Births in North Carolina

Data analysis on North Carolina Births

Michael Bergstrom - Mattew Cupich - Nickolas Diaz

Introduction

This project is a statistical analysis on a study of child births in North Carolina using the techniques, calculations, and tests learned in Statistics 1. First, a summary of the data will be taken by analyzing each variable individually. Then, confidence intervals will be found for key variables in the study. Finally, hypothesis tests will be conducted on those key variables and conclusions will be made about everything.

The summary of the data is a basic analysis of each variable, finding numerical data for quantitative variables, and proportional data for categorical variables. This information will be important when finding the confidence intervals and hypothesis tests later.

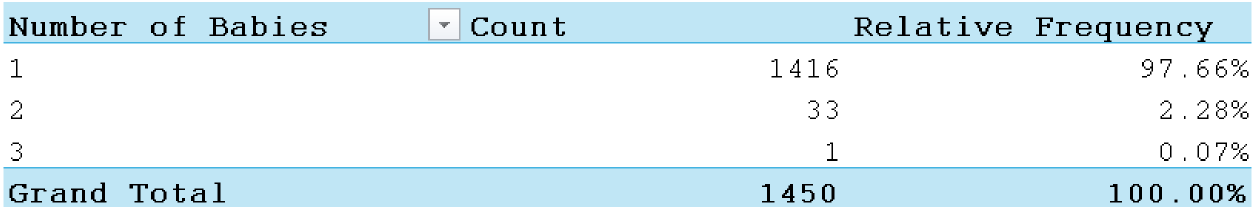
The confidence intervals and hypothesis are a part of inferential statistics, where data from a sample is used to make inferences about the population. Each of these tests come with a verbal conclusion that explains what was found mathematically in words. For instance, using a 90% confidence interval, we could conclude that we are 90% confident that the population parameter is within the interval. This will be important to draw general conclusions about the data about correlations and associations.

Data Summary

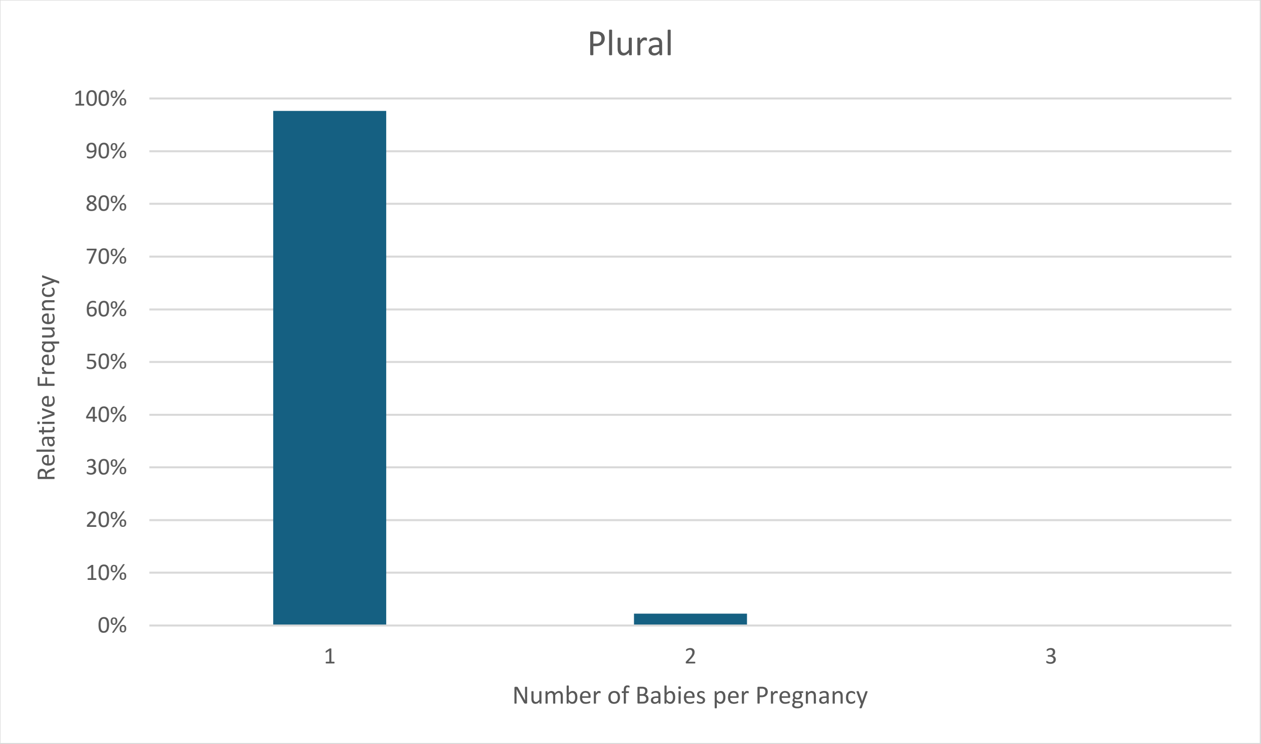
The data is from the North Carolina State Center for Health and Environmental Statistics. It is a sample of 1,450 birth records. The data includes information on the number of children born, the length of the pregnancy, the weight of the child, the mother's age and race, the child's gender; the data also evaluates whether the mother drank or smoked during the pregnancy, and if the child had a low birth weight.

