Impact of Maternal Smoking on Infant Birthweight

Author 1 and Author 2

2024-10-10

Header

Author Contributions

Author 1: Worked on questions #1, #3, and #5, and created the data analysis template for the homework. Additionally, Author 1 worked on the advanced analysis question, creating the visualization for the problem and describing it.

Author 2: Worked on questions #2 and #4, and did additional analysis for identifying outliers. Also created R file and did code documentation.

Use of GPT

ChatGPT was used as a substitute for documentation for R. Since we were unfamiliar with R, we asked ChatGPT how to use R in certain methods in order to find and filter out conditions in the dataset. We additionally used GPT to analyze reasoning and to confirm what we thought was correct about the dataset, as well as to identify extra questions that could be answered for our advanced analysis.

Introduction

The data provided is a Child Health and Development Studies dataset, which consisted of all pregnancies that occurred from 1960-1967 among women with the **Kaiser Health Plan** in Oakland, CA. Some important things to note are that all 1236 babies in the dataset are boys, there are no twins, and all lived at least 28 days. It's important to keep in mind that this is not classified as a simple random sample of all pregnancies, because the conditions just posed cannot be proven to be a totally random sample of all babies born to mothers. However, we are still studying this data beause it still should be a decent representation of differences in weight between babies born to mothers who smoked during pregnancy and those who didn't, even if it is not totally representative of all babies.

Main Research Questions

1. What are the numerical distributions of the birth weight for babies born to women who smoked versus those who didn't smoke?
2. Is there a significant difference in these two distributions? If so, what type of conclusion can be reached?
3. What percentage of babies born between these two groups (non-smoking mothers and smoking mothers) are considered low-birth-weight babies? Is there a difference?
4. How does the reliability of the three types of comparisons - numerical, graphical, and incidence - change based on our data, and which was the best?

Outline

The remainder of the report will go through a basic analysis of the data, including our cleaning methods, basic analysis on various variables in the study, and more. Additionally, we will analysis the questions posed above, along with the conclusions that we came up with in our data. We will also pose an advanced analysis question based on the relationship between the gestation period of a pregnancy and see how the relationship between a child's birthweight and their mother's smoking-status are intertwined.

Basic Analysis

1. Data Processing and Summaries

Methods

In order to analyze our data, we first have to understand it. To do that, we read the data and got a basic summary of each variable to see what we were working with. Through this, we noticed that the variables 'parity' and 'smoke' were both binary variables, representing a True/False statement.

```
data <- read.table("babies.txt", header = TRUE)

bwt_description <- summary(data$bwt)
bwt_description
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    55.0   108.8   120.0   119.6   131.0   176.0
```

```
gestation_description <- summary(data$gestation)
gestation_description
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   148.0   272.0   280.0   286.9   288.0   999.0
```

```
parity_description <- summary(data$parity)
parity_description
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.2549  1.0000  1.0000
```

```
age_description <- summary(data$age)
age_description
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   23.00   26.00   27.37   31.00   99.00
```

```
height_description <- summary(data$height)
height_description
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   53.00   62.00   64.00   64.67   66.00   99.00
```

```
weight_description <- summary(data$weight)
weight_description
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      87     115     126     154     140     999
```

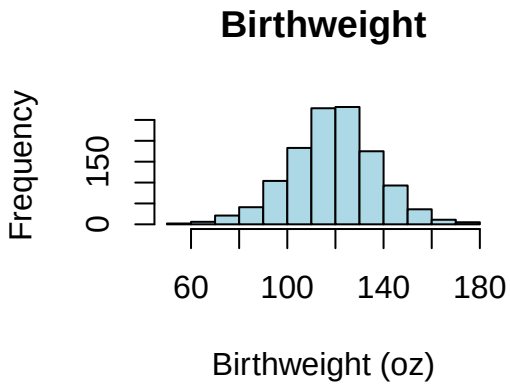
```
smoke_description <- summary(data$smoke)
smoke_description
```

