Project 1: Report

Group 17 Members:

Yaoxin Li - Introduction, Data, Investigation (Analysis), Theory

Ye Yint Win - Introduction, Background

Ruotian Gao - Investigation (Analysis), Theory

Introduction

As we know, infant deaths are related to low birth weight of babies. According to New York Times, smoking by pregnant mothers may result in fetal injury, premature birth, and low birth weight. The fetal rate of babies is highly important, and smoking seems to affect it. In other words, smoking is relevant to the health of babies. Therefore, we want to investigate whether the babies' birth weight is related to whether the mother smokes or not. To be specific, we are going to investigate the problem of the difference in weight between babies born to mothers who smoked during pregnancy and those who did not smoke.

As we mentioned, the big question and the purpose of this project is to investigate the difference in weight between babies born to mothers who smoked during pregnancy and those who did not smoke. However, we cannot simply investigate the big question in a straightforward manner but by dividing it into smaller questions. And by separating our big question into five smaller questions in this project, we can investigate the big question scientifically. Specifically, we will summarize numerically and graphically the two distributions of birth weight of babies born to mothers who smoke and born to those who do not, and, meanwhile, we will use graphical methods to compare the two distributions. Then, we will investigate the frequencies of low birth weight babies for two different groups, and we will evaluate how reliable our estimates are. Furthermore, we are going to assess the importance of our numerical, graphical, and incidence findings. Eventually, we will relate our results to the relevant studies.

For convenience, the smaller questions will be answered is listed:

- Summarize numerically the two distributions of birth weight for babies born to women who smoked during their pregnancy and for babies born to women who did not smoke during their pregnancy.
- Use graphical methods to compare the two distributions of birth weight.
- Compare the frequency, or incidence, of low-birth-weight babies for the two groups.
- Assess the importance of the difference we found in three types of comparisons (numerical, graphical, incidence).
- Summarize our findings and relate them to other studies.

Data

The data set, babies23.txt, is provided, and it contains 1236 observations and 23 columns, variables. Each observation can be considered as a description of a pregnancy.

Furthermore, since there are many variables in this data set, and many of them are not relevant to the questions that we are seeking. And therefore, the first sensible approach is to do data cleaning and we will only be talking about variables related to the questions and our analysis. In this data set, plurality columns indicate the number of babies born in a single pregnancy, and in this data set, all observations are all single fetus. Outcome variables illustrate whether the baby lives at least 28 days, and, in this data set, all observations have **1** in the outcome column which means all the baby lives at least 28 days. On the other hand, the genders included in this data set are all males. For two most important variables, birth weight of babies is recorded in **wt** columns, and the smoke status of mother is recorded in the **smoke** column. More specifically, for smoke status, **0** means never smoked; **1** means smoke during pregnancy; **2** means smoke until current pregnancy; **3** means once smoked; and **9** means unknown. Actually, there are 10 observations with 9, unknown smoke status, we mutually agreed to exclude them from our analysis. For the analysis, we are focusing on the two groups with smoke status of 0 and 1 which are never smoke and smoke during pregnancy.

For these two variables, birth weight of babies and smoke status of mothers, birth weight is a numerical discrete variable, and smoke status is a categorical discrete variable. To be specific, since birth weight is rounded to whole numbers in ounces, they are discrete instead of continuous. Meanwhile, for smoke status, as it is categorical and categorized by integers, it is categorical discrete variable.

Background

We have always heard or read in the news that cigarette smoking or smoking in general during pregnancy relates to the reduced birth weight for newborn babies. According to the article *Cigarette Smoking in Pregnancy: Its Influence on Birth Weight and Perinatal Mortality*, published by British Medical Journal, smoking during pregnancy “increased the late fetal plus neonatal mortality rate by 28% and reduced birth weight by 170g”[1] which is equivalent to 6 ounces in imperial units based on 16,994 births. Even though smoking could be an underlying factor, there could be other confounding factors that potentially lead to reduced birth weight in babies such as race, mother’s age, malnutrition, secondhand(passive) smoking, alcohol consumption, usage of illicit drugs and contract to Rubella (commonly known as German measles or red rash) during pregnancy.

