Step 1 : Describing Data and Auditing

```
/* Generated Code (IMPORT) */
/* Source File: File BIRTH.xlsx */
/* Source Path: /home/u54770142/BAN100StatisticsforAnalytics */
/* Code generated on: 2/16/21, 9:28 PM */

%web_drop_table(WEBWORK.BIRTH);


FILENAME REFFILE '/home/u54770142/BAN100StatisticsforAnalytics/File BIRTH.xlsx';

PROC IMPORT DATAFILE=REFFILE
    DBMS=XLSX
    OUT=WEBWORK.BIRTH;
    GETNAMES=YES;
    SHEET="BIRTH";
RUN;

PROC CONTENTS DATA=WEBWORK.BIRTH; RUN;


%web_open_table(WEBWORK.BIRTH);
```

The CONTENTS Procedure

Data Set Name	WEBWORK.BIRTH	**Observations**	50000
Member Type	DATA	**Variables**	10
Engine	V9	**Indexes**	0
Created	02/16/2021 21:28:32	**Observation Length**	80
Last Modified	02/16/2021 21:28:32	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	31
First Data Page	1
Max Obs per Page	1635
Obs in First Data Page	1596
Number of Data Set Repairs	0
Filename	/home/u54770142/.sasstudio/webwork/RS10402/birth.sas7bdat
Release Created	9.0401M6
Host Created	Linux
Inode Number	8869778948
Access Permission	rw-r--r--
Owner Name	u54770142
File Size	4MB
File Size (bytes)	4194304

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Label
2	Black	Num	8	BEST.	Black
4	Boy	Num	8	BEST.	Boy
7	CigsPerDay	Num	8	BEST.	CigsPerDay
3	Married	Num	8	BEST.	Married
5	MomAge	Num	8	BEST.	MomAge
10	MomEdLevel	Num	8	BEST.	MomEdLevel
6	MomSmoke	Num	8	BEST.	MomSmoke
8	MomWtGain	Num	8	BEST.	MomWtGain

All Variables are numeric so we have to determine if any is bi-categorical to be able to run T-test.

The CONTENTS Procedure

Alphabetic List of Variables and Attributes					
#	**Variable**	**Type**	**Len**	**Format**	**Label**
9	Visit	Num	8	BEST.	Visit
1	Weight	Num	8	BEST.	Weight

```
*Auditing the Data Base Variables and its values to determine type of variable:
Categorical or not? an dIf yes, how many categories.;

Proc PRint Data=webwork.birth (obs=51);
Run;
```