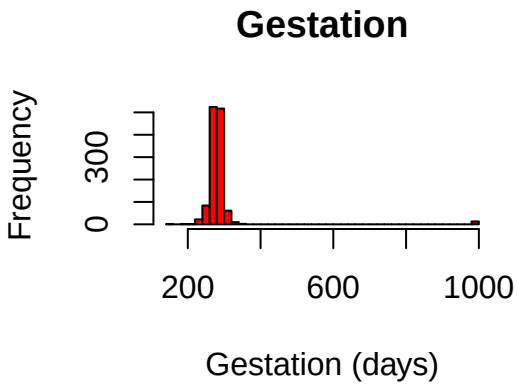
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.4644  1.0000  9.0000
```

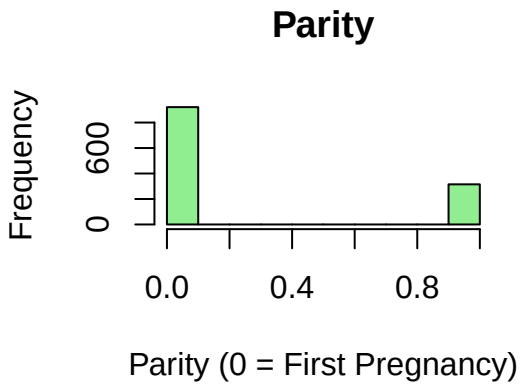
We also looked through the type of each variable, to see if we were mostly working with numerical, categorical, or a mix between the two. As seen below, all of the variables are of type 'int', meaning most are numerical variables. However, as we pointed out above, since 'smoke' and 'parity' were binary variables, these are categorical.

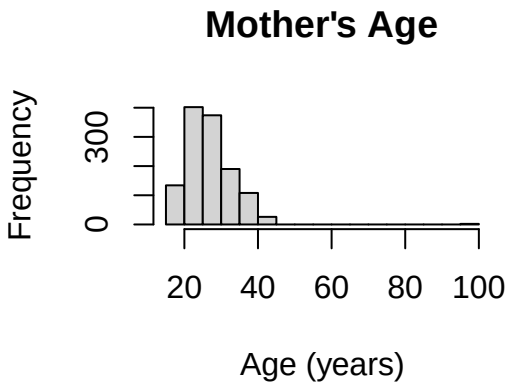
```
## 'data.frame':    1236 obs. of  7 variables:
##  $ bwt      : int  120 113 128 123 108 136 138 132 120 143 ...
##  $ gestation: int  284 282 279 999 282 286 244 245 289 299 ...
##  $ parity   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age      : int  27 33 28 36 23 25 33 23 25 30 ...
##  $ height   : int  62 64 64 69 67 62 62 65 62 66 ...
##  $ weight   : int  100 135 115 190 125 93 178 140 125 136 ...
##  $ smoke    : int  0 0 1 0 1 0 0 0 0 1 ...
```

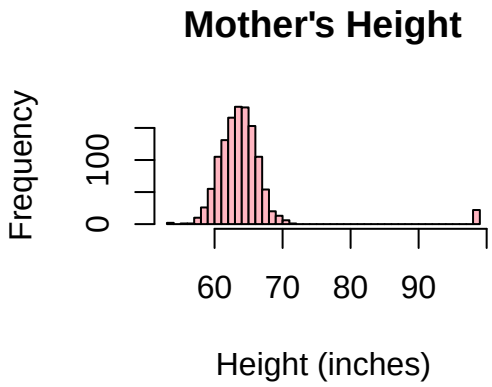
We also plotted histograms of our data to observe what type of distributions we were working with. Through this, we were able to notice that some of the histograms that we graphed had what we interpreted as outliers, which were 'gestation', 'age', 'height', 'weight', and 'smoke'. Otherwise, for a variable like 'birthweight', we noticed it was roughly normal, while parity was bimodal (since it is a binary variable).

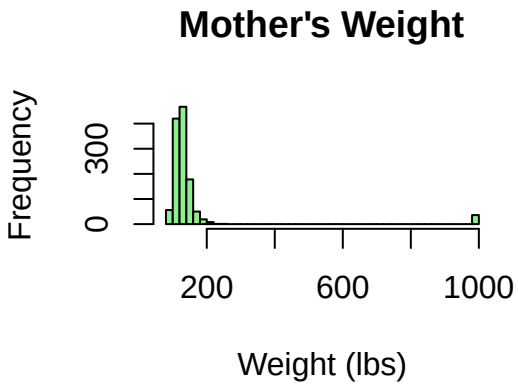


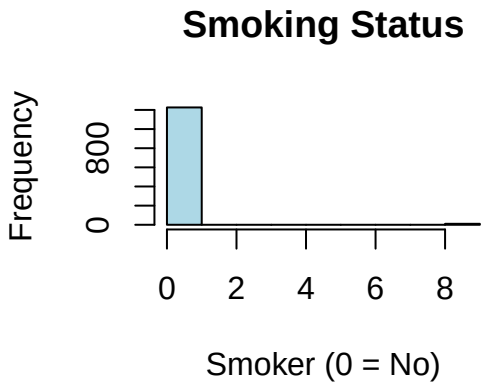












In order to clean the dataset, we handled null values and outliers. Counting nulls in the dataset revealed that there were **no null values**. Outliers in the smoking status variable can be removed because any values other than 0 and 1 make no sense in context. A hypothesis test is conducted to determine if the actual observed skewness is unusual enough for us to reject the **null hypothesis**: **the observed data, including the outliers, are generated by a Gaussian distribution**. We use a **significance level of 0.05**. This process is done for each numeric variable (so excluding smoking status and parity).