Variable	Label	Description
Plural	Plurality (number of children born of the pregnancy)	Number of children born in the pregnancy.
Weeks	Length of gestation (completed weeks of gestation)	The number of completed weeks of gestation.
Tounces	Weight of the child (ounces)	The weight of the child in ounces.
Mage	Age of the mother (years)	The age of the mother in years.
Marital	Marital status of the mother	1 = Married, 2 = Not married.
Racemom	Race of the mother	0 = Other Non-white, 1 = White, 2 = Black, 3 = American Indian, 4 = Chinese, 5 = Japanese, 6 = Hawaiian, 7 = Filipino, 8 = Other Asian or Pacific Islander.
Sex	Gender of the child	1 = Male, 2 = Female.
Drink	Alcohol consumption during pregnancy	0 = Mother did not consume alcohol during pregnancy, 1 = Mother did consume alcohol during pregnancy.
Smoke	Smoking status during pregnancy	0 = Mother did not smoke during pregnancy, 1 = Mother did smoke during pregnancy.
Low	Low birth weight status of child	0 = Infant was not low birth weight, 1 = Infant was low birth weight.
Primie	Premature birth status of child	0 = Infant was not born prematurely, 1 = Infant was born prematurely Premature defined as 36 weeks or earlier

1. Number of children born from the pregnancy

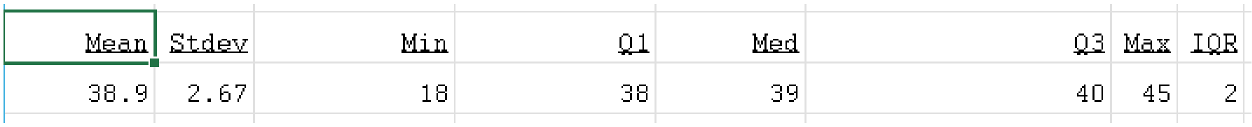


Number of Babies	Count	Relative Frequency
1	1416	97.66%
2	33	2.28%
3	1	0.07%
Grand Total	**1450**	**100.00%**

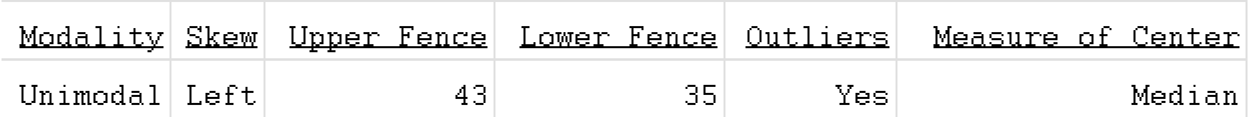


Most pregnancies have only one baby. Having two or three babies are unusual events.

