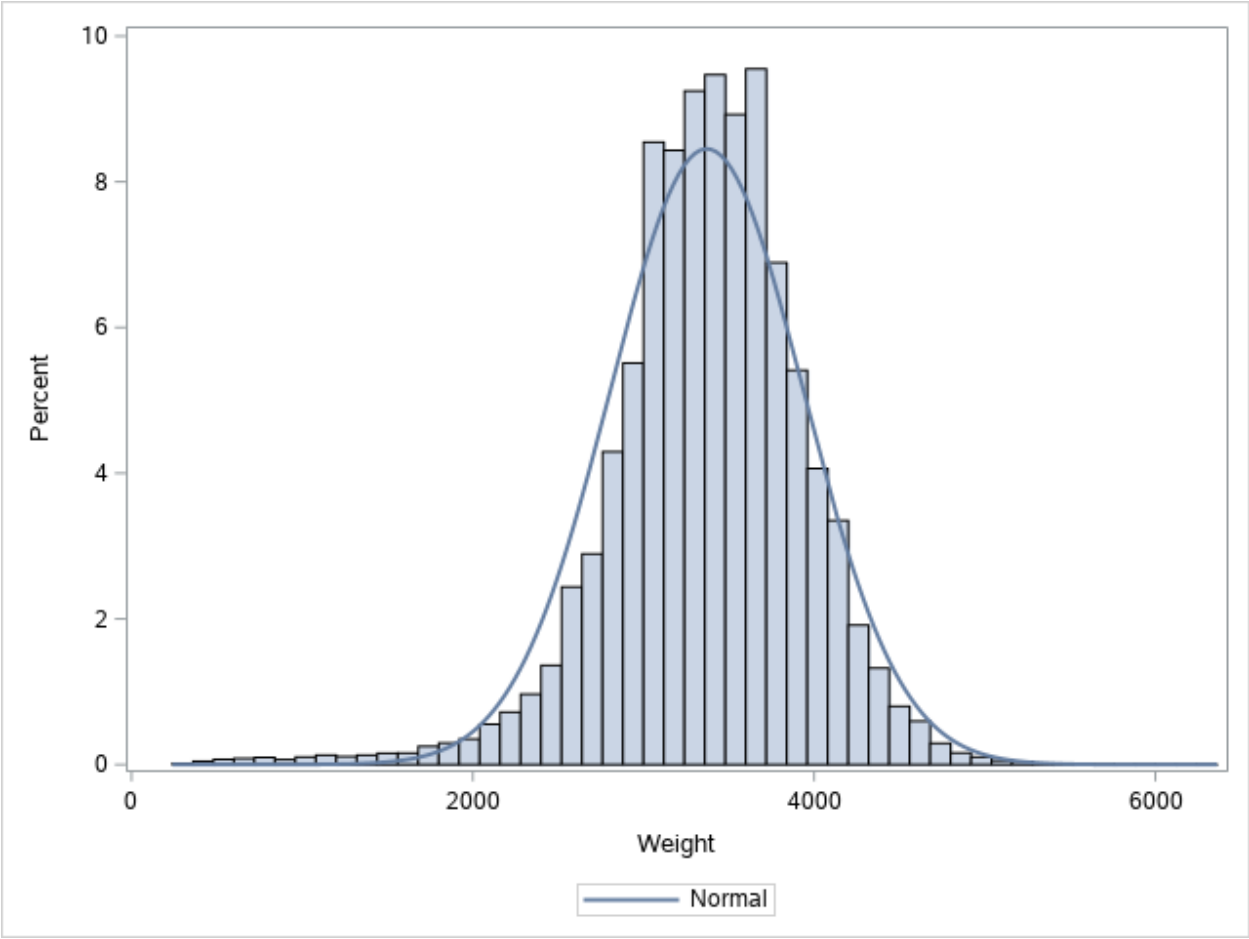
BAN 100 Assignment 2: Hypothesis Testing - Birth Dataset

In this analysis, we will be exploring a dataset called **Birth** that contains data about the weight of babies at birth and other variables that may have an effect on these weights.

The dataset consists of 10 variables (Weight, Black, Married, Boy, MomAge, MomSmoke, CigsPerDay, MomWtGain, Visit, MomEdLevel) as well as 50,000 records (rows).

