BIVARIATE ANALYSIS

(DESCRIPTIVE STATISTIC ANALYSIS)

Maryam Najimigoshtasb

Analyzing dependencies, influences or relations of variables either IV with DV or DV with DV are of importance. In this document we concentrate on bivariate of our IV and categorical independent variable in the case of one or two samples level.

We are trying to predict weight of new born infant on the basis of some attributes. Before that one of the stages is to see, Do these attributes make a significant differences or have impact on weight?

In this specific document we are trying to answer four following questions ?

1)- Are there any differences between the weight of babies with white mother or black mother?
2)-Are there any any differences between the weight of babies who their mothers are married or single?
3)-Does the gender of bay make any difference in the weight of babies?
4)-Does mother smoking make any differences on the weight of babies?

On the base of these questions we make our hypothesis and we try to apply proper statistic test on them.

However there are other variable we need to check their relation or their dependency with weight. For instance, we need to use **correlation** between weight and mothers age(momage) , number of cigars smoked by per day by mom (CigsPerDay) and Mom's gain weight(MomWtGain) to see are they In one direction of increase or decrease? or they are not related at all. Moreover we can apply **multicollinearity**, **simple regression and multiple regression**. And on the basis of these we will select our important variable for our model.

On the other hand, we need to use **ANOVA** test for weight and mothers' level of educations(MomEdLevel) an visit.

In all the slides we have the **null hypothesis** in which we assume means in the samples are same among the groups of our categorical variable and **alternative hypothesizes** are inequality of means.

As we don't have the variance we use **t-test** and in which the test will consider two test for **equality and non equality of variance** for each sample.

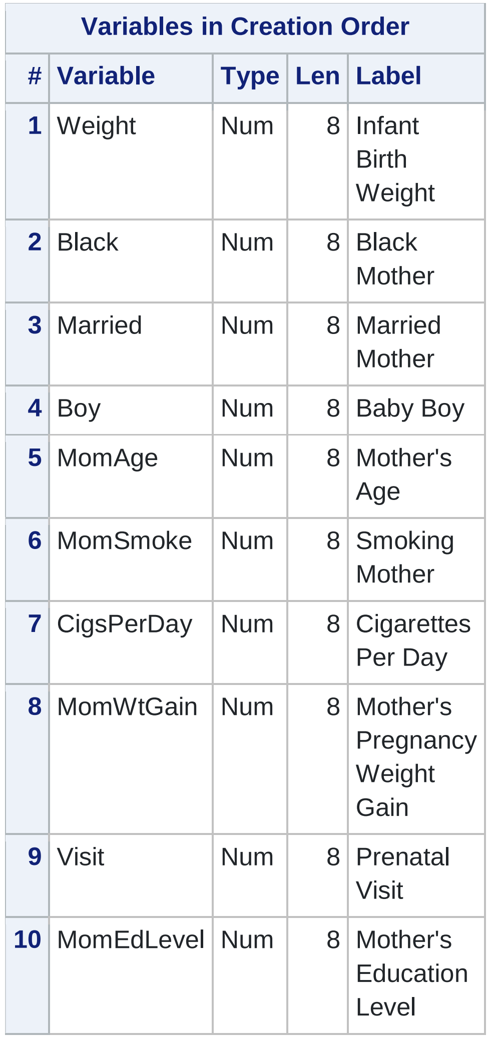
If the we fail to reject the hypothesis it means the categorical variable dose not make any differences on weight then we might delete or not considering them in our model. However, I believe we need to to more investigations.

Before starting the test we are going to put a brief descriptions on the variable.

Among the 10 independent variables, *Black, Married, Boy*, and *MomSmoke* are binary variables. For these variables, the mean represents the proportion in the category. The two continuous variables, MomAge and MomWtGain, are centered at their medians, which are 27 and 30, respectively.

There are four *levels of maternal education*. High School, Some College, College and Less Than High School which are respectively formatted to 0, 1,2,3.