Obs	Weight	Black	Married	Boy	MomAge	MomSmoke	CigsPerDay	MomWtGain	Visit	MomEdLevel
1	4111	0	1	1	-3	0	0	-16	1	0
2	3997	0	1	0	1	0	0	2	3	2
3	3572	0	1	1	0	0	0	-3	3	0
4	1956	0	1	1	-1	0	0	-5	3	2
5	3515	0	1	1	-6	0	0	-20	3	0
6	3757	0	1	0	3	0	0	0	3	2
7	2977	1	0	1	-5	1	5	5	3	0
8	3884	0	0	0	-5	0	0	0	3	2
9	3629	0	1	0	6	0	0	-5	3	0
10	3062	0	1	1	-1	0	0	6	3	2
11	4026	0	1	1	-2	1	4	22	3	1
12	3642	0	1	1	-6	0	0	-1	3	0
13	2296	0	0	1	0	0	0	7	3	0
14	2665	0	0	1	1	1	10	-6	3	1
15	2948	0	1	1	1	0	0	10	3	1
16	3467	0	1	0	7	0	0	15	3	0
17	3430	1	1	1	-4	0	0	-6	3	0
18	4139	0	1	0	2	0	0	-2	3	1
19	3657	0	1	0	6	0	0	15	3	0
20	2722	0	1	0	5	0	0	12	3	1
21	4026	0	1	1	1	0	0	0	3	0
22	4224	0	1	1	7	0	0	0	3	2
23	3459	0	1	1	6	0	0	-1	3	0
24	3544	0	1	1	1	0	0	0	3	2
25	4054	0	1	1	-5	0	0	21	3	2
26	3380	0	1	1	7	0	0	0	3	0
27	3430	0	0	0	0	0	0	-7	3	2
28	4054	0	1	0	1	1	15	5	3	1
29	3232	0	1	1	-5	0	0	-6	3	0
30	3045	0	1	1	5	0	0	-13	3	0
31	3575	0	1	1	-2	0	0	-8	3	1
32	3544	0	0	0	-6	0	0	-10	1	0
33	3760	0	1	1	6	0	0	-21	1	0
34	3572	0	1	1	10	0	0	5	3	1
35	4536	1	0	1	3	0	0	-1	3	3
36	2783	0	1	0	5	1	20	-23	3	0
37	3295	0	1	1	6	0	0	-29	3	2
38	2551	0	1	0	7	0	0	3	3	1
39	3374	1	1	1	7	0	0	-10	3	2
40	3147	0	1	0	10	0	0	-20	3	2
41	3544	0	1	1	1	0	0	15	3	1
42	3827	0	1	1	3	0	0	15	3	2
43	3458	0	1	1	8	0	0	-18	3	1
44	3997	0	1	1	-1	0	0	4	3	0
45	3147	0	1	0	10	0	0	8	3	1
46	3350	0	1	0	-2	0	0	-6	3	2
47	3856	0	0	0	-3	0	0	19	1	0
48	4281	0	1	1	-1	0	0	29	3	2
49	3686	0	1	0	4	0	0	25	3	1
50	3657	0	1	0	8	0	0	10	3	0
51	3714	0	0	1	2	0	0	25	1	1

Althougth the following variables are numeric, they have only 2 values (Boolean)so they can be considered categorical:
Boy
Black
Married
MomSmoke

```
*Relationship between variables weight and Boy.
Assumption: If baby is a Boy, Boy=1 and if baby is NOT baby, baby=0.
I'll run an unpaired t-test to find out if the weight mean is the same for babies that are born
 boy and the ones born girls.
 H0:u1=u2
 H1:u1<u2;

 *Title 'Conducting a two-sample t-test and demonstrating Weight is normally distributed';
 ODS graphics on;
 Proc ttest data=webwork.birth;
   class Boy;
   Var Weight;
 Run;
 ODS graphics off;
```

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		

Mean and variance differences assuming equal and unequal variances in populations

