Seneca College

Assignment 1

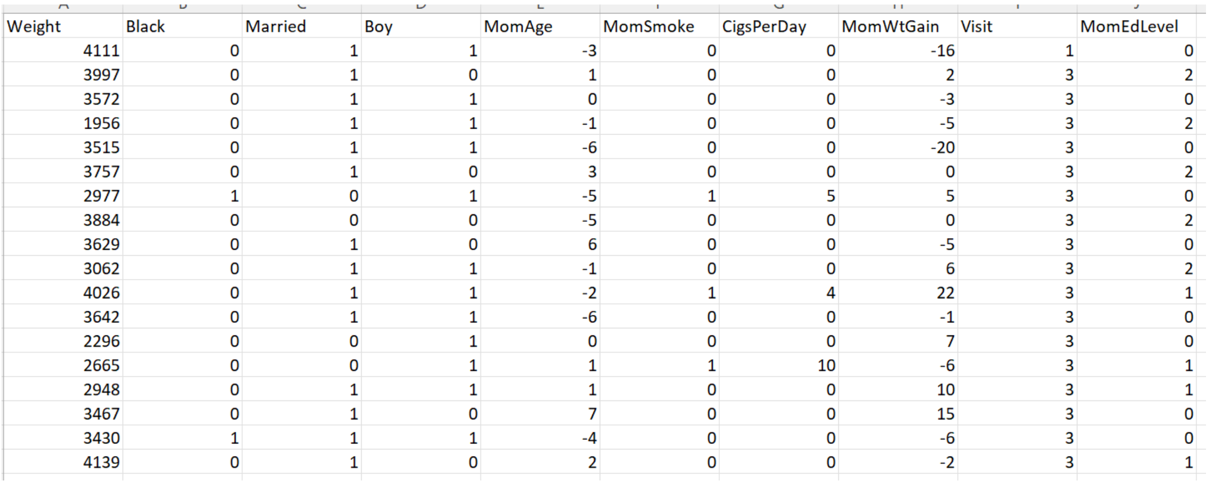
Statistical Analysis Technical Deck:

BIRTH dataset

Student ID: 153898226

Student Name: Ching Kiu Chau

Early Considerations:



Weight	Black	Married	Boy	MomAge	MomSmoke	CigsPerDay	MomWtGain	Visit	MomEdLevel
4111	0	1	1	-3	0	0	-16	1	0
3997	0	1	0	1	0	0	2	3	2
3572	0	1	1	0	0	0	-3	3	0
1956	0	1	1	-1	0	0	-5	3	2
3515	0	1	1	-6	0	0	-20	3	0
3757	0	1	0	3	0	0	0	3	2
2977	1	0	1	-5	1	5	5	3	0
3884	0	0	0	-5	0	0	0	3	2
3629	0	1	0	6	0	0	-5	3	0
3062	0	1	1	-1	0	0	6	3	2
4026	0	1	1	-2	1	4	22	3	1
3642	0	1	1	-6	0	0	-1	3	0
2296	0	0	1	0	0	0	7	3	0
2665	0	0	1	1	1	10	-6	3	1
2948	0	1	1	1	0	0	10	3	1
3467	0	1	0	7	0	0	15	3	0
3430	1	1	1	-4	0	0	-6	3	0
4139	0	1	0	2	0	0	-2	3	1

Upon analysis, the columns “Black”, “Married”, “Boy”, “MomSmoke” have values of 0 and 1. Excluding “Black” as it is not mentioned in the question, “Married”, “Boy”, and “MomSmoke” are feasible to conduct the t-tests as it clearly separates the weight into two groups.

Metadata/ Data Dictionary

Variable	Description
Weight	**Weight of the infant at birth, in grams (g)**
Married	**Marital status of the mother of the infant, with 0 referring to unmarried and 1 referring to married.**
Boy	**Gender of the infant, with 0 referring to girl and 1 referring to boy.**
MomAge	Mothers between the ages of 18 and 45. The MomAge variable is centered at the mean age at 27. Thus, MomAge=-7 means 27-7 and the mother is 20 years old. MomAge=5 means 27+5 and that the mother was 32 years old
MomSmoke	**Smoking habits of the mother, with 0 referring to mothers of the infant who do not smoke and 1 referring to mothers who smoke**
CigsPerDay	Number of cigarettes smoked per day by Mother
MomWtGain	Mother’s pregnancy weight gain in pounds (lbs)
Visit	Number of prenatal visits
MomEdLevel	Mother’s Education Level. 0=No high school graduation, 1 = High school graduate, 2= Obtained bachelor’s degree, 3= obtained postgraduate degree

