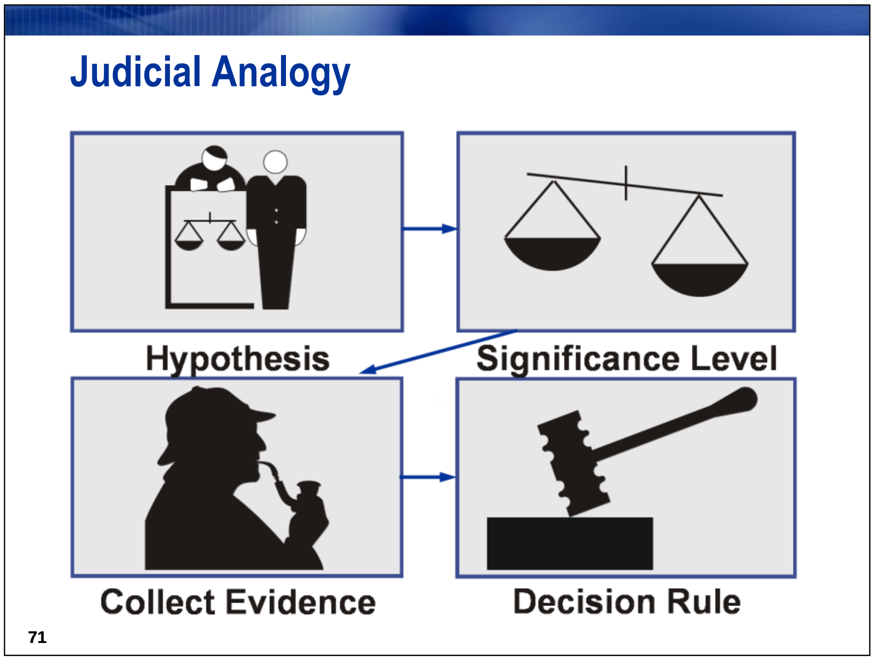
SASEG 5 - Exercise – Hypothesis Testing

(Fall 2015)

<u>Sources</u> (adapted with permission)**-**

T. P. Cronan, Jeff Mullins, Ron Freeze, and David E. Douglas Course and Classroom Notes
Enterprise Systems, Sam M. Walton College of Business, University of Arkansas, Fayetteville
Microsoft Enterprise Consortium
IBM Academic Initiative
SAS® Multivariate Statistics Course Notes & Workshop, 2010
SAS® Advanced Business Analytics Course Notes & Workshop, 2010
Microsoft® Notes
Teradata® University Network

Hypothesis Testing



In a criminal court, you put defendants on trial because you suspect they are guilty of a crime. But how does the trial proceed?

Determine the null and alternative hypotheses. The *alternative* hypothesis is your initial research hypothesis (the defendant is guilty). The *null* is the logical opposite of the alternative hypothesis (the defendant is not guilty). You generally start with the assumption that the null hypothesis is true.

Select a *significance level* as the amount of evidence needed to convict. In a criminal court of law, the evidence must prove guilt “beyond a reasonable doubt”. In a civil court, the plaintiff must prove his or her case by “preponderance of the evidence.” The burden of proof is decided on before the trial.

Collect evidence.

Use a *decision rule* to make a judgment. If the evidence is

- sufficiently strong, reject the null hypothesis.
- not strong enough, fail to reject the null hypothesis. Note that failing to prove guilt does not prove that the defendant is innocent.

Statistical hypothesis testing follows this same basic path.

Types of Errors

You used a decision rule to make a decision, but was the decision correct?

	"TRUTH"	
YOUR DECISION	H_0 Is True	H_0 Is False
Fail to Reject Null	Correct	Type II Error
Reject Null	Type I Error	Correct

- Probability of a Type I error = α
- Probability of a Type II error = β
- Probability of Correct Rejection = $(1 - \beta)$ = Power

77

Recall that you start by assuming that the coin is fair.