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

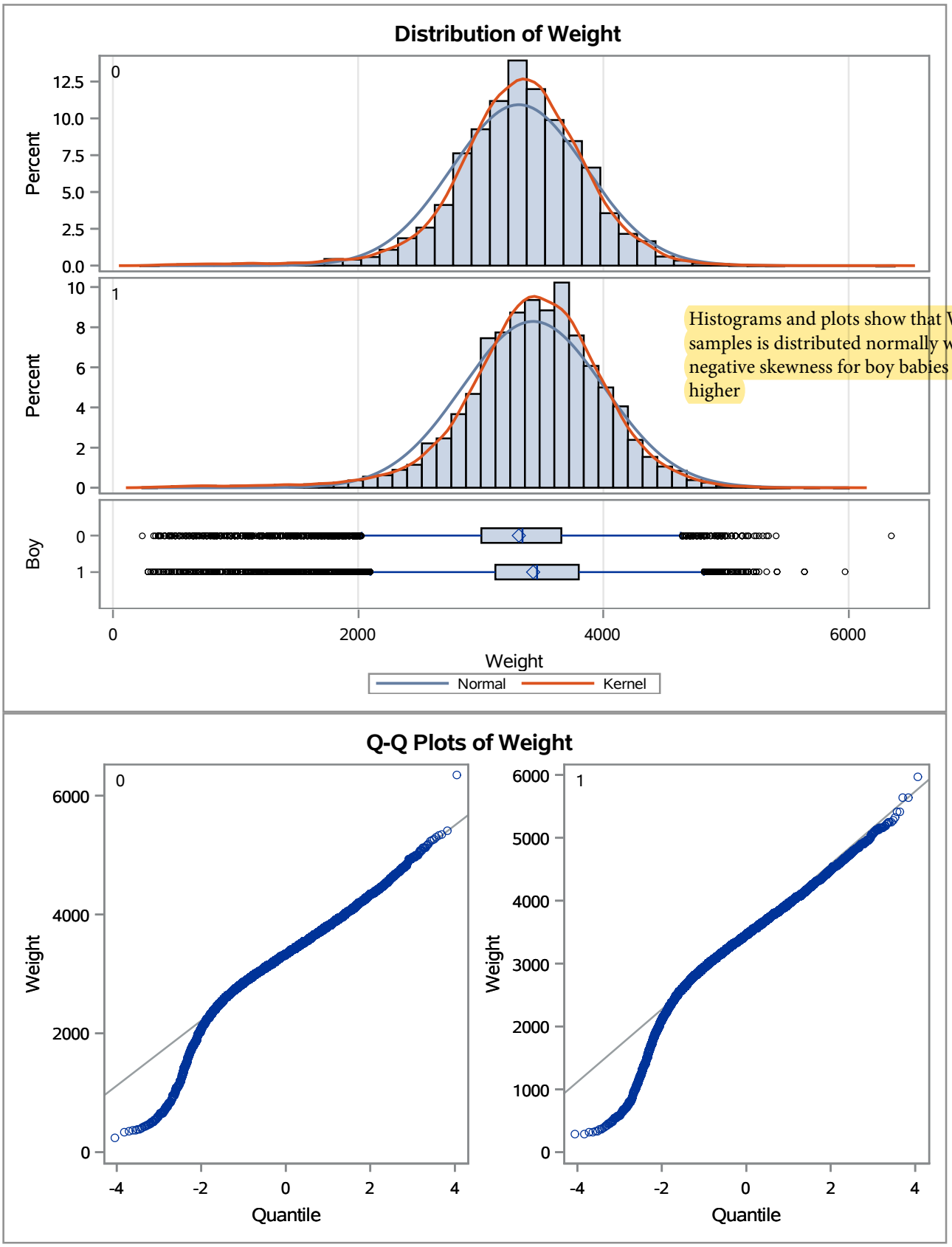
P value assuming equal variances is less than Confidence Interval 5%.

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

P value is less than confidence interval so we reject the null hypothesis that populations come from populations with equal variances.

The TTEST Procedure

Variable: Weight (Weight)



Histograms and plots show that Weight samples is distributed normally with slightly negative skewness for boy babies -weight is higher

```
*Relationship between variables weight and black.
Assumption: If mother is black, black=1 and if mother is NOT Black, black=0.
I'll run an unpaired t-test to find out if the weight mean is the same for babies that are born
 from a black mother and the ones born from non-black mothers.
 H0:u1=u2
 H1:u1<u2;



 ODS graphics on;
 *Title 'Conducting a two-sample t-test and demonstrating Weight is normally distributed';
 Proc ttest data=webwork.birth;
   class Black;
   Var Weight;
 Run;
 ODS graphics off;
```

The TTEST Procedure

Variable: Weight (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		

Mean and variance differences assuming equal and unequal variances in populations