```
## Observed skew for gestation: 8.923922
```

```
## The p-value for the gestation hypothesis test is: 0

## Observed skew for age: 2.583689
## The p-value for the age hypothesis test is: 0

## Observed skew for height: 4.853647
## The p-value for the height hypothesis test is: 0

## Observed skew for weight: 5.423906
## The p-value for the weight hypothesis test is: 0
```

Since gestation, age, height, and weight all had significantly unusual skews, we reject the null hypotheses, giving us justification to remove these outliers from the dataset as shown below.

```
##   bwt gestation parity age height weight smoke
## 1 120       284      0  27     62    100     0
## 2 113       282      0  33     64    135     0
## 3 128       279      0  28     64    115     1
## 5 108       282      0  23     67    125     1
## 6 136       286      0  25     62     93     0
```

Analysis

Visual inspection of the variables' histograms gave insight that some variables had very large skews. Focusing on the portions of the histograms without extreme outliers, all distributions of **numeric variables** seemed roughly normal. To determine if the outliers were unusual enough for us to remove them from the dataset, we conducted a hypothesis test by simulating data from a Gaussian distribution 1000 times to get 1000 different sample skewness metrics. The observed skewness was compared to these simulated skewness coefficients with a significance level of **0.05**. Each numeric variable that we conducted the hypothesis test on resulted in a p-value of **0.0**, allowing us to reject the null hypothesis that the data and the outliers all came from one Gaussian distribution. Because almost all data in each variable looks normal with the exception of the outliers, and since it is unlikely that the outliers came from the same data generating normal distribution, we deemed it reasonable to remove these outliers from the dataset.

In order to generalize findings from this dataset to the general population, we must discuss its sampling method. The dataset does not qualify as a simple random sample. Assuming that we are interested in the general population, a simple random sample would mean that all births that occurred during the sample's time frame had an equal chance of being selected. Since only pregnancies that occurred under the Kaiser Health Plan in Oakland, CA were included in the data, this clearly is not the case. Additionally, only boy births were recorded with no cases of twins in the data. All recorded births resulted in babies that lived at least 28 days. All of these factors show that not all types of pregnancies are proportionally represented in the sample (particularly girl births). There is no apparent sampling method. This means we can not generalize findings from this analysis to the general population because there are too many differences between subjects represented in our data and actual individuals in the world. Our sample is not representative.

Conclusions

The goal of this section was to understand the data, data types, outliers, missing values, and then clean the data. Data types were found to be either numeric or categorical (binary). Outliers were identified, deemed significantly unusual, and removed from the dataset. The only numeric variable without significantly unusual outliers was birthweight. No missing values were found.

2. Numeric Summary of Birthweights Between Smokers and Non-Smokers

Methods

Let's find the minimum, maximum, mean, median, and standard deviation values of birthweight for these two groups.

