Smoking Mother and their Baby's Birth Weight

Author contributions

Lu Xu and Zehui(Barry) Zhang separate the work equally. We write the code and analyze data together. Any time one of us comes up with some great ideas, the other one will help to record and edit them. And find out what is missed.For the whole analysis part, we do help each other together and write and edit the doc.. Hard to distinguish the exact amount of work we did. For the introduction and conclusion, Lu writes the introduction part and Barry writes the conclusion part. Generally, Barry did the code part more, and Lu did the doc.

1.1 Introduction

The *Child Health and Development Studies* reports 1236 babies that were born between 1960 and 1967 in the Kaiser Health Plan and within the country of Oakland in California. In this data set, all observations are in the same gender, boys, and limited to single births. It means there are no twins, and, moreover, all these boys have been born at least 28 days. But not all data points are valid ones. Some missing values are shown as "9s" in the original data set.

The purpose of this analysis is to compare the difference in weight between babies born to mothers who did not smoke during pregnancy to those who did smoke during pregnancy. Also, determine whether those differences will impact the weight of the baby or not.

This analysis will show that smoked mothers during their pregnancy could affect the babies in the low birth weight. In other words, smoking baby mothers during pregnancy have a higher probability of having a low birth weight baby. This conclusion will be shown by the order of numerically, graphical, and proportional analysis. Then assess the importance of the difference. The more cigarettes mothers smoked the higher weight loss will be caused in their babies.

Index

2 Basic Analysis

2.1 Data processing and Numerical analysis

Methods

Generally, anytime, so as to do research, the first thing we have to do is to clean the data ---- remove some invalid data points and break up the original data set to two subsets(smoked group and non smoked group). Then by using R, we can easily get some special values from the cleaned two subsets. To be more specific, we can find mean, mode, median, standard deviation, first quantile, etc.. After that we can also check the normality of data by 68-95-99.7 rules, skewness and kurtosis. In R we can create two helper functions to get skewness and kurtosis coefficients. While, in order to make the analysis easier, values are standardized when looking for the skewness and kurtosis coefficients.

Analysis

Invalid data: In the original data sets, there are 1236 observers. In the column of smoke, it is pretty straightforward that 1 implies this certain overvator, the mother, smoked during pregnancy. The number 0 under smoke column implies non smoked during pregnancy. However, some “9” appeared among all “0s and 1s”. “9s” are representing some missing information that studies do not get from those babies.And totally, among 1236 observers, only 10 data points were missed. So it does not matter removing them.

	Mean	Median	Mode	Standard deviation
Smoked group	114.1	115	115	18.1
Non-Smoked group	123	123	129	17.4

Note: 484 observers in Smoked group, 742 observers in Non Smoked group.

Mean, Median, Mode: By the chart above, it apparently shows that the non smoked group has higher birth weight than the smoked group. Each of them have several number differences from range 8 to 14. Mean represents the average behavior of a data set, Median is the middle value, and Mode is the number occurred most frequently in the data set. All these three basic and most important statistical measures all induce the same result, that baby birth weight is litter if the baby mother smokes during pregnancy. What’s more, combining with the help standard deviation, a smaller standard deviation means a higher coherenness of the data. Thus non smoked groups in general enhance the behavior that they have a heavier baby than smoked mothers.

	% of values within 1 units of center	% of values within 2 units of center	% of values within 3 units of center	skewness	kurtosis
Smoked Group	70.45	94.62	99.79	-0.033	2.97
Non-Smoked Group	65.36	92.05	98.78	-0.187	4.02

Normality: By R, we find some values collected above. Since the data has been standardized before finding these values. So based on the 68-95-99.7 rule, these two data sets are likely normally distributed. More specifically, differences here for the smoked group are 2.45, -0.38, 0.09; for non-smoked group, -2.64, -2.5, -0.92 are differences respectively. Therefore, the Smoked group is more likely normally distributed than the Non-smoked group. What's more, skewness and kurtosis coefficients also proves this.