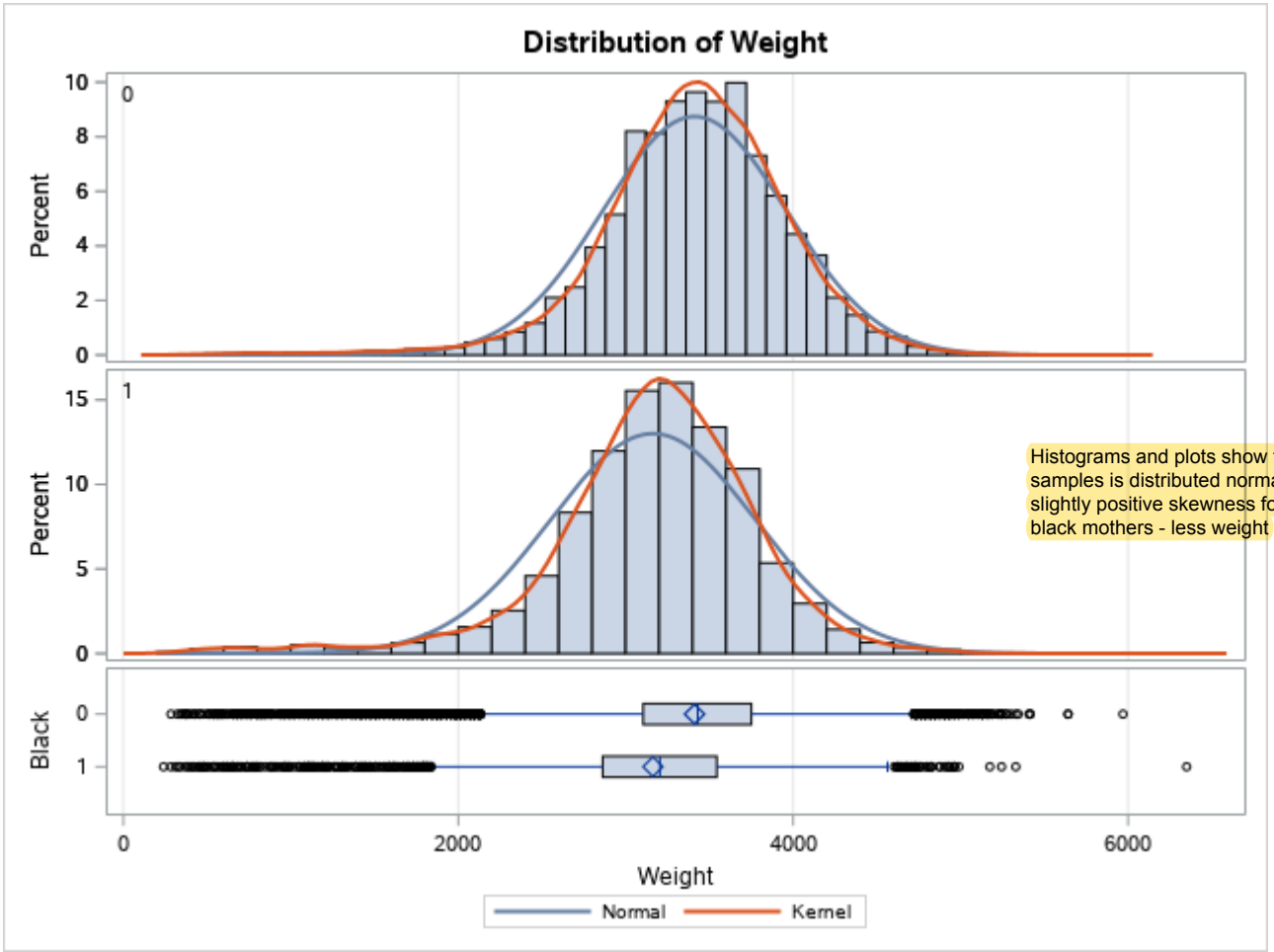
Black	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		3411.2	3406.0	3416.5	547.6	543.9	551.4
1		3162.7	3149.3	3176.0	613.7	604.4	623.3
Diff (1-2)	Pooled	248.6	235.3	261.8	558.9	555.5	562.4
Diff (1-2)	Satterthwaite	248.6	234.2	262.9			