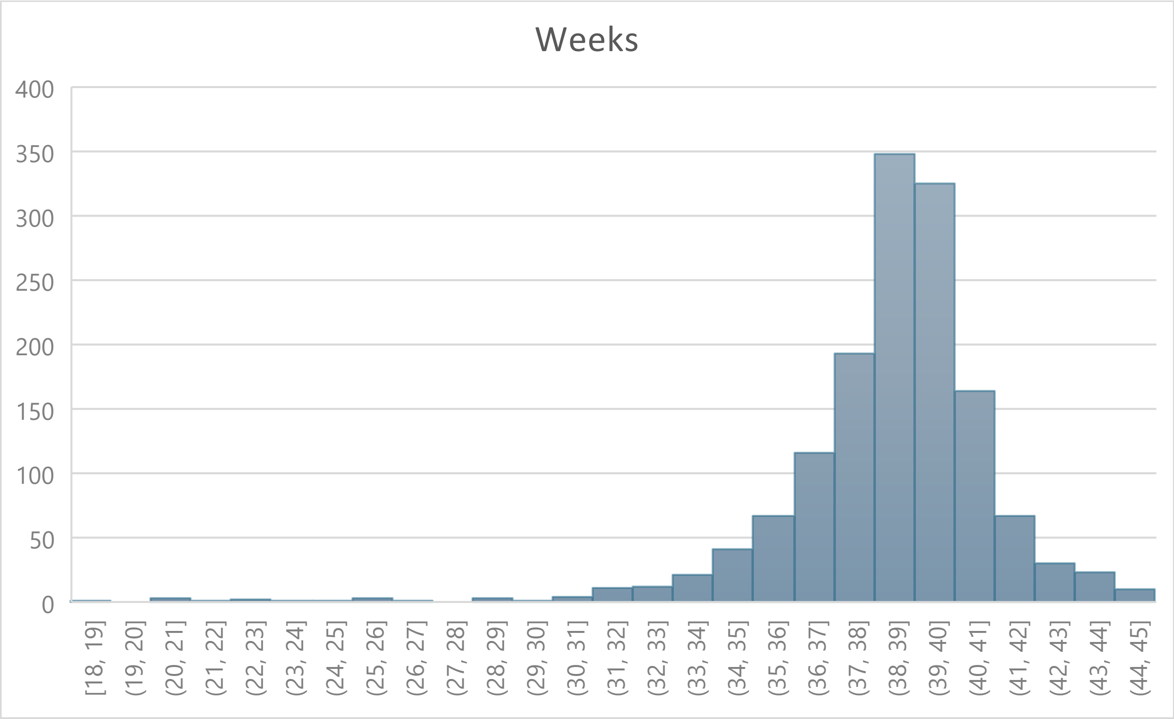
2. Number of weeks of gestation



Mean	Stdev	Min	Q1	Med	Q3	Max	IQR
38.9	2.67	18	38	39	40	45	2

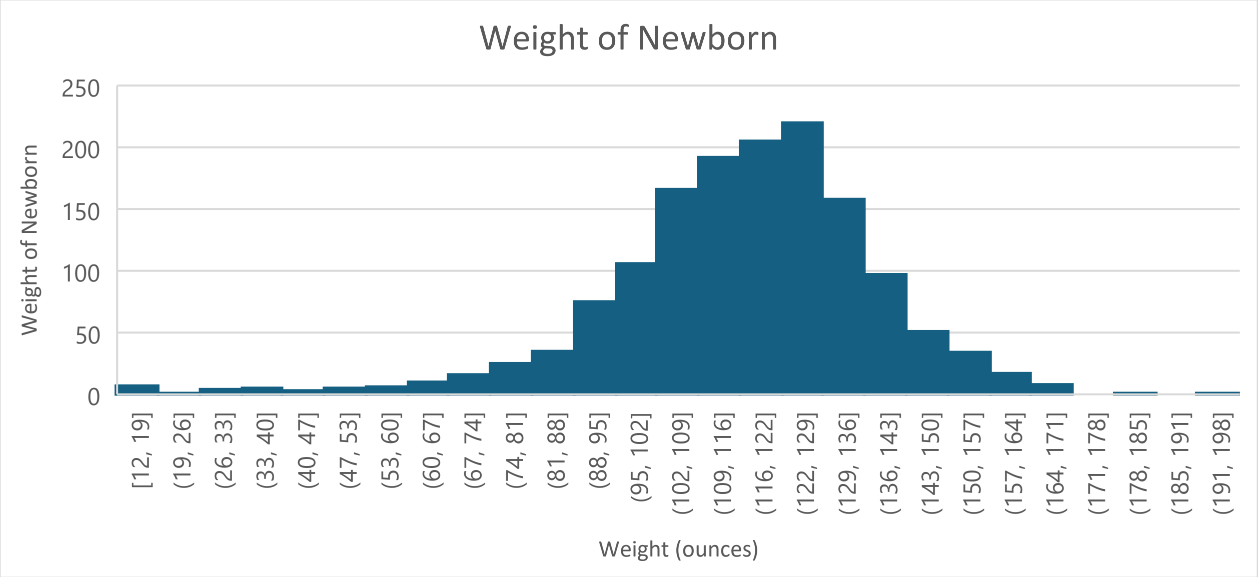


Modality	Skew	Upper Fence	Lower Fence	Outliers	Measure of Center
Unimodal	Left	43	35	Yes	Median



This data is left skewed and has outliers, so the median is the best measure of center. The median is 39 weeks.

3. Weight of the child

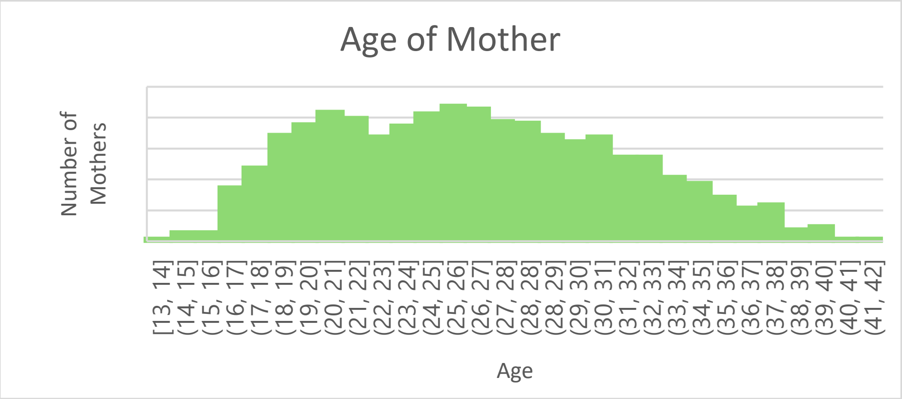


	Weight (ounces)
Mean	115.9
Standard Deviation	22.21
Minimum	12
Q1	105
Median	118
Q3	130
Max	194

An appropriate measure of center for this would be the median, as there are several outliers in the sample. The histogram appears to be bimodal, being semi-skewed to the left, with the peaks at 107.37 ounces and 121.97 ounces.

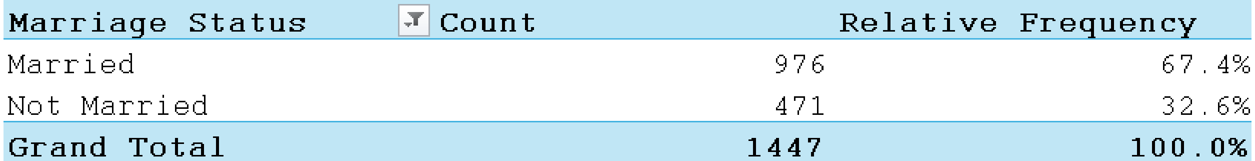
4. Age of mother

	Age of Mother
Mean	25.9
Standard Deviation	6.00
Min	13
First Quartile	21
Median	26
Third Quartile	30
Max	42

