The Effect of Maternal Smoking on Infant Birth Weight and Health

Math 189 - Investigation #1

Professor Schwartzman

Spring 2020

Arthur Chang - A14410373

Raya Kavosh - A14826756

Siddharth Saha - A15572442

Contributions

Raya - Formulated research questions and methodology based on background information and relevant studies. Aided in analysis and interpretation of graphs, tests, and results. Structured introduction and conclusion.
Arthur - Tested and analyzed normality of groups (interpreted skewness, Shapiro-wilk and kurtosis). Wrote code to generate all graphical comparison of the data set and ensure they are easily understandable.
Siddharth - Wrote an initial body for the report and came up with initial methods of graphical and statistical analysis(graphical analysis was then distributed to the rest of the team). Wrote all code for the z tests as well as conducted the incidence analysis for our 2nd research question

Introduction

Epidemiological studies indicate that smoking during pregnancy is responsible for a decrease in birth weight and that smoking mothers are twice as likely as non-smoking mothers to have a low-birth-weight baby, defined as below 2500g. It is also known that babies with lower maturity, measured by gestational age and birth weight, have lower survival rates.

Contradictingly, data from Child Health and Development Studies (CHDS) collected of all pregnancies that occurred between 1960 and 1967 among women under the Kaiser Health Plan in the San Francisco region led to the unexpected finding that babies of smokers do not, in fact, have a higher death rate than babies of non-smokers.

While the association between birth weight and infant mortality is strong and well-established in the field of epidemiology, the cause of this relationship is unclear as it may be influenced by many confounding variables. We want to know if maternal smoking in particular leads to higher rates of infant mortality through its effects on birth weight.

We will be using an enlarged portion of the CHDS dataset to determine if there is a meaningful difference between the birth weights of babies born to smokers and those born to non-smokers, as well as whether or not this difference has consequences for the health of the baby.

Background Knowledge

While some babies may remain in utero for 42 weeks (294 days), the typical gestation period for an infant is 40 weeks (280 days). A delivery before 37 weeks is considered preterm (259 days). Most newborns range from 5.5 to 8.8 pounds (88-140.8 oz).

It is thought that smoking reduces birth weight and gestational age by reducing the amount of oxygen that can reach the fetus. This in turn may cause the placenta to expand and break away from the uterine wall, resulting in preterm delivery and/or fetal death.

Given this information, we would expect to find a significant difference between the birth weights of babies from smoking mothers and nonsmoking mothers.

Research Questions

I. Given our dataset, is there a significant difference between the birth weights of babies from smoking mothers and nonsmoking mothers?

To answer this, we use various methods of visualization to compare the distributions of birth weight for babies from smokers and nonsmokers, then run a 2-sample z-test between the two populations, grouping by gestational age, to verify a difference in their distributions.

II. Is there sufficient evidence that birth weight is related to infant mortality, considering confounding variables?

We run z-tests on a few more thresholds of low weight so that we can see how much smoking affects a babies' survival chances. However, to declare a generalizable and meaningful relationship between birth weight and infant mortality, further data is required and possible confounds must be addressed.

Data

Our dataset consists of 1236 single-birth, male babies that lived for at least 28 days after birth. The following features were recorded for each baby:

- Birth Weight (ounces) - birth weight of the baby
- Gestation (days) - number of days fetus was in utero before delivery
- Parity (count) - number of times mother has given birth to fetus beyond 24 weeks
- Age (years) - age of mother
- Height (inches) - height of mother
- Weight (pounds) - weight of mother prior to pregnancy
- Smoke (binary) - 0 if mother is nonsmoker, 1 if mother is smoker

The values 9, 99 and 999 present in the columns indicate "unknown".

Analysis

Is the difference in baby weights between smokers and non smokers just a coincidence?