```
##            Statistic Smokers NonSmokers
##              Minimum   58.00      55.00
##              Maximum  163.00     176.00
##                 Mean  113.82     123.09
##               Median  115.00     123.00
##  Standard Deviation   18.30      17.42
```

We also examine the quartiles of birthweights for both groups.

```
## [1] "Smoker Mothers:"
```

```
## Q1: 101 Q2: 115 Q3: 126
```

```
## [1] "Non-Smoker Mothers:"
```

```
## Q1: 113 Q2: 123 Q3: 134
```

Analysis

Comparing median and mean values can tell us about the skews of birthweights for both smoker and non-smoker mothers. Looking at the smoker-moms' children, since the median of their children's birthweight is larger than their mean (by about 1.2 ounces), the data is most likely skewed left. For the non-smoker moms, since their mean is almost the same as the median (only a 0.08 ounce difference), the distribution is roughly symmetric.

Comparing means and quartiles reveals that birthweights tend to be higher for mothers that don't smoke in this dataset. The mean birthweight from smoker mothers is about 113.82 ounces, while the mean from non-smoker mothers is 123.09 ounces.

Additionally, looking at the quartiles and standard deviations can tell us about the width of the underlying distributions that generate the data as well as variation. Interquartile range of birthweights in the smoker group is **25**, while the IQR is only **21** for non-smokers. Additionally, the standard deviation is about 0.87 ounces higher for the smoker group. Both of these pieces of evidence tell us that we observe more variation in birthweights from mothers that smoke in this dataset.

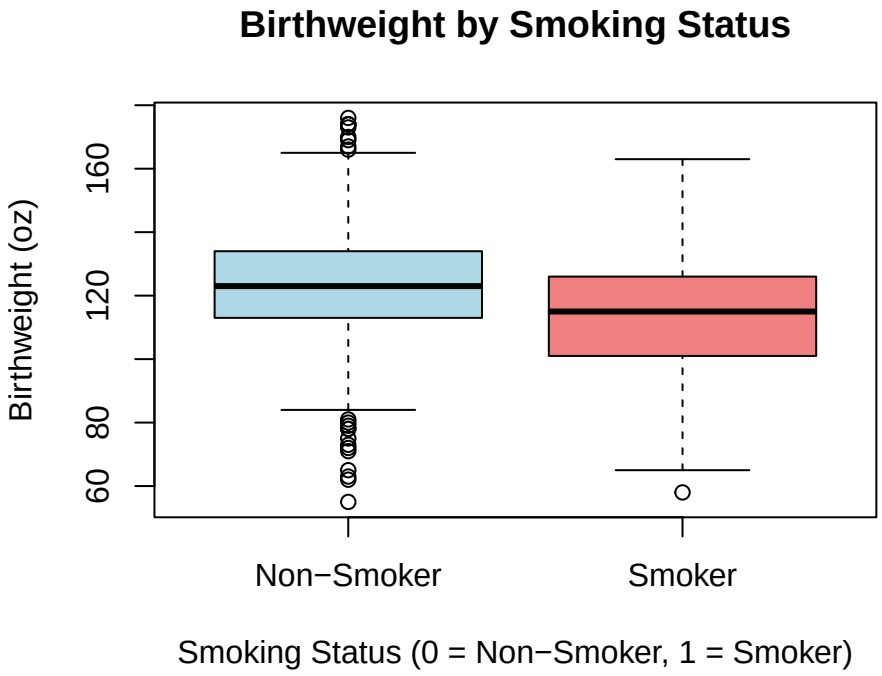
Conclusions

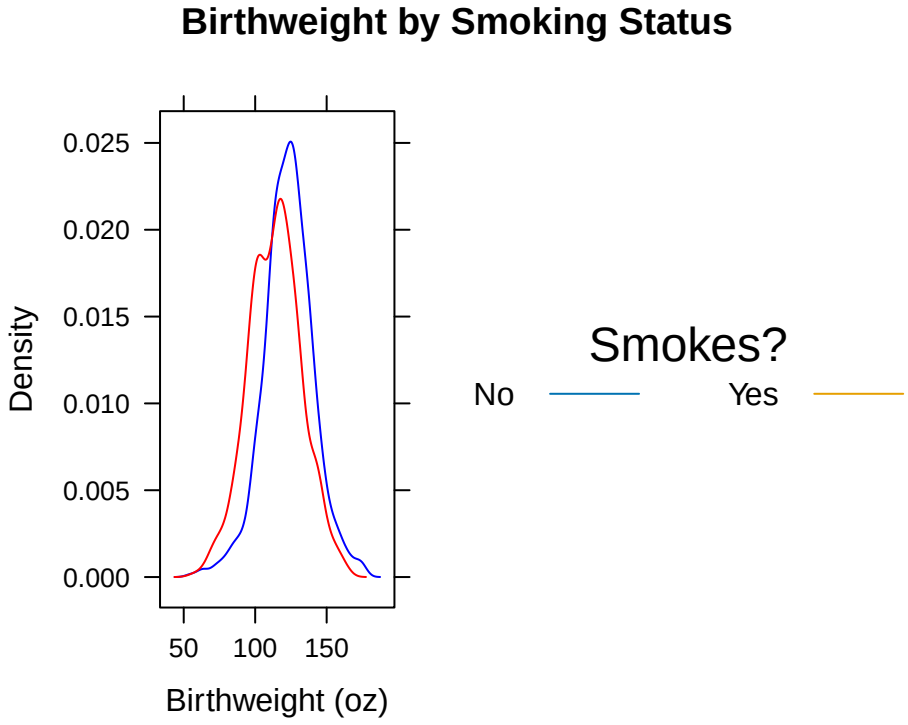
The goal of this section was to numerically summarize the distributions of birthweight for smoker and non-smoker mothers. Birthweights for the smoker group were found to be slightly skewed left, while the non-smoker group had a more symmetric distribution of birthweights. In general, birthweights were observed to be higher for the non-smoker group. More variation in birthweights were found in the smoker group.

