# Math 189 Project 1

Jonathan Chen (A11323343) Applied Mathematics/Political Science(Fourth Year)
Yihang Cheng (A92039418) Math/Computer Science (Second year)
Haoshen Hong (A92083316) Applied Mathematics (Second year)
Zihao Zhou (A91146985) Math-Computer Science (Third year)
Ruoning Guan(A92013043) Applied Mathematics(Second year)

## 1.Introduction

This report summarizes several statistical modelling and analysis associated with the survey conducted by Child Health and Development Studies. The purpose of this report is to investigate the potential impact of smoking during pregnancy by pregnant mothers. The data set used in this research is the file "babies.txt", containing 1236 objects, each has 7 features.

Section 2 gives a list of specific statistical observation relevant to the analysis including the mean, variance and skewness; Section 3 introduces background information regarding to the study conducted by CHDS, briefly summarizing the observations of the study and confounders that could influence results of the study; Section 4 uses various statistical approach, including diagrams, to interpret the relationship between smoking by pregnant women and weight of newborns; Section 5 discusses the theories behind statistical techniques used in the analysis, as Section 6 concludes the analysis with a conclusion.

## 2.Data

The data in babies.txt are represented by 7-feature vectors. Features include birth weights, gestation period, mother's age, height, pre-pregnancy weight, as well as whether the baby is the first born and mother's smoking record. In babies23.txt, data are specified with additional features and detailed categorizations. Data were collected from the CHDS data, which consisted of pregnancies occurred between 1960 and 1967 among women in Kaised Health Plan in Oakland, California with prenatal care in San Francisco area. The study is comprised of 1236 male babies, with factors such as different gender, twin status, and factors that may impact weight of the babies removed.

Information available to us via data matrix include mother's smoking status, which is a regular categorical observation, as well as baby's weight in ounces, a continuous, numerical variable that we considered to be relating to mother's smoking status.

## 3.Background

Further research suggests that many confounders could mask the influence of smoking on babies' weight at birth. For instance, a statistic analysis report from 1990 made a conclusion that race and

ethnicity has a correlation with smoking and babies' weight at birth.[1] Such possible factors within the context of this analysis could potentially include mother's weight, mother's age and gestation period. Those confounding factors might mask the influence of smoking on babies' born weight.

Research has shown that typical gestational period is 40 weeks and the relationship of birthweight and gestation time is not entire linear.[2] The growth between 252 days (36 weeks) and 297 days (42 weeks) is linear but there is a slowing down for growth in weight after 297 days. Also, babies who are born prior to 37 weeks are called preterm delivery and their weight are significantly less than those born after 37 weeks. Thus it is essential to consider the gestation period as a confounder.

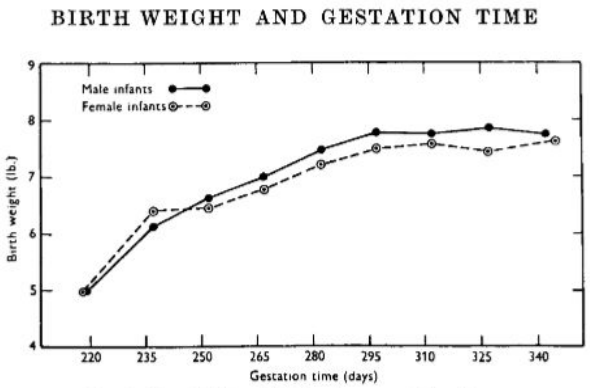

Fig. 2. Mean birth weight for given gestation time.

Likewise, maternal age constituted another confounder that was observed in our study. According to an analysis published in 2000, the difference in birthweight between smoking mothers and non-smoking mothers "increased with maternal age from 182g (<20 years of age) to 232g (35+ years of age).[3] This study indicates that the difference in weight of newborns among smoking and non-smoking mothers could partially be attributed to the age of mothers.

Based on the data and background information, we have adopted a null hypothesis: there exist no relationship between the newborn's weight and their mother's smoking status.

**4.Investigation**

Prior to categorizing the data, outlying entries of babies with mother smoking status unknown are removed from the data set.

---

[1] Petitti, Diana B., and Charlotte Coleman. "Cigarette Smoking and the Risk of Low Birth Weight: A Comparison in Black and White Women." *Epidemiology* 1, no. 3 (1990): 201-05.

