Maternal Smoking and Infant Death

Joshua Castro, Kevin Elkin, Yuka Sakamaki, Sneheil Saxena, Tony Wu

Abstract -

I. INTRODUCTION

In the United States alone, cigarette smoke accounts for more than 480,000 deaths per year, including more than 41,000 deaths resulting from secondhand smoke exposure [4]. Numerous studies published by the CDC have concluded that smoking cigarettes is directly related to causing and increasing your risk for cancer, lung diseases, strokes, heart disease, diabetes, and chronic obstructive pulmonary disease (COPD) [2]. However, much of what we know about smoking and its various health risks only pertain to the user and others around them; with more medical technology available and a better understanding of scientists are starting to study the repercussions smoking while pregnant can have both on the user and unborn child.

According to the CDC, smoking and inhaling tobacco products while pregnant has been linked to causing congenital birth defects in infants. About 1 in 10 women (10.9%) who gave birth in 2014 smoked 3 months prior to getting pregnant, and approximately 1 in 4 of these women (24.2%) did not smoke during pregnancy [6] Maternal smoking was associated with a 27% excess risk of gastrointestinal abnormalities, including, but not limited to, issues with the throat, colon, esophagus, bile ducts, intestine, gallbladder, and liver [3]. Additionally, smoking has been found to decrease the amount of oxygen available to an unborn baby, increase an infant's heart rate, increase the chance of a premature birth, and increase the risk a baby is born with low birth weight [6].

This study will investigate the relationship between maternal smoking and birth weight by using a variety of statistical approaches and methods to best and most accurately analyse the data set we have gathered for use. The conclusions and findings in this report will allow for expecting mothers to make a more insightful and informed decision on whether or not smoking while pregnant is appropriate.

II. DATA

There are N = 1,236 observations on each of the characteristics of the infant, during gestation and postnatal stages, and physical and economic characteristics of the parents. Data gathered was controlled to that of male, single infants, all of whom lived to at least 28 days located within the San Francisco-Oakland Bay Area region of California.

The data we will examine are both discrete and categorical. Information collected regarding the gestation period and the infant weight are considered discrete as they are quantified characteristics. A variable of discrete nature that we will also consider is the total income of the parents. The main categorical data we will consider is the smoking status of the mother, as our primary goal is to uncover any associations between maternal smoking and the survivability of the infant.

Normal distributions about different means of infant weight between both smoking and non-smoking mothers provide slight indications of the survivability and health of the infant post birth considering how infant survivability has some connection to weight. As noted in *Figure 1*, we require further investigation to establish any sort of connection between the habit of the parent and the health of the child.

III. BACKGROUND

In finding a relationship between neonatal infant mortality and maternal smoking, we must first discuss fetal development to then better understand how the problem at hand has developed. The gestational process as a whole is a forty week process on average. However, seeing how this describes the average, the range of the gestation period can be shortened to 37 weeks, in which the infant would be considered a preterm delivery, or lengthened up to 42 weeks. It is worth nothing that in a prior study, infants of preterm delivery status have higher mortality rate within the first 5 years of life [17]. We will primarily consider preterm delivery and its connection to the birth weight of the infant, as the weight gain of developing fetus is interrupted due to the earlier delivery affects the mortality rate, along with other considerations to be discussed.

Preterm delivery stunts the growth of the infant, as it is known that the infant gains 0.2 pounds per week during the final stage, building upon it's 5 to 5.5 pound, 45 centimeter long body in its 32nd week, finally settling within the range of 5.5 to 8.8 pounds. It has been previously studied that increase in infant gestation period has association of decrease in mortality rate postnatal [11], where, according to the study, the mortality rate of those with shorter gestation period starts near 50% and decreases to nearly 13% as the gestation period is closer to what is considered regulation. With this in mind, we shall disregard the gestation period factor to limit the variability of our results. Our study will strictly consider the birth weight of the infants born at term, in which we can differentiate underweight infants as those born under 5.5 pounds.

We also choose to examine how the socioeconomic status of the family could possibly factor into the overall health of the infant. Socioeconomic status is broken down into categories such as educational levels attained, total income, and occupation, and is used as a measure of social standing or class of an individual or group of people [1]. A prior study established an association between socioeconomic status and health of the offspring, where infants with lower socioeconomic backgrounds of tended to report, in adulthood, poorer quality of health. Additionally, the study placed particular emphasis on the economic background the offspring is brought into, in which lower economic backgrounds also resulted in poorer quality of health as an adult [16]. Seeing how socioeconomic status has association with the resulting health of the infant in a later stage, we wish to investigate whether or not socioeconomic status could be used to predict the resulting health of the infant. Using the information previously discussed, we will accomplish this investigation by distinguishing whether or not there is a pattern between the birth weight of the child and the socioeconomic status of the parent(s), total income of the parent(s). We restrict our study to just refer to the income as there is no formal equation to give a score for socioeconomic status. We do this since the level of education attained becomes relatively arbitrary compared to the income of the individual.

We will carry out this investigation by performing numerous mathematical formulas with the data to calculate statistical descriptions of the data. Specifically, we shall examine the minimum, maximum, quartiles, median, mean, skewness, and kurtosis of the dataset to provide insight as to what to expect from the graphs. We will then verify these findings with histograms, quantile-quantile plots, and box plots to graphically understand our dataset. From this, we will move on to understand the incidence or frequency of underweight infants as a result of maternal smoking to lead us to a conclusion regarding our first question. As for our second question, we will use the data given on the parent's income, in which we can distinguish if infants born with certain income scores find general association with birth weight. We will accomplish our analysis of this question using strategies similar to which we used to solve the first question.

IV. HYPOTHESIS

This study proposes two questions, in which we formulate the two hypotheses as follows:

I. What is the difference in weight between babies born to mothers who smoked during pregnancy and those who did not? Is this difference important to the health of the baby?”

A. Our null hypothesis is maternal smoking has no direct effects on the average birth weight of the infant. Using the knowledge acquired from the literary discussion and background, our hypothesis is maternal smoking will cause a reduction in the average birth weight of the infant.