The 50th percentile of birth weight of babies is 123 ounces (7 pounds, 7 ounces) and babies who are born weighing less than 88 ounces (5 pounds, 8 ounces) are considered low birth weight and one of the main causes of low birthweight is due to premature birth (born with less than 37 weeks gestation). According to the Centers for Disease Control and Prevention (CDC), smoking during pregnancy could lead to babies’ slow growth before birth and born too small even after a full-term pregnancy, and born prematurely.[2] One of the suggestions that is informed to women are expecting a baby is to quit smoking before they get pregnant or to stop any time during pregnancy. With all the assumptions and possibilities, we will be focussing on the data set of women who smoke, never smoked, smoked until pregnancy and last not but least, who once smoke to analyze how smoking during pregnancy has an impact to answer our big question of the difference in weights between babies born to mothers who smoked during pregnancy compared to those who did not.

[1] Butler N. R., Goldstein H., Ross E. M.. Cigarette Smoking in Pregnancy: Its Influence on Birth Weight and Perinatal Mortality Br Med J 1972; 2 :127

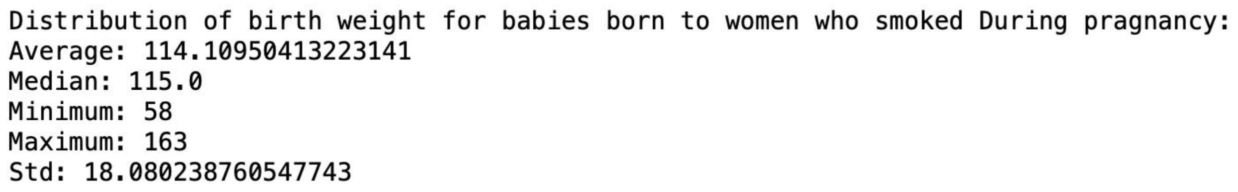
[2] https://www.cdc.gov/tobacco/campaign/tips/diseases/pregnancy.html

Investigation (Analysis)

First, we did some data cleaning to only keep the two columns we need from the 3rd dataset: “wt” and “smoke”. Then, we used numpy and pandas to calculate the numeric distribution of two groups.

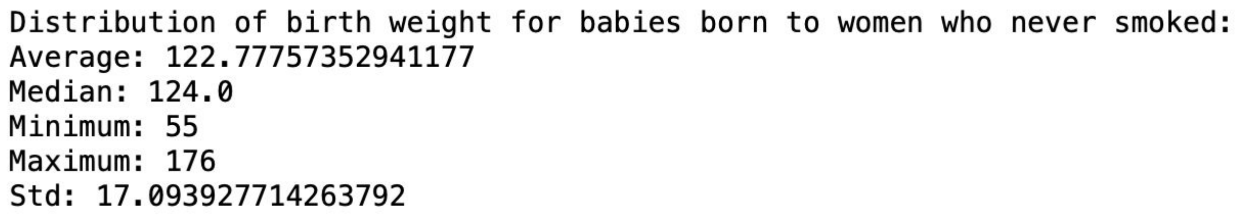
- Summarize numerically the two distributions of birth weight for babies born to women who smoked during their pregnancy and for babies born to women who did not smoke during their pregnancy.