. Likewise, there are four *levels of prenatal medical care of the mother*. No Visit, Second Trimester, Last Trimester and First Trimester which are respectively formatted to 0, 1,2,3



Variables in Creation Order

#	Variable	Type	Len	Label
1	Weight	Num	8	Infant Birth Weight
2	Black	Num	8	Black Mother
3	Married	Num	8	Married Mother
4	Boy	Num	8	Baby Boy
5	MomAge	Num	8	Mother's Age
6	MomSmoke	Num	8	Smoking Mother
7	CigsPerDay	Num	8	Cigarettes Per Day
8	MomWtGain	Num	8	Mother's Pregnancy Weight Gain
9	Visit	Num	8	Prenatal Visit
10	MomEdLevel	Num	8	Mother's Education Level

Weight -black

Are there any differences between the weight of babies with white mother or black mother?

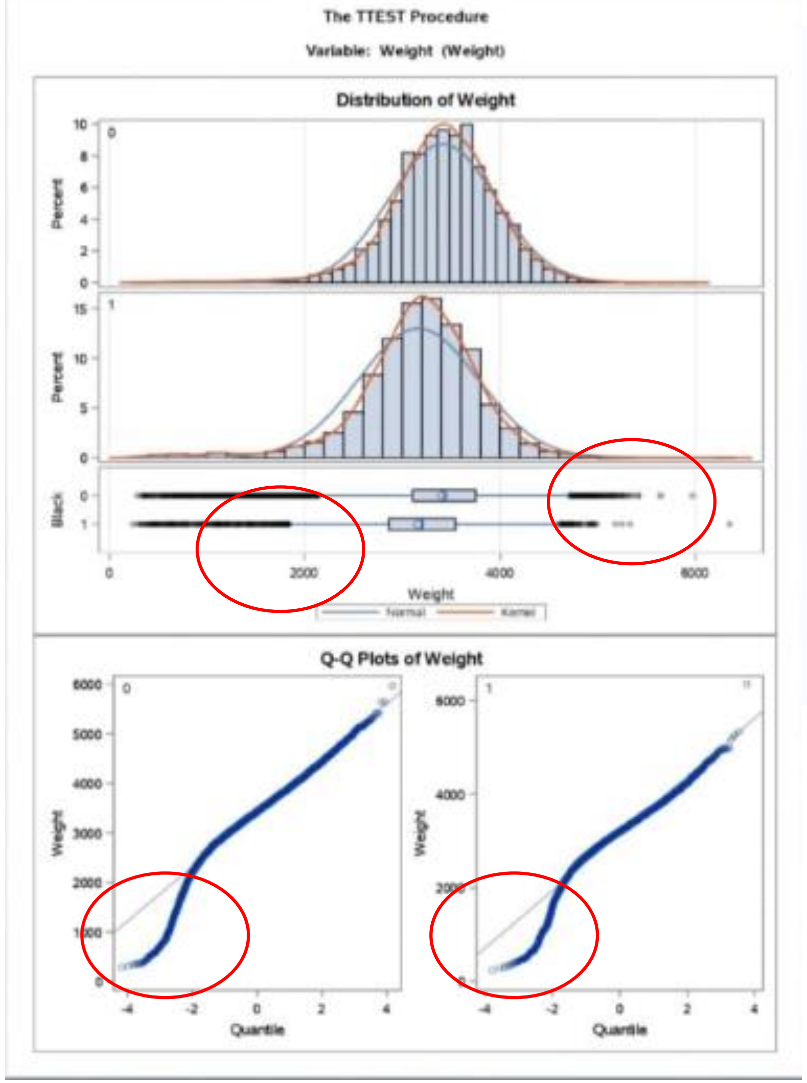
$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases}$$

As we did not have the variance t-test considered both equal and unequal variances case.(pooled, Satterthwaite)

We can see here **F-test on** equal variance is lower than the default level of significant 0.05 so we reject the equality of variance so we look at the case of Satterthwaite and this giving us the **p-value of <.0001** which is lower than $\alpha = .05$, in which we reject the equality of mean.
Looking at the distribution of the black and not black(white)we can see that on the each side of the whisker box lot there are **outliers** which means our **distribution** is **not quite normal** which more clear with qqplot.
the **confidence interval** does not include the 0 which the value of our hypothesis and this another sign for rejecting the hypothesis.