II. Can the income score of the parent(s) be used as a predictor of the birth weight of the infant?

A. The research and findings given in our background lead us to believe the income score of the parent(s) will be a predictor of the birth weight of the infant.

Statistical Analysis-

V. NUMERICAL COMPARISONS

TABLE I: A table outlining basic numerical statistics and data of baby birth weights from both the distributions of mothers who did not smoke while pregnant and mothers who smoked while pregnant

	Did Not Smoke During Pregnancy	Smoked During Pregnancy
Sample Size	742.00	484.00
Minimum	55.00	58.00
1st Quartile	113.00	102.00
Median	123.00	115.00
Mean	123.00	114.10
3rd Quartile	134.00	126.00
Maximum	176.00	163.00
Standard Deviation	17.39869	18.0985
Skewness	-0.1869841	-0.03359498
Kurtosis	4.03706	2.988032

Table I. above contains a numerical representation of the two distributions of mothers who did not smoke during pregnancy and mothers who did smoke during pregnancy. It is important to note that the original data had four separate distributions: mothers who never smoked and had a baby, mothers who smoked through pregnancy and had a baby, mothers who smoked up until pregnancy and had a baby, and mothers who quit smoking and had a baby. However, for the purpose of this report, our analysis will primarily be conducted around two categories: mothers who did not smoke during pregnancy and mothers who did smoke during pregnancy. Since this report intends to draw a conclusion as to whether or not smoking during pregnancy has an effect on the birthweight of infants, it was deemed acceptable to combine the original four distributions into two.

It must be noted that there may be underlying confounding factors that affect the birthweight of babies if the mother gave up smoking directly before pregnancy or if the mother was a prior smoker; however, this study is only set to analyze if smoking during pregnancy affects the birthweight of an infant as compared to non maternal smoking.

The *Did Not Smoke During Pregnancy* distribution contains all the data in the three following distributions: mothers who never smoked and had a baby, mothers who smoked up until pregnancy and had a baby, and mothers who quit smoking and had a baby (smoked = ‘0’, ‘2’, ‘3’). The *Smoked During Pregnancy* distribution contains the data in the distribution of mothers who smoked through pregnancy and had a baby (smoked = ‘1’). *Table I.* shows that there is a larger population of mothers who did not smoke during pregnancy as compared to those who did; additionally, the 1st quartile, median, mean, and 3rd quartile of babies birth weights from mothers who did not smoke is higher as compared to mothers who did smoke during pregnancy

VI. GRAPHICAL COMPARISONS

Fig. 1: Two overlapping histograms showing the distributions of the birth weights of babies born to non-smoking (in red) and smoking (in cyan) mothers

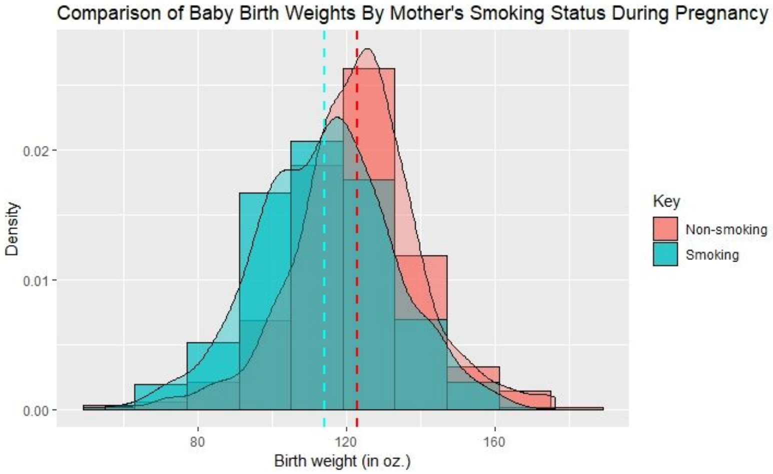
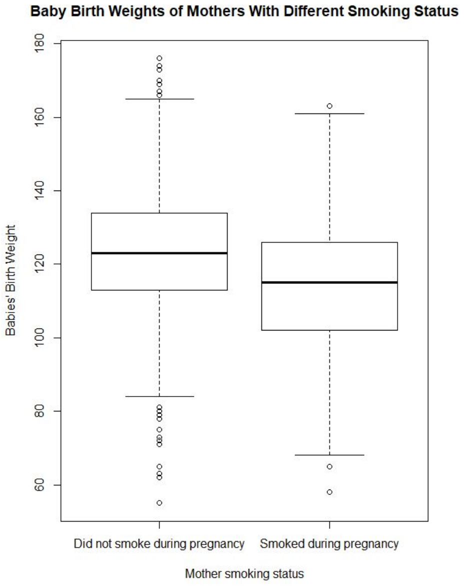


Figure 1 above shows the distribution of babies birth weights for mothers who smoked during their pregnancy (blue) and the distribution of babies birth weights for mothers who did not smoke during their pregnancy (red).

The distribution of babies birth weights for mothers who smoked during their pregnancy is both a bimodal and symmetric distribution. The distribution of babies birth weights for mothers who did not smoke during their pregnancy is both a unimodal and symmetric distribution. Furthermore, *Figure 1* shows a density curve mapped onto each histogram (smoking and non smoking mothers) which more accurately depicts the difference in birth weight between babies that come from mothers who smoked and did not smoke during pregnancy. The mean of birth weights that come from mothers who smoked is below 120 ounces while the mean of birth weights that came from mothers who did not smoke is above 120 ounces (refer to *Table I.* for exact values of the mean).

Fig. 2: A boxplot modeling the two distributions mothers who smoked and did not smoke while pregnant. The boxplot shows the shape of the distribution, its central value, and its variability



The boxplot in *Figure 2* shows the standardized infant birth weights of mothers who did not smoke and mothers who did smoke during pregnancy. *Figure 2* shows that the median, 1st quantile, and 3rd quartile for the standardized infant birth weights of non smoking mothers is higher compared to that of smoking mothers. Additionally, there exist numerous outliers for an infant's birth weight when the mother does not smoke during her pregnancy (left box).