For the numerical distribution of the birth weights of babies born to women who smoked during pregnancy, we got:



```
Distribution of birth weight for babies born to women who smoked During pragnancy:
Average: 114.10950413223141
Median: 115.0
Minimum: 58
Maximum: 163
Std: 18.080238760547743
```

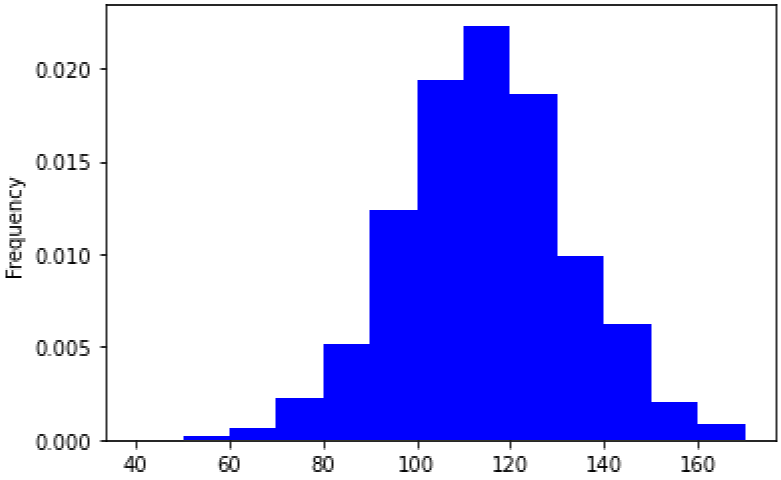
For the never smoked group, we have :



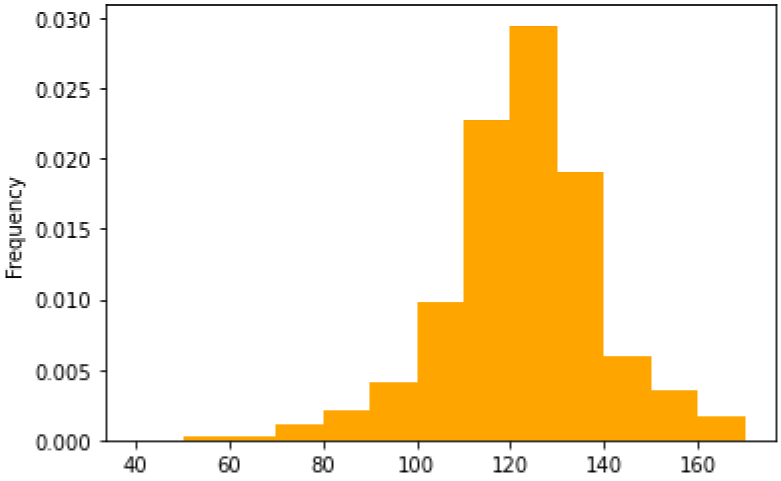
```
Distribution of birth weight for babies born to women who never smoked:
Average: 122.77757352941177
Median: 124.0
Minimum: 55
Maximum: 176
Std: 17.093927714263792
```

From the comparison of two averages from two groups, we can clearly see that the birth weight of the baby born to women who smoked are drastically lower than the babies born to never smoked women. It is an indication that smoking would have negative effects on embryogenesis and cause lower birth weight.

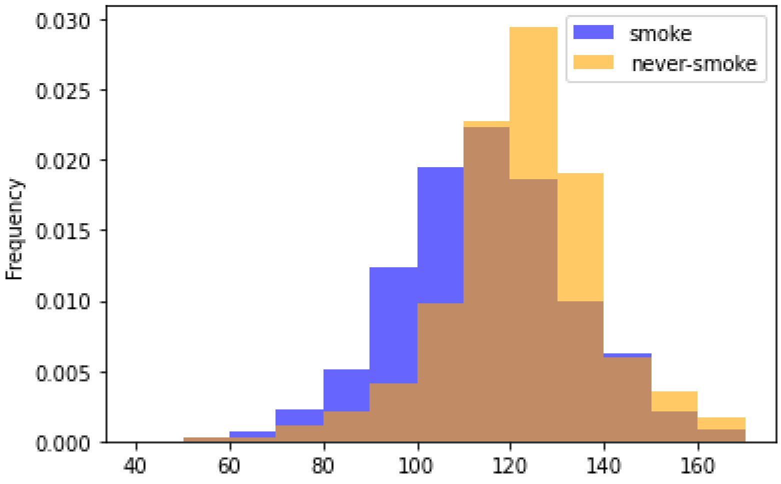
- Use graphical methods to compare the two distributions of birth weight.
  - Histogram of birth weight of babies born to smoking mothers



