Two Samples *t* -Test in SAS® Enterprise Guide™

In this lesson, you are going to conduct an unpaired t test (a.k.a. *t* test for independent groups) that you will use to compare means between two groups. When there one continuous variable and one categorical variable with only two groups, the null hypothesis is that mean scores for group 1 and group 2 come from populations with equal means. How likely is it to get a difference of the point difference we observe with our sample if the null hypothesis were true? The answer, provided by a two-sample *t* test, is used to determine to whether reject the null to conclude that the group mean differences are statistically significant or fail to reject the null and say that there is not have enough evidence to reject a hypothesis of equality as the chance of getting the observed difference is nothing but by chance, the probability that the difference occurred by chance is higher than the traditional 5%.

There are some assumptions that need to be met before performing the two-sample t-test. One is that the values in each group are normally distributed. The other assumption is that the variances are equal in the two groups. There are tests to check this assumption and SAS Enterprise Guide™ provides output for us to check the findings.

Dataset Description

For this example, you will be revisiting the BIRTH data.

Babies with low birth weights (defined to be less than 2500 grams) are a concern because of their potential medical problems. Health researchers used this data to identify possible contributing factors to low birth weight and recommend strategies to reduce the number of low-birth-weight babies. There are 189 cases with 11 variables including:

- ID: Identification Code
- LOW: Low Birth Weight (0: birth weight>=2500g, 1: birth weight<2500g)
- AGE: Age of the Mother in Years
- LWT: Weight in Pounds at the Last Menstrual Period
- ETH: Ethnicity Category (1, 2, 3)
- SMOKE: Smoking Status During Pregnancy (1: yes, 0: no)
- PTL: History of Premature Labor (0: none, 1: one, and so on)
- HT: History of Hypertension (1: yes, 0: no)
- UI: Presence of Uterine Irritability (1: yes, 0: no)
- FTV: Number of Physician Visits During the First Trimester (0: none, 1: one, and so on)
- BWT: Birth Weight in Grams

I. **Test the Assumption of Normalcy**

For this example, we will be investigating the question “Does a baby’s birth weight (BWT) differ among smoking mothers versus non-smoking mothers (SMOKE)?”.

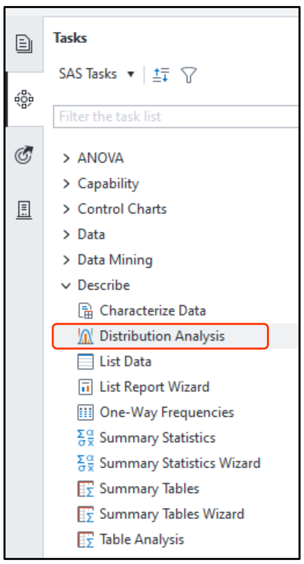
Therefore, the hypothesis statements are:

H_0 = The mean BWT with Non-smoking mom is equal to the mean BWT with smoking mom.

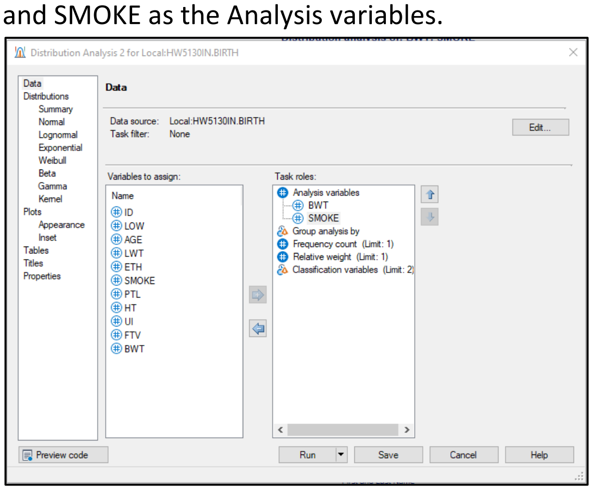
H_A = The mean BWT with Non-smoking mom is not equal to the mean BWT with smoking mom.

In this question, we have one continuous variable, BWT, and one categorical variable that has only two groups (SMOKE). Therefore, we can use the *t*-test to identify if there is a difference between these two groups.