Fig. 3: A bar graph comparing the proportion of babies born with a low birth weight below 88 ounces to mothers who smoked during pregnancy to mothers who did not smoke during pregnancy

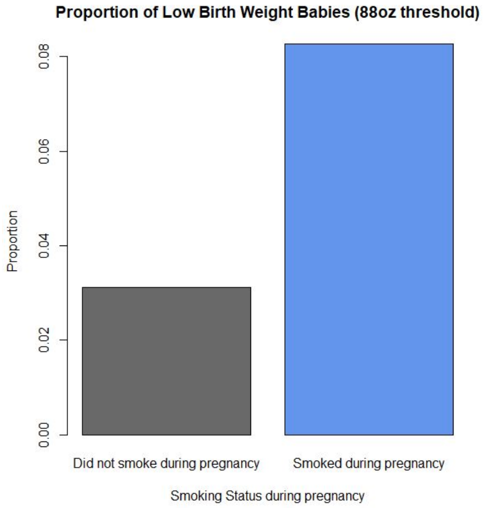


TABLE II: A table that shows the proportion of babies who are born with a low birth (less than 88 ounces) rate for women who smoked during pregnancy and women who did not smoke during pregnancy

	Did not smoke during pregnancy	Smoked during pregnancy
Proportion of low birth weight (under 88oz)	0.02941	0.07438

Figure 3 shows the proportion of babies who are born with a low birth weight (less than 88 ounces) for each of the distributions. *Table II.* indicates the exact proportion of low birth weight babies found in each distribution. The data shows that the proportion of low birth weight babies for mothers who did not smoke (0.02941) is lower as compared to the proportion of babies who were born from mothers who smoked throughout pregnancy (0.07438).

To better understand the significance and accuracy of this difference, other bar graphs were constructed with different thresholds to see if there was any unusual groupings around the 88 ounce threshold. *Figure 4* and *Figure 5* below show the proportion of low birth weight babies for mothers who did not smoke and did smoke during pregnancy with differing threshold values.

Fig. 4: A bar graph comparing the proportion of babies born with a low birth weight below 86 ounces to mothers who smoked during pregnancy to mothers who did not smoke during pregnancy

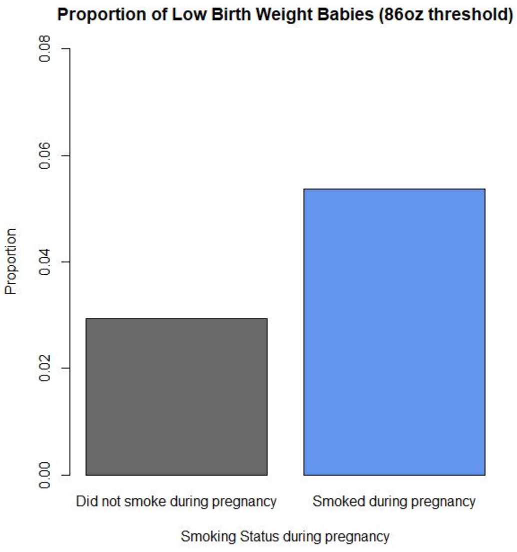


Fig. 5: A bar graph comparing the proportion of babies born with a low birth weight below 88 ounces to mothers who smoked during pregnancy to mothers who did not smoke during pregnancy

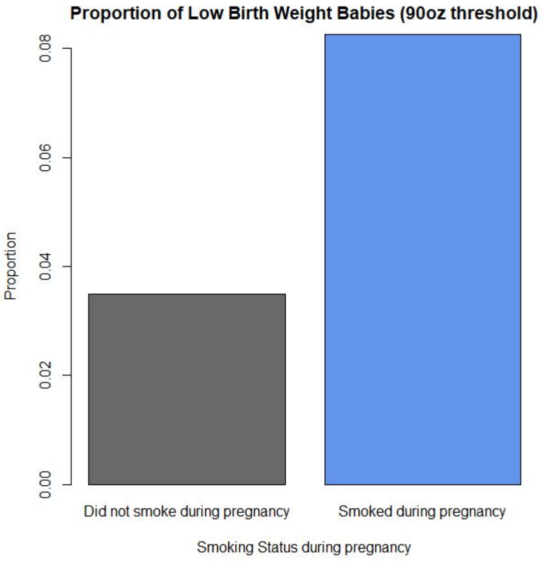


TABLE III: A table that shows the proportion of babies who are born with a low birth rate for women who smoked during pregnancy and women who did not smoke during pregnancy with varying thresholds used

	Did not smoke during pregnancy	Smoked during pregnancy

Proportion of babies (under 86oz)	0.0294118	0.0537190
Proportion of babies (under 88oz)	0.0294118	0.0743801
Proportion of babies (under 90oz)	0.0349265	0.0826446

Figure 4 and *Figure 5* show that changing the threshold by ± 2 ounces has minimal effect on the proportion of babies that are born with a low birth weight. Thus, it is determined that there are no unusual clusterings of birth weights that surround the original threshold of 88 ounces. The respected proportions for each threshold are shown in *Table III*.