  - Histogram of birth weight of babies born to non-smoking mothers

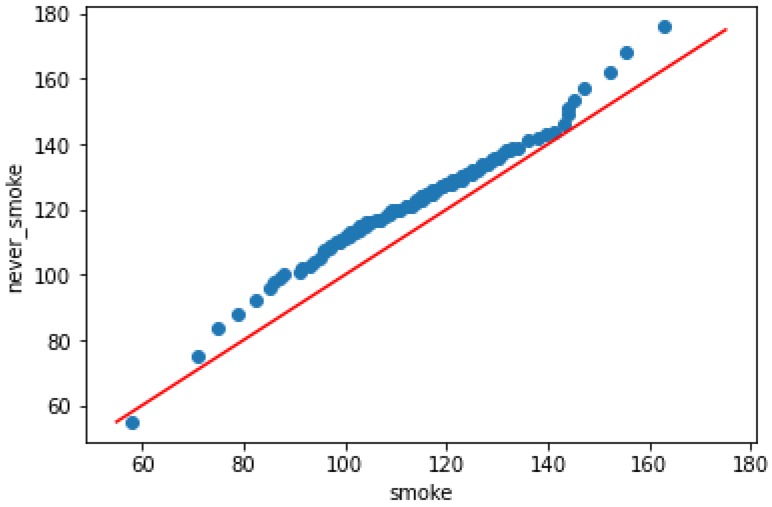


- Overlaid histogram for smoking and non-smoking mothers



- For these histograms, we use density histogram instead of frequency histogram, since the sample size of two groups are different. As we can see, two distributions both look like bell-shaped and seemingly normal distributions.
- Same to the numerical summaries, histograms demonstrates that the birth weight of the baby born to women who smoked are drastically lower than the babies born to never smoked women.
- The variation of the smoke group is greater than the never-smoke group.
- In conclusion, the distribution of two groups are pretty similar, although there are some slight differences.

- We use the quantile-quantile plot as our graphical method to compare the distributions of two groups of birth weight.
- As a result, we got the quantile-quantile plot below

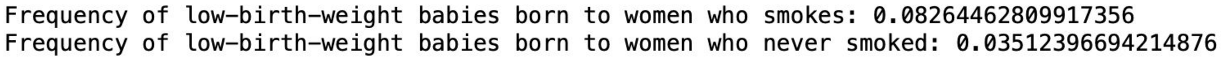


- As the graph shows, the q-q plot is close to the straight line with slope 1 and intercept 0. In other words, the distribution of two groups of birth weight are roughly the same. To be specific, the intercept of the actual line seems a little bit higher than 0, and this means there is a shift between two distributions. Meanwhile, the slope of the actual q-q plot is changing, and this indicates that the scales of the two distributions are not identical all the time. Even though there are some slight differences, the two distributions seem quite similar.
- For a quantile-quantile plot, our null hypothesis is that the distribution of two samples are the same. As the results shown above, we cannot reject the null hypothesis, since the plot has a roughly linear relationship but no intercept at 0

and the slope is not constant 1. In other words, two distributions of two groups are the same shape, but have different means and standard deviations.

- Compare the frequency, or incidence, of low-birth-weight babies for the two groups. How reliable do you think your estimates are? That is, how would the incidence of low birth weight change if a few more or fewer babies were classified as low birth weight?