Variable	Description	Type		Unit
Weight	Weight of infant at birth	Numeric	Discrete	Grams
Black	If the mother is black	Character	Nominal	1=True 0=False
Married	If the mother is married	Character	Nominal	1=True 0=False
Boy	If the baby is boy	Character	Nominal	1=True 0=False
MomAge	Mother's Age - Values Unclear	Numeric	Discrete	
MomSmoke	If the mother smokes	Character	Nominal	1=True 0=False
CigsPerDay	Number of cigarettes per day if she smokes	Numeric	Discrete	
MomWtGain	Weight gained by the mother during pregnancy	Numeric	Discrete	Grams
Visit	Number of parental visits	Numeric	Discrete	
MomEdLevel	Education level of the mother	Character	Ordinal	0=No Education 1=High School 2=Undergraduate 3=Graduate

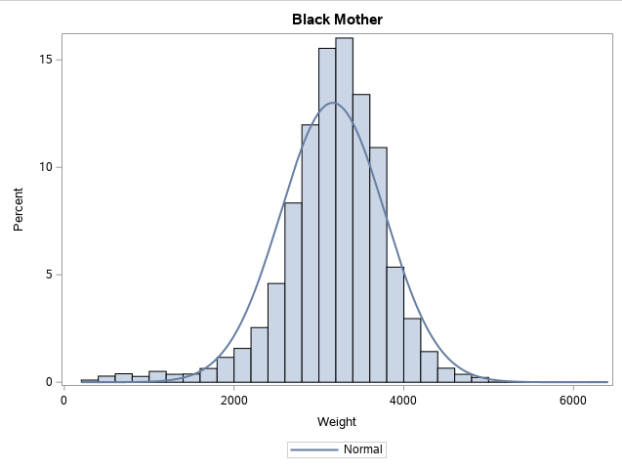
First, let's have a look at the weight variable, the graph below shows that the data is distributed close to a normal distribution. Where the mean for new born babies is 3370.8 grams, Standard deviation is 566.4, the data ranges between 240 grams and 6350 grams. And finally 50% of the data are between 3062 grams and 3720 grams.

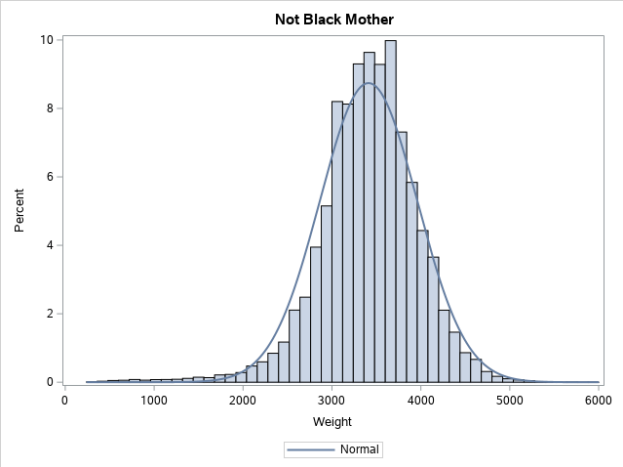


Analysis Variable : Weight Weight								
Mean	Median	Mode	Minimum	Maximum	Range	Lower Quartile	Upper Quartile	Std Dev
3370.8	3402.0	3402.0	240.0	6350.0	6110.0	3062.0	3720.0	566.4

Black Mother Variable

When the data is separated into two samples, one for babies who have black mothers and the other for babies who don't then the data is split into 41,858 and 8142 records respectively. Both of them follow a normal distribution.





Analysis Variable : Weight Weight				
N	Mean	Std Dev	Minimum	Maximum
8142	3162.7	613.7	240.0	6350.0

Analysis Variable : Weight Weight				
N	Mean	Std Dev	Minimum	Maximum
41858	3411.2	547.6	284.0	5970.0

The weight mean for babies with black mothers is 3162.7 grams and the standard deviation is 613.7 while the weight mean for babies with non black mothers is 3411.2 grams and the standard deviation is 547.6. From these numbers, the null hypothesis would be that there is no difference in babies weight whether their mother is black or not.