Fig. 6: A Q-Q Plot comparing the theoretical distribution of the nonsmoking mothers dataset to the observed findings

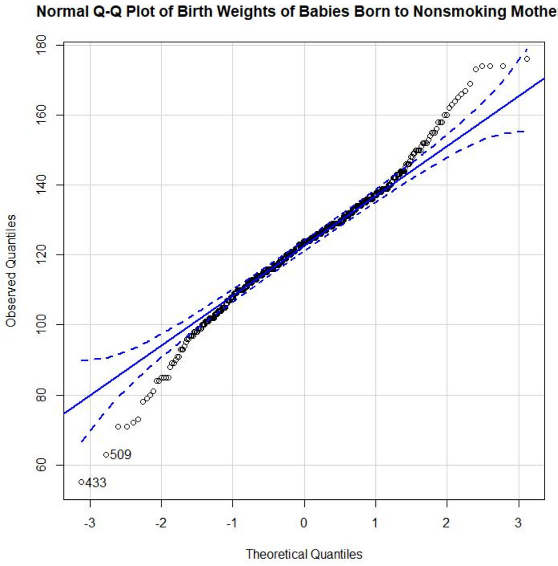


Fig. 7: A Q-Q Plot comparing the theoretical distribution of the smoking mothers dataset to the observed findings

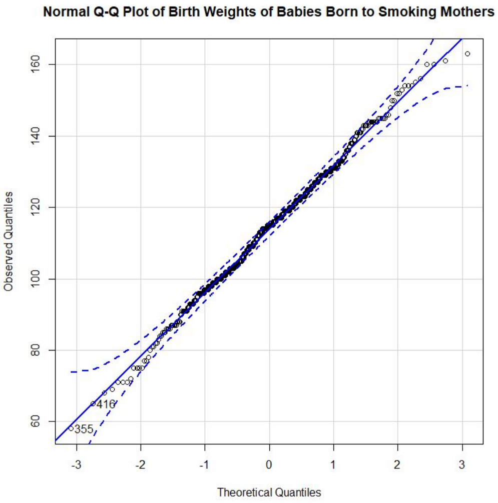
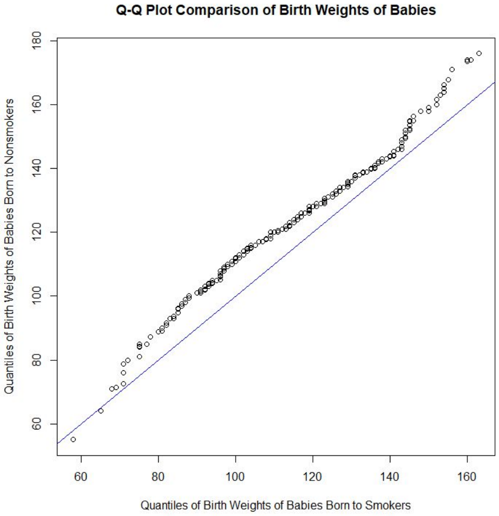


Fig. 8: A Q-Q Plot comparing the observed findings of the birthweights of the smoking mothers dataset to the observed findings of the birthweights of the smoking mothers dataset.



In examining *Figure 6*, we see that the observed findings of the nonsmoking mothers dataset relatively fits the normal distribution and is similar to the theoretical findings near the median of the dataset. Notice the deviation from this theoretical line near the tails of the observed findings, suggesting the distribution was "heavy-tailed," suggesting the outliers occurred frequently..

However, in examining *Figure 7*, we see that the observed findings of the smoking mothers dataset is more identical to the normal distribution than its counterpart. This is seen as the observations remain close to the theoretical line with the exception of the fluctuations as the observations tended away from the mean. The fluctuations also suggest the middle of the distribution is symmetric, and we can clearly see this when referring to *Figure 1*.

As we compare the distributions to each other as described in *Figure 8*, we first notice that the Q-Q plot tends above the theoretical line. This suggests that the quantiles of the smoking mothers dataset generally occur before the quantiles of the nonsmoking mothers dataset, and this is again supported by *Figure 1*.

Fig 9: A histogram showing the skewness coefficients of the normal distribution from MonteCarlo simulation and skewness of birth weights of babies of mothers who smoked during pregnancy and mothers who didn't smoke during pregnancy, taken by bootstrapping

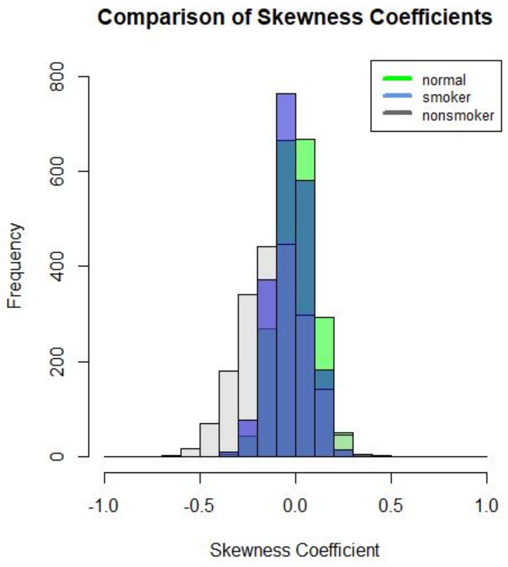
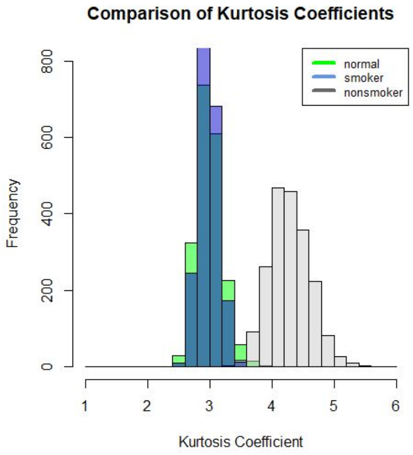


Fig 10: A histogram showing the kurtosis coefficients of the normal distribution from MonteCarlo simulation and skewness of birth weights of babies of mothers who smoked during pregnancy and mothers who didn't smoke during pregnancy, taken by bootstrapping



To tell if the distribution of both populations are normal the kurtosis and skewness of each population are compared to those of a Monte Carlo normal distribution. The mothers who smoked while pregnant and the mothers who did not smoke while pregnant were resampled using bootstrapping. The distributions were then overlaid and the results show that there is a rather high probability that suggest the distribution of our two populations is normal due to the kurtosis coefficient being close to 3 and a skewness coefficient being close to 0.

TABLE V: Frequency of Low Birth Weight Babies

Welch Two Sample t-test	Mean difference in birth weights	P-value
Birth weights of babies of smoking mothers vs nonsmoking mothers	-8.6681	$4.816e^{-15}$

From the information calculated, we see that the p-value is significantly less than 0.05, so we can reject the null hypothesis; this means that our hypothesis of maternal smoking affecting birth weight holds true and there exists some underlying association between maternal smoking and non smoking birth weights.