To answer the above question we first clean our dataset. After dropping rows where smoke is 9 and replacing the other unknown values with the means of their respective columns we get the below numerical summaries

Statistics of Smokers Baby Birth Weight

	bwt

Statistics of Non-smokers Baby Birth Weight

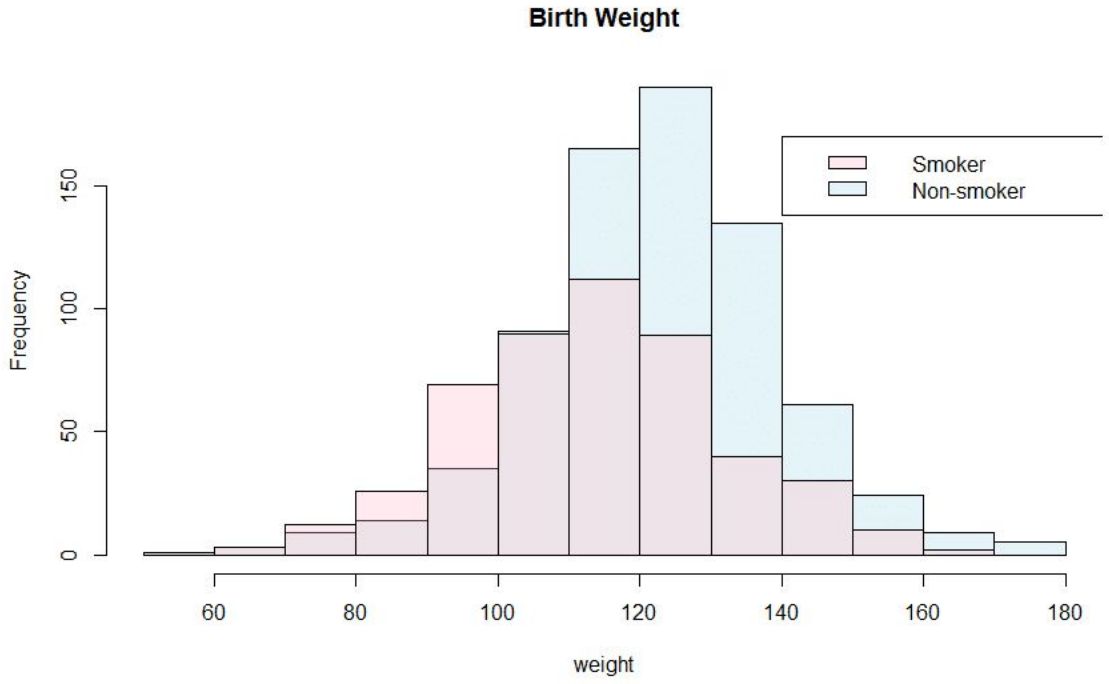
	bwt

Min	58
1st Quartile	102
Median	115
Mean	114.1
3rd Quartile	126
Max	163

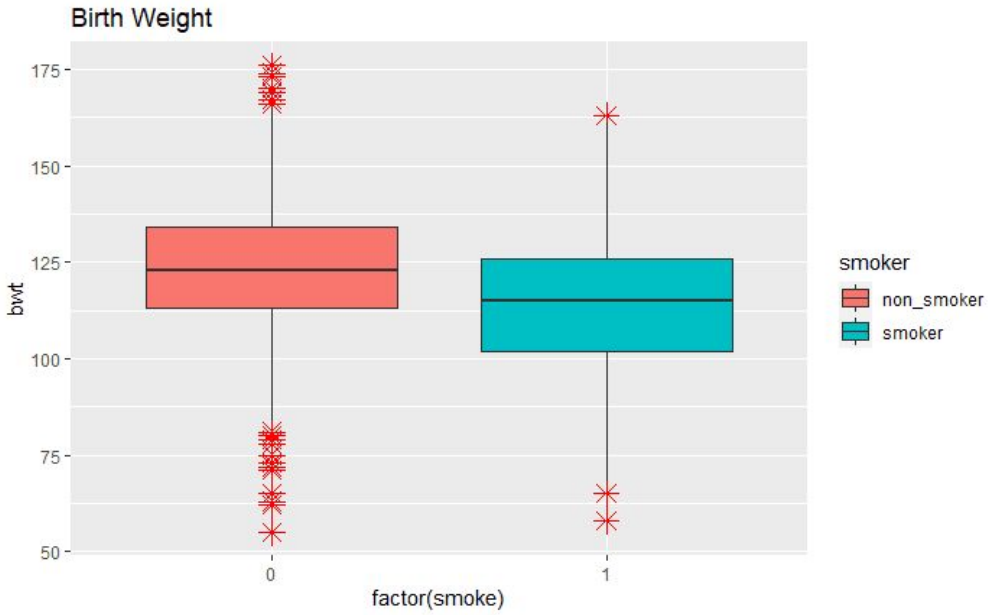
Min	55
1st Quartile	113
Median	123
Mean	123
3rd Quartile	134
Max	176

From the statistics above we can see that babies of smokers, on average, do have a lower weight than the babies of non smokers. However, the lowest baby weight of a smoker is higher than the lowest baby weight of a non smoker. Since most of the summary statistics show that babies of smokers generally have lower weights than the babies of non smokers, we can rule the minimum out as an outlier.