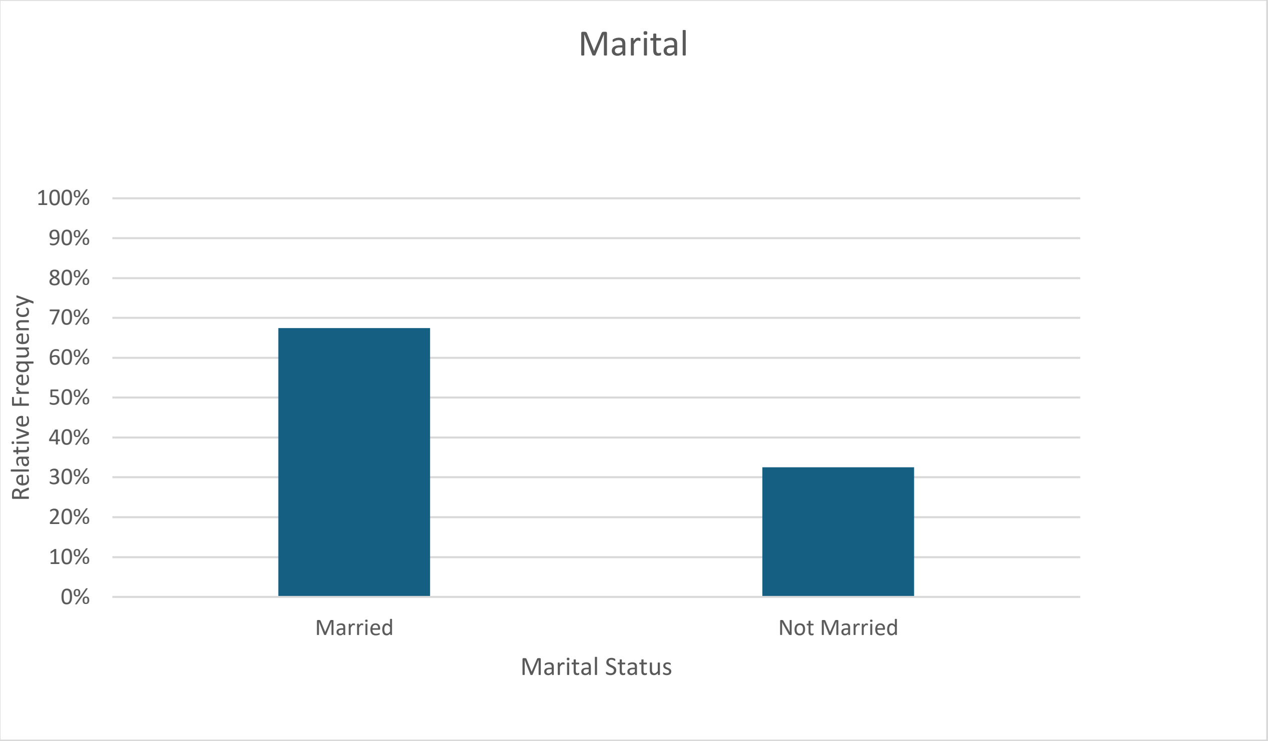


The skew of the age of the mothers is skewed a little towards the right. The histogram is unimodal, and its peak is around 25 to 26 years of age. There are no outliers in the data. Since the data does not have any outliers and is not very skewed, the mean would be the most appropriate way to measure the data.

5. Marital status of the mother

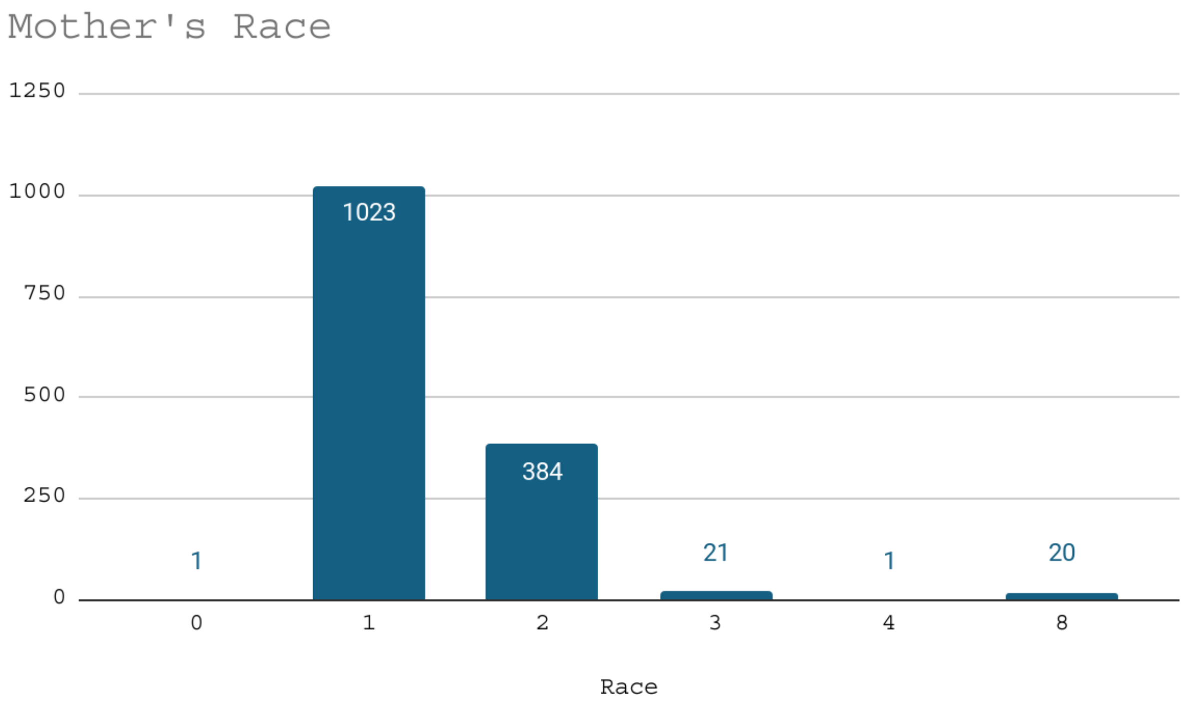


Marriage Status	Count	Relative Frequency
Married	976	67.4%
Not Married	471	32.6%
Grand Total	**1447**	**100.0%**



More women who are pregnant are married, but there is still a good amount of single mothers.

6. The race of the mother

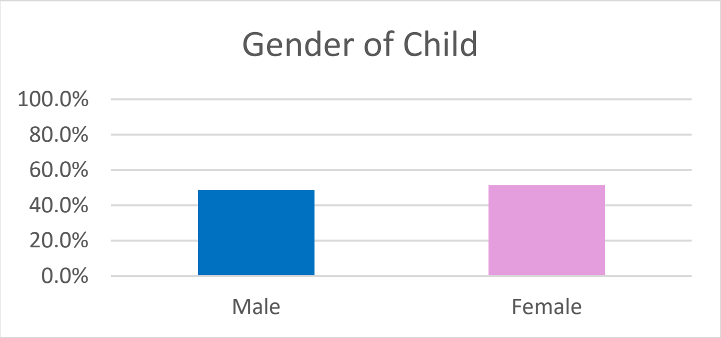


Mother's Race	Frequency	Relative Frequency
Other Non-white (0)	1	0.1%
White (1)	1023	70.5%
Black (2)	384	26.5%
American Indian (3)	21	1.4%
Chinese (4)	1	0.1%
Japanese (5)	0	0%
Hawaiian (6)	0	0%
Filipino (7)	0	0%
Other Asian (8)	20	1.4%

Despite there being a category for them, there were no Japanese, Hawaiian, or Filipino ethnicity births. White is the most common race for the mother.