1. T-test on Weight variable and Married variable

Beginning with the t-test between infant weight and married variable, the null and alternative hypothesis must be first stated:

Step 1: Setting the Null and Alternative Hypothesis

[Null Hypothesis] H0:

$\mu_{unmarried} = \mu_{married}$ (The mean weight of infant birth for married woman is the same as the weight of infant birth for unmarried woman)

[Alternative Hypothesis] H1:

$\mu_{unmarried} \neq \mu_{married}$ (The mean weight of infant birth for married woman is different from the weight of infant birth for unmarried woman)

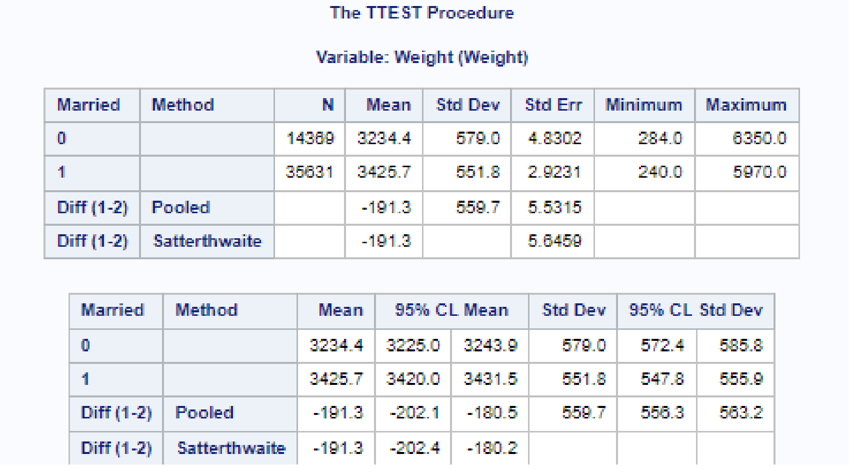
Step 2: Choosing the significance level

Choosing a significance level of α = 0.05 as it is common practice.

Default alpha of t-test in SAS is 0.05.

Step 3: Checking Assumptions

In the Married variable, there are two groups 0 and 1, referring to unmarried and married. This variable is the independent variable in the t-test. The dependent variable Weight contains the numeric value which refers to the grams of the baby. The values in the married group do not affect the values in the unmarried group, thus the sample is independent.



The TTEST Procedure

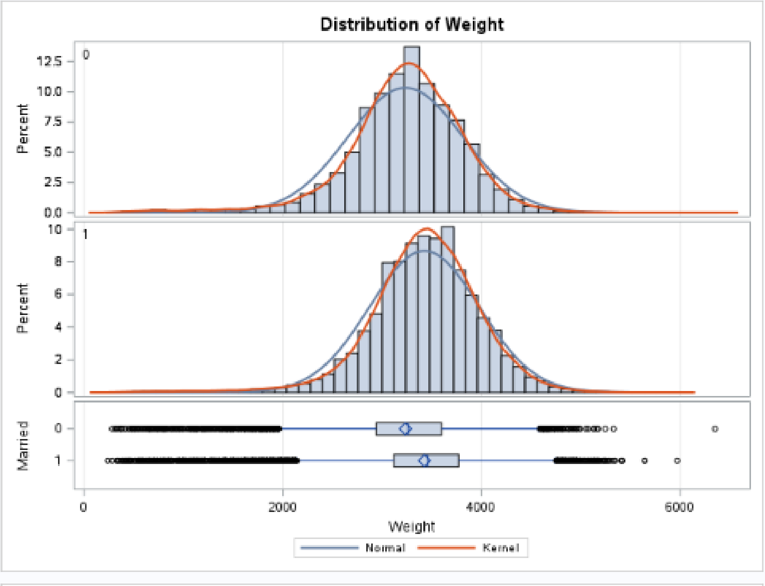
Variable: Weight (Weight)