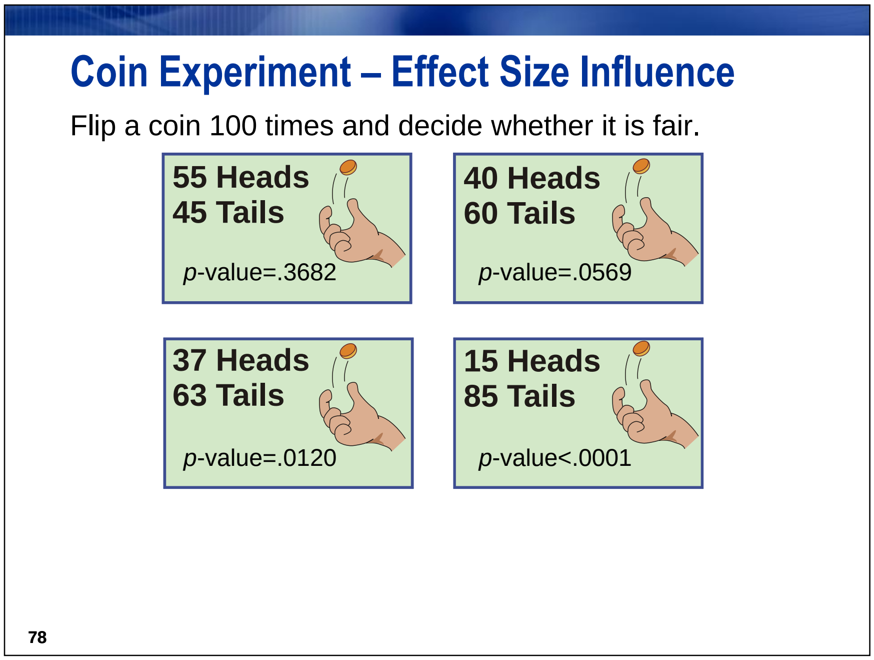
The probability of a Type I error, often denoted α, is the probability that you reject the null hypothesis when it is true. It is also called the ***significance level*** of a test. In the

- legal example, it is the probability that you conclude the person is guilty when he or she is innocent
- coin example, it is the probability that you conclude the coin is not fair when it is fair.

The probability of a Type II error, often denoted β, is the probability that you fail to reject the null hypothesis when it is false. In the

- legal example, it is the probability that you fail to find the person guilty when he or she is guilty
- coin example, it is the probability that you fail to find the coin is not fair when it is not fair.

The power of a statistical test is equal to $1–\beta$, where β is the Type II error rate. This is the probability that you correctly reject the null hypothesis.



The *effect size* refers to the magnitude of the difference in sampled population from the null hypothesis. In this example, the null hypothesis of a fair coin would suggest 50% heads and 50% tails. If the true coin flipped were actually weighted to give 55% heads, the effect size is 5%.

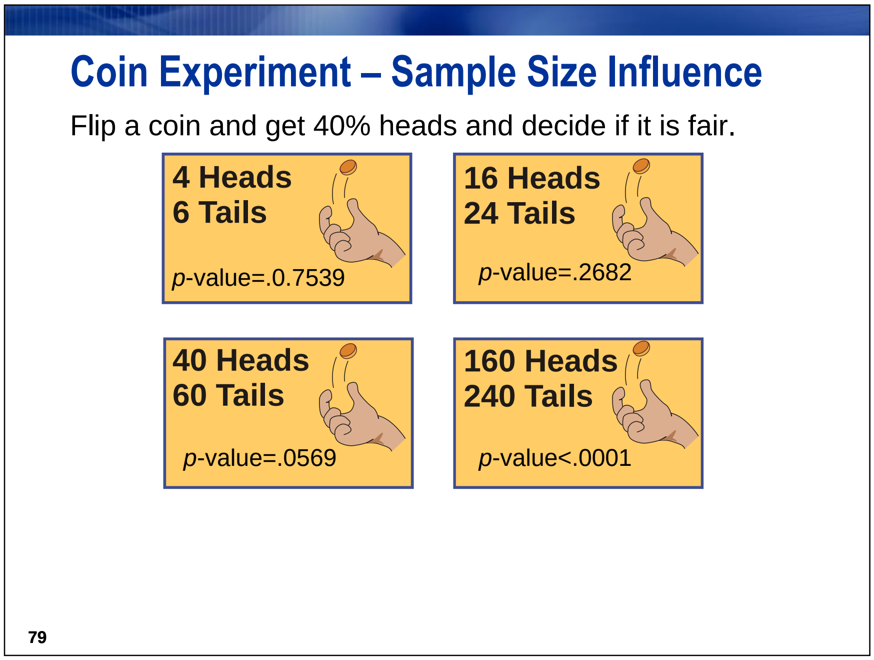
If you flip a coin 100 times and count the number of heads, you do not doubt that the coin is fair if you observe exactly 50 heads. However, you might be

- somewhat skeptical that the coin is fair if you observe 40 or 60 heads
- even more skeptical that the coin is fair if you observe 37 or 63 heads
- highly skeptical that the coin is fair if you observe 15 or 85 heads.

In this situation, the greater the difference between the number of heads and tails, the more evidence you have that the coin is not fair.

