand if the datasets were skewed and the kurtosis was analyzed to understand if the datasets differed from a normal distribution. The mean and median were analyzed to understand the central tendency of the data and if there were any lopsided outliers. The standard deviation was calculated to determine how far from the mean allowed for data points to have less than one z-score. Lastly, the 5-point summary was analyzed to understand the sections of the data. After analyzing the descriptive statistics, side-by-side box and whiskers plots showing birthweights of the newborns of smoking and non-smoking mothers were created and analyzed to determine if this visualization show a difference in the two sets of birthweights. Finally, a quantile plot was created for each dataset and was evaluated to determine the variation between each dataset and the normal distribution. If the datapoints strayed from the line in the plot, the distribution strayed from the normal distribution. Using all of these statistical tests, a conclusion was made regarding whether or not there was a correlation between smoking and non-smoking mothers.

Results

Birthweights of babies born from non-smoking mothers tend to have a central tendency around 110-130oz and tend to have many babies born in this range, creating a seemingly tall histogram (Figure 1). Meanwhile, birthweights of babies born from smoking mothers tend to have a central tendency around 90-130oz and tend to have a more gradual transition to the extremes of the x-axis, creating a seemingly shorter histogram (Figure 2). When density plots of smoking mothers and non-smoking mothers are overlayed, it shows a clear difference in the frequency of weights for the two groups, even though there is overlap, with smoking mothers having babies with lower birthweights and non-smoking mothers having babies with higher birthweights (Figure 3).
Figure 1: A histogram showing the frequency of birthweights of babies of non-smoking mothers.

Figure 2: A histogram showing the frequency of birthweights of babies of smoking mothers.
Smoking mothers’ average baby birthweight is smaller than that of their non-smoking counterpart in both the mean and median values. The standard deviation of smoking mothers is slightly higher than that of non-smoking mothers. The skewness of both datasets is slightly negative in the smoking mother data but is more negative in the non-smoking mother data with a value of -0.187 compared to the -0.034 value from the smoking mother data, meaning that the non-smoking mother’s data distribution is skewed slightly left compared to the normal distribution. The kurtosis of the smoking mother data is approximately 3, while the kurtosis of the non-smoking mother data is approximately 4 meaning that the smoking mother data is very close to the normal distribution, while the non-smoking mother data is noticeably different from the normal distribution. The 5-point summary shows the minimum of the non-smoking mother data to be 3oz lower than the smoking data’s, while the maximum is 13oz higher, showing that the non-smoking mother data covers a wider range of values. The values of the first quartile, median, and third quartile are all higher in the non-smoking mother data than the smoking mother data by about 9oz. (Figure 4)
<table>
<thead>
<tr>
<th>Non-Smoking Mothers</th>
<th>Smoking Mothers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>123</td>
</tr>
<tr>
<td>Median</td>
<td>123</td>
</tr>
<tr>
<td>St Dev</td>
<td>17.4</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.19</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>4.04</td>
</tr>
<tr>
<td>Minimum</td>
<td>55</td>
</tr>
<tr>
<td>Quartile 1</td>
<td>113</td>
</tr>
<tr>
<td>Median</td>
<td>123</td>
</tr>
<tr>
<td>Quartile 3</td>
<td>134</td>
</tr>
<tr>
<td>Maximum</td>
<td>176</td>
</tr>
</tbody>
</table>

Figure 4: A table comparing the birthweights of smoking and non-smoking mothers through descriptive statistics.

The interquartile range for birthweights of babies of non-smoking mothers is noticeably higher than that for the smoking mothers, and there are more outliers for the non-smoking mothers baby birthweight data. These outliers are closer to the interquartile range than in the non-smoking mother plot, meaning that the non-smoking mother data has a much stronger central tendency than the smoking mother data such that it allows for outliers closer to the mean (Figure 5).

Figure 5: Side-by-side box and whisker plots comparing the birthweights of smoking and non-smoking mothers.