Married	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		14369	3234.4	579.0	4.8302	284.0	6350.0
1		35631	3425.7	551.8	2.9231	240.0	5970.0
Diff (1-2)	Pooled		-191.3	559.7	5.5315		
Diff (1-2)	Satterthwaite		-191.3		5.6459		

Married	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		3234.4	3225.0	3243.9	579.0	572.4	585.8
1		3425.7	3420.0	3431.5	551.8	547.8	555.9
Diff (1-2)	Pooled	-191.3	-202.1	-180.5	559.7	556.3	563.2
Diff (1-2)	Satterthwaite	-191.3	-202.4	-180.2			

In the t-test procedure, we can see that there are 14,369 unmarried woman in the population with a mean infant weight of 3,234g. There are also 35,631 married woman with a mean infant weight of 3,425. There is a mean infant weight difference of about 191 grams. Both datasets exceed 30 datapoints, thus allowing the t-test assumption to hold true.

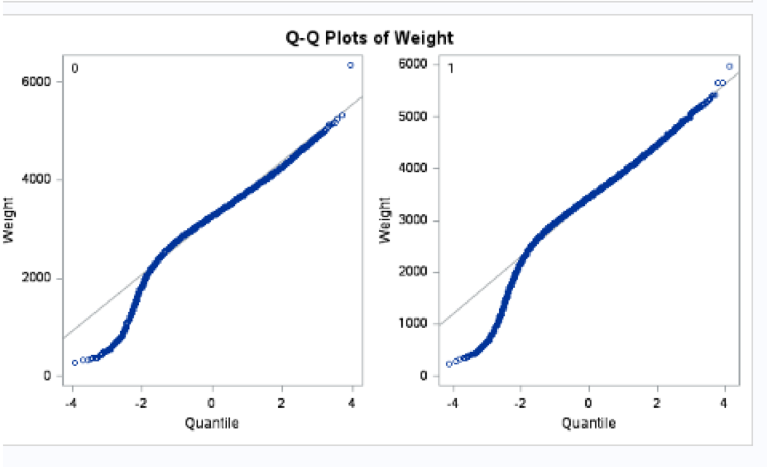
The standard deviation of the infant weight of unmarried woman is 579, and the standard deviation of infant weight of married woman is 551.



.

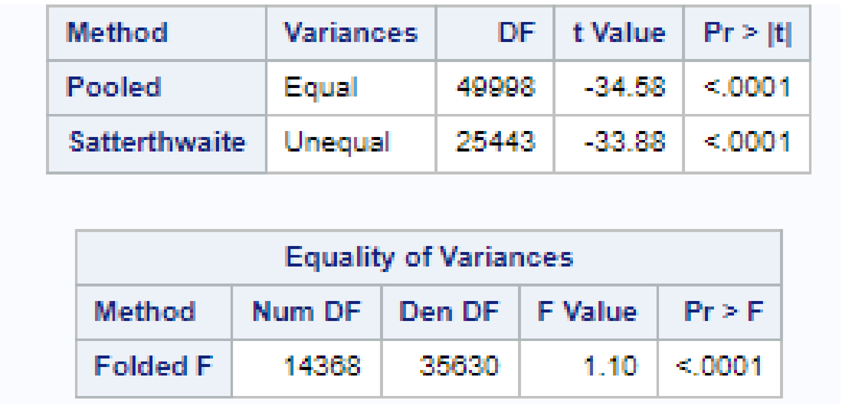
The histogram 0 shows the distribution of weight for unmarried woman while the histogram 1 shows the distribution of weight for married woman. They are both bell-shaped and symmetrical, which are signs that there is normality in the data.

The boxplots below the histogram show a similar width of the boxplot and thus similar interquartile ranges, which indicates equal variation. Also, the mean (symbol) and the median (line in boxplot) align, indicating a normal distribution.



Looking at the q-q plots, we can see that the data points for infant weight of unmarried and married woman roughly follows a normal distribution, with the values in the middle lining up with the line indicating normal distribution. Both married and unmarried have small deviation at the left side of the q-q plot, but mostly follow the normal distribution.

Step 4: Conclusion



Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	49998	-34.58	<.0001
Satterthwaite	Unequal	25443	-33.88	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	14368	35630	1.10	<.0001

With the t-test following the assumptions of normal distribution, independent samples, and has a more than 30 datapoints in each group, the t-test procedure can be conducted.