VII. ADDITIONAL INVESTIGATION

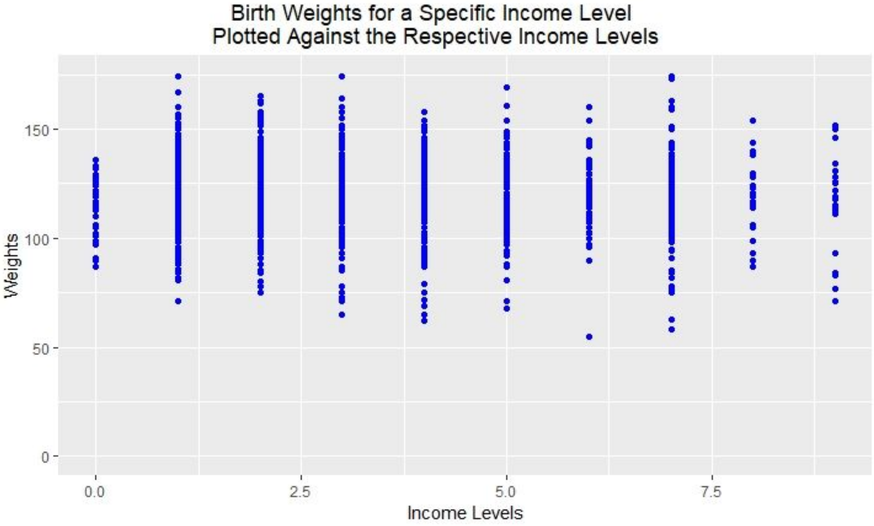
As mentioned previously, no widely used formula for a socioeconomic score that objectively reflects reality to a non-trivial extent currently exists, which is why we restricted our investigation to essentially seeing if there we could predict the birth weight of a baby based on its parents' income levels. We could have perhaps built a linear model for the socioeconomic score where education and income levels are independent variables with corresponding weights but our problems with that would be:

a) Level of education attained and income levels are generally not independent; and
b) We'd be assigning importance to how much those factors matter somewhat arbitrarily.

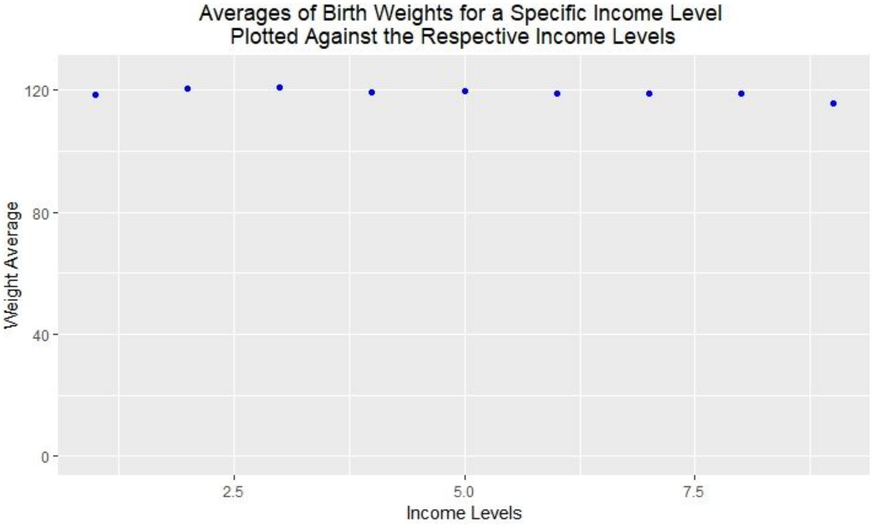
Another issue with the data currently available is that we essentially just have categorical data for the incomes, where the incomes are organized into 9 different "levels", with the highest level representing the highest incomes. This makes creating a model to predict the relationship even harder.

That said, we can check what the correlation between the birth weights of the babies and the income level of its parents' is. The correlation coefficient, for two random variables X and Y can be defined as corr(X, Y) = $\frac{Cov(X, Y)}{\sigma_X \sigma_Y}$, where Cov(X, Y) is the covariance of X and Y, and $\boldsymbol{\sigma}_X$ and $\boldsymbol{\sigma}_Y$ are the standard deviations of X and Y respectively. The value of corr(X, Y) lies between -1 and 1, where the closer it is to either end, the stronger the correlation and potential dependence between X and Y. A value of 0 indicates that there is no correlation between the two variables.

Plotting all of the birth weights of the babies against the respective income levels of the parents, no immediate correlation can be observed:



Even after taking into account that our view of the birth weights is distorted, the general range of the birth weights is actually very similar, lying mostly between 140 and 100 oz. Plotting the mean of the birth weights from each category shows us this:



Almost all of the averages are basically around the same value, which is 119 ounces. The correlation coefficient value between the averages and the income levels is -0.627, which indicates non-trivial dependence between the two, except this value isn't very meaningful as we only have 9 data points here.

The correlation coefficient value between all of the birth weights and the income levels is -0.0033, which is incredibly close to 0, and thus indicates that there isn't a strong correlation between the two variables, and we thus cannot predict birth weight based on income levels, at least for this dataset.

VIII. IMPORTANCE OF THE DIFFERENCE

The numerical statistics suggest that babies born to mothers who smoked during pregnancy have relatively lower birth weights in the population median, mean, 1st quartile, and 3rd quartile. This is indicative of some significant shift between the two distributions.

The graphical statistics are also supportive of a shift between these two distributions and further reinforce this idea; the boxplot shows much of the data from the numerical statistics in a graphical representation and also highlights any outliers that occur in the data set. The histogram drawn shows the density curve for each specific distribution allowing us to further our understanding of the shift between the two distributions.

The individual Q-Q plots for the smoking and nonsmoking mothers datasets both follow similar distribution. This reinforces the idea that since the two datasets are alike in distribution, we look toward the statistical information gathered, specifically the median, mean, and their respective quartiles, for analysis regarding the differences between the two.