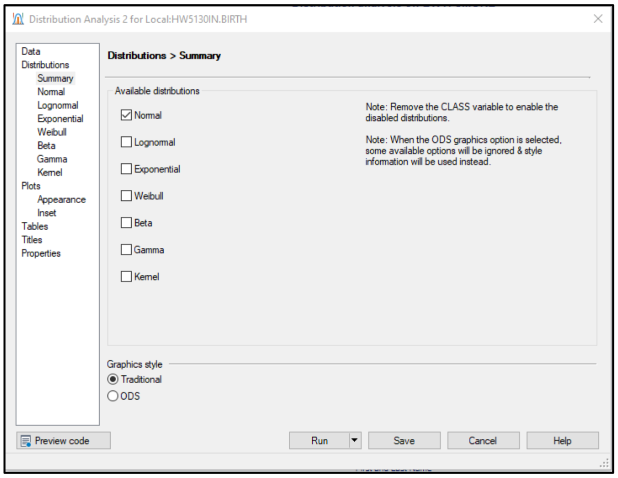
1. First, lets test the normality assumption. Select Tasks, expand Describe, and then select Distribution Analysis.



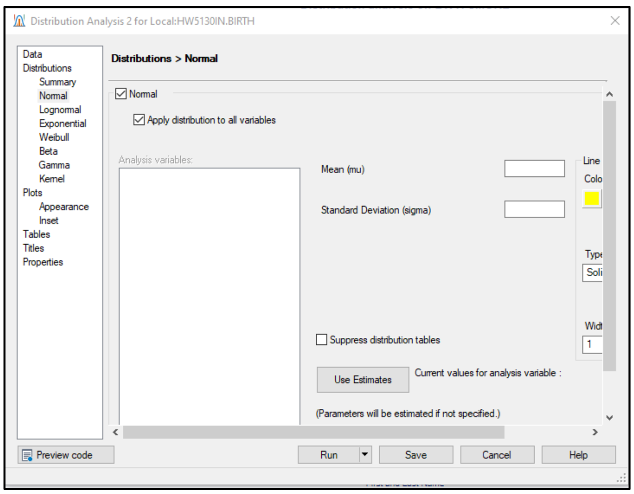
2. Under Data, make sure that the Birth file is selected as the source. Identify the variables BWT and SMOKE as the Analysis variables.



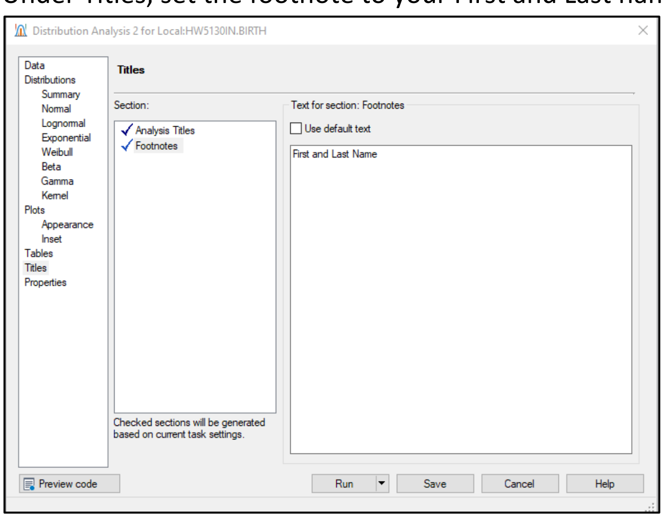
3. Under Distributions: Summary, select Normal.



4. Under Distributions: Normal, select Normal and Apply distribution to all variables.



5. Under Titles, set the footnote to your First and Last name. Click Run.



6. The resulting output is as follows. The output provides several tests for normality, all of which compare the scores in the sample to a normally distributed set of scores with the same mean and standard deviation using the null hypothesis that "the sample distribution is normal". All of these tests resulted in p values greater than the standard level of 0.05. Therefore, we failed to reject the null hypothesis and determine that it is appropriate to treat this distribution as if it were normally distributed:

   <u>The Kolmogrov-Smirnov test</u> is an empirical distribution function in which the theoretical cumulative distribution function of the test distribution is contrasted with the empirical distribution function of the data. This test is highly sensitive to extreme values and tends to have lower power. This test is not recommended when parameters are estimated from the data, regardless of sample size.