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

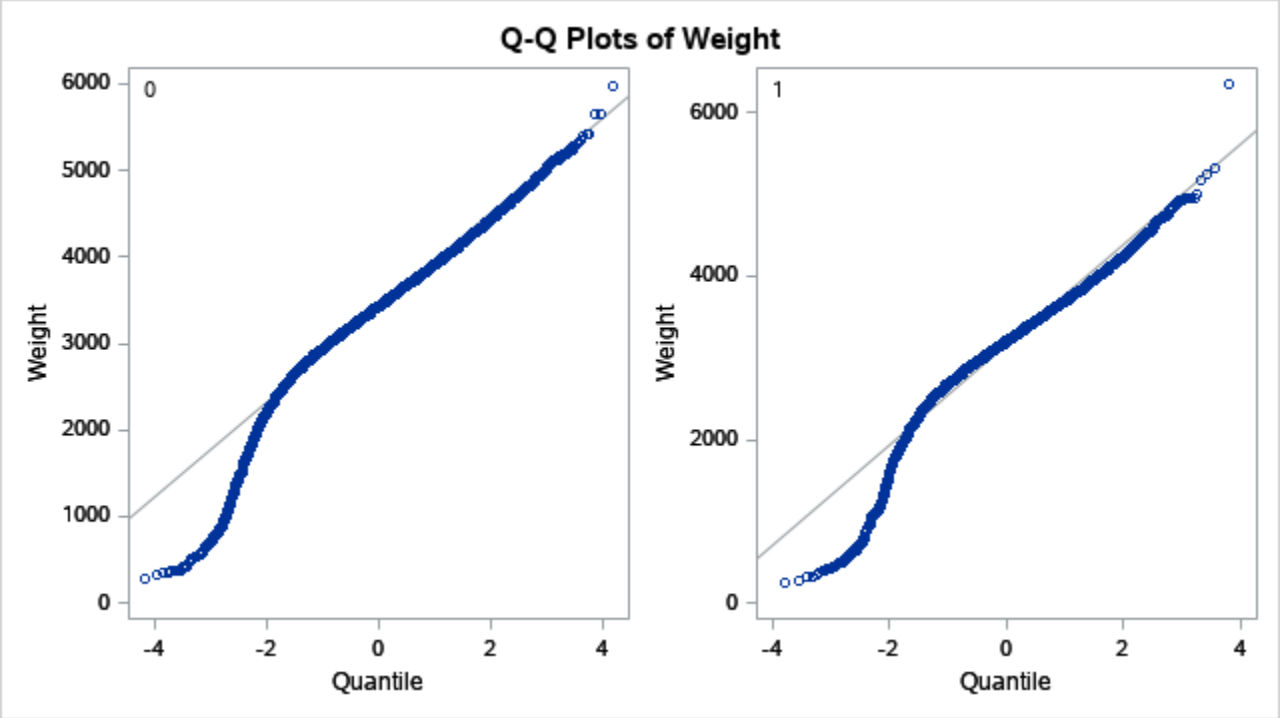
P value assuming equal variances is less than Confidence Interval 5%.

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

P value is less than confidence interval so we reject the null hypothesis that populations come from populations with equal variances.



Histograms and plots show that Weight samples is distributed normally with slightly positive skewness for babies with black mothers - less weight


Q-Q Plots of Weight
0
1
Weight
Quantile
6000
5000
4000
3000
2000
1000
0
-4
-2
0
2
4

```
*Relationship between variables weight and Married.
Assumption: If mother is married, married=1 and if mother is NOT married, married=0.
I'll run an unpaired t-test to find out if the weight mean is the same for babies that are born
 from married mothers and the ones born from non-married mothers.
 H0:u1=u2
 H1:u1<u2;

 ODS graphics on;
  *Title 'Conducting a two-sample t-test and demonstrating Weight is normally distributed';
 Proc ttest data=webwork.birth;
   class Married;
   Var Weight;
 Run;
 ODS graphics off;
```

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		

Mean and variance differences assuming equal and unequal variances in populations