The Q-Q plots also denote that the two datasets both follow the normal distribution, but the findings from the incidence or frequency study suggest that individuality in each dataset's robust statistical features. The incidence study clearly suggests that the population of underweight infants is higher in the smoking mothers dataset than its counterpart. Hence, these differences lead us to the ideas in the discussion and conclusion.

IX. DISCUSSION

There are many improvements that can be made to this analysis to better reach a more definitive and evident conclusion. The analysis conducted was based around an observational study thus limiting the conclusions and findings we could draw from analysis. In future studies, various confounding factors should be taken into account. Additionally, this study sought out to answer the question as to whether or not smoking during pregnancy has any effect on the birth weight of babies; thus, prior smoking habits and conditions were not accurately assessed and could be confounding factors affecting the birth weight of an infant.

Moveover, the dataset was only inclusive of male birth rates and thus can only infer the association between smoking and male birth weights. Additionally, other substance abuse was not taken into account (i.e. alcohol, illegal drugs and substances, medical prescriptions, etc.) [3] These potential confounding factors could have influenced the data and birth weights of babies thus leading to an inaccurate finding. Pre Existing medical conditions and second hand smoke was not taking into account either. To gather more accurate data and eliminate (some) confounding factors a controlled experiment would need to be conducted.

X. CONCLUSION

Our investigations has resulted in our rejecting the null hypothesis, so we then find that maternal smoking does have an effect on the birthweight of the single-infant, male child. The numerical summary clearly suggested the statistical findings of the smoker's distribution are less than that of the nonsmoker's distribution. This suggestion was reinforced as the graphical summary consistently depicted the distribution of infants affected by maternal smoking was visually and analytically less than that of its counterpart. Hypothesis testing denoted there was no significance in the null hypothesis, supporting our alternate hypothesis.

As for the second question, our findings resulted in finding no correlation between the income score and the birth weight of the baby. This disagrees with the previous findings stated within the background, but we find that this disagreement is due to the limitations provided by the data. We were not able to calculate some sort of true score of socioeconomic status, in which we were hindered from finding correlation to support the previous findings.

Despite these findings, it is worth noting that because the data was not produced from a controlled experiment we may only infer there is an association between smoking during pregnancy and low birth weight and cannot establish a causal relationship. Thus, we can say that smoking while pregnant is strongly linked to low birth weight.

XI. METHODS AND THEORY

1. Normal Distribution

The normal distribution with parameter μ and $\sigma^2 > 0$ is given by the familiar bell-shaped probability.

$$\phi\left(x; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty.$$

The density function is symmetric about the point μ, and the parameter 2 is the variance of the distribution. The case $\mu = 0$ and $2 = 1$ is referred to as the *standard normal distribution*. If X is normally distributed with **mean** μ and **variance** $\sigma^2 > 0$, then $Z = (X - \mu)/ \sigma$ has a standard normal distribution. By this means, probability statements about arbitrary normal random variables can be reduced to equivalent statements about standard normal random variables. The standard normal density and distribution functions are given respectively by

$$\phi(\xi) = \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2,} \quad -\infty < \xi < \infty,$$

and

$$\Phi(x) = \int_{-\infty}^{x} \phi(\xi) \mathrm{d}\xi, \quad -\infty < x < \infty.$$

<u>Notation</u>

Mean μ (= $\overline{X}$) is given by the formula

$$\mu\ (= \overline{X}) = \frac{\Sigma Xi}{n}$$

where i = 1, 2, 3, …, n

Variance σ^2 is given by the formula

$$\sigma^2 = \frac{\Sigma(Xi - \overline{X})^2}{n-1}$$

where i = 1, 2, 3, …, n

Central Limit Theorem

The *Central Limit Theorem* explains in part the wide prevalence of the normal distri- bution in nature. A simple form of this aptly named result concerns the partial sums, $Sn = \xi 1 + \cdots + \xi n$ of independent and identically distributed summands $\xi 1$, $\xi 2$, . . . having finite means $\mu = E[\xi k\ ]$ and finite variances σ^2 = Var[ξk]. In this case, the Central Limit Theorem asserts that

$$\lim_{n\to\infty} \Pr\left\{\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\right\} = \Phi(x) \quad \text{for all } x.$$

The precise statement of the theorem's conclusion is given by the above equation. Intuition is sometimes enhanced by the looser statement that, for large n, the sum Sn is approximately normally distributed with mean $n\mu$ and variance $n\,\sigma^2$. In practical terms we expect the normal distribution to arise whenever the numerical outcome of an experiment results from numerous small additive effects, all operating independently, and where no single or small group of effects is dominant. [12]

2. Skewness

In probability theory and statistics, asymmetry or distortion is an indicator of the probability distribution asymmetry of a real-valued random variable. The value of a degree may be positive or negative and may not be defined. If the degree is negative, the probability density function has a long tail on the left side of the function and a median containing more data on the right side. When the degree is positive, it has a long tail on the right side of the probability density function and indicates that the data are distributed more on the left side. If the mean and median are the same, the degree is zero.

Skewness coefficient

Skewness coefficient is the average of the third power of the standardized data.

$Skewness\ coefficient\ = \frac{1}{n}(\sum_{i=1}^{\infty} \frac{Xi-\overline{X}}{S})^{\ 3}$ where S is the standard deviation, with

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ [19]

3. Kurtosis

Kurtosis is a measure of the degree of sharpness of the probability distribution. It is used to measure how intensively the observations are centered. If the kurtosis value (K) is close to 3, the scatter is close to the normal distribution. (K < 3), the distribution can be judged to be a flat distribution more gently than the normal distribution, and if the kurtosis is a positive number larger than 3 (K> 3), the distribution can be considered to be a more pointed distribution than the normal distribution.

Kurtosis coefficient

Kurtosis coefficient is the average of the fourth power of the standardized data.