First, the equality of variances must be tested, and this is done by comparing the Pr >F value with the significance level of **α = 0.05,** where the significance level chosen is common practice.

[Null Hypothesis] H0:

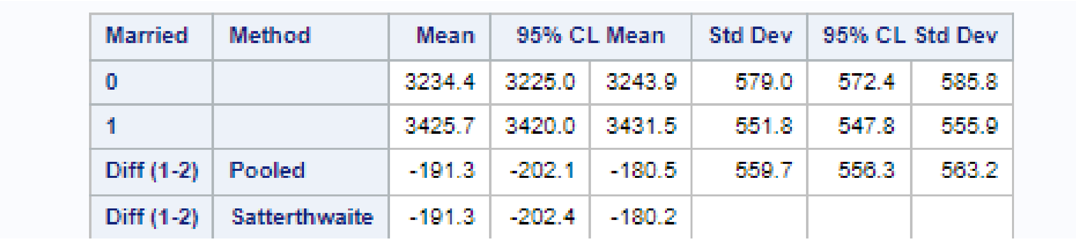
The variances between infant weight of unmarried and married woman are the same.

[Alternative Hypothesis] H1:

The variances between infant weight of unmarried and married woman are different.

As the Pr> F value is <0.0001, which is less than α = 0.05, there is a significant difference between the variances, and the null hypothesis H0 is rejected. Thus we will be using the Satterthwaite section showing unequal variances.

The p-value read in Pr > |t| is <0.0001, which is less than α = 0.05, showing a significant difference of infant birth weight between the two groups unmarried and married. The null hypothesis at Step 1 ($\mu_{unmarried} = \mu_{married}$) is rejected.



Married	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		3234.4	3225.0	3243.9	579.0	572.4	585.8
1		3425.7	3420.0	3431.5	551.8	547.8	555.9
Diff (1-2)	Pooled	-191.3	-202.1	-180.5	559.7	556.3	563.2
Diff (1-2)	Satterthwaite	-191.3	-202.4	-180.2			

In conclusion, there is a significant difference in weight of infant birth between married and unmarried mothers. This indicates there is a relationship between infant birth weight and the married variable.

The average infant birth weight for married mothers is 191 grams heavier than unmarried mothers, with a 95% confidence level of married mothers having an infant birth weight of 180.2 to 202.4 grams more than unmarried mothers, as seen in the Satterthwaite 95% CL Mean Diff (1-2) in the table above.

2. <u>T-test on Weight variable and Boy Variable</u>

Comparing the infant birth weight and boy variable, the null and alternative hypothesis must be first stated:

<u>Step 1: Setting the Null and Alternative Hypothesis</u>

[Null Hypothesis] H0:

$\mu_{boy} = \mu_{girl}$ (The mean weight of infant birth for a baby boy is the same as the weight of infant birth for a baby girl)

[Alternative Hypothesis] H1:

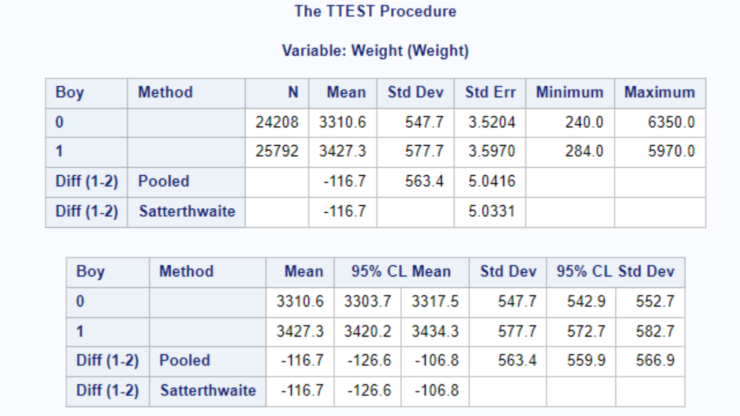
$\mu_{boy} \neq \mu_{girl}$ (The mean weight of infant birth for a baby boy is different from the weight of infant birth for a baby girl)

<u>Step 2: Choosing the significance level</u>

Choosing a significance level of α = 0.05 as it is common practice.