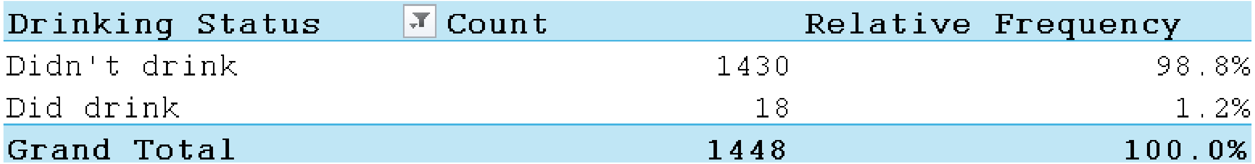
7. Gender of the child

Id	Gender	Frequency	Percentage
1	Male	707	48.8%
2	Female	743	51.2%
	Total	**1450**	

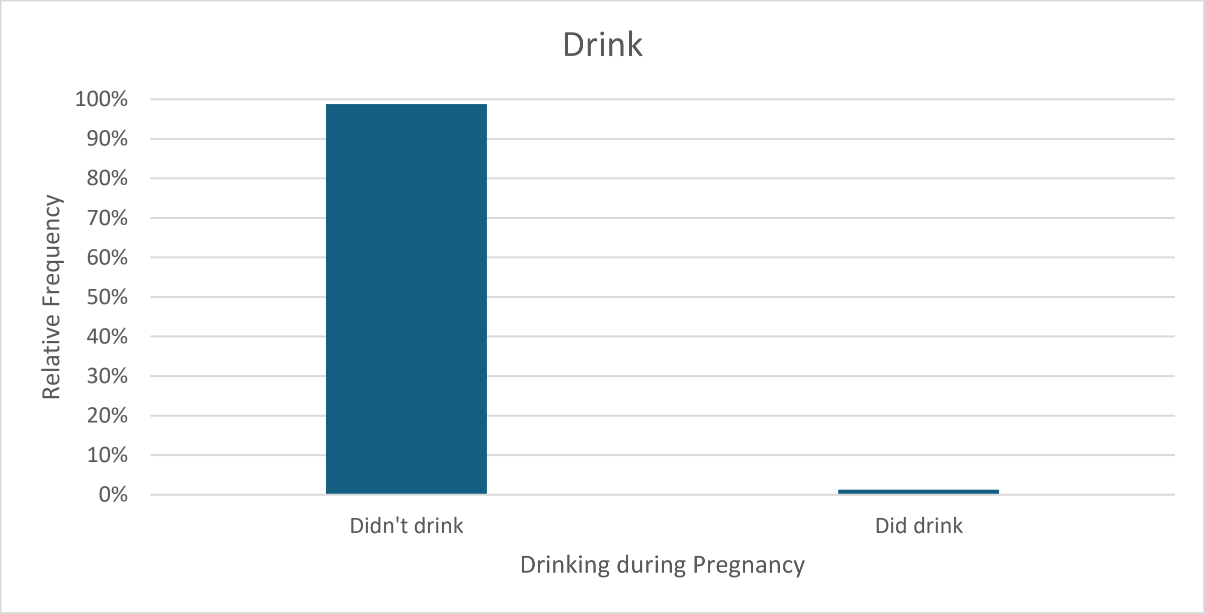


There are 2.4% more females being born than males. With 1,450 data sets, the data should be even, so there may be an additional factor that causes more females than males.

8. Pregnancy drinking status

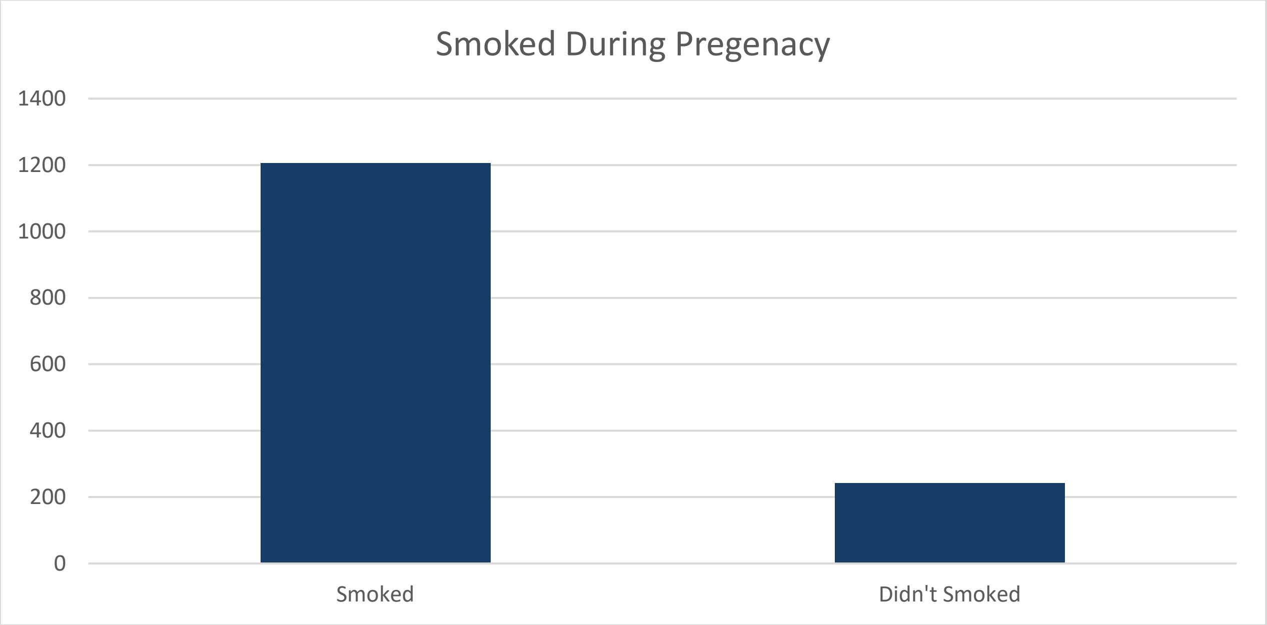


Drinking Status	Count	Relative Frequency
Didn't drink	1430	98.8%
Did drink	18	1.2%
Grand Total	**1448**	**100.0%**



Only 1.24% of mothers drink alcohol during their pregnancy. The other 98.76% didn't drink.