The normal distribution line is most accurately followed by the data of the birthweights of babies of smoking mothers. This line slightly deviates from the normal distribution line in the center, but only for brief periods along the x-axis. Overall, it follows the line very accurately, showing that the distribution of the data is quite
close to the normal distribution (Figure 6). However, the data of the birthweights of babies of non-smoking mothers falls off the line noticeably towards the beginning and end of the graph, pointing towards a difference between the distribution of the dataset and the normal distribution, supporting the results gathered previously explaining skewness and kurtosis (Figure 7). These plots show that the data from smoking mothers is more random around its mean, since an added datapoint is more likely to disperse to outer values than it would be in the data of non-smoking mothers who is more likely to have an added point fall in the middle.

![Quantile plot of Non-Smoking Mother Data](image)

Figure 6: A quantile plot comparing the distribution of birthweights of non-smoking mothers to the normal distribution.
Figure 7: A quantile plot comparing the distribution of birthweights of smoking mothers to the normal distribution.

Discussion & Conclusions

The results show us that there is a relationship between baby birthweight and their mother’s smoking habits during pregnancy. The overlayed density plots (smooth histograms, Figure 3) show us visually that smoking mothers have babies with lower birthweights than babies of mothers who do not smoke during pregnancy. However, the descriptive statistics (Figure 4) also tell us that babies whose mothers smoked during pregnancy have, on average, a lower birthweight than babies whose mothers did not smoke during pregnancy through the use of metrics. The side-by-side box and whisker plots (Figure 5) show us the same results – that the average birthweight of babies from smoking mothers is lower than that of non-smoking mothers.

Regarding the distribution of the data, the side-by-side box and whisker plots (Figure 5) tell us more than just the fact that most of the birthweights of babies from non-smoking mothers are higher than that of babies of smoking mothers. They tell us a key fact about the distribution of each dataset; the data from the plots shows us that there is a stronger central tendency in the data of birthweights of babies from non-smoking mothers than there is in the data of birthweights of babies from smoking mothers. This point to the fact that the data of birthweights of babies from smoking mothers is more randomly distributed around a lower mean, meaning that if a baby is born from a smoking mother then it’s birthweight will be lower and more random (less likely to be the average) than if the mother did not smoke. To further support this idea, the quantile plots of each dataset (Figures 6 and 7) give more insight the distribution. The quantile plot of the non-smoking mother data shows how that the data of birthweights of babies from non-smoking mothers strays from the line of a normal distribution, supporting the idea that this data has a strong central tendency (i.e., It is more likely for a baby to have the average birthweight if the mother does not smoke during pregnancy). On the other hand, the quantile plot of the smoking mother data shows that the data of birthweights of babies from smoking mothers follows the line of a normal distribution almost exactly, supporting the idea that this data is more randomly distributed around the mean (i.e., There is more variability around the mean birthweight if the mother smokes during pregnancy). This variability in the possible birthweight of their baby should serve as another warning to mothers considering smoking during pregnancy since a low birthweight puts the newborn at much more risks of death.
Taking all of the previous information into account, the results produced would support the hypothesis that a mother smoking during pregnancy leads to a lower birthweight of the smoking mother’s newborn. However, there is much room for improvement in this study to the convenience sampling that occurred. This data was pulled from only Kaiser hospitals in San Francisco. This convenience sample leaves room for much deviation between the results of this experiment and the results if this experiment was run on a random sample from the general population. Furthermore, the study was only conducted on mothers with a prenatal plan (Kaiser Foundation Health Plan). This selective sample disregards mothers who are unable to afford or unwilling to obtain a prenatal plan. This selective convenience sample could have created results that differ significantly from that of the general population of pregnant mothers. Taking this bias into account, the results of this study cannot be expanded to a general population and can therefore not support the hypothesis that a mother smoking during pregnancy leads to a lower birthweight of the smoking mother’s newborn. However, this can support a more specific hypothesis that, in a Kaiser hospital in San Francisco among mothers in the Kaiser Foundation Health Plan, a mother smoking during pregnancy leads to a lower birthweight of the smoking mother’s newborn. This reformed hypothesis has very little practicality, so, to fix this, the data should be broadened to include data from around the world in order for the initial hypothesis to be supported without being specified.