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

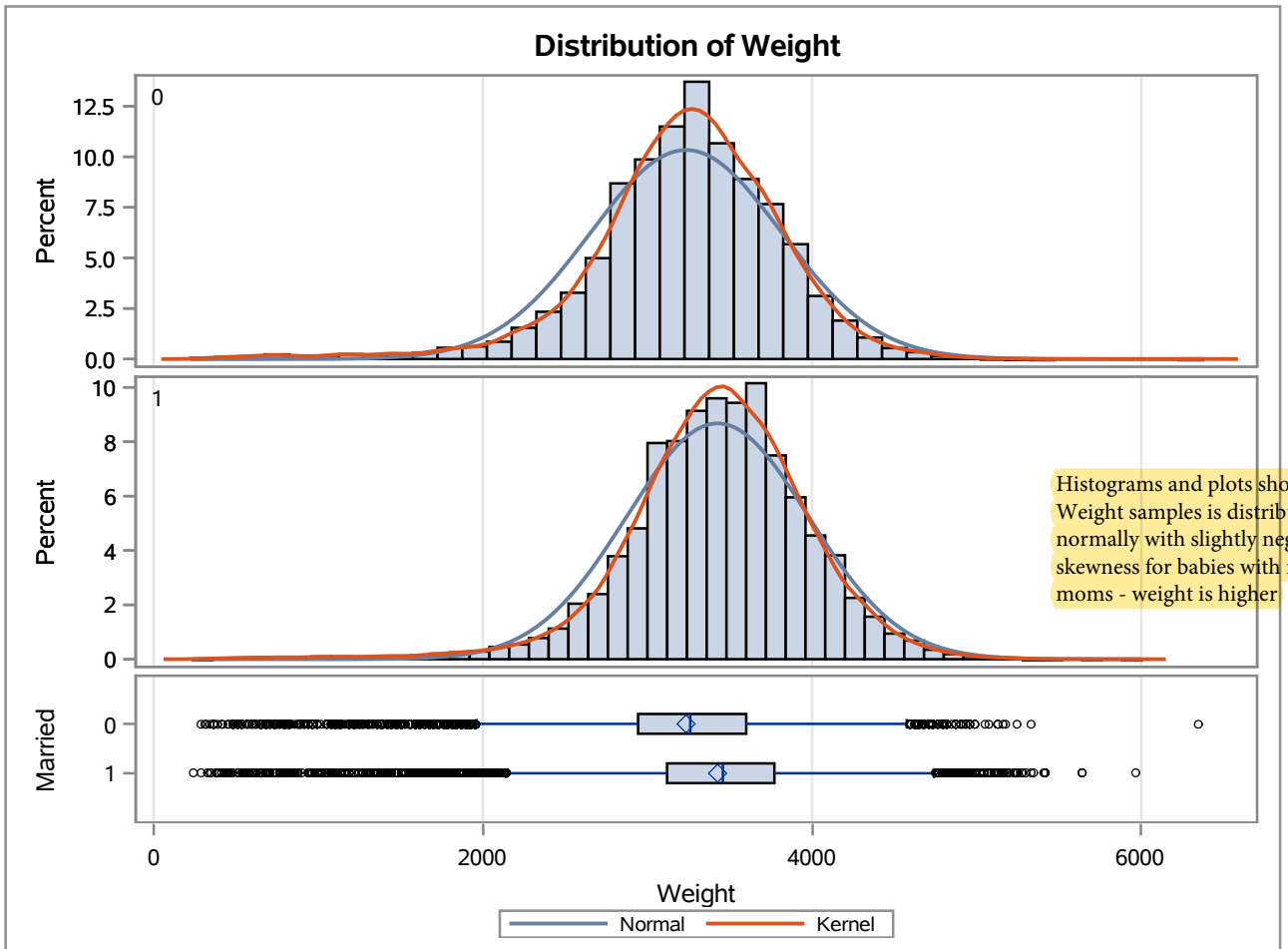
P value assuming equal variances is less than Confidence Interval 5%.