9. Pregnancy smoking status

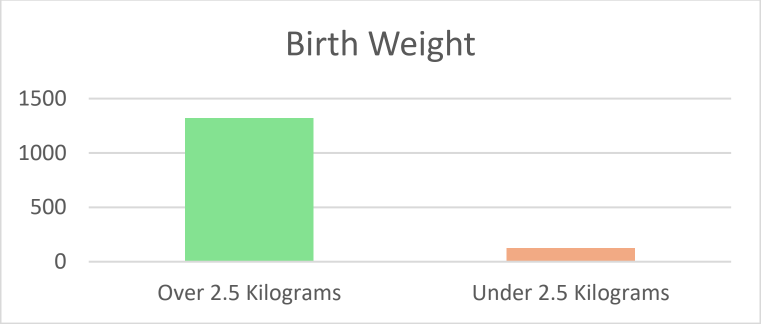


Id	Status	Frequency	Percentage
0	Smoked	1206	83.3%
1	Didn't Smoked	242	16.7%
	Total	**1448**	

It would seem that it’s more likely for a mother to not smoke, but some may still be addicted, given that 16.7% who still partook.

10. Weight status of child

Id	Weight Status of Child	Frequency	Percentage
0	Over or equal to 89 Ounces	1322	91.3%
1	Under 89 Ounces	126	8.7%
	Total	**1448**	



Most babies born were over 89 ounces. The birth weight may indicate the health of the baby.

Confidence Intervals

The purpose of this section is to use the data from the previous section to make estimations of the population of pregnancies of the world. This will be done using confidence intervals. A confidence interval is a range of numbers of which a chosen population parameter is estimated to be within. A confidence interval goes with a confidence level of a certain percentage. For each parameter, we will have confidence intervals of 90%, 95%, 98%, and 99%. In order to be able to accurately find a confidence interval two requirements must be met: the sample taken is a simple random sample, and the number of successes and failures are both greater than or equal to ten. The number of successes is found by multiplying the total number of subjects observed by the sample probability of success. The number of failures can be found by subtracting the number of successes from the number of subjects.

1. Low Birth Weight

   Definition:

   An infant has a low birth weight if they weigh less than 89 ounces.

   Requirements:

   a. Random Sample? Yes.

   b. $n\hat{p} = 1450 \times \frac{126}{1450} = 126 \geq 10$

   c. $n(1 - \hat{p}) = 1450\left(1 - \frac{126}{1450}\right) = 1322 \geq 10$

90% Confidence Interval:

$\hat{p} \pm z^{*}\sqrt(\frac{\hat{p}(1-\hat{p})}{n})$ where $z^{*} = 1.65, \hat{p} = \frac{126}{1450}, and\ n = 1450$

The confidence interval is:

0.0869 ± 0.0122 or (0.0747, 0.0991).

This means that we are 90% confident that the true value of the population proportion of babies with a low birth weight is within the interval (0.0747, 0.0991).

95% Confidence Interval:

$\hat{p} \pm z^{*}\sqrt(\frac{\hat{p}(1-\hat{p})}{n})$ where $z^{*} = 1.96, \hat{p} = \frac{126}{1450}, and\ n = 1450$

The 95% confidence interval is within 0.07249866 and 0.101534489.

This interval of values means that we are 95% confident that the true value of the population proportion of babies with a low birth weight is between the interval of 7.3% to 10.1%.

98% Confidence Interval:

$\hat{p} \pm z^{*}\sqrt(\frac{\hat{p}(1-\hat{p})}{n})$ where $z^{*} = 2.33, \hat{p} = \frac{126}{1450}, and\ n = 1450$

The confidence interval is:

0.0869 ± 0.0172 or (0.0697, 0.1041).

This means that we are 98% confident that the population proportion of babies with a low birth weight is within the interval (0.0697, 0.1041).

99% Confidence Interval:

$\hat{p} \pm z^{*}\sqrt(\frac{\hat{p}(1-\hat{p})}{n})$ where $z^{*} = 2.58, \hat{p} = \frac{126}{1450}, and\ n = 1450$

The point estimate, or the probability of a child having a low birth weight, is 0.0869, and we have a margin of error of 0.0191. With this information, we are 99% confident that the proportion of babies with a low birth weight in North Carolina is between 0.0678 and 0.106.

Why the 99% confidence interval is larger than the 90% interval:

The 99% confidence interval has a larger range because in order to be more confident, more values must be included to compensate for the increase in the percent confidence. Mathematically it is larger because the value for z* is larger for a 99% confidence interval, and since it is multiplied by the square root in the confidence interval, it makes the overall number increase in value.

2. Smoking Status During Pregnancy

Requirements:

a. Random Sample? Yes.

b. $n\hat{p} = 1448 \times \frac{242}{1448} = 242 \geq 10$

c. $n(1-\hat{p}) = 1448\left(1 - \frac{242}{1448}\right) = 1206 \geq 10$

90% Confidence Interval:

$\hat{p} \pm z^{*}\sqrt{}(\frac{\hat{p}(1-\hat{p})}{n})$ where $z^{*} = 1.65, \hat{p} = \frac{242}{1448}, and\ n = 1448$

The confidence interval is:

$0..1671 \pm 0.0162$ or (0.1509, 0.1833).

This means that we are 90% confident that the true value of the population proportion of mothers who smoked during their pregnancy is within the interval (0.1509, 0.1833).

95% Confidence Interval:

$\hat{p} \pm z^{*}\sqrt{}(\frac{\hat{p}(1-\hat{p})}{n})$ where $z^{*} = 1.96, \hat{p} = \frac{242}{1448}, and\ n = 1448$

The 95% confidence interval is within 0.147910117 and 0.186344026.

This interval of values means that we are 95% certain that the true value of the population proportion of mothers who smoked during their pregnancy is between the interval of 14.8% to 18.6%

98% Confidence Interval:

$\hat{p} \pm z^{*}\sqrt{}(\frac{\hat{p}(1-\hat{p})}{n})$ where $z^{*} = 2.33, \hat{p} = \frac{242}{1448}, and\ n = 1448$

The confidence interval is:

$0..1671 \pm 0.0228$ or (0.1443, 0.1900).

This means that we are 98% confident that the true value of the population proportion of mothers who smoked during their pregnancy is within the interval (0.1443, 0.1900).