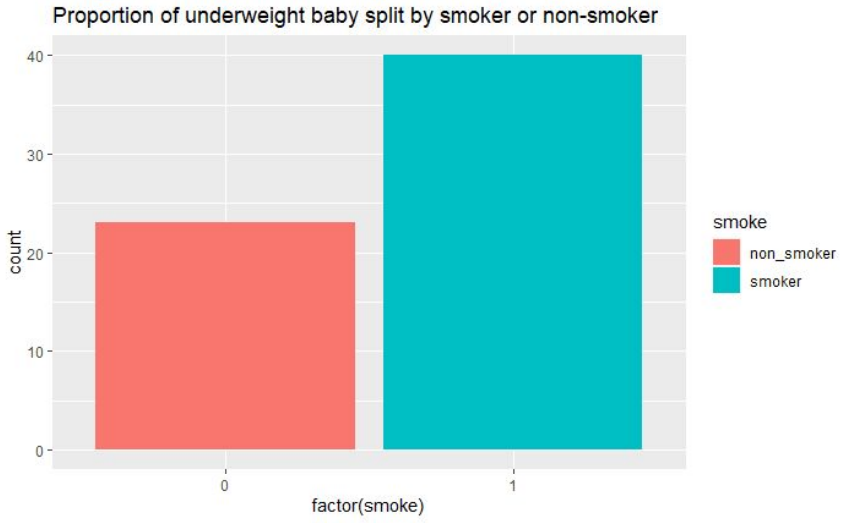
Moreover, the mean and median are almost the same for both smokers and non smokers, indicating that there may be a symmetric distribution. This can be verified through a histogram:



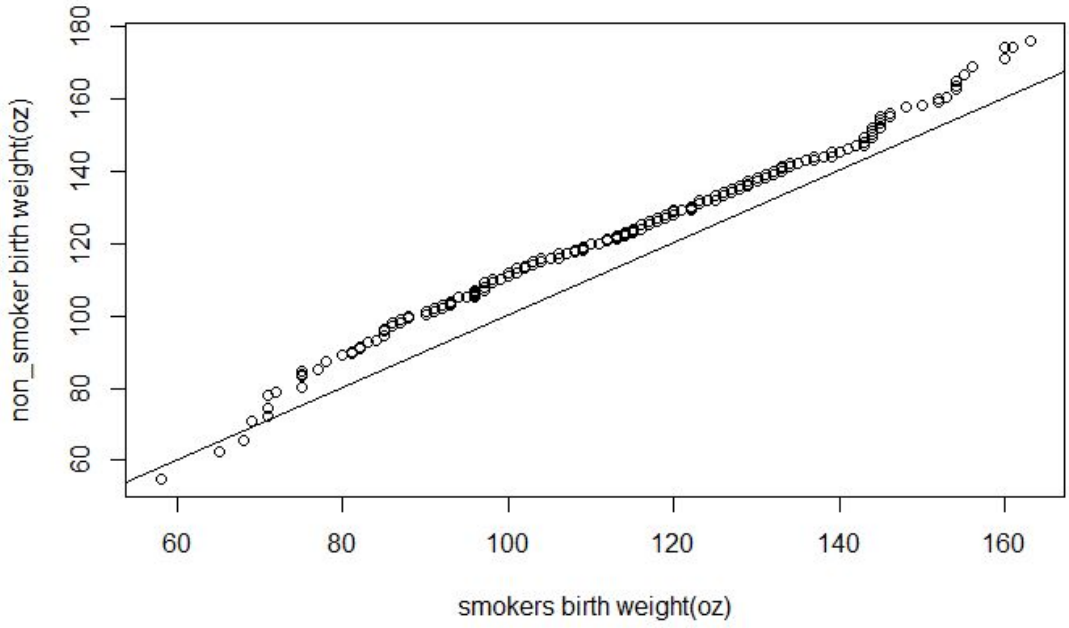
The histogram shows that both of the distributions are roughly symmetric. However, there are almost double the occurrences of babies at higher weights for non smokers then there are for smokers. It also shows that the smokers tend to have more lower weight babies than the non-smokers.