3. Graphically Summarize Birthweights Between Smokers and Non-Smokers

Methods

To visualize the distributions of birthweights between both groups, we generate density plots and box plots.





Analysis

Inspecting the boxplot reveals that the non-smoking group has a higher minimum, first quartile, median, third quartile, and maximum birthweight than the smoking group. Additionally, the range and interquartile range of the birthweight distributions are much smaller for non-smokers, suggesting less variance among the babies of non-smoking mothers. Both of these findings agree with the earlier numerical summaries of both distributions. However, the boxplots display far more outlier birthweights for the non-smoking mothers, which is counter-intuitive because of the lower variance and interquartile range observed in the non-smoking group.

Looking at the density plot, the distribution looks unimodal for non-smokers and almost bimodal for smokers. The second, smaller peak in the density plot for smokers occurs at a lower birthweight. The width of the density plot appears to be higher for the smoking group, which agrees with the earlier finding that the smoking group has a greater standard deviation of birthweights. The distribution of birthweights for the non-smoking group looks tighter and taller, suggesting more values closer to the mean.

Conclusions

The goal of this section was to visualize the distributions of birthweights for smoker mothers and non-smoker mothers, and compare them. Findings from the previous section, like higher variation and lower average birthweights for the smoker group, were visually confirmed. Visualization revealed new findings as well. Surprisingly, more outlier birthweights were detected in the non-smoking group. The smoking group has a bimodal distribution.

4. Incidence Comparison of Low-Birth-Weight Between Smokers and Non-Smokers

Methods

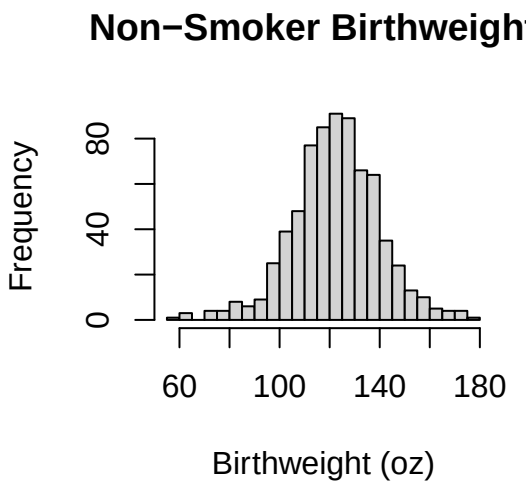
We calculate the percentage of babies that weigh under 100 ounces to women who used to smoke.

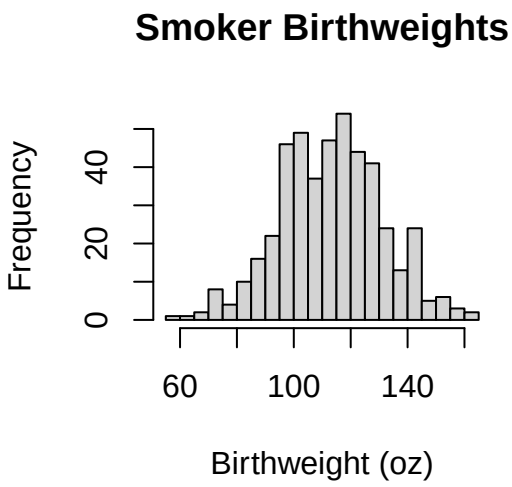
```
## low_bwt_smoker: 21.56863 %
```

We calculate the percentage of babies that weigh under 100 ounces to women who did not smoke.