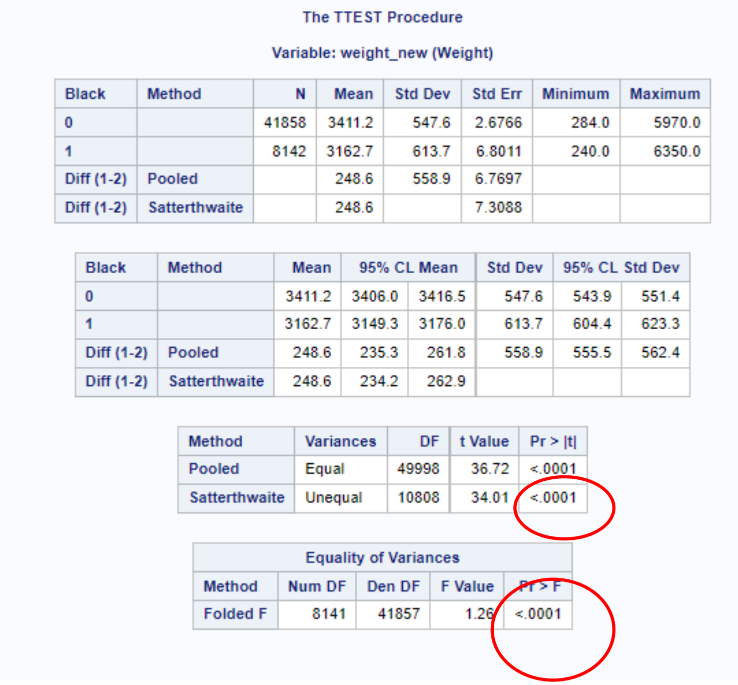
All these mean that black mother or white mother make a difference on the babies' weight. In fact there is a significant difference in average weight of babies who have black mother or white mother.

I suppose as the variance are not equal it means the mother's color has a significant contribution.



The TTEST Procedure

Variable: Weight (Weight)



The TTEST Procedure

Variable: weight_new (Weight)

Black	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		41858	3411.2	547.6	2.6766	284.0	5970.0
1		8142	3162.7	613.7	6.8011	240.0	6350.0
Diff (1-2)	Pooled		248.6	558.9	6.7697		
Diff (1-2)	Satterthwaite		248.6		7.3088		

Black	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		3411.2	3406.0	3416.5	547.6	543.9	551.4
1		3162.7	3149.3	3176.0	613.7	604.4	623.3
Diff (1-2)	Pooled	248.6	235.3	261.8	558.9	555.5	562.4
Diff (1-2)	Satterthwaite	248.6	234.2	262.9			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	49998	36.72	<.0001
Satterthwaite	Unequal	10808	34.01	<.0001

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	8141	41857	1.26	<.0001

Married -black

Are there any any differences between the weight of babies who their mothers are married or single?

$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$

As we did not have the variance t-test considered both equal and unequal variances case.(pooled, Satterthwaite)

