Xueru Xie, Yichi Zhang, Jinsong Yang

Professor Schwartzman

MATH 189 Report 1

19 April 2020

Baby Birth Weights From Non-Smoking and Smoking Mothers

In week 2, the class examined data of infant death tied to premature births. Using the data given by Professor Schwartzman, we decided to see whether baby birth weights of non-smoking and smoking mothers have the same distribution. Our null hypothesis would be that maternal smokers had no effect on the weight of their birth child. Contrarily, the alternative hypothesis would be that maternal smokers did affect the birth weights of their babies such that they weighed less compared to non-smokers' babies. The data that we were given contained all the pregnancies between 1960 and 1967 in Kaiser Health Plan of San Francisco region. During that time, these data were collected by researchers because smoking was popular in society. While some people were enjoying cigarettes, some were concerned with the harm of smoking. The bigger question behind our experiment was to find out if smoking caused babies to have less birth weights. Unfortunately, our experiment could not answer this question based on the data given because it was merely an observational study. We could not manipulate pregnant mothers to conduct a randomized controlled trial in order to test the causation. Although our data contained many information about pregnant women, we did not have everyone's data as a whole. Thus, we moved away from using hypothesis testing when choosing the model. Using the data, we applied permutation testing to answer our question. Our result indicated that these two

groups, maternal non-smokers and maternal smokers, did not come from the same distribution. For the rest of this paper, we are going to talk about how we applied our model to answer our "big question".

Data

After opening our dataset in R, we noticed that it contained 7 columns and 1236 rows. The 7 columns were: birth weight of the infant (continuous), gestation day (continuous), parity (binary), age of the mother (continuous), height of the mother (continuous), weight of the mother (continuous), and smoking status (binary) with 0 being not smoke or 1 being smoked.

```
In [203]:
df <- read.table('babies.txt',head=TRUE)
head(df,10)
```

A data.frame: 10 × 7

	bwt	gestation	parity	age	height	weight	smoke
	<int>	<int>	<int>	<int>	<int>	<int>	<int>
1	120	284	0	27	62	100	0
2	113	282	0	33	64	135	0
3	128	279	0	28	64	115	1
4	123	999	0	36	69	190	0
5	108	282	0	23	67	125	1
6	136	286	0	25	62	93	0
7	138	244	0	33	62	178	0
8	132	245	0	23	65	140	0
9	120	289	0	25	62	125	0
10	143	299	0	30	66	136	1

The first step of our data analysis was data cleaning. While we assumed that the smoking status columns contained binary numbers, it was false after we examined closely. We noticed that there were numbers 0, 1, and 9 in the column.

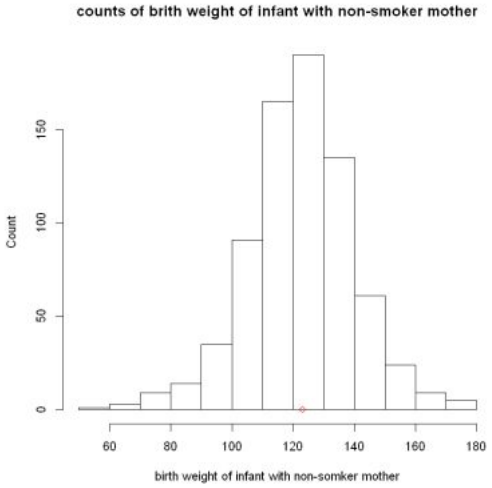
```
In [3]:
unique(df$smoke)
```

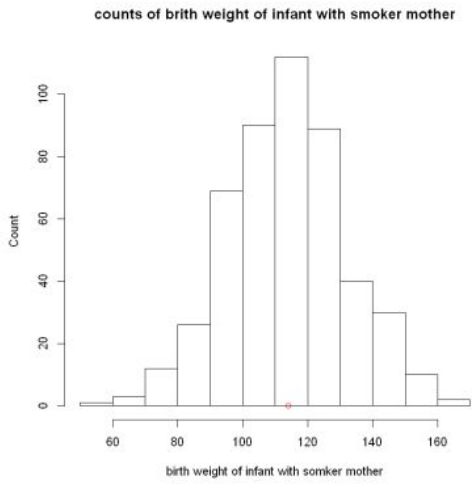
```
0· 1· 9
```

We did not understand the meaning of 9 in this context, and there were not a lot of rows that had the value 9, so we decided to delete these rows in order to test our hypothesis, although this removal might cause bias in our experiment. Since our focus was the relationship between birth weight and the status of smoking, other 5 columns were unnecessary in our research. Therefore, we decided to drop those columns in R. The next step would be checking for the abnormality of our given data. We checked for any NaN values and unusual data types, it turned out that there were not any. Thus, our process of data cleaning was finished.