99% Confidence Interval:

The point estimate, or the probability of a mother smoking during pregnancy, is 0.1669, and we have a margin of error of 0.0253. With this information, we are 99% confident that the proportion of mothers smoking during pregnancy in North Carolina is between 0.1416 and 0.1922.

Sample size needed for a margin of error of 2%:

For a 90% confidence interval, in order to have a margin of error of 2%, you would need a sample size of 948.

For a 95% confidence interval, in order to have a margin of error of 2%, you would need a sample size of 1337.

For a 98% confidence interval, in order to have a margin of error of 2%, you would need a sample size of 1890.

For a 99% confidence interval, in order to have a margin error of 2%, you would need a sample size of 2314.

Why a smaller sample size gives a tighter estimate of the unknown population proportion:

As the sample size increases, the margin of error decreases, therefore decreasing the range of values in the confidence interval and giving a tighter estimate of the unknown proportion.

3. Premature Birth

A birth is premature when gestation takes less than 37 weeks.

Requirements:

a. Random Sample? Yes.

b. $n\hat{p} = 1450 \times \frac{174}{1450} = 174 \geq 10$

c. $n(1-\hat{p}) = 1450\left(1-\frac{174}{1450}\right) = 1276 \geq 10$

90% Confidence Interval:

$\hat{p} \pm z^{*}\sqrt{(}\frac{\hat{p}(1-\hat{p})}{n})$ where $z^{*} = 1.65, \hat{p} = \frac{174}{1450}, and\ n = 1450$

The confidence interval is:

0.12 ± 0.0141 or (0.1059, 0.1341).

This means that we are 90% confident that the true value of the population proportion of babies that are born prematurely is within the interval (0.1059, 0.1341).

95% Confidence Interval:

$\hat{p} \pm z^{*}\sqrt{(}\frac{\hat{p}(1-\hat{p})}{n})$ where $z^{*} = 1.96, \hat{p} = \frac{174}{1450}, and\ n = 1450$

The 95% confidence interval is within 0.10327354 and 0.13672646.

This interval of values means that we are 95% confident that the true value of the population proportion of babies that are born prematurely is between the interval of 10.3% to 13.7%

98% Confidence Interval:

$\hat{p} \pm z^{*}\sqrt{(}\frac{\hat{p}(1-\hat{p})}{n})$ where $z^{*} = 2.33, \hat{p} = \frac{174}{1450}, and\ n = 1450$

The confidence interval is:

0.12 ± 0.0199 or (0.1001, 0.1399).

This means that we are 98% confident that the population proportion of babies that are born prematurely is within the interval (0.1001, 0.1399).

99% Confidence Interval:

The point estimate, or the probability of a premature birth, is 0.1193, and we have a margin of error of 0.022. With this information, we are 99% confident that the proportion of premature births in North Carolina is between 0.0973 and 0.1413.

Hypothesis Tests

A hypothesis test is a form of statistical inference where a null hypothesis is assumed to be true and it is tested whether or not there is evidence against it. There are two different types of hypothesis tests: a proportion test and a mean test. Of those tests, you can either use one proportion or mean, or two proportions or means for the test. More than two proportions or means can be used in a hypothesis test, but for this example two will be the highest number of them. The test statistic is the z-score of the sample proportion or mean given when the sampling distribution is approximately normal. With that information, the P-value, or the probability that another sample is at least as extreme as the given sample, can be found with technology or the 68, 95, 99.7 rule. In this section, hypothesis tests will be conducted on the variables premature birth, weeks of gestation, and low birth weight.

1. Weeks of Gestation

This is a one-mean hypothesis test to determine if the population mean number of weeks of gestation is below 39 weeks.

Hypotheses:

H_0: $\mu_0 = 39$ weeks

H_a: $\mu_0 < 39$ weeks

Requirements:

1) Random sample? Yes.
2) $n = 1450 \geq 30$

Test:

The level of significance or $\alpha = 0.05$ for the test.

The equation for a t test of one mean is: $t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$.

For this variable: $\bar{x} = 38.9, \mu_0 = 39, s = 2.67, and\ n = 1450.$

The test statistic $t = -1.4262$ and the P-value $P = 0.077 > 0.05 = \alpha.$

The test results are not statistically significant because the P-value is greater than the level of significance. This means that there is not enough evidence to conclude that the population mean number of weeks of gestation is below 39 weeks. Therefore, we do not reject the null hypothesis.

2. Low Birth Weight

Hypotheses:

For low birth weight, we have a single proportion test. The null hypothesis is $H_0: p = 0.06$, and the alternative hypothesis is $H_a: p > 0.06$.

Requirements:

The data was obtained via randomization and it passes the np test: ($np = 126 > 10$ and $n(1 - p) = 1324 > 10$).

Test:

The results of said test is $z = 4.313$, and the associated p value is 0.000008053. At the confidence level of 5%, we can safely say that low birth weight rate is above 6%.

3. Smoking Status and Birth Weight

Hypotheses:

H_0: μ1 = μ2

H_a: μ1 < μ2

Requirements:

1) the data is obtained by randomization

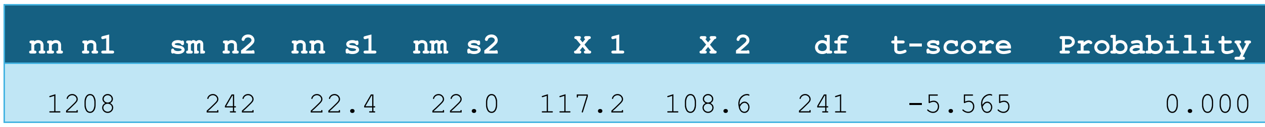
2) Two independent samples: birth weight and smoking mothers