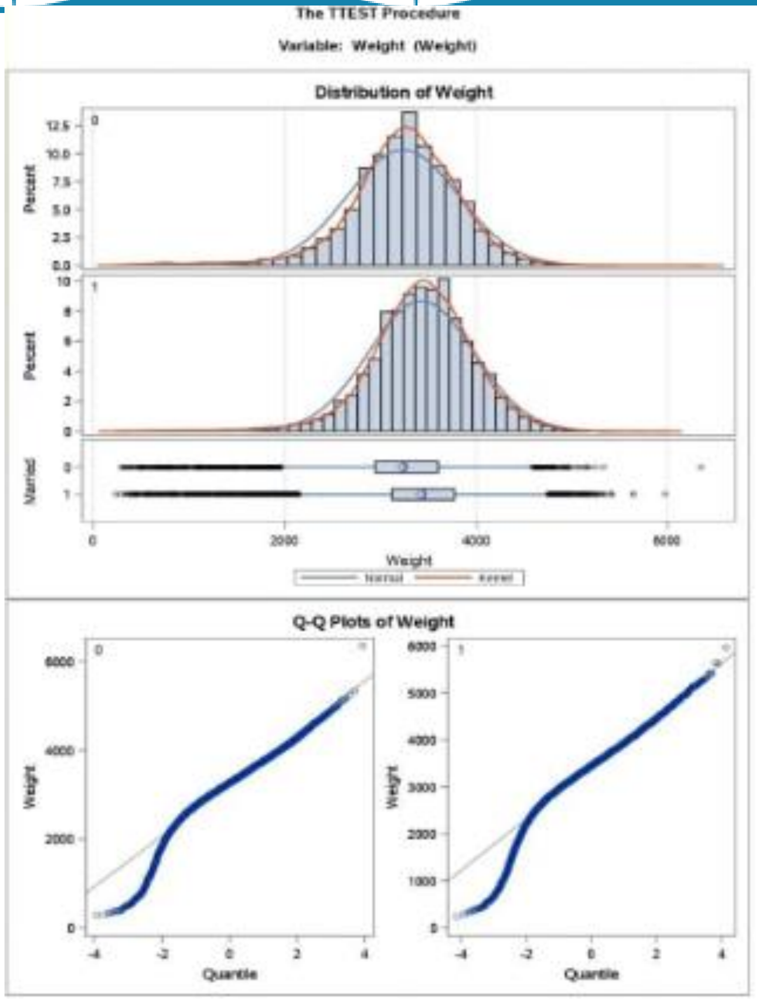
We can see here F-test on equal variance is lower than the default level of significant 0.05 so we reject the equality of variance so we look at the case of Satterthwaite and this giving us the p-value of $<.0001$ which is lower than $\alpha = .05$, in which we reject the equality of mean.
Looking at the distribution of the Married and single Mother we can see that on the each side of the whisker box lot there are **outliers** which means our **distribution** is **not quite normal** which more clear with qqplot.

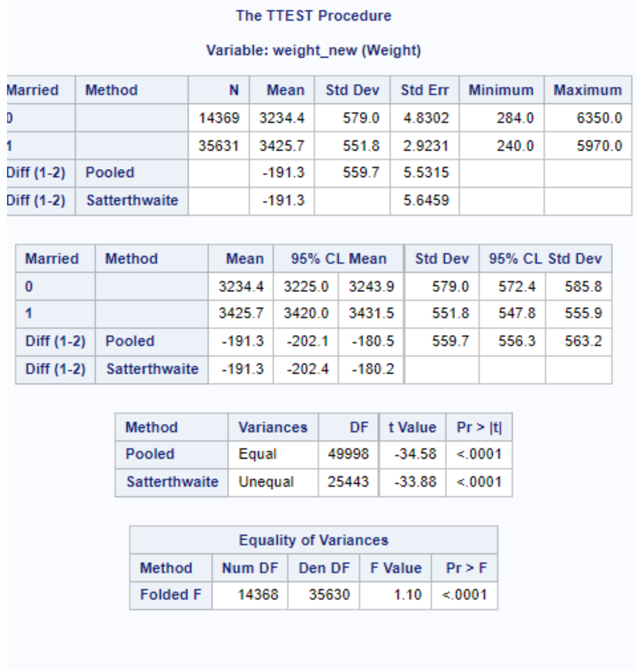
the **confidence interval** does not include the 0 which the value of our hypothesis and this another sign for rejecting the hypothesis.

All these mean that the Married and single Mother make a difference on the babies' weight. In fact there is a significant difference in average weight of babies who have the Married or single Mother

All these means Married will affect on the babies' weight.

I suppose as the variance are not equal it means the Status of mother has a significant contribution.