<u>Step 3: Checking Assumptions</u>

In the Boy variable, there are two groups 0 and 1, referring to baby girl and baby boy. This variable is the independent variable in the t-test. The dependent variable Weight contains the numeric value which refers to the grams of the baby. The values in the baby boy group do not affect the values in the baby girl group, thus the sample is independent.



The TTEST Procedure

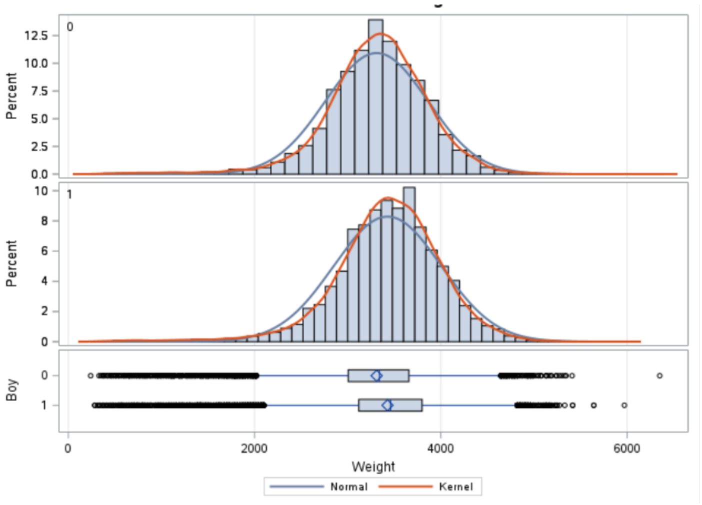
Variable: Weight (Weight)

Boy	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		24208	3310.6	547.7	3.5204	240.0	6350.0
1		25792	3427.3	577.7	3.5970	284.0	5970.0
Diff (1-2)	Pooled		-116.7	563.4	5.0416		
Diff (1-2)	Satterthwaite		-116.7		5.0331		

Boy	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		3310.6	3303.7	3317.5	547.7	542.9	552.7
1		3427.3	3420.2	3434.3	577.7	572.7	582.7
Diff (1-2)	Pooled	-116.7	-126.6	-106.8	563.4	559.9	566.9
Diff (1-2)	Satterthwaite	-116.7	-126.6	-106.8			

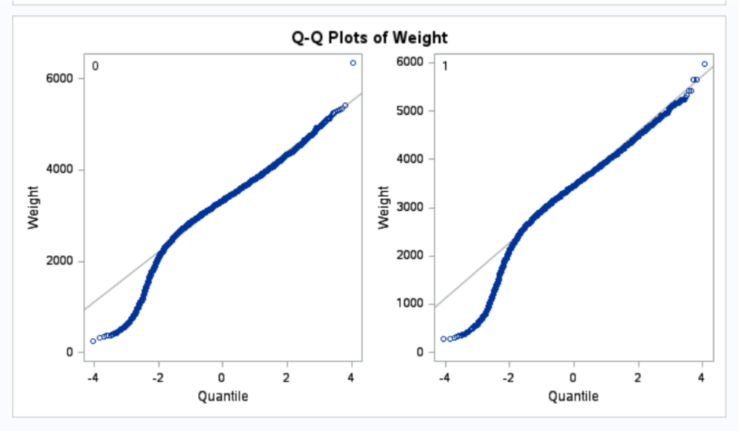
In the t-test procedure, we can see that there are 24,208 baby girls in the population with a mean infant weight of 3,310g. There are also 25,792 baby boys with a mean infant weight of 3,427. There is a mean infant weight difference of about 117 grams. Both datasets exceed 30 datapoints, thus allowing the t-test assumption to hold true.

The standard deviation of the infant weight of baby girls is 547, and the standard deviation of infant weight of baby boys is 577.



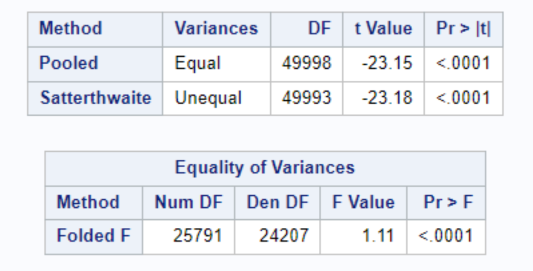
The histogram 0 shows the distribution of weight for baby girls while the histogram 1 shows the distribution of weight for baby boys. They are both bell-shaped and symmetrical, which are signs that there is normality in the data.