[2] Mary.N Carn Birth Weight and L.S.Penrose "Gestation Time in Relation to Maternal Age, Parity and Infant Survival", Blackwell Publishing Ltd/University College London Published,1951

[3]HAUG, K., IRGENS, L. M., SKJÆRVEN, R., MARKESTAD, T., BASTE, V. and SCHREUDER, P. (2000), Maternal smoking and birthweight: effect modification of period, maternal age and paternal smoking. Acta Obstetricia et Gynecologica Scandinavica, 79: 485–489. doi:10.1034/j.1600-0412.2000.079006485.x

The histogram, boxplot and statistic information of the babies' birth weights of smoking and non smoking mothers are presented as following:

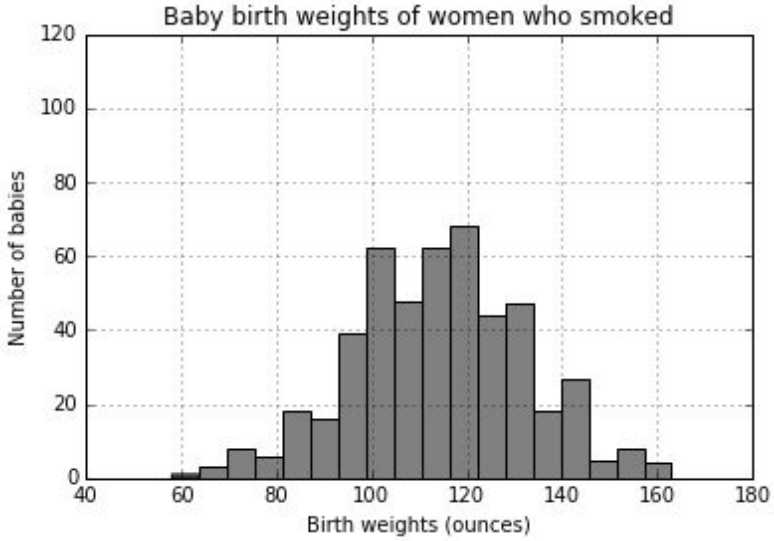


Figure - Histogram of babies weights of smoker mother

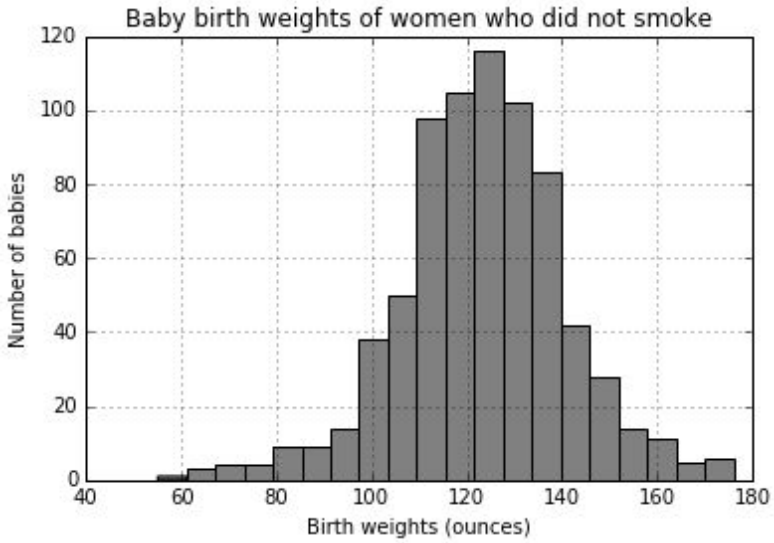


Figure - Histogram of babies weights of nonsmoker mother

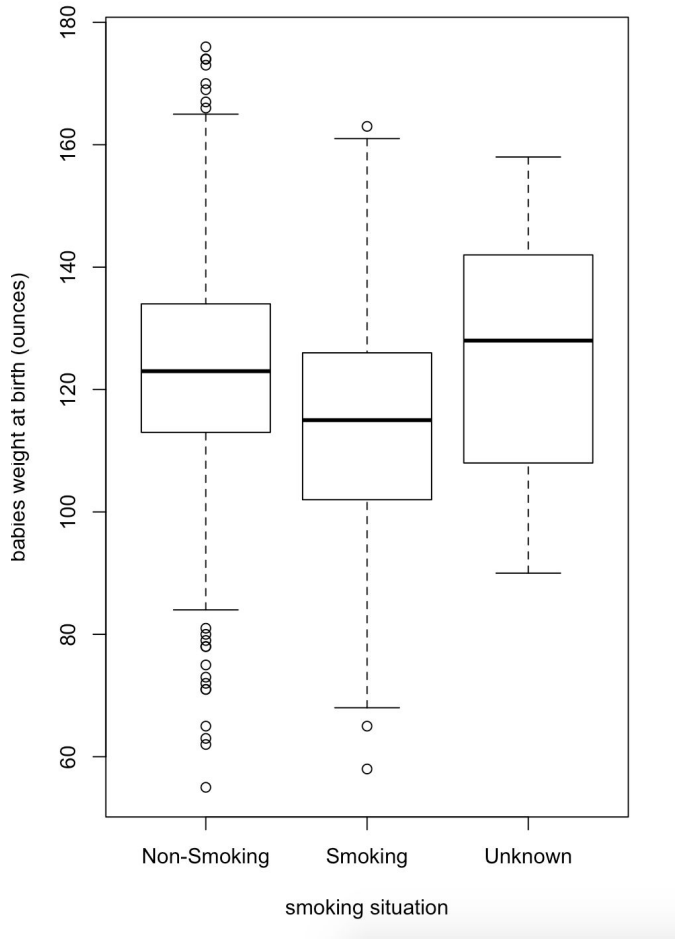


Figure - Boxplot of babies weights of smoker and nonsmoker mother

| Samples | Count | Min | Max | Mean | Variance | Skewness |
|---|---|---|---|---|---|---|
| Babies' weights of smoker mother | 484 | 58 | 163 | 114.11 | 327.57 | -0.03 |
| Babies' weights of non smoker mother | 742 | 55 | 176 | 123.05 | 302.71 | -0.19 |

Figure - Statistic information about babies weights of smoker and nonsmoker mother

From the chart and diagrams above, we know there are more samples of non smoker mother, and the distribution of babies' weights of non smoker mother is more spread-out, has larger mean than that of smoker mother, since the sample size is large, it's sample mean and variance can represent the population mean and variance. Furthermore, the distribution of babies' weights of non smoker mother has nearly-normal distribution while that of smoker mother is negatively-skewed.