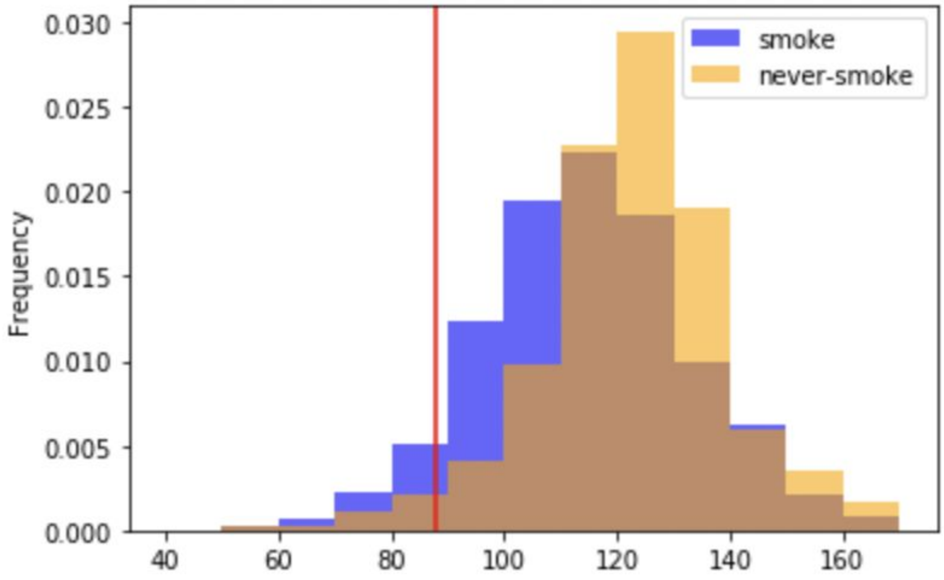
According to the lecture, the cut off of the definition of low-birth-weight babies is 2500 grams, which is 88.18 in ounces. So using this information, I calculated the frequency of the low-birth-weight babies of the women smokes and who never smoked:



```
Frequency of low-birth-weight babies born to women who smokes: 0.08264462809917356
Frequency of low-birth-weight babies born to women who never smoked: 0.03512396694214876
```

Numerically, we can see that the frequency of the low-birth-weight babies born to women who smoked during pregnancy is twice as higher as the babies born to women who never smoked.

For graphical comparison, we made a histogram plot:



In this plot, the yellow bars represent the frequency of birth weight of the babies born to women who never smoked and the blue bars represent the frequency of birth weight of the babies born to women who smoke during pregnancy, and the red vertical line is the cutoff of the low-birth-weight babies. As we can see, the frequency of low-birth-weight babies among the mothers who smoked during pregnancy is much higher than the ones never smoked.

To prove our estimate, we randomly chose 300 samples from each group and calculate the frequencies again, we got:

```
Frequency of low-birth-weight babies born to women who smokes: 0.07333333333333333
Frequency of low-birth-weight babies born to women who never smoked: 0.023333333333333334
```

So the frequency of low-birth-weight babies born to women who smoke is still much higher than the never smoked group although both frequencies dropped a little bit. Thus, our estimate is reliable.

- Assess the importance of the difference you found in your three types of comparisons (numerical, graphical, incidence). Summarize your findings and relate them to other studies.

First, the numerical data comparisons can show specific details of the datasets like mean, std, min and max, but we cannot directly see the distribution of the whole population. With graphics, we can easily spot the differences between different groups which can help a lot to estimate the result. For incidence analysis, we can get the accurate difference between each group, thus we can know how the rate varies in different groups.

To conclude, all of our findings from the data show that there is a great correlation between babies' birth weight and whether the baby is born to a woman who smokes during pregnancy. By comparing the frequency, we get the result that a baby born to a woman who smokes during pregnancy has a 5% chance of having low birth weight. When compared to other studies, for instance, CDC once stated that "One in every five babies born to mothers who smoke during pregnancy has low birth weight"[3], which is the 20% of the babies born to women who smoke would have low birth weight. But the frequency we have is 8%, which is dramatically lower than the report from CDC. I think the difference between the results is due to having different samples. Since the data we have only have 1236 observations while CDC's report are from thousands of observations. And another possible reason is that we might have different cutoffs for the definition of low-birth-weight babies. Nevertheless, although the numbers are different, we both conclude the result that smoking can drastically increase the probability of having a low-birth-weight baby, which indicates that our findings are reliable.