This boxplot of baby weights of smokers and non-smokers helps reinforce the histogram and numerical summaries above. The non smokers tend to have a much higher birth weight. In fact, baby weights below 87.5oz are more than likely occurrences of outliers in the non smoker group, in contrast to the smoker group.



The above bar chart shows that the proportion of low weight babies is much higher in mothers who smoke than mothers who do not smoke.



The qq plot above demonstrates the distribution for both smokers and non-smokers are similar based on the line. However, they have different means and standard errors. To get a better idea of the distribution, we run skewness and kurtosis tests.

	Skewness	Kurtosis
Smokers	-0.03359498	2.988032
Non-smokers	-0.1869841	4.03706

The skewness values we get indicate that our data is slightly left skewed. While this does not affect the normality of the smokers group (they maintain a kurtosis value close to 3) it does affect the normality of the non-smokers group. This change in normality for the non-smokers group could be because of various factors such as confounding variables and bias present inside our sample. We run normality tests below to confirm our findings

```
        Shapiro-Wilk normality test

data:  non_smokers$bwt
W = 0.98812, p-value = 1.017e-05
```

```
        Shapiro-Wilk normality test

data:  smokers$bwt
W = 0.99711, p-value = 0.5542
```

The above normality shows babies under smoker mom have approximately normal distribution (p-value > 0.05). Contrastly, non-smoker's babies do not have a normal distribution. (p-value < 0.05).

It is possible that babies could have had a lower weight due to other factors. For example, their lower weight could be attributed to them spending less time in the womb and therefore having a shorter gestation period. To control for this, we group the data according to the quartile of gestational age they belong to. This allows us to more clearly verify whether the baby's weight actually depends on the mother's smoking habits.

We then ran a 2-sample z-test using the proportions of babies classified as low weight in each gestation age group for smokers and non smokers. This allows us to verify that the findings from the various visualizations above are not just because of a coincidence but have statistical significance. We use the alternative hypothesis that the proportion of low weight babies in the smoker group is greater than the proportion of low weight babies in the non-smoker group. Our null hypothesis is that there is no such difference and that the proportion of low weight babies in both groups are roughly equal. With the threshold for being classified as low weight being under 2500 grams we get:

Gestational Age	199-250	250-302
P-Value	0.0376	0.0007

Both of the p-values we get are extremely low. Using a significance level of 0.05, we reject the null hypothesis. We have found statistically significant evidence that the proportion of low weight babies in the smoker group is greater than the proportion of low weight babies in the non-smoker group.

It is important to note that we cannot conclude with certainty that smoking causes the mother to have low weight babies. This test simply implies a correlation between smoker status for mothers and birth weight. The cause of this correlation is unclear, as it may be caused by a multitude of confounding factors including those we have in our dataset, such as the mother's weight and age. Adjusting for all of the confounding factors in the study is difficult given the data we have access to. A stricter data collection process may limit the confounding factors, and is crucial for obtaining a better understanding of the correlation. Moreover, all of our data is collected from one hospital in one city, which may negatively impact the generalizability of the findings.

How does this relate to a baby's health?

When analyzing the relationship between birth rate and mortality in their 1992 study, Wilcox and authors advocated grouping babies by their gestational age, or by their relative birth weight. This is to account for the right-shifted mortality curve of smokers (relative to nonsmokers) caused by the tendency of babies born to smokers to be smaller. If they had been smaller, but otherwise healthy, the standardized curves should have coincided, but they did not. For this reason, they found that the mortality rate of smokers was higher.