```
## low_bwt_nonsmoker: 7.552448 %
```

We generate histograms for both groups to visually inspect how to proportion of babies considered “underweight” changes as we change the threshold value.





Different threshold values are tested to see how the percentages of underweight babies change for both groups.

```
## Babies under 120 ounces:
## Smokers: 0.6253
## Non-smokers: 0.4042
##
## Babies under 80 ounces:
## Smokers: 0.0327
## Non-smokers: 0.0154
##
## Babies under 60 ounces:
## Smokers: 0.0022
## Non-smokers: 0.0014
##
## Babies under 58 ounces:
## Smokers: 0.0000
## Non-smokers: 0.0014
```

Analysis

Inspecting the histograms of birthweights for smokers and non-smokers shows that as we decrease the threshold for what is considered an "underweight" birth will decrease the percentage of underweight births in both groups. This is because fewer data points will be considered underweight as we lower the threshold.

The histograms also reveal that the distributions look more similar in the tail ends and differ more in the central areas near the median. This means that as we decrease the threshold for "underweight" births, the percentages of underweight births in both groups start to look more similar. To confirm this, we re-calculated the percentages using thresholds of 120 oz, 80 oz, 60 oz, and 58 oz. The 120 oz threshold reveals the largest difference in percentage of "underweight" births between the two groups with the smoker group having 62.5% and the non-smoker group having 40.4%. The original 100 oz threshold results in 21.6% for smokers and 7.6% for non-smokers. As the threshold decreases further, the percentages for both groups start to look more similar with a threshold of 60 oz resulting in 0.2% for smokers and 0.1% for non-smokers. When a threshold of 58 oz is set, the percentage of underweight births actually becomes **larger** for non-smokers. Thus, changing the threshold can completely reverse the comparison of underweight birth percentages between the two groups. As the threshold decreases, differences between the groups become more difficult to detect.

Conclusions

The goal of this section was to compare the frequency of underweight births between smoking mothers and non-smoking mothers. With a threshold of 100 oz, it was found that about **21.6%** of smoking mothers' births were underweight and **7.6%** of non-smoking mothers' births were underweight. As the threshold decreases, the difference between these two percentages becomes smaller and the comparison becomes more difficult to make.

5. Evaluate the Three Comparison Methods (Numerical, Graphical, Incidence)

Analysis

Numerical comparisons give exact values for statistics, allowing for very precise measurements of the data that we have. Comparing the statistics for both groups of mothers is very straightforward. On the other hand, it's more difficult to determine the exact shape and distribution of data, since we only used quartiles, averages, and standard deviations, which is much more coarse than viewing a full density plot.

The graphical approach revealed that the distribution for birthweights from smoking mothers was bimodal, while the distribution for non-smoking mothers was unimodal. This information is important to the comparison as it adds more context to how much variation was occurring in the smoking group. A disadvantage of using a graphical approach is the lack of quantitative metrics to compare distributions. Statistics like standard deviations, means, and medians are objective descriptions of distributions, but graphs can leave room for interpretation.