Epidemiological studies define that  a low-birth-weight baby as baby with weight lower than 2500g (88.2 ounces). The probability that a smoker mother has a low-birth-weight (P(low-weight | smoker)) is 0.0826, and that of nonsmoker mother (P(low-weight | nonsmoke)) is 0.0310. Then a Pearson's independence test is applied to test whether this difference is significant.

In order to explore whether smoking status of mother has association with fetal low weight, we set a null hypothesis that these two variables are independent. When we consider babies who are under 88.2 oz (2500g) as underweight babies, the chi-square test for independence returns P value 0.0001082. Since P value is much less than 0.05, we could reject the null hypothesis confidently that underweight babies are independent from the smoker mothers.

```
> table(data.cleanWtClnSmk$smoke, babyUnderweight.ind)
   babyUnderweight.ind
    FALSE TRUE
  0   719   23
  1   444   40
> chisq.test(data.cleanWtClnSmk$smoke, babyUnderweight.ind)

        Pearson's Chi-squared test with Yates' continuity correction

data:  data.cleanWtClnSmk$smoke and babyUnderweight.ind
X-squared = 14.987, df = 1, p-value = 0.0001082
```

Figure - Result of independence test with low-birth-weight as 88 ounces

To have a strong a robust research, we lowered the definition of low-birth-weight slightly and run the Chi-square test again. The birth weight is still highly related with the smoking situation of the mothers ($P < 0.01$). The result remains the same.

```
> babyUnderweight.ind <- data.cleanWtClnSmk$bwt<=86
> table(data.cleanWtClnSmk$smoke, babyUnderweight.ind)
   babyUnderweight.ind
    FALSE TRUE
  0   721   21
  1   454   30
> chisq.test(data.cleanWtClnSmk$smoke, babyUnderweight.ind)

        Pearson's Chi-squared test with Yates' continuity correction

data:  data.cleanWtClnSmk$smoke and babyUnderweight.ind
X-squared = 7.5118, df = 1, p-value = 0.00613
```