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

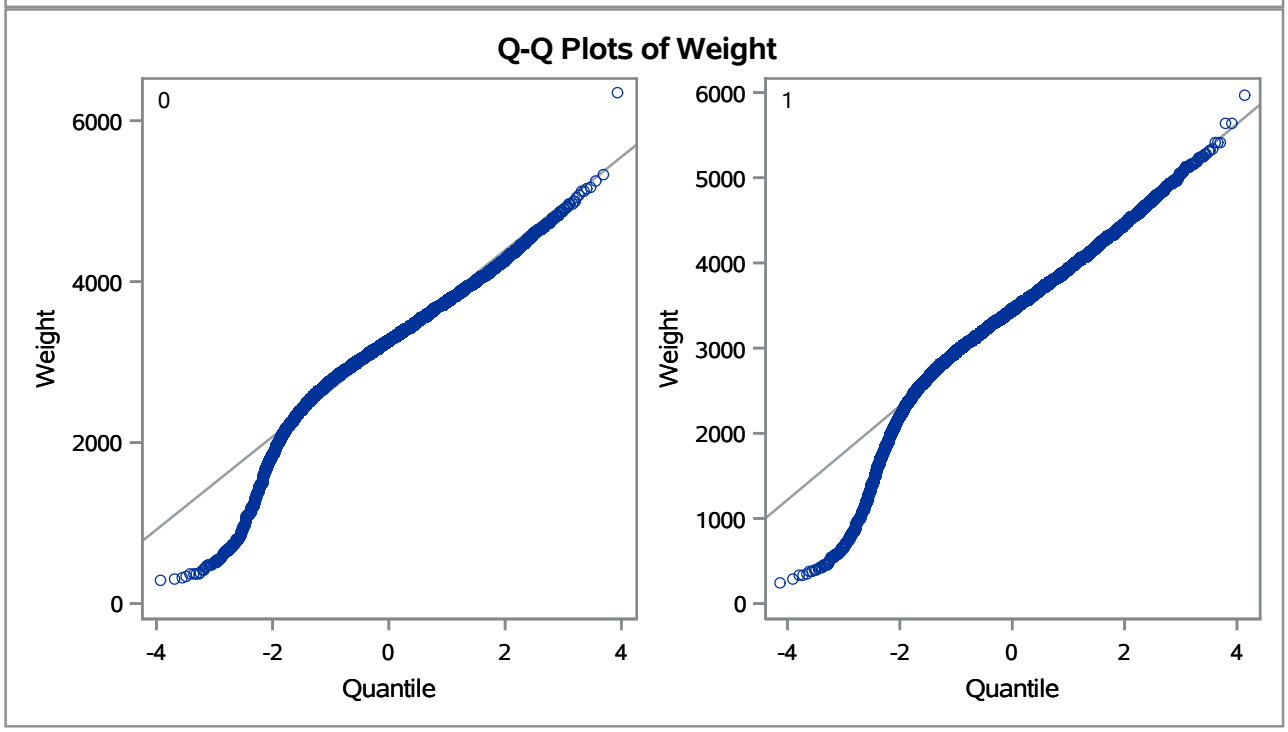
P value is less than confidence interval so we reject the null hypothesis that populations come from populations with equal variances.

The TTEST Procedure

Variable: Weight (Weight)



Histograms and plots show that Weight samples is distributed normally with slightly negative skewness for babies with married moms - weight is higher



```
*Relationship between variables weight and MomSmoke.
Assumption: If mother smokes, smoke=1 and if mother does NOT smoke, MomSmoke=0.
I'll run an unpaired t-test to find out if the weight mean is the same for babies that are born
 from smoker mothers and the ones born from non-smoker mothers.
 H0:u1=u2
 H1:u1<u2;

 ODS graphics on;
  *Title 'Conducting a two-sample t-test and demonstrating Weight is normally distributed';
 Proc ttest data=webwork.birth;
   class MomSmoke;
   Var Weight;
 Run;
 ODS graphics off;
```

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		

Mean and variance differences assuming equal and unequal variances in populations