	Skewness (484 observations)	Kurtosis(484 observations)	Skewness (742 observations)	Kurtosis(742 observations)
Min	-0.469	2.373	-0.355	2.480
Max	0.491	4.232	0.405	3.914

So as to find the region for both coefficient with generality, we can repeat the experiment of finding skewness and kurtosis of normal distribution 10000 times. And in order to reduce the error and increase the generality, we did it by using the same number of observations. It is clear that both skewness of these two groups are included in the general region. It means that both groups have symmetric distribution. Same thing happened on the kurtosis of the Smoked group. But here 4.02 is not located in the general Kurtosis region of 742 observations. It does not mean that Non-smoked group is not normally distributed. But as likely as the smoked group.

Conclusion

But simply dig the data numerically, so far the data shows baby birth weight is lighter if the mother smoked during pregnancy, compared with non smoked mothers. And roughly, both subsets behave like normal distributions. But non-smoked group has a weaker normality than smoked group.

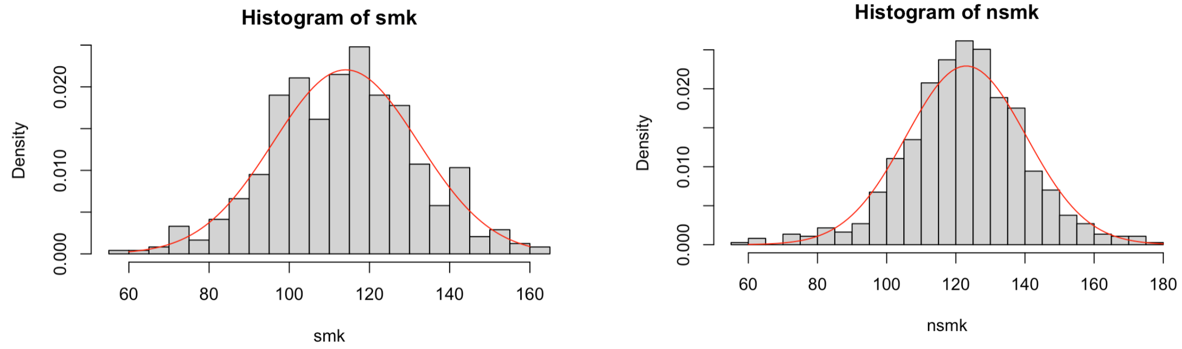
2.2 Distribution of birth weight

Methods

By using R, we can create the histograms for both groups, and the normal curves with the same means and standard deviations within a single graph. Intuitively show the characteristics of the data, and check if it is likely normally distributed. Then to be more accurate, we can compare the behavior of the 45 degree line and the Quantile-Quantile plot.

Analysis

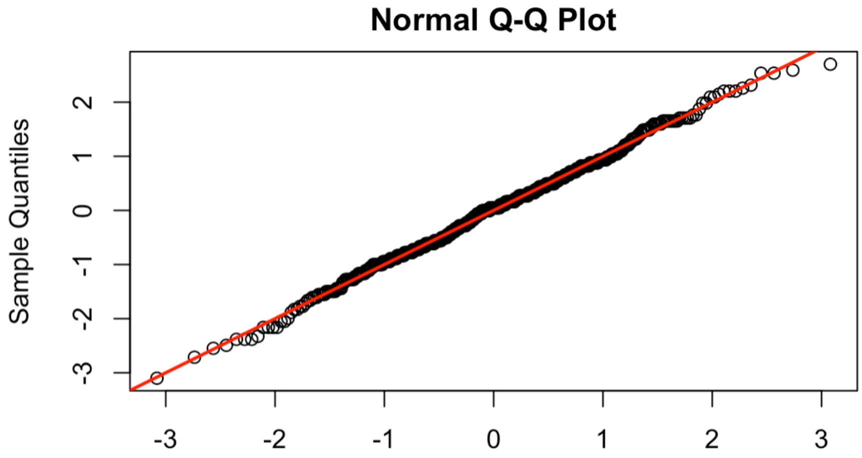
Histograms:

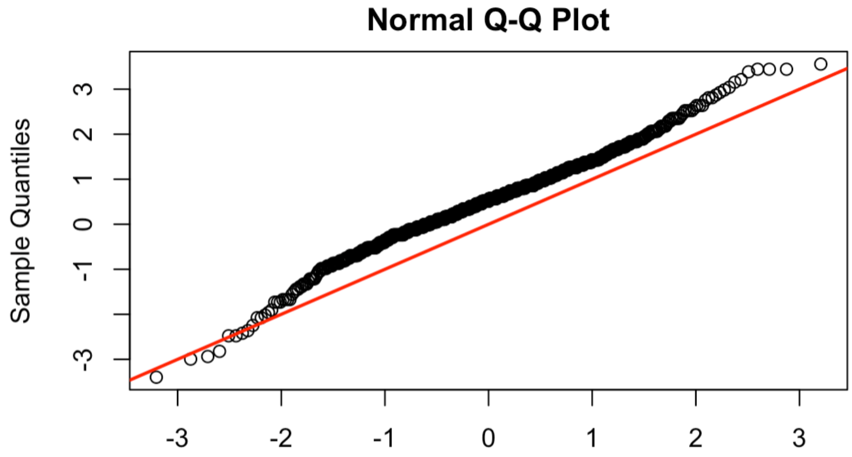


NOTE: smk represent smoked group, nsmk represent non-smoked group

The red curve is the normal curve which has the same mean and the standard deviation of the data set. It shows what a normal distribution looks like under the certain condition. The less difference between the red curve and the histogram, the higher normality the data set has. Looking straightforward from these two histograms, both of them are unimodal and roughly symmetric. For the histogram of smoked data set, the real peak is roughly equal to the peak of the normal curve. And it has a mode at about 115. While means in this data set, the most frequent baby birth weight occurred is about 115. This matches the mode we found in the prior question. Outliers, here, are seldom, which are less than 60. Based on the skewness we get above, and the histogram, we can say the data is normal, and distributed symmetrically. Roughly, the majority of data representing the birth weight is in the region from about 90 to 145. While for the other one, non-smoked group, comparing with the normal curve, the real peak is higher than the normal one, and few outliers occurred around 60. The mode here is at about 125 to 130, which is the same as the value we get before. The skewness is also too tiny to be ignored here. Thus it is roughly normally distributed.

QQ Plot(Quantile-Quantile Plot):





The QQ plot provides a graphical means of comparing the data distributions to the normal. The plot here used standardized data points from two subsets made before respectively. The above one is for the Smoked group, and the bottom is for the non-smoked group. Red line is the 45 degree line with slope 1, which helps to compare the normality between two data sets. Since the data points are standardized, so if the distribution is normal, the qq plot should tend to have intercept 0 and slope 1. In other words, the plot should be close to the red line. We can see that there is no strong skewness for both data distributions. They roughly form a straight line which indicates both data sets have an approximate normal distribution. And specifically, the data in the first plot, in the smoked data set, almost located exactly on the red line. This behavior indicates high normality, approximately normal distribution. Compared with the second plot, we can say that the data in smoked sets is more normally distributed than the other.

Conclusion

Based on these four graphs, we find out the both data, smoked data set and unsmoked data set, have roughly the normal distribution. More specially, data in smoked sets is more normally distributed than the other.

2.3 Comparison of incidence

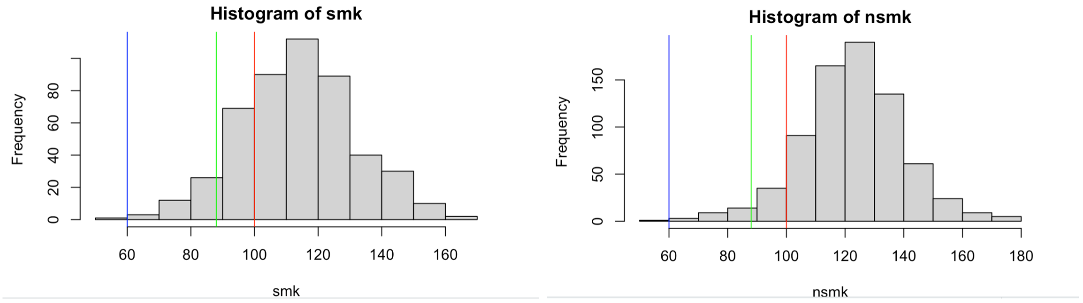
Methods

Based on the *March of Dimes*, “low birth weight is when a baby is born weighing less than 5 pounds, 8 ounces.” So in this part we define low birth weight as 88 ounces. By R, find the proportion of smoked and non-smoked mother, and the proportion of low-birth-weight babies. Then change the value of definition of low-birth-weight, see how incidence changes.

