Baby Weights of Smoking vs Non-Smoking Mothers

Nick Collins

UMass Dartmouth

January 30, 2022

Abstract-

Pregnant mothers are often told not to smoke during pregnancy because it can be harmful to the baby. The data used for the analysis has the weights of newborns and whether the mother is smoking, which can provide some insight when the weights are broken up into groups based off the smoking status. I used Python to perform the analysis and used different libraries like pandas, scipy and statsmodel. Outputting summary statistics, histograms, box and whisker plots and quantile plots was the best and easiest way to analyze the weights for the two different groups. We want to see if there is some statistical evidence that shows mothers who smoke during pregnancy have differences in birth weights from the mothers who do not smoke.

Introduction-

The data we were given has the birthweights of babies and whether the mothers smoke or not. Doctors and medical professionals say that mothers should stay away from things that can be harmful to the babies' health. Smoking is one of those key things and has been recorded to have harmful effects on the babies. Looking at this data will give us some insight into if and how babies' weights are affected by the smoking status of the mother.

Methods-

The raw data is a csv file with two columns. The first column is labeled "bwt" and is the numerical weight of the babies. The second column is labeled "smoke" and is represented by the integers 0,1 or 9. If there is a zero that means the mom does not smoke now, a one means the mother does smoke now and a nine means the smoking status is unknown. Since the three types of mothers were in the same dataset, I had to separate them into their own data frames. To do this project I mainly used pandas for all of the tasks and variable creation. To separate the variables I used the groupby method in pandas. I grouped by the column "smoke" so this way all the similar values would be stored in one group. After separating them into one group it was easy to take the summary statistics of the two groups. There were three overall groups, but the instructions only asked for the analysis of the smoking vs nonsmoking groups. Pandas made it very easy to take the mean, median, standard deviation, skewness and the five-point summary. There are built in functions in pandas that will take care of this. To get the kurtosis I had to use the library scipy. For some reason, the kurtosis value I got through pandas wasn't accurate, but the value obtained through scipy was the right answer. The quantile plots were generated through the library statsmodel.

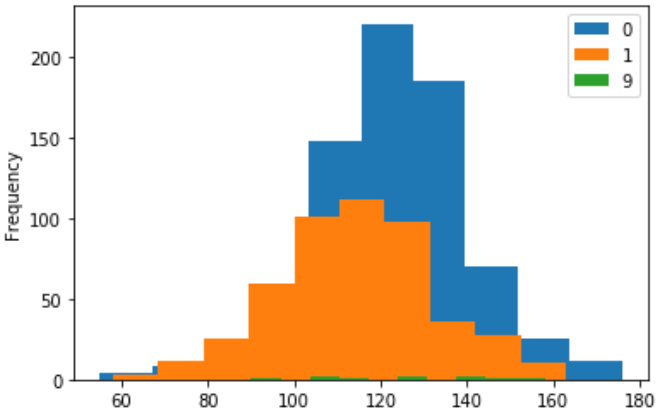
Results-

The first task was to output some basic summary statistics on the two groups of smoking and nonsmoking mothers.

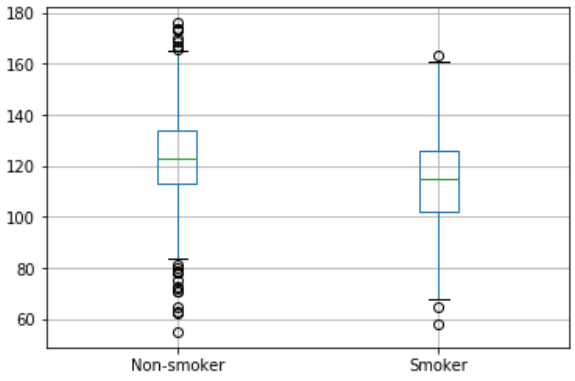
```
Descriptive statistics of birth-weights for non-smoking mothers:
Mean =  123.05
Median =  123
Standard deviation =  17.4
Kurtosis =  4.037060312433822
Skewness =  -0.18736306526595664
Minimum =  55
Maximum =  176
Quartile 1 =  113.0
Quartile 3 =  134.0
```

```
Descriptive statistics of birth-weights for smoking mothers:
Mean =  114.11
Median =  115
Standard deviation =  18.1
Kurtosis =  2.988032478793404
Skewness =  -0.033699506713282625
Minimum =  58
Maximum =  163
Quartile 1 =  102.0
Quartile 3 =  126.0
```