The boxplots below the histogram show a similar width of the boxplot and thus similar interquartile ranges, which indicates equal variation. Also, the mean (symbol) and the median (line in boxplot) align, indicating a normal distribution.



Looking at the q-q plots, we can see that the data points for infant weight of baby girls and baby boys roughly follows a normal distribution, with the values in the middle lining up with the line indicating normal distribution. Both baby girls and baby boys have small deviation at the left side of the q-q plot, but mostly follow the normal distribution.

Step 4: Conclusion



Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	49998	-23.15	<.0001
Satterthwaite	Unequal	49993	-23.18	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	25791	24207	1.11	<.0001

With the t-test following the assumptions of normal distribution, independent samples, and has a more than 30 datapoints, the t-test procedure can be conducted.

First, the equality of variances must be tested, and this is done by comparing the Pr >F value with the significance level of **α = 0.05,** where the significance level chosen is common practice.

[Null Hypothesis] H0:

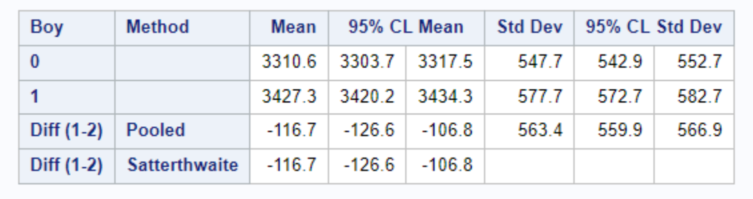
The variances between infant weight of baby boys and baby girls are the same.

[Alternative Hypothesis] H1:

The variances between infant weight of baby boys and baby girls are different.

As the Pr> F value is <0.0001, which is less than α = 0.05, showing there is a significant difference between the variances, and the null hypothesis H0 is rejected. Thus, we will be using the Satterthwaite section showing unequal variances.

The p-value read in Pr > |t| is <0.0001, which is less than α = 0.05, showing a significant difference of infant birth weight between the two groups baby boy and baby girl. The null hypothesis at Step 1 ($\mu_{boy} = \mu_{girl}$) is rejected.



Boy	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		3310.6	3303.7	3317.5	547.7	542.9	552.7
1		3427.3	3420.2	3434.3	577.7	572.7	582.7
Diff (1-2)	Pooled	-116.7	-126.6	-106.8	563.4	559.9	566.9
Diff (1-2)	Satterthwaite	-116.7	-126.6	-106.8			

In conclusion, there is a significant difference in weight of infant birth between baby boys and baby girls. This indicates there is a relationship between infant birth weight and the Boy variable.

The average infant birth weight for baby boys is 117 grams heavier than baby girls, with a 95% confidence level of baby boys having an infant birth weight of 106 to 126 grams more than baby girls, as seen in the Satterthwaite 95% CL Mean Diff (1-2) in the table above.

3. T-test on Weight variable and MomSmoke Variable

Comparing the infant birth weight and MomSmoke variable, the null and alternative hypothesis must be first stated:

Step 1: Setting the Null and Alternative Hypothesis

[Null Hypothesis] H0:

$\mu_{Smoke} = \mu_{NoSmoke}$ (The mean weight of infant birth for a mom who smokes is the same as the weight of infant birth for a mom who doesn't smoke)

[Alternative Hypothesis] H1:

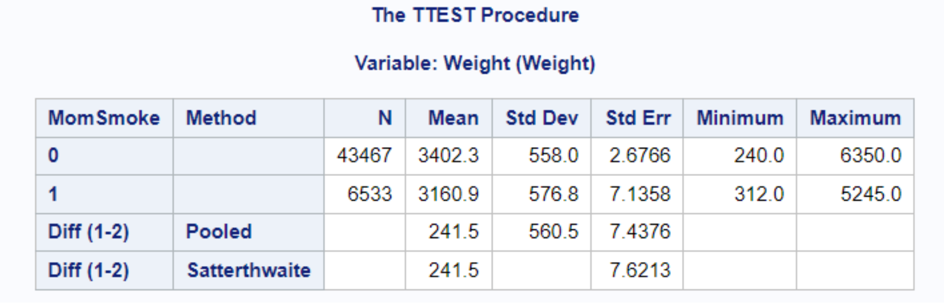
$\mu_{boy} \neq \mu_{girl}$ (The mean weight of infant birth for a mom who smokes different from the weight of infant birth for a mom who doesn't smoke)

Step 2: Choosing the significance level

Choosing a significance level of α = 0.05 as it is common practice.