Analysis

	LBW = 88 (# of babies)	LBW = 60 (# of babies)	LBW = 100 (# of babies)	Overall proportion
Smoked	7.44 (36)	0.20 (1)	20.6 (100)	39.16
Non smoked	2.96 (22)	0.13 (1)	7.5 (56)	60.03

Note: All numbers here is percentage %



Smk represent the group of smoked mothers, and nsmk is the group of non smoked mothers

Trends and Reliability: Based on the above table we gained by R, it is pretty clear and straightforward that no matter what low-birth-weight is, the ratio, or the incidence, of low birth weight babies whose mother smokes during pregnancy is higher than those who do not smoke. In other words, the low-birth weight babies occurred more frequently among mothers who smoked. Generally, this tendency does not change with the changes of value of low-birth-weight. What's More, with the change of the value of low-birth-weight, the number of babies classified as low birth weight also changes (area). When LBW changes from 60 to 100, in both sets, the number of babies classified as low-birth-weight increases as well. So this shows that the number of babies classified as low birth weight does not affect incidence of low birth weight. Further, the value also does not affect the incidence. Finally, this also shows us that this data and defining 88 as low birth weight is reliable. Generally, mothers who smoke during pregnancy have a higher possibility having a low birth weight baby.

Conclusion

In this part of analysis, we find out the incidence of low birth weight babies. And we find out that in general, smoked mothers have a higher probability of having a low birth weight baby. This is not affected by the values of LBW and number of babies classified as low birth weight.

2.4 Asses the importance and summarize

Methods
Comparing the data set we have been analyzed in 2.1 to 2.3. See whether there are other values that could affect the data or whether the data are significant differences or not. Search for other studies online and compare the result to our data set.

Analysis
In numerical summary, we find that mean, median, and mode are all larger than those in non-smoker mothers compared to the smoker smokers. Also the standard deviation shows that the non-smoked group has less standard deviation compared to the smoked group which means that the non-smoked group data are more clustered closed around the mean. When we are using the 68-95-99.7 rule to check the area under the curve within 1,2,3 units of its center we find out that the Smoked group is more likely normally distributed than the Non-smoked group because there percentage are more nearly to the 68-95-99.7 rules. Also it is similar to the skewness and kurtosis coefficient. From those three points, we cannot really tell whether babies born to mothers who smoked will have lighter weight than those who did not smoke during pregnancy. Especially, the second analysis tells us that the smoked group is more likely to be normally distributed which means the numerically summaries are not really helpful to understand how the trend goes.

From the graphical summaries, we have used the histograms and quantile-quantiles plot to see the trend of the data. Using the red line to represent the sample normal mean and standard deviation, we could see that the non-smoked groups are more likely to be a normal distribution than the graph of smoked groups. Graphical actually provided more clearly summaries than the numerical summary because both graphs are showing the same information that the non-smoked groups are more nearly to the normal distribution than the other one but I think there is a better way to give more information.

Therefore, an incident summary came out. We are using the *March of Dimes* to find out the low birth weight is 88 ounces. Then random choice low birth weight to equal to 60 and 100 ounces. We could see that except for the low birth weight equal to 60 ounces, others show that there are more babies with smoked mothers than non-smoked mothers. The chance or we could say that the frequency of babies below the average happens more on the mother who smoked during pregnancy.

From *smoking during pregnancy and harm reduction in birth weight: a cross-sectional study*, they are coming out with a result of mean birth weight being lower than the mother who does not smoke during pregnancy. As the number of cigarettes increases the birth weight will go in the opposite direction. In this study, it said 6 to 10 cigarettes per day will make the birth weight 320 g lower in infants and 11 to 40 cigarettes per day will lower 423g birth weight in infants. By using the middle number of cigarettes they smoked per day we find out that around 18 to 40g lower in infants by one cigarette they smoked.