The TTEST Procedure

Variable: weight_new (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			

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

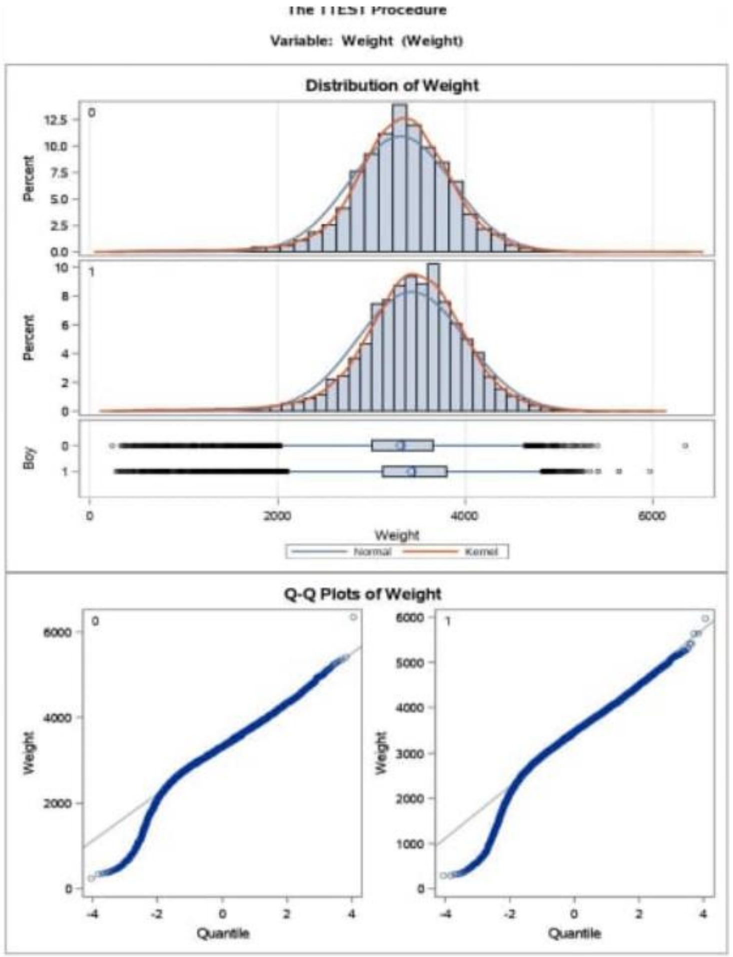
boy -black

Does the gender of bay make any difference in the weight of babies?

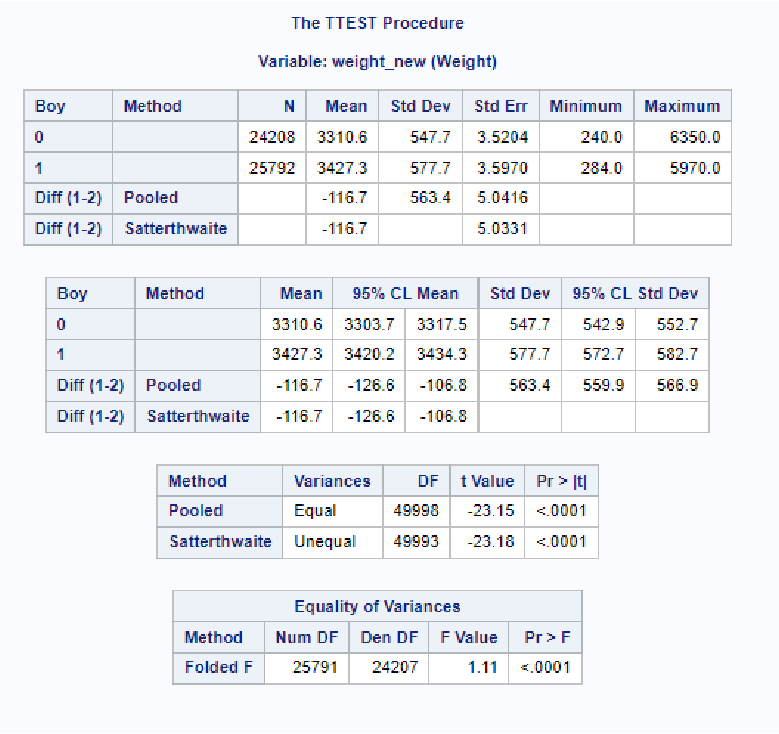
$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases}$$