Step 3: Checking Assumptions

In the MomSmoke variable, there are two groups 0 and 1, referring to a mother who doesn't smoke and mother who smokes. This variable is the independent variable in the t-test. The dependent variable Weight contains the numeric value which refers to the grams of the baby. The values in the Mother that smokes group do not affect the values in the mother does not smoke group, thus the sample is independent.



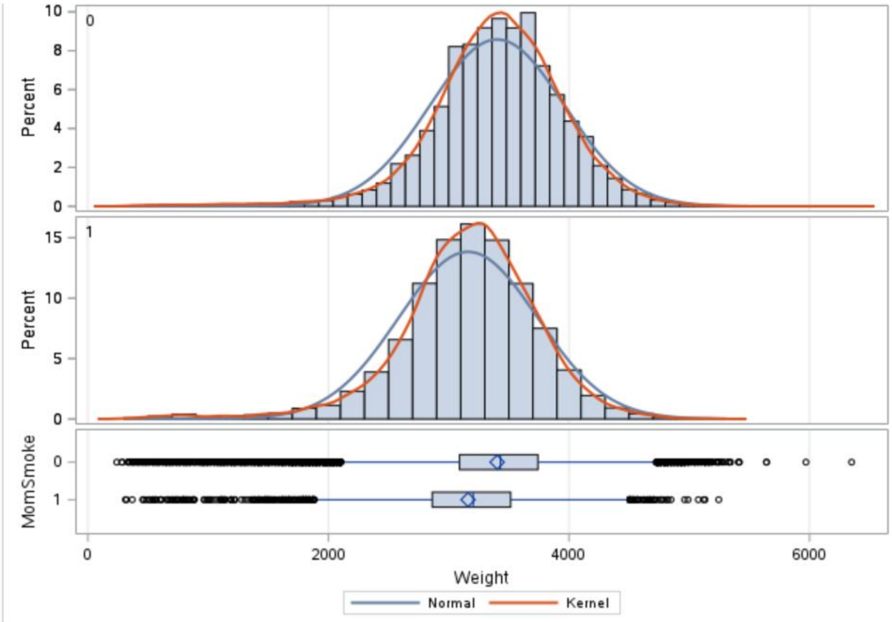
The TTEST Procedure

Variable: Weight (Weight)

MomSmoke	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		43467	3402.3	558.0	2.6766	240.0	6350.0
1		6533	3160.9	576.8	7.1358	312.0	5245.0
Diff (1-2)	Pooled		241.5	560.5	7.4376		
Diff (1-2)	Satterthwaite		241.5		7.6213		

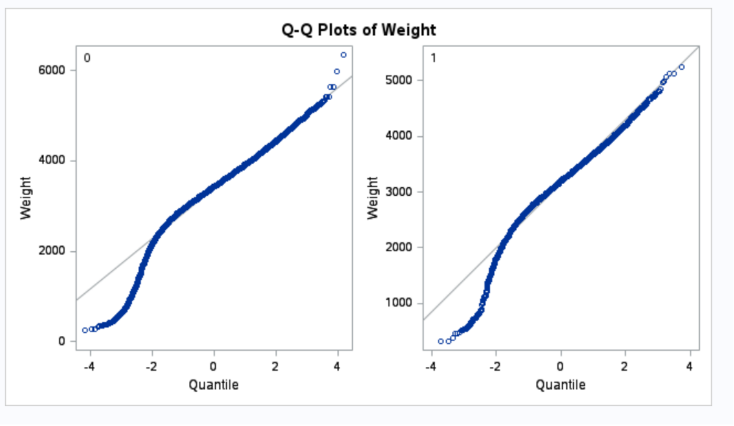
In the t-test procedure, we can see that there are 43,467 mothers who don't smoke with a mean infant weight of 3,402grams. There are also 6,533 mothers who smoke with a mean infant weight of 3,160. There is a mean infant weight difference of about 242 grams. Both datasets exceed 30 datapoints, thus allowing the t-test assumption to hold true.