Conclusion

Based on the method we used to test out we find out that smoked mothers during their pregnancy could affect the babies in the low birth weight. It definitely harms their baby's weight and reduces their weights. The more they are smoked the more reduction will affect their babies weight.

<u>3 Advanced Analysis</u>

Brief introduction and analysis question

After the comprehensive analysis the difference between in weight babies born to mothers who smoked during pregnancy and those who did not. We conclude that, in general, smoked mothers have a higher probability of having a low birth weight baby. So smoking or not is indeed a variable that affects babies weight. So what's more? In this analysis, I will dig out the effect of age. Will the mother's age also affect the baby's weight?

Method

Separating ages into 3 groups, 18-30, 31-40, and41+. Then find the mean, median, mode and incidence of low-birth-weight babies for three groups, and check the normality by skewness.

Analysis:

	mean	median	mode	incidence(# babies)	skewness
18-30 (839 obs)	119.3	120	123	4.05%(34)	-0.07647
31-40 (287 obs)	118.8	119	121	7.31%(21)	-0.16930
41+ (28 obs)	120.2	120	119	3.37%(1)	0.03889

The first thing rushes out here is the number of the observations in the third group. It is pretty reasonable that elder pregnant women are seldom in real life. Since the sample size here is too tiny to be analyzed, in the following analysis, we only care about the first two groups. For mean, median and mode here, all three values are larger in 18-30 groups. Which means, in general, in this case, the majority baby weight is larger if the mother is younger. But these three values are only slightly higher than the group of 31-40. To be more precise, mean in the younger group is only 0.419% larger than the middle age group. So we can just ignore this difference. What's next, the same thing happened on the incidence of low birth weight babies. 4.05% is slightly lower than the 7.31%. But this difference is not strong enough. We can only conclude that age may be a variable that slightly affects the baby's weight. Finally, the skewness of both groups are approximately close to 0, which means they are roughly normally distributed.

Conclusion

Different from the result we get from the above analysis, we can only conclude that age may slightly affect the low-birth-weight. Elder mothers have slightly higher chances to have a low-birth-weight baby than younger mothers. The impact of age is not as strong as smoke or not.

4 Conclusion

4.1 Conclusion summary

In this analysis, we comprehensively analyzed the data set which is composed of 1236 babies. Specifically we find out the underline relationship between baby weight and the smoke behavior of their mothers. And furthermore, we also make an analysis to find out if age is a variable that could affect the baby's weight or induce a low birth weight in Advanced Analysis. With the help of R, we could comprehensively look for and compare the characteristics then get the conclusion. To be more specific, we did the analysis in order of numerical, graphical, and proportional(incidence). And then we assess the importance of the difference from these three types of comparisons. We can conclude that smoked mothers during their pregnancy could affect the babies in the low birth weight. In other words, smoking during pregnancy negatively affects the baby's weight. And also, we find out that age is not a variable that affects the birth weight heavily. We did more than the goal setted at the beginning by looking at other studies to help us insure our result.

4.2 Discussion

Our tests are only based on the boy babies, so we think we need to also test the girl babies. This biased data set may probably induce an unfair conclusion. Somehow we think the food might also be some factors that have an impact on our data.The eating habit and the food mother daily eat could possibly affect the health condition of the mother. This is also quite a significant consideration. If smoked mothers during their pregnancy eat more fast food than usually this might affect our data result. Mothers in bad health may affect the baby's growth during pregnancy. Moreover, the original data set has a nice sample size. It is appropriate to make the analysis and gain some conclusions.However, in the advanced analysis, the sample size for the third group(age 40+) is only 28. It is too tiny to do the analysis.

5 Reference

Kataoka, Mariana Caricati, et al. “Smoking during Pregnancy and Harm Reduction in Birth Weight: a Cross-Sectional Study.” *BMC Pregnancy and Childbirth*, BioMed Central, 12 Mar. 2018, bmcpregnancychildbirth.biomedcentral.com/articles/10.1186/s12884-018-1694-4.

Dolan, Siobhan. “Low Birthweight.” *Home*, Mar. 2018, www.marchofdimes.org/complications/low-birthweight.aspx.