$Kurtosis\ coefficient\ = \frac{1}{n}(\sum_{i=1}^{\infty} \frac{Xi-\overline{X}}{S})^{\ 4}$

where S is the standard deviation. [9]

4. Simulation Studies

The term, Simulation Studies is that a computer experiment that generates data using pseudorandom sampling method. The greatest advantage of simulation research is the ability to understand the behavior of statistical methods because "truth" is known in the data generation process. This allows you to take into account the attributes of methods such as offsets. In widely used simulation studies, in many cases insufficient design, analysis and presentation are done. This manual provides the rationale for the use of simulation studies and recommendations for design, implementation, analysis, reporting and presentation. In particular, the guide provides a systematic approach to planning and reporting simulation studies, including goals, mechanisms of data generation, estimation, method definition.

5.Quantile

Quantiles are points taken at regular intervals from the distribution function of a random variable.

The quantile term was used for the first time by Kendall in 1940. The quantile of order p of a distribution (with 0 < p <1) is the value of the variable x_p that marks a cut so that a proportion p of values of the population is less than or equal to x_p . For example, the quantile of order 0.36 would leave 36% of values below and the quantile of order 0.50 corresponds to the median of the distribution.

Quantiles are often used by groups that divide the distribution equally; understood these as intervals that comprise the same proportion of values. The most used are:

1. The quartiles, which divide the distribution into four parts (correspond to the 0.25, 0.50 and 0.75 quantiles);

2. The quintiles, which divide the distribution into five parts (correspond to the quantiles 0.20, 0.40, 0.60 and 0.80);

3. The deciles, which divide the distribution into ten parts;

4. The percentiles, which divide the distribution into one hundred parts.

In the calculation of quantiles with distributions of continuous variable (for example, with grouped data) it can be easily achieved that the parts in which the distribution is divided are exactly the same.

However, in discrete variable distributions we must be satisfied that these parts are approximately equal. Unfortunately, there is no consensus on how to carry out this approach, there being nine different methods in the scientific literature, which lead to different results. Therefore, when calculating any quantile of data not grouped by means of calculator, software or manually, it is essential to know and indicate the method used.

The function assigned to each p by the cut point x_p, that is, the value of the quantile of order p, is called the quantile function. [14]

6. Q-Q plots (quantile-quantile plots)

The Q-Q plot is a statistical graphical way to compare by plotting the two probability distributions against each other. First, the set of intervals of quantiles is selected. Point (x, y) on the plot corresponds to one of the fractions (y coordinate) of the second distribution plotted against the same fraction (x coordinate) of the first distribution. Therefore, the line is a parametric curve with parameters, and this parametric curve is the concatenation of the quantiles.

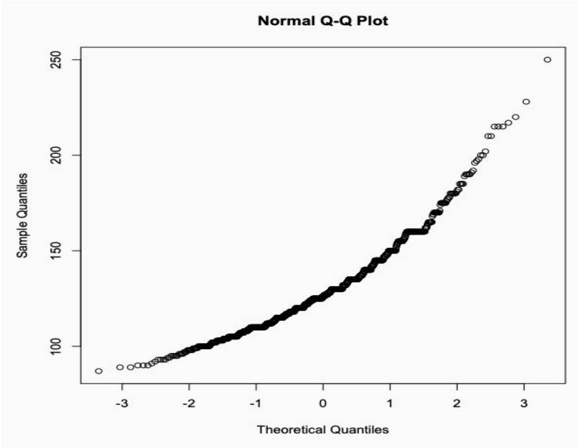
If the two distributions being compared are similar, the points in the Q-Q plot are located near the line y = x. When the distribution is in a linear relationship, the points in the Q-Q plot are distributed almost on a straight line, but they are not necessarily on the straight line y = x. The Q-Q plot can be used to estimate the type of local distribution.

We compare the shape of the distribution using the Q-Q plot and judge how the characteristics such as position, scale, skewness etc. are similar in how they are similar in the two distributions. The Q-Q plot can be used to compare sets or theoretical distributions of data. Using Q-Q plots to compare two data can be seen as a nonparametric approach to comparing their underlying distributions. Q-Q plots are generally a more powerful approach than the general approach of comparing histograms of two samples, but required more skills to interpret from the diagram. The Q-Q plot is used to compare the data set to the theory / model. This allows graphical evaluation of "goodness of fit" as well as numerical summary. The Q-Q plot is also used to compare two theoretical distributions to each other. Since the Q-Q plot compares distributions, it is not necessary to observe the values as a pair as in a scatterplot, and it is not necessary to equalize the numbers of the values of the two groups being compared.

Normal Q-Q plots

A probability plot in which the expected value when observed values follow a normal distribution is taken on the Y axis and the observed value itself is taken on the X axis.

The percentile (cumulative probability) is obtained from the order in which the observation values are arranged in ascending order, and the expected value is predicted by using the inverse function of the probability density function of the normal distribution. If the plots are aligned, it is considered that observed values follow a normal distribution.



[15]

7. T-test

The t test is a name of a statistical test method that utilizes that statistics follow the t distribution when assuming that the null hypothesis is correct. It is a parametric test method that assumes that the population follows a normal distribution and uses that the t distribution does not depend directly on the original average or standard deviation (but depends on the degree of freedom). It is used for testing whether there is a significant difference in mean between two sets of specimens. One of statistical hypothesis tests. In the Japanese Industrial Standard, it is defined as "a statistical test that assumes that the test statistic follows the t distribution under the null hypothesis." The t-test is also called the Student's t-test.

One sample T-test

One sample T-test is for testing the null hypothesis that the population mean value μ is equal to a specific value μ_0 .

$$t = \frac{\overline{X}-\mu}{s/\sqrt{n}}$$

Notation

$\overline{X}$: *Sample mean*

s : *Standard deviation*

μ : *Population mean value*

The t test can be broadly divided as follows.

1. A test of whether the means are equal, assuming that both populations follow a normal distribution.
2. If specimens are paired, that is, if there is a special relationship between each member of a set of specimens and another set of specific members (for example, when investigating the same person twice before and after, When comparing between husband and wife, etc).
3. If the samples are independent and it can be assumed that the variances of the two groups being compared are equal (assumption of equidistribution).
4. When the specimens are independent and equal variance can not be assumed (heterogeneous dispersion). This is exactly called Welch's t test.