Finally, the incidence comparison allowed for a more well-defined question to drive the comparison between smoking mothers and non-smoking mothers. Concrete questions like these are simpler to answer because the answers are quantitative and target a specific difference between distributions rather than a vague one. However, our incidence comparison relied on the definition of an "underweight" birth, which had room for interpretation, leading to inconsistent analysis depending on what was chosen for the underweight threshold. For instance, if we defined low-birth-weight as a birthweight under 80 ounces, the difference between the two groups would seem negligible. If low-birth-weight was defined as below 100 ounces, however, the difference would suddenly seem significant.

Conclusions

The goal of this section was to evaluate the strengths and weaknesses of the three different approaches to compare smoking and non-smoking mothers. The numerical approach was found to be more concrete and objective, but lacked the granularity to completely capture shapes of distributions. The graphical approach captured shapes of distributions well, but lacked the quantitative measures of a numerical approach. The incidence comparison posed a concrete, relevant question, but the threshold for an "underweight" birth was left up to interpretation, leading to inconsistent analysis and results.

Advanced Analysis

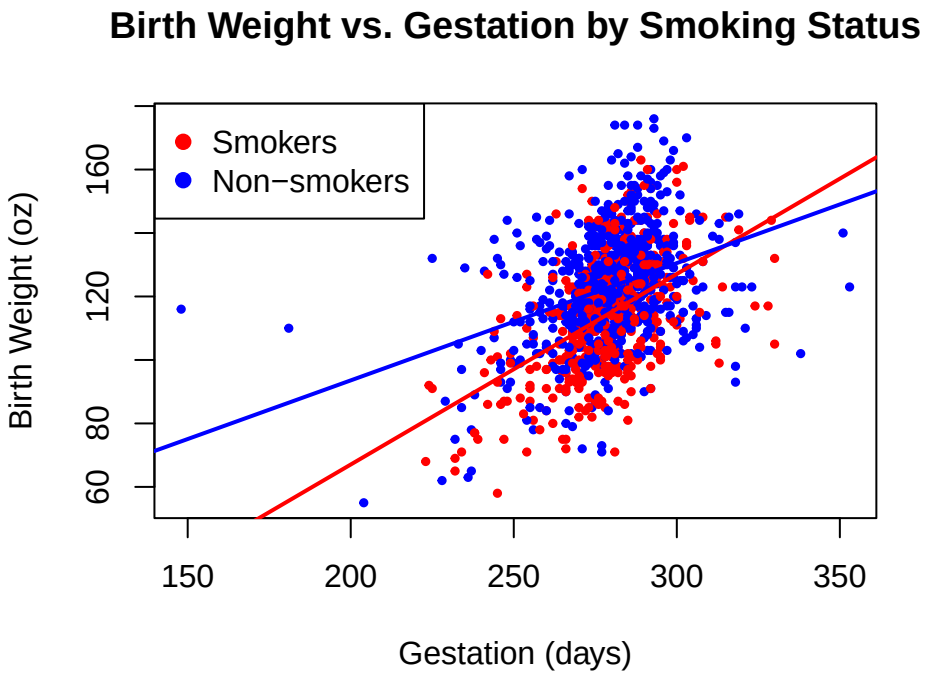
Additional Research Question

How does smoking's impact on a baby's birth weight vary depending on how long the pregnancy last?

Methods

The method we used to analyze the relationship between one of the other variables in the dataset, "gestation", and the babies' weights and parents' smoking status was by using a scatterplot. For starters, we wanted to

visually see if there was a relationship between the gestation period and the birth weight, so we used these as our x and y-axes, respectively. In order to identify the difference between smoking mothers and non-smoking mothers, we colored the dots to indicate its status. Additionally, we created a regression line for each group of mothers to visualize the relationship separately for smokers and non-smokers.



Analysis

As a general trend, both smokers and non-smokers show a positive correlation between gestation and birth-weight, indicating that the longer a gestation period is, the baby's birth weight typically increases.

We also see that the two regression lines (for smoking moms and non-smoking moms) have different slopes. It's shown that the line associated with the smoking moms has a steeper slope than than the line associated with non-smoking mothers. This suggests that for smokers, birth weight increases more rapidly with longer gestation periods for smokers.

We can also notice that over the same gestation periods, babies born to smokers tend to have lower birth weights than their respective counterparts, as observed by the fact that the red line is consistently lower than the blue line for most of the gestation period.

Another thing that can be noticed is the variability of the gestation periods, specifically for non-smoking mothers. Looking through the data, gestation periods for mothers ranged from 150-350 days, while smoking mothers had a much smaller range, typically falling from a range of 220-330 days.

Finally, the last thing to notice from this graph is that there are significantly less mothers that smoke than those that did not.

Conclusions