Similarly to the way we grouped by gestational age in the z-tests above to verify a significant difference in weight distributions, we run z-tests on a few more thresholds of low weight so that we can see how much smoking affects a babies' survival chances. If the p-values remain low, it means that smoking will almost always affect a baby's survival chance. However, if we reach a threshold where smoking no longer has an impact, it would imply that the threshold of 2500g needs to be well researched so that we can be sure of what threshold we should use to determine a baby's health. Setting the thresholds at around 2300g and 2700 grams we get:

Gestational Age	199-250	250-302
P-value for 2300g	.131	.301
P-value for 2700g	.047	.0000006

We notice that upon decreasing the threshold, the p-value increases drastically. However, the p-value reduces sharply among the higher gestational age groups. This discrepancy in p-values for various gestational age groups may be explained by the variation in the number of data points contained in each gestational group as displayed below:

	199-250	250-302
Non smoker	30	665
Smoker	24	437

The above demonstrates that we need more data for the gestational age groups besides the 250-302 group. In fact, we had to exclude two groups due to them having no low weight babies. For more details, refer to the Appendix to find more details.

Conclusion

In order to find out if there is a significant difference between the birth weights of babies from smoking mothers and nonsmoking mothers, we used various numerical statistics and visualizations to compare the distributions. We initially observed that there were almost double the occurrences of babies at higher weights for the nonsmoker group than for the smoker group. The smokers generally appeared to have a higher proportion of low weight babies than the non-smokers.

This was verified through our 2-sample z-test using the proportions of babies classified as low weight in each gestational age group for smokers and non smokers. We found statistically significant evidence that the proportion of low weight babies in the smoker group is greater than the proportion of low weight babies in the

non-smoker group. We then ran another z-test on more thresholds of low weight to further analyze how much smoking actually affects mortality rate.

The grouping methodology used in our z-tests resulted from consideration of related prior studies. In particular, the Wilcox and authors' 1992 study of the relationship between birth rate and mortality advocated grouping babies by their gestational age, or by their relative birth weight to adjust for the right shifted curve of smokers' baby weights. Even after standardizing the groups, they found that the mortality rate of smokers' babies was higher.

In our z-test, the thresholds of low-weight used caused a fluctuation of p-values compared to the first test. However, this was likely due to the uneven distribution of data points in each group. Some groups had several hundred samples while some had to be dropped due to having no low-weight babies. For this reason, the z-test results cannot be used to draw any meaningful conclusions between smoking and mortality rate.

This is an example of one of the major drawbacks of our data, which is that it is severely limited in scope. There are not enough samples to accurately represent different groups. Additionally, all of our data was collected from one hospital in one city, which may further negatively impact the generalizability of the findings. Finally, there are several other possible confounds that could be causing the trends we observed, such as the mother's age, lifestyle, health, etc.

While there is a significant difference between the birth weights of babies of smokers and nonsmokers, further data collection with stricter guidelines and exploration is required to prove a specific association between smoking and mortality rate.

Appendix

Numerical summaries of smoker and non smokers:

Smoker

	bwt	gestation	parity	age	height	weight	smoke
Min	58	223	0	15	53	87	1
1st Quartile	102	271	0	22	63	112	1
Median	115	279	0	26	64	125	1
Mean	114.1	278.1	.25	26.73	64.11	127.9	1
3rd Quartile	126	286	.25	30	66	140	1
Max	163	330	1	43	72	215	1

Non smoker

	bwt	gestation	parity	age	height	weight	smoke

Min	55	148	0	17	56	89	0
1st Quartile	113	273	0	23	62	115	0
Median	123	281	0	27	64	127	0
Mean	123	280.3	.2615	27.56	64.02	130.2	0
3rd Quartile	134	289	1	31	66	140	0
Max	176	353	1	45	71	250	0

Running 2-sample z-tests:

We omit the 148-199 gestational age groups and 302-353 gestational age groups due to the few number of low weight babies in those categories affecting their reliability in the z test

The 1st table below gives us number of low weight babies in each gestational category(as defined by threshold of 88.2 ounces/2500g) while the second table just gives us the number of babies we have for each gestational category:

smoke <int>	(148,199] <dbl>	(199, 250] <dbl>	(250,302] <dbl>	(302, 353] <dbl>
0	0	8	15	0
1	0	13	27	0

smoke <int>	(148,199] <dbl>	(199, 250] <dbl>	(250,302] <dbl>	(302, 353] <dbl>
0	2	30	665	45
1	0	24	437	23

Where 0 indicates not smoking and 1 indicates smoking in the smoke column