A test of whether the average of the population according to the normal distribution is equal to a specific value.

Test on whether the slope of the regression line is significantly different from 0.

Independent two-samples T test

It is used to test the null hypothesis of whether the mean value μ_1 of the first population is equal to the mean value μ_2 of the second population. In other words, test the null hypothesis of whether.

$\mu_1 - \mu_2 = 0$.

Equal sample size and equal variance

For explanation, two specimens of sample sizes n and m, $x_1 , ..., x_m$ and $y_1 , ..., y_n$ are normal distributions N (μ_1 , σ_1^2) and N (μ_2 , σ_2^2) (σ_1 , σ_2 are unknown). At this time, we perform a test on the difference of the population mean of two samples, δ, and set the hypothesis as follows.

Null hypothesis $H_0 : \delta = \mu_0$

Alternative hypothesis $H_1 : \delta \neq \mu_0$

Here, the difference d = $\overline{x}$ - $\overline{y}$ between the average values x - ~ N (μ_1 , $\frac{\sigma_1^2}{m}$) and y - ~ N (μ_2 , $\frac{\sigma_2^2}{n}$) of the two specimens is also normal distribution N ((μ_1 - μ_2), σ^2 ($\frac{1}{m} + \frac{1}{n}$)) is used. [19]

Acknowledgements

[1] APA. “Socioeconomic Status.” *American Psychological Association*, American Psychological Association, www.apa.org/topics/socioeconomic-status/index.aspx.

[2] Bernstein, I. M., MD, Mongeon, J. A., MS, Badger, G. J., MS, & Solomon, L., PhD. (2005, November). Maternal Smoking and Its Association With Birth Weight : Obstetrics & Gynecology. Retrieved January 29, 2019, from https://journals.lww.com/greenjournal/Fulltext/2005/11000/Maternal_Smoking_and_Its_Association_With_Birth.15.aspx

[3] CDC. (2017, September 29). Tobacco Use and Pregnancy | Reproductive Health | CDC. Retrieved from https://www.cdc.gov/reproductivehealth/maternalinfanthealth/tobaccousepregnancy/index.htm

[4] CDC. (2018, February 20). Smoking & Tobacco Use. Retrieved January 31, 2019, from https://www.cdc.gov/tobacco/data_statistics/fact_sheets/fast_facts/index.htm

[5] “Central Limit Theorem.” *Wikipedia*, Wikimedia Foundation, 13 Dec. 2018, en.wikipedia.org/wiki/Central_limit_theorem.

[6] Curtin, S. C., MA, & Matthews, T., MS. (2016, February 10). *Smoking Prevalence and Cessation Before and During Pregnancy: Data From the Birth Certificate, 2014*[Scholarly project]. Retrieved February 1, 2019.

[7] Hernandez-Diaz, S., Schisterman, E. F., & Hernan, M. A. (2006, August 24). Birth Weight "Paradox" Uncovered? Retrieved January 31, 2019, from https://academic.oup.com/aje/article/164/11/1115/61454

[8] Kleinman, J., Pierre, M., Madans, J., Land, G., & Schramm, W. (1988, February 01). EFFECTS OF MATERNAL SMOKING ON FETAL AND INFANT MORTALITY. Retrieved from https://academic.oup.com/aje/article/127/2/274/95153

[9] “Kurtosis.” *Wikipedia*, Wikimedia Foundation, 27 Jan. 2019, en.wikipedia.org/wiki/Kurtosis.

[10] Lowe, C. R. (1970, January 01). Effect of Mothers' Smoking Habits on Birth Weight of their Children. Retrieved January 31, 2019, from https://link.springer.com/chapter/10.1007/978-94-011-6621-8_40

[11] Lubchenco, L, et al.*Neonatal Mortality Rate: Relationship to Birth Weight and Gestational Age*.

[12] “Normal Distribution.” *Wikipedia*, Wikimedia Foundation, 29 Jan. 2019, en.wikipedia.org/wiki/Normal_distribution.

[13] Pinksy, M, and S Karlin. *An Introduction to Stochastic Modelling*. Vol. 4.

[14] “QQ Plot.” *Wikipedia*, Wikimedia Foundation, 19 Dec. 2018, en.wikipedia.org/wiki/Q–Q_plot.

[15] “Quantile.” *Wikipedia*, Wikimedia Foundation, 1 Feb. 2019, en.wikipedia.org/wiki/Quantile.

[16] Rahkonen, O, et al. “Past or Present? Childhood Living Conditions and Current Socioeconomic Status as Determinants of Adult Health.” *NeuroImage*, Academic Press, 15 June 1998, www.sciencedirect.com/science/article/pii/S0277953696001025.

[17] Romero, R., Dey, S. K., & Fisher, S. J. (2014, August 15). Preterm labor: One syndrome, many causes. Retrieved February 2, 2019, from http://science.sciencemag.org/content/345/6198/760/tab-pdf

[18] “Skewness.” *Wikipedia*, Wikimedia Foundation, 1 Feb. 2019, en.wikipedia.org/wiki/Skewness.

[19] “Student's t-Test.” *Wikipedia*, Wikimedia Foundation, 17 Jan. 2019, en.wikipedia.org/wiki/Student's_t-test.

[20] Wang, X., MD, MPH, ScD, Zuckerman, B., MD, & Colleen, P., BA. (2002, January 09). Maternal Cigarette Smoking, Metabolic Gene Polymorphism, and Infant Birth Weight. Retrieved January 28, 2019, from https://jamanetwork.com/journals/jama/fullarticle/194545

[21] Ward, C., Lewis, S., & Coleman, T. (2007). Prevalence of maternal smoking and environmental tobacco smoke exposure during pregnancy and impact on birth weight: Retrospective study using Millennium Cohort. Retrieved January 30, 2019, from http://www.readbag.com/cpd-screening-nhs-uk-choicestoolbox-docs-antenatal-c