3) n1,n2 ≥ 30; nn, nonsmokers 1208 ≥ 30 and sm, smokers 242 ≥ 30

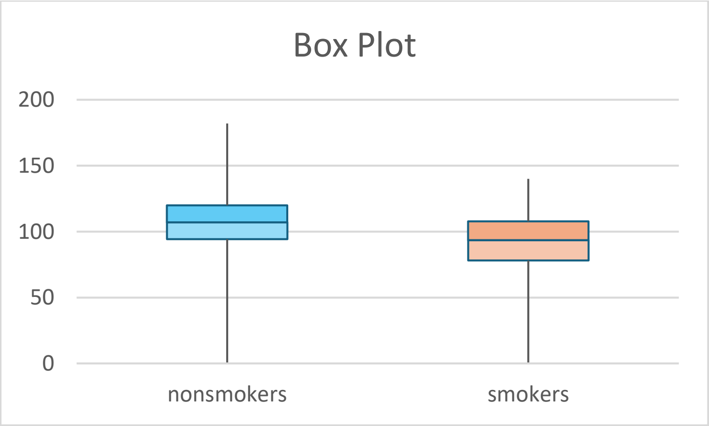
Test:

Test Statistic

$$t = \frac{\bar{x}1 - \bar{x}2}{\sqrt{\frac{s1^2}{n1} + \frac{s2^2}{n2}}},$$



nn n1	sm n2	nn s1	nm s2	X 1	X 2	df	t-score	Probability
1208	242	22.4	22.0	117.2	108.6	241	-5.565	0.000



Explanation of P-Value:

The results are statistically significant at the 5% level of significance because

$P = .00 \leq .05 = a$. Therefore, we reject the null hypothesis. There is sufficient evidence to conclude that the mean birth weight of smoking mothers at 108.6 is lower than the mean birth weight for non-smoking mothers at 117.2 ounces. Conclusion: The mean birth weight for smoking mother giving birth is lower than if they were not to smoke.

4. Smoking Status and Premature Birth

This is a two-proportion hypothesis test to determine if the population proportion of babies born prematurely is higher for smoking mothers than non-smoking mothers.

Hypotheses:

H_0: $p_1 = p_2$ where p_1 is for smoking mothers and p_2 is for non-smoking mothers

H_a: $p_1 > p_2$

Requirements:

1) Random sample? Yes.

2) Two Independent Samples? Yes.

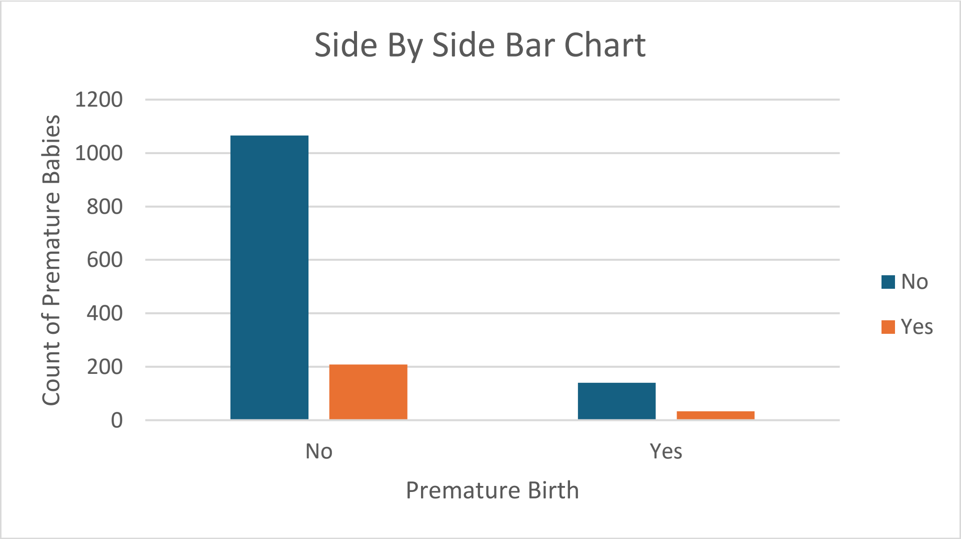
3) $n_1\hat{p}_1 = 242 \times \frac{33}{242} = 33 \geq 10, and\ n_2\hat{p}_2 = 1206 \times \frac{140}{1206} = 140 \geq 10.$

Test:

The level of significance or $\alpha = 0.05$ for the test.

The equation for a t test of one mean is: $z = \frac{\hat{p}_1-\hat{p}_2}{\sqrt{(\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right))}}$

For this variable: $\hat{p}_1 = \frac{33}{242}, \hat{p}_2 = \frac{140}{1206}, \hat{p} = \frac{173}{1448}, n_1 = 242, and\ n_2 = 1206.$



The test statistic $z = 0.8876$ and the P-value $P = 0.1874 > 0.05 = \alpha$.

The test results are not statistically significant because the P-value is greater than the level of significance. This means that there is not enough evidence to conclude that the population proportion of babies born prematurely whose mothers smoked is higher than the population proportion of babies born prematurely whose mothers did not smoke. Therefore, we do not reject the null hypothesis.

Conclusions

The data in this study provided some interesting results. Some interesting statistics from the data summary include: less than 2.5% of mothers had more than one baby, the gender of the babies was around 50% of both, and only 8.7% of babies were born with a low birth weight. Some significant conclusions of the confidence intervals are: the proportion of all babies with a low birth weight is between 7.5% and 9.9% with 90% confidence, the proportion of all mothers who smoked during their pregnancy is between 14.8% and 18.6% with 95% confidence, and the proportion of all babies who were born prematurely is between 9.7% and 14.1% with 99% confidence. Thankfully these proportions are relatively low and hopefully they decrease with more time and more medical advancements.

For the hypothesis tests, while we could not find sufficient evidence to claim that the proportion of babies born prematurely from mothers who smoked was larger than the proportion of babies born prematurely from mothers who did not smoke, we did find sufficient evidence that the mean birth weight of babies whose mothers smoked was lower than the mean birth weight of babies whose mothers did not smoke. This means that there might be some relationship between a mother smoking during their pregnancy and the weight of their baby. However, since this is an observational study, we cannot conclude that smoking is a causation of a lower birth weight in a child, nor can we conclude any causation in the data. Only associations can be concluded.