As we did not have the variance t-test considered both equal and unequal variances case.(pooled, Satterthwaite)
We can see here F-test on equal variance is lower than the default level of significant 0.05 so we reject the equality of variance so we look at the case of Satterthwaite and this giving us the p-value of <.0001 which is lower than $\alpha = .05$, in which we reject the equality of mean.
Looking at the distribution of gender of baby, being boy or girl, we can see that on the each side of the whisker box lot there are **outliers** which means our **distribution** is **not quite normal** which more clear with qqplot.
the **confidence interval** does not include the 0 which the value of our hypothesis and this another sign for rejecting the hypothesis.

All these mean that sex of baby make a difference on the babies' weight. In fact there is a significant difference in average weight of babies who are boy or girl


The TTEST Procedure

Variable: Weight (Weight)


The TTEST Procedure

Variable: weight_new (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			

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

momsmoke -black

Does mother smoking make any differences on the weight of babies?

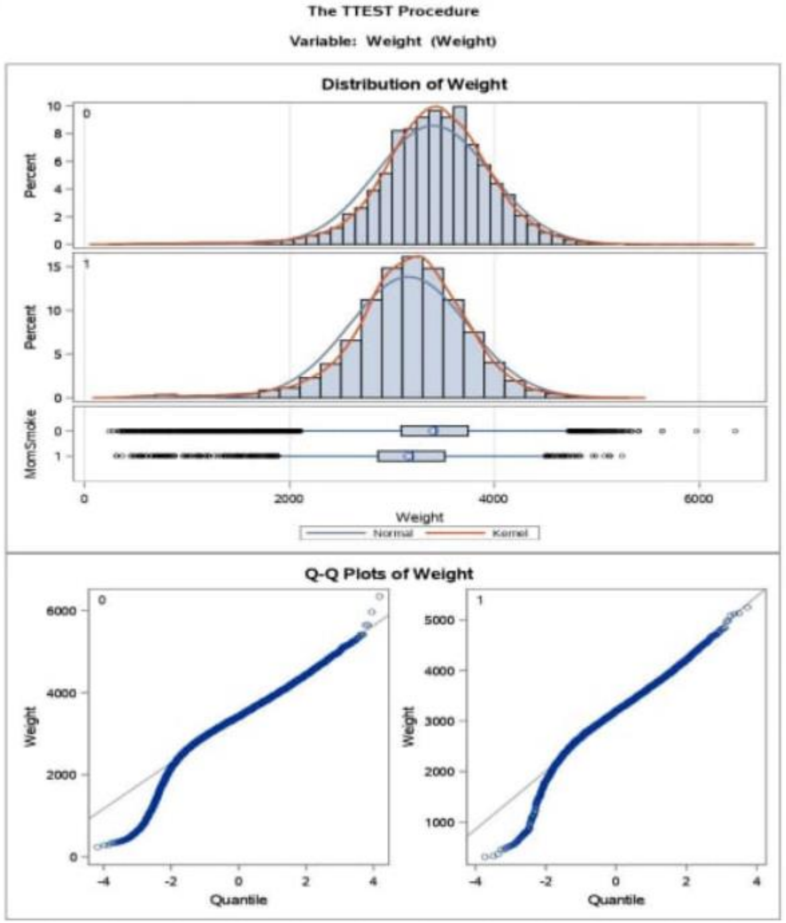
$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$

As we did not have the variance t-test considered both equal and unequal variances case.(pooled, Satterthwaite)

We can see here F-test on equal variance is lower than the default level of significant 0.05 so we reject the equality of variance so we look at the case of Satterthwaite and this giving us the p-value of <.0001 which is lower than $\alpha = .05$, in which we reject the equality of mean.