   <u>The Cramér-von Mises's test</u> is an empirical distribution function omnibus test for the composite hypothesis of normality. It uses the summed squared differences between observed and expected cumulative proportions as the test statistic.

   <u>The Anderson–Darling test</u> is a statistical test of whether a given sample of data is drawn from a given probability distribution. When applied to testing whether a normal distribution adequately describes a set of data, it is one of the most powerful statistical tools for detecting most departures from normality.

Distribution analysis of: BWT, SMOKE

The UNIVARIATE Procedure
Fitted Normal Distribution for BWT

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	2944.656
Std Dev	Sigma	729.0224

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.04348375	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.05842114	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.41567116	Pr > A-Sq	>0.250

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	1021.00	1248.70
5.0	1790.00	1745.52
10.0	1970.00	2010.38
25.0	2414.00	2452.94
50.0	2977.00	2944.66
75.0	3475.00	3436.37
90.0	3884.00	3878.94
95.0	3997.00	4143.79
99.0	4593.00	4640.62

First and Last Name

Distribution analysis of: BWT, SMOKE

The UNIVARIATE Procedure
Fitted Normal Distribution for SMOKE

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	0.391534
Std Dev	Sigma	0.48939

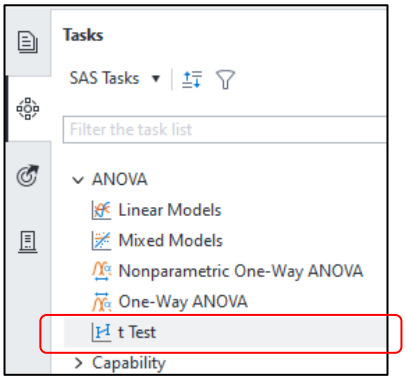
Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.3968235	Pr > D	<0.010
Cramer-von Mises	W-Sq	6.0597543	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	35.9292138	Pr > A-Sq	<0.005

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	0.00000	-0.74696
5.0	0.00000	-0.41344
10.0	0.00000	-0.23564
25.0	0.00000	0.06145
50.0	0.00000	0.39153
75.0	1.00000	0.72162
90.0	1.00000	1.01871
95.0	1.00000	1.19651
99.0	1.00000	1.53003

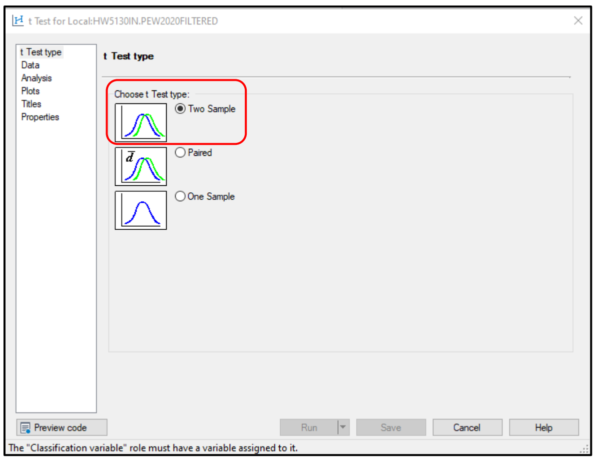
First and Last Name

II. Conduct a Two-Sample *t*-Test.

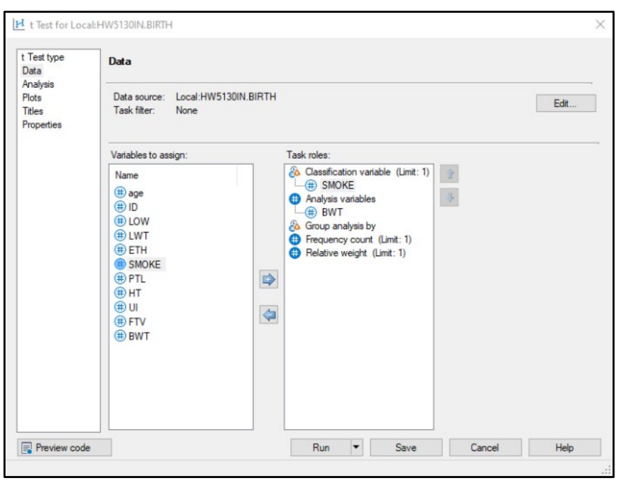
1. Select Tasks, expand ANOVA, and then select *t*-Test.