$H_0 : \mu_{\text{Black Mother}} - \mu_{\text{Non Black Mother}} = 0$

$H_1 : \mu_{\text{Black Mother}} - \mu_{\text{Non Black Mother}} \neq 0$

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

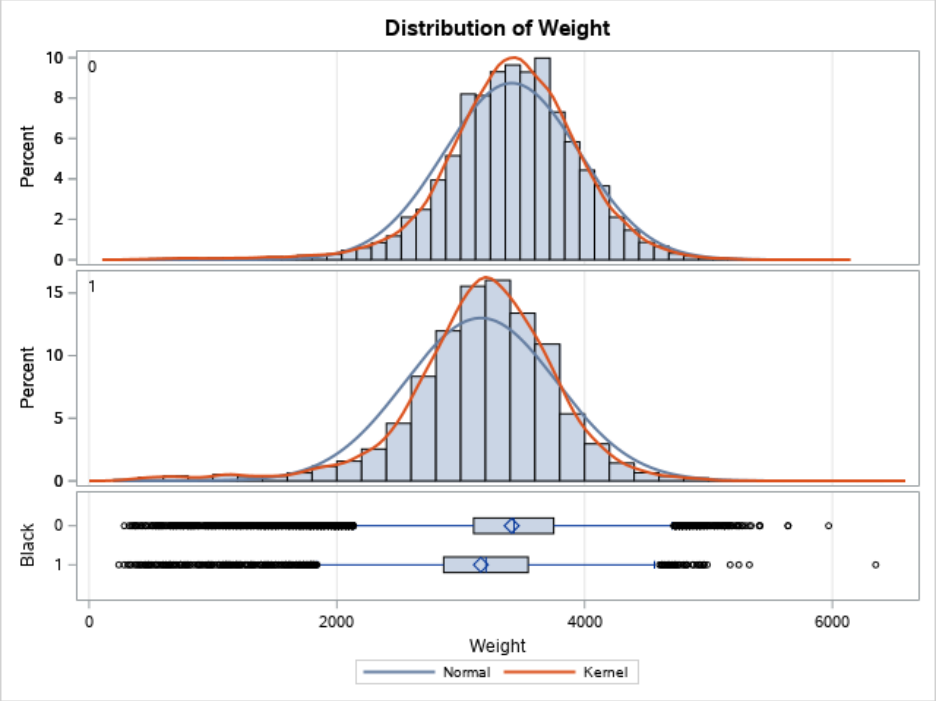
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	49998	36.72	<.0001
Satterthwaite	Unequal	10808	34.01	<.0001

Since the data split into two independent two samples, I have used the SAS t-test procedure with 95% confidence level. and it turns out that both data sets have unequal variances ($p < 0.0001$ which is lower than α).

Therefor, I picked the Satterthwaite method and from it we realize that the t value is 34.01 and the p value is <.0001 which is less than α so we reject the null hypothesis.

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			

The difference between the means of babies with black and non black mothers is 248.6 grams for the dataset. We are 95% confident that babies with black mothers have different weight mean than the babies with non black mothers within the confidence interval of 234.2 grams and 262.9 grams for any samples taken.



Married Variable

For whether the mother was married or not, the data is distributed into two samples, the first one Married Mothers have 35,631 records, a mean of 3425.7 grams and standard deviation of 551.8

Analysis Variable : Weight Weight				
N	Mean	Std Dev	Minimum	Maximum
35631	3425.7	551.8	240.0	5970.0

Unmarried mothers have 14,369 records, a mean of 3234.4 grams and a standard deviation of 579.

Analysis Variable : Weight Weight				
N	Mean	Std Dev	Minimum	Maximum
14369	3234.4	579.0	284.0	6350.0

In order to determine whether the mean weight of married mothers is the same as the mean of weight for unmarried mothers we test the following hypothesis using the same two tail t-test as before.

$H_0 : \mu_{\text{Married Mother}} - \mu_{\text{Unmarried Mother}} = 0$

$H_1 : \mu_{\text{Married Mother}} - \mu_{\text{Unmarried Mother}} \neq 0$

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

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

The Sas proc ttest returns that both samples have unequal variance ($p < 0.0001$ which is lower than α) and therefore we choose the Satterthwaite method.

The t Value for unequal variance is -33.88 and p Value is <.0001 thus we reject the null hypothesis and assume with 95% confidence that being married has a significant effect on the weight of newborn babies.