Through this graph, a few conclusions can be reached:

- Smoking is correlated with lower birth weight. The data shows a clear distinction between smokers and non-smokers in terms of their child's birth weight. Since the red line consistenly falls below that of the blue line for most gestation periods, it implies that on average, smoking is associated with lower birth weights.
- Gestation length is positively correlated with birth weight. This means that the longer a gestation period is, the higher the birth weight of a baby usually is.
- Non-smokers show a slower increase in birth weight compared to smokers. There can be many speculations about this, but looking at this graph, it seems that smokers typically have lower birth weights in the same gestation period as non-smokers, leading to their line needing to start at a lower area and climb back up to catch up to other babies that . This could also potentially mean that non-smokers typically have as much of a deviation of a child's birth weight as smokers do, since their slope is smaller, with many of their birth weights being concentrated, while children of smoking mother's have more variability in terms of their birth weight.

Conclusions and Discussion

Summary of Findings

The data reveals distinct trends in the birth weights of babies in relation to their mothers' smoking habits. Both non-smoking and smoking mothers show roughly normal distributions for birth weights, but babies born to smoking mothers tend to weigh less on average compared to those born to non-smoking mothers. Additionally, the variability in birth weights among non-smoking mothers is greater, with their distribution closely resembling a Gaussian curve. Furthermore, smoking mothers have a significantly higher likelihood—over 14% more—of having a baby with a low birth weight (under 100 ounces) compared to non-smoking mothers, highlighting a notable disparity in infant health outcomes between the two groups. Finally, our numerical analysis reveals that at every quantile, babies born to non-smoking mothers weigh more than those born to smoking mothers.

This analysis underscores the impact of smoking on birth weight, with clear evidence of increased risk for low-birth-weight babies among smoking mothers

Discussion

Although there are many warnings everywhere about the dangers of things like alcohol or smoking during pregnancy, many still doubt that doing these things would harm a child. However, our data and our analysis shows a clear trend that smoking can significantly harm a baby's health, specifically towards their birth weight, which may potentially lead towards other health problems in the future.

At the same time however, our analysis is not conclusive. This data cannot be considered a simple random study of all mothers in the world who have had children, because this data is only from a limited period of time (1960 - 1967), only contains male children, and does not have any twins. This data is also from a very specific part of the world (Oakland, CA) that may not be representative of the entire population that we want to study.

Additionally, there may be other confounding factors, like socioeconomic status, access to healthcare, nutrition, and other variables that could potentially influence this study. One question that could be posed is regarding the future implications of smoking on a child. Although this dataset only contains the birthweight of a child, we'd like to see if there are any long term health implications on children whose mothers smoked.