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			

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

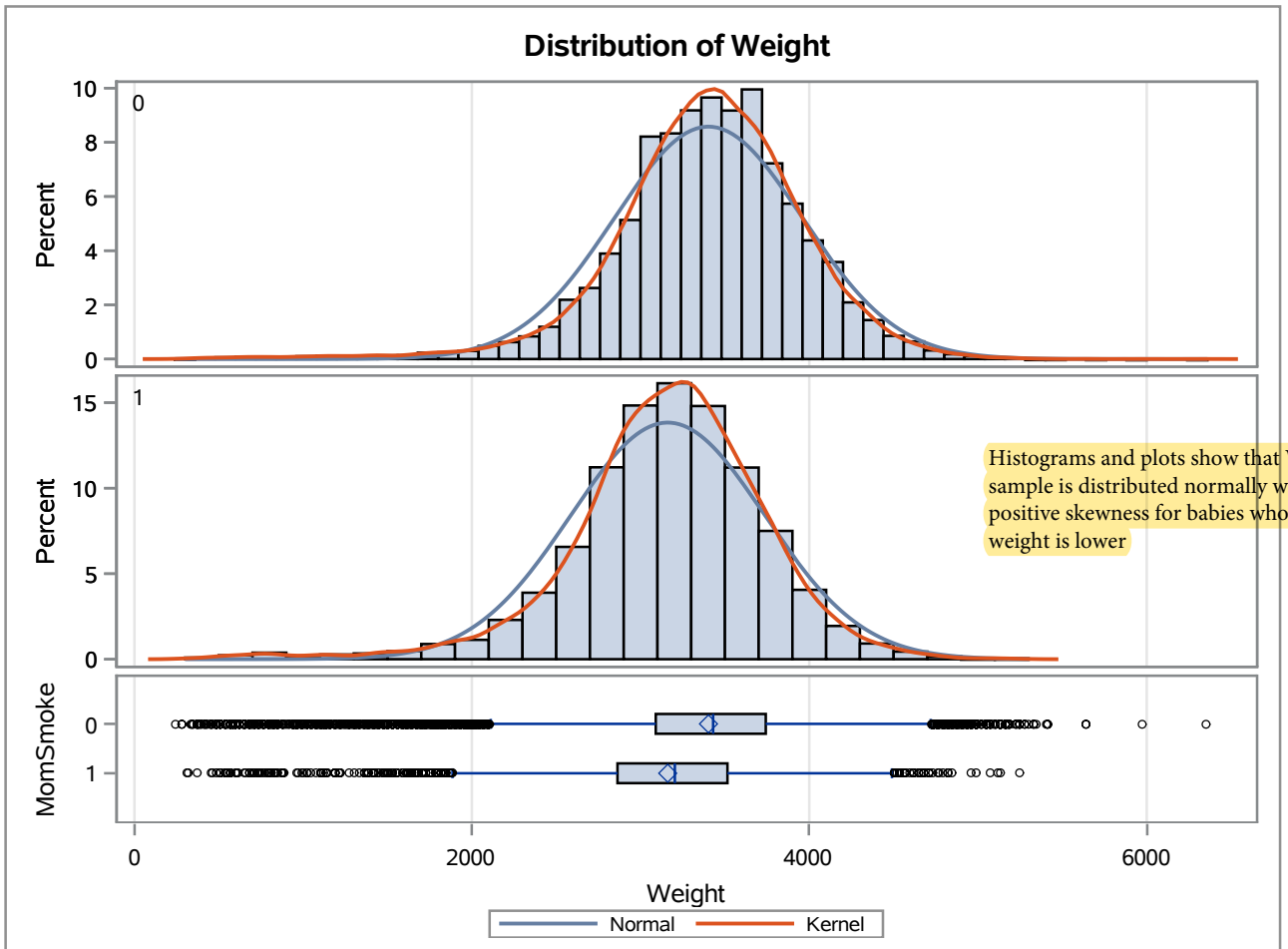
P value assuming equal variances is less than Confidence Interval 5%.

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

P value is less than confidence interval so we reject the null hypothesis that populations come from populations with equal variances.

The TTEST Procedure

Variable: Weight (Weight)



Histograms and plots show that Weight sample is distributed normally with slightly positive skewness for babies who smoke - weight is lower