The difference between the means of babies with Married and unmarried mothers is -191.3 grams for the dataset. We are 95% confident that babies with Married mothers have different weight mean than the babies with unmarried mothers within the confidence interval of -202.1 grams and -180.4 grams for any samples taken.

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			

Boy Variable

Babies born as boys had 25,792 records, a mean of 3,427.3 grams and a standard deviation of 577.7

Analysis Variable : Weight Weight				
N	Mean	Std Dev	Minimum	Maximum
25792	3427.3	577.7	284.0	5970.0

While babies born as girls had 24,208 records, a mean of 3310.6 grams and a standard deviation of 547.7

Analysis Variable : Weight Weight				
N	Mean	Std Dev	Minimum	Maximum
24208	3310.6	547.7	240.0	6350.0

In order to determine whether the mean weight for boys is the same as the mean of weight for girls we test the following hypothesis using the same two tail t-test as before.

$H_0 : \mu_{Boys} - \mu_{Girls} = 0$

$H_1 : \mu_{Boys} - \mu_{Girls} \neq 0$

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

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

The Sas proc ttest returns that both samples have unequal variance and therefore we choose the Satterthwaite method ($p < 0.0001$ which is lower than α).

The t Value for unequal variance is -23.18 and p Value is <.0001 thus we reject the null hypothesis and assume with 95% confidence that being married has a significant effect on the weight of newborn babies.

The difference between the means of babies that are boys and girls is -116.7 grams for the dataset. We are 95% confident that babies that are boys have different weight mean than the babies who are girls within the confidence interval of -126.6 grams and -106.8 grams for any samples taken.

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			

Is the difference in weight mean for boys higher than for girls?

To answer this question we need to make the following hypothesis and test it using a lower sided t-test.

$H_0 : \mu_{Boys} - \mu_{Not\ Boys} = 0$

$H_1 : \mu_{Boys} - \mu_{Not\ Boys} > 0$

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

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

The Sas proc ttest returns that both samples have unequal variance and therefore we choose the Satterthwaite method ($p < 0.0001$ which is lower than α).

The t Value for unequal variance is -23.18 and p Value is <.0001 thus we reject the null hypothesis and assume with 95% confidence that the weight of boys at birth have a higher weight mean than girls height mean at birth.

The difference between the means of babies that are boys and girls is -116.7 grams for the dataset. We are 95% confident that babies who are boys have different weight mean than the babies who are girls within the confidence interval of -Infty grams to -106.8 grams for any samples taken.

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	-Infty	-108.4	563.4	559.9	566.9
Diff (1-2)	Satterthwaite	-116.7	-Infty	-108.4			

MomSmoke Variable

For whether the mother is a smoker or not the data is separated into two sets. 6,533 records for babies with smoking mothers, 3160.9 grams for the mean and a standard deviation of 576.8.

Analysis Variable : Weight Weight				
N	Mean	Std Dev	Minimum	Maximum
6533	3160.9	576.8	312.0	5245.0

43,467 records are for babies with Non Smoking mothers, a mean of 3402.3 grams and a standard deviation of 558.

Analysis Variable : Weight Weight				
N	Mean	Std Dev	Minimum	Maximum
43467	3402.3	558.0	240.0	6350.0

In order to determine whether the mean weight of Smoking mothers is the same as the mean of weight for Non Smoking mothers we test the following hypothesis using the same two tail t-test as before.

$H_0 : \mu_{\text{Smoking Mother}} - \mu_{\text{Non Smoking Mother}} = 0$

$H_1 : \mu_{\text{Smoking Mother}} - \mu_{\text{Non Smoking Mother}} \neq 0$

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

The Sas proc ttest returns that both samples have unequal variance ($p < 0.0004$ which is lower than α) is lower than a and therefore we choose the Satterthwaite method .

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

The t Value for unequal variance is 31.68 and p Value is <.0001 thus we reject the null hypothesis and assume with 95% confidence that being married has a significant effect on the weight of newborn babies.

The difference between the means of babies with Smoking and Non Smoking mothers is 241.5 grams for the dataset. We are 95% confident that babies with Smoking mothers have different weight mean than the babies with Non Smoking mothers within the confidence interval of 226.5 grams and 256.4 grams for any samples taken.

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			