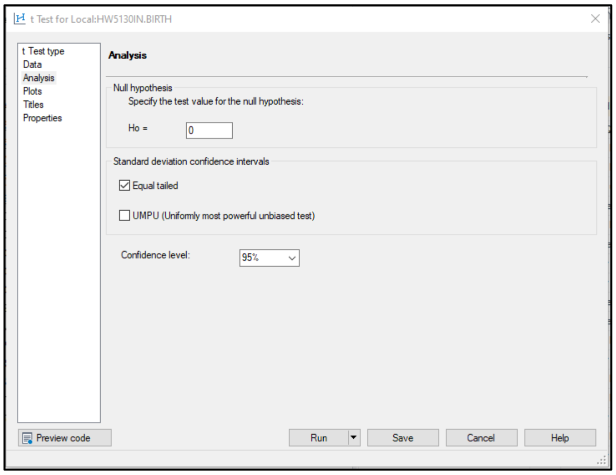
2. In the *t*-Tests window under *t* Test type, select Two Sample.



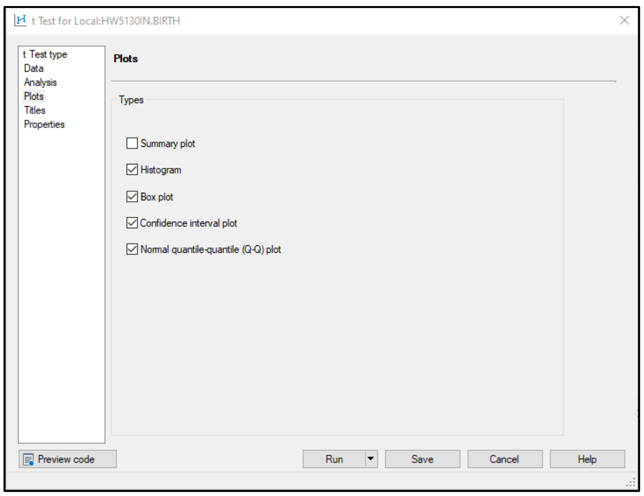
3. Under Data, select the Birth file. Select BWT as the Analysis variable, and Smoke as the Classification variable.



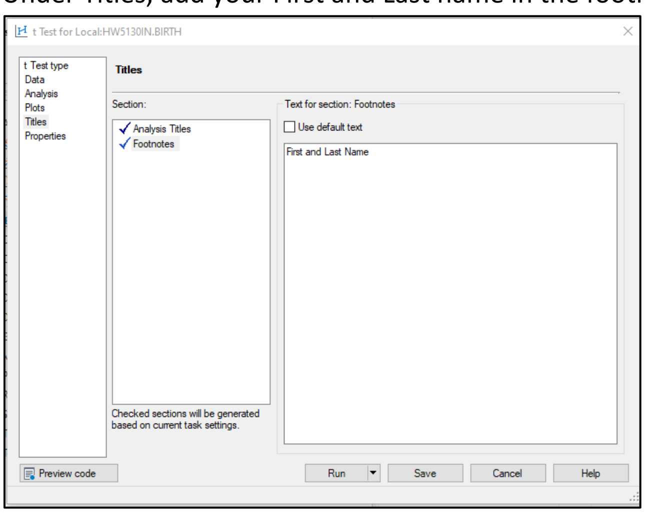
4. Go to Analysis and specify that $H_0 = 0$, meaning that the null hypothesis is that there is not a difference between the groups. Keep this analysis at the 95% Confidence level.