[3] https://www.cdc.gov/tobacco/basic_information/health_effects/pregnancy/index.htm

Theory

- **Histogram**

Histogram is an estimator of probability density function, and it provides the distribution of data. In other words, histogram reconstructs the probability density function through data points, babies weights in our projects. To be specific, our data points, babies weights, are observed data which is realization of random variables that are independently and identically distributed. In other words, random variables, X_1, …, X_n, have the same distributions (X). Thus, the histogram helps us to reconstruct the probability density function of random variables (X), and we can use the observed data, x_1, …, x_n, to estimate the quantities at the population level.

Another important part for histogram is bins. The ratio of observation within the bin is equal to density estimate times the length of the bin. Theoretically, since histogram is the estimator of probability density function, we want our number of bins to minimize both the variance and bias of the histogram. Nevertheless, increasing the number of bins decreases the bias, while increasing the number of bins increases the variance. Therefore, we want to trade-off the number of bins to minimize bias and variance of the histogram, and the optimal number of bins is where the bias and variance are of equal value.

- **Average (Mean)**

  Mean is the center of the data distribution. To be specific, data, x_1, …, x_n, are realization of random variable X, and data have the same distribution as X. Thus, the P(less than mean) of the data is equal to the P(less than mean) for the random variable X. For population, the average of the population is defined as: $\mu = E[X]$

  As μ is at the population level, we cannot know its true value. But we can estimate it by our estimator the x^{-} which is defined as: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}(x_i)$

- **Standard Deviation**

  Standard deviation measures how far individual value may vary from the center of the distribution. The standard deviation is defined as: $\sigma = \sqrt{\frac{\sum_{i=0}^{n}(x_i - \bar{x})}{n-1}}$

- **Skewness Coefficient**

  The skewness coefficient defined as: $\text{skewness} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s}\right)^3$

  It measures the skewness of the distribution, and for normal distribution skewness = 0.

- **Kurtosis Coefficient**

  Kurtosis coefficient illustrates how pronounced is the peak of the distribution, and it defined as: $\text{kurtosis} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s}\right)^4$

  For normal distribution, the kurtosis = 0.

- **Simulation study**

  We can use simulation study to check the normality of distribution. Generally, we can generate pseudo random values from a given distribution, and then check the similarity of the simulated distribution with the distribution of observed data. To be specific, for checking normality, we calculate the kurtosis coefficient of observations, and simulate data from normal distribution. The size of simulations is just equal to the size of observations. Then, we repeat the simulation many times, and we plot the histogram of simulations' kurtosis coefficients. Eventually, determine whether the observed distribution is normally distributed according to the histogram.

- **Quantile-Quantile Plot**

  We use a quantile-quantile plot as our graphical method to compare the distributions of two groups. To be specific, a quantile-quantile plot shows how two distributions are

$$
\begin{aligned}
&p \in [0, 1] \\
&C(p) = (F_x^{-1}(p), F_y^{-1}(p))
\end{aligned}
$$

  similar. The quantile-quantile plot is defined as:

  We can get the curve of the plot by plotting the two groups of data, all x_i and all y_i, into the plot. As a result, if we get a straight line with intercept 0 and slope 1, the two distributions are identical. Otherwise, they are not the same distribution. Specifically, non-zero intercept means there is a shift for two distributions, and not one slope indicates there is scale change among two distributions.

  In our analysis, we actually get quantiles of distribution of smoke mothers and non-smoke mothers. Then, we made the plot and got the q-q plot among two distributions.