The standard deviation of the infant weight of mothers who don't smoke is 558, and the standard deviation of infant weight of mothers who smoke is 577.



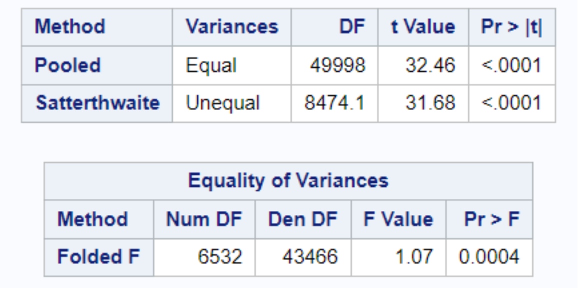
The histogram 0 shows the distribution of weight for infants from mothers who don't smoke while the histogram 1 shows the distribution of weight for infants from mothers who smoke. They are both bell-shaped and symmetrical, which are signs that there is normality in the data.

The boxplots below the histogram show a similar width of the boxplot and thus similar interquartile ranges, which indicates equal variation. Also, the mean (symbol) and the median (line in boxplot) align, indicating a normal distribution.



Looking at the q-q plots, we can see that the data points for infant weight from mothers who don't smoke and mothers who smoke roughly follows a normal distribution, with the values in the middle lining up with the line indicating normal distribution. Both groups have small deviation at the left side of the q-q plot.

Step 4: Conclusion



Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	49998	32.46	<.0001
Satterthwaite	Unequal	8474.1	31.68	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	6532	43466	1.07	0.0004

With the t-test following the assumptions of normal distribution, independent samples, and has more than 30 datapoints in each group, the t-test procedure can be conducted.

First, the equality of variances must be tested, and this is done by comparing the Pr >F value with the significance level of **α = 0.05,** where the significance level chosen is common practice.

[Null Hypothesis] H0:

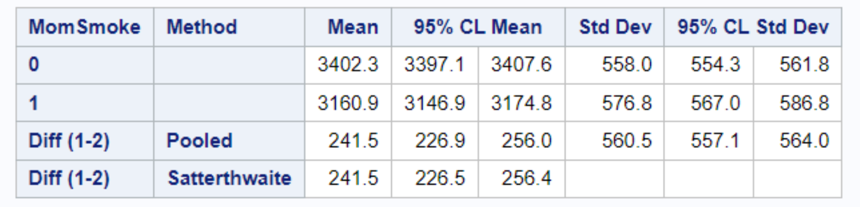
The variances between infant weight from mothers who smoke and don't smoke are the same.

[Alternative Hypothesis] H1:

The variances between infant weight from mothers who smoke and don't smoke are different.

As the Pr> F value is 0.0004, which is less than α = 0.05, showing there is a significant difference between the variances, and the null hypothesis H0 is rejected. Thus, we will be using the Satterthwaite section showing unequal variances.

The p-value read in Pr > |t| is <0.0001, which is less than α = 0.05, showing a significant difference of infant birth weight between the two groups mothers who smoke and mothers who don't smoke. The null hypothesis at Step 1 ($\mu_{Smoke} = \mu_{NoSmoke}$) is rejected.



MomSmoke	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		3402.3	3397.1	3407.6	558.0	554.3	561.8
1		3160.9	3146.9	3174.8	576.8	567.0	586.8
Diff (1-2)	Pooled	241.5	226.9	256.0	560.5	557.1	564.0
Diff (1-2)	Satterthwaite	241.5	226.5	256.4			

In conclusion, there is a significant difference in weight of infant birth between mothers who smoke and mothers who don't smoke. This indicates there is a relationship between infant birth weight and the MomSmoke variable.

The average infant birth weight for mothers who don't smoke is 242 grams heavier than infants from mothers who smoke, with a 95% confidence level of mothers who don't smoke having an infant birth weight of 226 to 256 grams more than from mothers who smoke, as seen in the Satterthwaite 95% CL Mean Diff (1-2) in the table above.