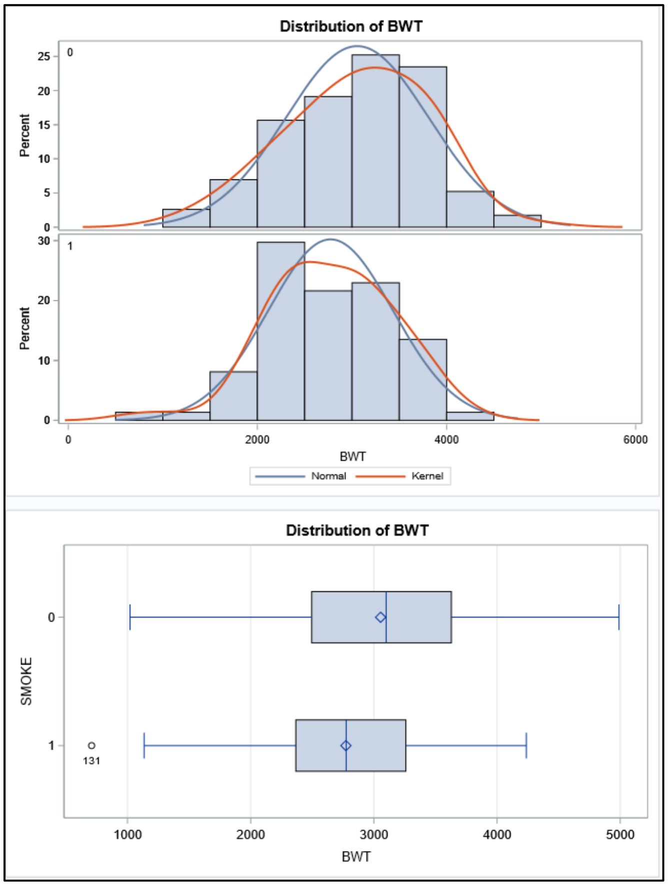
5. Since we are assessing if the assumptions necessary for a *t*-test, select test for normality, under plots, we will select the Histogram, Box plot, Confidence interval plot, and Normality (Q-Q) plot.



6. Under Titles, add your First and Last name in the footnote.



7. Click Run. The determination that this data can be treated as if it were normally distributed after evaluating the histograms. Here, we see that the histograms of the data are relatively close to the normal distribution, shown with the blue curved line.



8. The results from the two-samples *t*-test are shown below.

SMOKE	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		115	3055.0	752.4	70.1625	1021.0	4990.0
1		74	2773.2	660.1	76.7322	709.0	4238.0
Diff (1-2)	Pooled		281.7	717.8	107.0		
Diff (1-2)	Satterthwaite		281.7		104.0		

SMOKE	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		3055.0	2916.0	3193.9	752.4	666.1	864.6
1		2773.2	2620.3	2926.2	660.1	568.2	787.7
Diff (1-2)	Pooled	281.7	70.6927	492.7	717.8	651.8	798.7
Diff (1-2)	Satterthwaite	281.7	76.4668	487.0			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	187	2.63	0.0092
Satterthwaite	Unequal	170	2.71	0.0074

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	114	73	1.30	0.2290

Under the null that the mean group differences is zero, we have an observed 281.7 differences between newborns with non-smoking moms (SMOKE=0) vs with smoking counterparts (SMOKE=1).

How likely to get this 281.7 differences in our sample while the true difference is 0, when the null is correct? We need standard error of our point estimate, sample mean differences. We have two numbers for this: 107 with Pooled method (under the assumption of equal group variances) and 104 with Satterthwaite method under the assumption of unequal group variances.

As for the t-statistics, we have two numbers: 2.63 under Pool method and 2.71 under the Satterthwaite method under the assumption of unequal group variances. Which one to choose? It does not matter in our case as both t-values are high with *p*-values less than indicating that the group differences we observe (281.7) cannot be just by chance if the null was true. Technically, you would need to check the test of equal variance result to decide whether you will use the Pooled or Satterthwaite *t*-values. According to the equal variance test, under the null, variances are equal, and our test has a *p* value of 0.2290 which is higher than 0.05, so it is not significant. It means it is very likely to have the observed variance differences while they are same in the population. Therefore, we will use pooled *t*- test results based on the equality of variance test.

The pooled *t* value is 2.63 and *p* value is 0.0092, which is less than 0.05. Therefore, our test statistics is statistically significant. We reject the null and conclude that:

> “Yes, baby’s birth weight (BWT) differs among smoking mothers versus non-smoking mothers (SMOKE).”