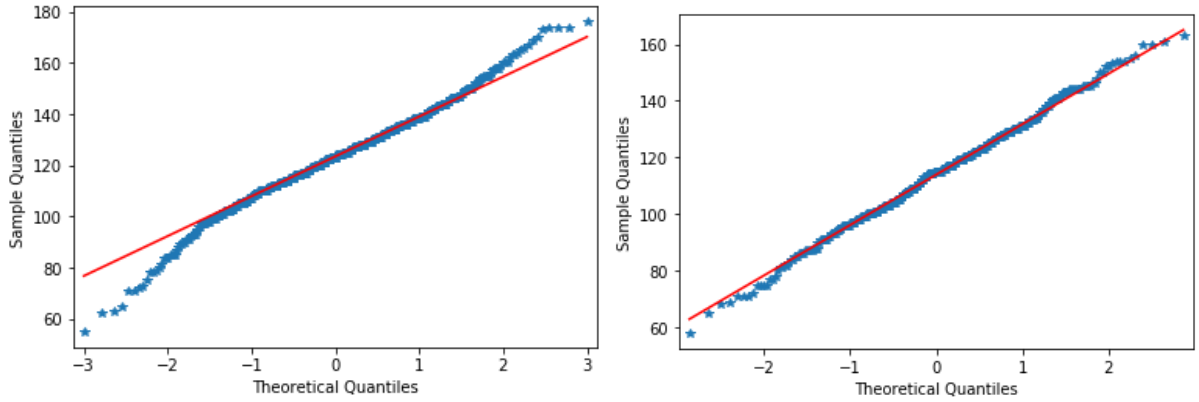
Above we can see the different statistics for both groups of mothers. The statistics for smoking mothers, on average, are less than nonsmoking mothers. When plotting the two data sets on a histogram we can also see the differences in the two groups.



The blue(0) is nonsmoking mothers, orange(1) is smoking mothers and green(9) is the unknown status of a mother. The x-axis is the bodyweight of the babies and the y-axis is the count for each group.



In this box and whisker plot we have the two groups and we can see that the nonsmoking mothers have a higher average and a smaller quartile range. The smoking mothers have just the opposite and they don't have the outliers that the nonsmoking mothers have, which is interesting.



The first quantile plot is the plot for the nonsmoking mothers and the second is the smoking mothers. The first plot the tails are deviating from the line on both ends, while in the second plot the points lie roughly on the line that we would expect.

Conclusion-

After doing some analysis on the dataset on baby weights for smoking and nonsmoking mothers we see noticeable differences in the analysis. In the summary statistics we see the nonsmoking mothers have higher statistics in almost every category. This makes sense because we would expect the babies to be healthier and more nourished if the mother doesn't smoke. The maximum value we see for nonsmoking mothers is 176, while the smoking mothers maximum is only 163. This is a large difference and is more evidence that nonsmoking mothers have healthier, heavier babies. From the histogram we see that the nonsmoking mothers is shifted right meaning that on average the babies are heavier. An interesting graph is the box and whisker plot. I was somewhat surprised at this graph. The mean and two quartiles of nonsmoking mothers make sense because they are higher than the smoking mothers which is what we would expect. However, there are many more outliers for nonsmoking mothers. Another interesting thing about the outliers is how many of them are below "minimum" calculated by taking quartile one minus 1.5 times the IQR(interquartile range). I would have thought it would have been the opposite with the smoking mothers having more outliers under the "minimum" value because on average they have lighter babies. We also see that the first quartile for nonsmoking mothers is about the same as the average for the smoking mothers. This means the 25th percentile of nonsmoking mothers is about the same as the average for smoking mothers. The last plot was the quantile plots for each group. In these plots we see the nonsmoking tails vary from the line we expect to see. This makes sense because the kurtosis is higher than three, so the graph makes sense. The smoking

mothers graph stays close to the line and this also checks out because the kurtosis is right around three, so we wouldn’t expect much deviation. Overall, the data backs up the claim that mothers shouldn't smoke during pregnancy.

References-

"Maternal Smoking and Infant Health I:" *Stat Labs Data Page*, https://www.stat.berkeley.edu/~statlabs/labs.html#babiesI.

Appendix-

Page 24 problem 19

$$n = \sum x_i - c$$

C = 1/n $\sum x_i$

We can rearrange these terms to get the original mean we want to solve for