References


Appendix

R Code

```r
# Project 1: Birth Weights of Babies
data <- read.table("babies1.data", header = TRUE, sep="")
str(data)

nosmoke <- subset(data, smoke == 0)
nosmokedata <- nosmoke[, "bwt"]
smoking <- subset(data, smoke == 1)
smokingdata <- smoking[, "bwt"]

hist(nosmokedata, main="Non-smoking Mothers Histogram", xlab="Birthweight")
hist(smokingdata, main="Smoking Mothers' Histogram", xlab="Birthweight")

p1 <- hist(nosmokedata)
p2 <- hist(smokingdata)
a <- min(min(nosmokedata), min(smokingdata))
b <- max(max(nosmokedata), max(smokingdata))
plot(p1, col=rgb(0,0,1,1/4), main="", sub="", xlab="", xlim=c(a,b)) # first histogram
plot(p2, col=rgb(1,0,0,1/4), main="", sub="", xlab="", xlim=c(a,b), add=T) # second histogram
```
title(main= "Histograms of nonsmoke and smoke data", xlab= "Birthweight", ylab="Frequency")
#Overlaid smooth histograms
newData <- subset(data, smoke == 1 | smoke==0)
library(ggplot2)
'Legend' <- factor(newData$smoke, levels= c("0","1"), labels = c("Non-smoker", "Smoker"))
ggplot(newData, aes(x = bwt, fill = 'Legend')) + geom_density(alpha = 0.5) + labs(title=" Smoking habits of mothers and newborn birthweights", x = "Birthweight", y ="Density")
#4
mean(nosmokedata)
mean(smokingdata)
median(nosmokedata)
median(smokingdata)
sd(nosmokedata)
sd(smokingdata)
library(moments)
skewness(nosmokedata)
skewness(smokingdata)
kurtosis(nosmokedata)
kurtosis(smokingdata)
fivenum(nosmokedata)
fivenum(smokingdata)
#5
library(plyr)
'Non-Smoking' <- nosmokedata
Smoking <- smokingdata
combined<- cbind.fill(as.data.frame('Non-Smoking'),as.data.frame(Smoking))
boxplot(combined, main="Boxplot of Mother’s Smoking Habits and Newborn Birthweight", ylab="Birthweight (ounces)", col=c("royalblue2","red"))
#6
qnorm(nosmokedata, pch = 1, frame = FALSE, main="Quantile plot of Non-Smoking Mother Data")
qline(nosmokedata, col="steelblue", lwd = 2)
qnorm(smokingdata, pch = 1, frame = FALSE, main="Quantile plot of Smoking Mother Data")
qline(smokingdata, col="steelblue", lwd = 2)

Python Code (To Create Figure 4)

import pandas as pd
df = pd.DataFrame([[[123,114],[123,115],[174,181],[123,115]], [[-0.19, -0.03], [4.04, -2.09], [55.58], [113,102], [123,115],[134,126],[176,163]]]
df

Theory Question: Proof of the mean squared error

\[ \sum_{i=1}^{n} (X_i - \bar{c})^2 \] We expand this expression:  
\[ = \sum_{i=1}^{n} (X_i^2 - 2\bar{c}+\bar{c}^2) \sum_{i=1}^{n} (x_i + nc)^2 \] Which we must derive and find when it equals zero to minimize the error. 
\[ = 2\bar{c} - 2 \sum_{i=1}^{n} x_i = 0 \] So, simplifying: 
\[ t = (\sum_{i=1}^{n} x_i)/n = \bar{x} \] Or, the mean.