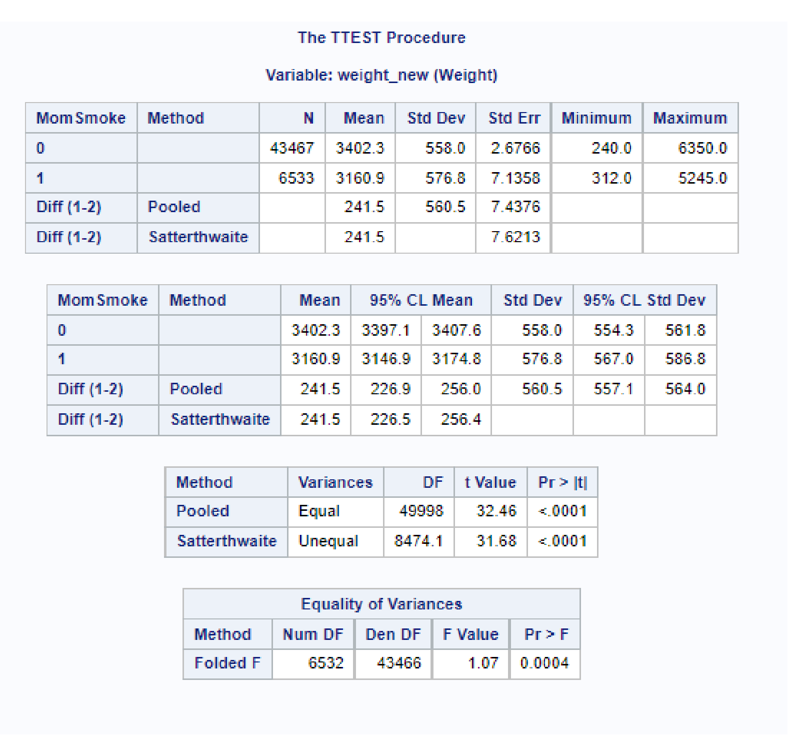
Looking at the distribution of whether mom smoking or not , we can see that on the each side of the whisker box lot there are **outliers** which means our **distribution** is **not quite normal** which more clear with qqplot.
the **confidence interval** does not include the 0 which the value of our hypothesis and this another sign for rejecting the hypothesis.

All these mean that whether mom smoking or not make a difference on the babies' weight. In fact there is a significant difference in average weight of babies who their mom smokes or not.



The TTEST Procedure

Variable: Weight (Weight)



The TTEST Procedure

Variable: weight_new (Weight)

Mom Smoke	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		

Mom Smoke	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			

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

Maryam Najimigoshtasb.

Mycodes

```
libname mylib '/home/u58699890/My practice/mylib';
filename Birth2 '/home/u58699890/My practice/mylib/File
BIRTH.xlsx';
proc import datafile= Birth2 out= mylib.birth replace dbms=xlsx ;
run;

data mylib.binew;
set mylib.birth;
rename Weight= weight_new;
run;

proc sort data=mylib.binew out= mylib.sortbirthblack;
by black;

run;

proc means data= mylib.sortbirthblack noprint;
var weight_new;
by black;
output out=mylib.w_black;
run;

proc print data=mylib.w_black;
run;

ods graphics on;
proc ttest data=mylib.w_black ;
class black;
var weight_new;
run;
ods graphics off;
```

```
proc sort data=mylib.binew out= mylib.sortbirthMaried;
by Married;

 run;
proc means data= mylib.sortbirthMaried noprint;
var weight_new;
by Married;
output out=mylib.w_married;
run;
proc print data=mylib.w_married; run;

 proc ttest data=mylib.w_married;
 class Married;
var weight_new;
run;
```

```
proc sort data=mylib.binew out= mylib.sortbirthms;
by MomSmoke;
run;
proc means data= mylib.sortbirthms noprint;
var weight_new;
by MomSmoke;
output out=mylib.w_momsmoke;
run;
proc print data=mylib.w_momsmoke; run;

 proc ttest data=mylib.w_momsmoke;
 class MomSmoke;
var weight_new;
run;
```

```
proc sort data=mylib.binew out= mylib.sortbirthboy;
by boy;
run;
proc means data= mylib.sortbirthboy noprint;
var weight_new;
by boy;
output out=mylib.w_boy;
run;
proc print data=mylib.w_boy; run;

 proc ttest data=mylib.w_boy;
 class boy;
var weight_new;
run;
```