We have also modified the definition of low-birth-weight slightly by and the outcome we got is the same as the previous one (P<0.01). Therefore we are more confident about our result.

```
> babyUnderweight.ind <- data.cleanWtClnSmk$bwt<=90
> table(data.cleanWtClnSmk$smoke, babyUnderweight.ind)
   babyUnderweight.ind
    FALSE TRUE
  0   715   27
  1   442   42
> chisq.test(data.cleanWtClnSmk$smoke, babyUnderweight.ind)

        Pearson's Chi-squared test with Yates' continuity correction

data:  data.cleanWtClnSmk$smoke and babyUnderweight.ind
X-squared = 13.07, df = 1, p-value = 3e-04
```

However, when we change the definition of underweight babies from 88 to 82, after the chi-square test, P value is 0.07449 which is larger than 0.05. Thus we cannot reject the null hypothesis that underweight babies are independent from the smoker mothers. Also we notice that lowering threshold causes P value to grow, and this might be caused by the nearly-normal distribution of baby weights with smoker mother, and the right-skewed distribution of baby weights with non smoker mother.

```
> babyUnderweight.ind <- data.cleanSmoke$bwt<82
> chisq.test(data.cleanSmoke$smoke,babyUnderweight.ind)

        Pearson's Chi-squared test with Yates' continuity correction

data:  data.cleanSmoke$smoke and babyUnderweight.ind
X-squared = 3.1813, df = 1, p-value = 0.07449
```

Figure - Result of independence test with low-birth-weight as 81 ounces

**5. Theory**

Mean and Median**:** Mean is the "average" of a set of data and median is the "middle" of the data. Usually they are used to describe the distribution of the data.

Variance and Standard Deviation **:** Variance is the expectation of the squared deviation of a random variable from its mean. Standard deviation describes how far each standardized individual value is positioned from the mean. Most of the data are located within 3 standard deviations from of the mean with the exception of outliers.

Skewness**:** Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. A perfectly normal distribution takes the shape of a bell curve with no skewness toward left or right. Skewness coefficient is calculated by taking the average of the third power of the standardized data:

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

Chi-square test: Relating to or denoting a statistical method assessing the goodness of fit between observed values and those expected theoretically; we use the result of the chi-square test to find P values, which in turn could be used to determine whether we can confidently reject the null hypothesis.

Boxplot: A boxplot describes the distribution of standardized data based on the distribution's maximum summary, minimum summary, mean, 25th percentile, and 75th percentile. The bottom and top of the box plot display the minimum and maximum values, respectively; the bottom and top of the box describes the 25 and 75 percentile respectively, while the line within the box describes the mean. Dots plotted outside of the maximum and minimum are considered outliers as they are unusually far from the mean.

Histogram: A histogram is a display of data using bars of various lengths. Within the context of this analysis, histograms are used to compare the distribution of weight between pregnant women who smoke and those who do not. The X-axis of the histograms describe the weight of babies in ounces and the Y-axis describe the number of babies.

**6. Conclusion/Discussion**

From the investigation, if we set the standard of low-weight baby as those that below 2500g (88.2 ounces), then the null hypothesis could be rejected, that is, we can confidently say that there exist some associations between mother's smoking status and newborns' low birth weight. However, there exist some limitations of these data that constrain the generalization of this conclusion.

One of the limitations is that voluntary questionnaire should be applied with care due to research biases such as non-response or faked responses. Wong and Koren has illustrated that women with fetal distress may under-report their smoking habits (P=0.04)[4]. Thus it is possible that birth weight of baby might not be associated with the smoking situation of their mothers. Other factors such as the data is too old to apply might also decrease the credibility of this investigation.

Therefore, whether smoking negatively impacts newborn's health in current world remains not very clear. For further exploration, we looked at how smoking situation of mothers may affect their babies in a long term perspective. Wickstrom demonstrates that exposure to smoke seems to leave a negative influence on the children which increased brisk for attention-deficit/hyperactivity disorder (ADHD) and

[4] Wong M, Koren G. Bias in maternal reports of smoking during pregnancy associated with fetal distress. Can J Public Health. Rev Can Sante Publ 2001; 92:109-112

conduct disorder (CD). He also points out that Animal studies show that nicotine--one of the main toxic chemicals in tobacco--does play a significant role in the pathogenic process[5].

[5] Wickstrom, R. "Effects of Nicotine During Pregnancy: Human and Experimental Evidence."Current Neuropharmacology 5.3 (2007): 213-22. Web.