Methods

Our data frame currently contained two columns: “bwt” (birth weight) and smoke (1 for yes, 0 for no). This implied that out of the 1236 rows, there were X amounts of maternal smokers that had their babies with birth weights, and Y being the amount of maternal non-smokers that had their babies with birth weights. We used plotted 2 histograms to see the distribution of the birth weight of infants with non-smoker and smoker mothers.





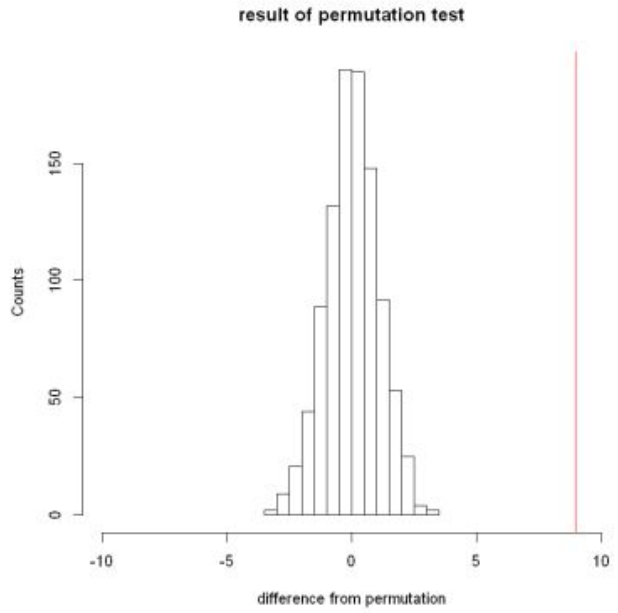
Based on the histograms above, it was obvious that the birth weight of infants with a smoker mother had less ounces compared to the birth weight of infants with non-smoker

mothers, on average. It turned out that the number was approximately 114 ounces vs. approximately 123 ounces. The difference between them was about 10 ounces, which was about 0.28 kg. This difference was the test statistic of our experiment. From a quick search on the Internet, we knew that the average weight of a baby was about 3.5 kg, so a 0.28 kg difference stood out to us. We began to question, was this number happened due to random chance, or there was a significant difference between the baby birth weights of smoking and non-smoking mothers? To answer this question, we applied a permutation test to find out. Before we decided to use the permutation test, we did consider using the hypothesis test. However, since both samples had fixed numbers, we could not simulate any new birth weights based on our data. Our plan for conducting a permutation test was to shuffle the column of birth weights. If we believed that the null hypothesis was that the difference of 10 ounces happened due to chance, then shuffling the birth weights would still result in a similar difference of 10 ounces.

Analysis

After we shuffled the birth weights, we calculated the difference of the average weights for non-smokers and smokers. After first shuffling, we found the difference to be about 1 ounce. In order to test our hypothesis, we needed to conduct this shuffling method many times until there would be a normal distribution. We also needed to consider the limit memory of R, so we did not repeat the shuffling too many times. Our shuffling number was determined to be 1000 at

the end. We created a histogram to aid our findings. Here is the result of 1000 shuffling below.



Results

From the histogram above, there was a normal distribution with the mean of approximately 0 ounces. This meant that after 1000 shufflings, the average differences were close to 0 ounces, with the extremes of about -4 to 4 ounces. During this 1000 simulations, none of the average differences were any close to our test statistics. This meant that no matter how many times we shuffled the column, the difference of baby birth weights between non-smoking and smoking mothers will never reach 9 ounces. From this, we could reject our null hypothesis and favor the alternative hypothesis, which was that the birth weights did not come from the same distribution. Formally, we would want to use a p-value to conclude our result. However, since none of the simulation was close to the test statistic, the p-value would become 0, and we would use it to reject the null hypothesis.

In conclusion. Using the given data and the method of permutation testing, we were confident to say that the birth weights of maternal non-smokers' babies did not come from the same distribution as the birth weights of maternal smokers' babies. This result did not mean or imply any causation between smoking and birth weights. This result only meant that there was a difference between the weights of the babies for two types of mothers. Our experiment was an observational study, meaning that the researchers like us did not manipulate the variables of the data. However, we could suggest performing experimental studies based on our result. We could suggest the need for a randomized controlled trial in order to find out whether smoking would cause lower birth weight. If there is a causation between smoking and lower birth weight, then researchers could use the result to promote no smoking advertisements to mothers before conceiving a child.

Appendix

In this report, student Xueru Xie wrote the rough draft of this report. Student Yichi Zhang had the responsibility to write all the R codes and export them into a PDF. Student Jinsong Yang polished the rough draft and turned it into the final draft. For the coding PDF, we conducted a histogram that combined two other histograms for our own purposes.