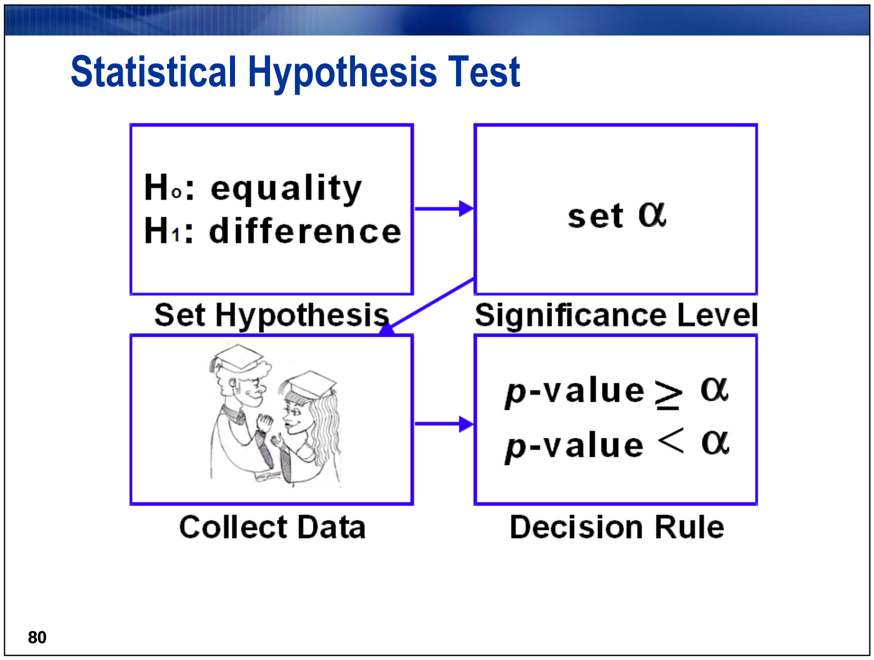
A *p-value* measures the probability of observing a value as extreme or more extreme than the one observed, simply by chance, given that the null hypothesis is true. For example, if your null hypothesis is that the coin is fair and you observe 40 heads (60 tails), the *p*-value is the probability of observing a difference in the number of heads and tails of 20 or more from a fair coin tossed 100 times.

A large *p*-value means that you would often see a test statistic value this large in experiments with a fair coin. A small *p*-value means that you would rarely see differences this large from a fair coin. In the latter situation, you have evidence that the coin is not fair, because if the null hypothesis were true, a random sample from it would not likely have the observed statistic values.



A *p*-value is not only affected by the effect size. It is also affected by the sample size (number of coin flips, *k*).

For a fair coin, you would expect 50% of *k* flips to turn up heads. In this example, in each case, the observed proportion of heads from *k* flips was 0.4. This value is different from the 0.5 you would expect under H_0. The evidence is stronger, the greater the number of trials (*k*) on which the proportion is based. As you saw in the section on confidence intervals, the variability around a mean estimate is smaller, the larger the sample size. For larger sample sizes, you can measure means more precisely. Therefore, 40% heads out of 400 flips would make you more sure that this was not just a chance difference from 50% than would 40% out of 10 flips. The smaller *p*-values reflect this confidence. The *p*-value here is assessing the probability that this difference from 50% occurred purely by chance.



In statistics,

1. the null hypothesis, denoted H_0, is your initial assumption and is usually one of equality or no relationship. For the test score example, H_0 is that the mean sum Math and Verbal SAT score is 1200. The alternative hypothesis, H_1, is the logical opposite of the null, namely here that the sum Math and Verbal SAT score is **not** 1200.

2. the significance level is usually denoted by α, the Type I error rate.

3. the strength of the evidence is measured by a *p*-value.

4. the decision rule is
   - fail to reject the null hypothesis if the *p*-value is greater than or equal to α
   - reject the null hypothesis if the *p*-value is less than α.

You never conclude that two things are the same or have no relationship; you can only fail to show a difference or a relationship.

Comparing α and the *p*-Value

In general, you

- reject the null hypothesis if *p*-value < α
- fail to reject the null hypothesis if *p*-value ≥ α.

81

It is important to clarify that

- α, the probability of Type I error, is specified by the experimenter before collecting data
- the *p*-value is calculated from the collected data.

In most statistical hypothesis tests, you compare α and the associated *p*-value to make a decision.

Remember, α is set ahead of time based on the circumstances of the experiment. The level of α is chosen based on the cost of making a Type I error. It is also a function of your knowledge of the data and theoretical considerations.

For the test score example, α was set to 0.05, based on the consequences of making a Type I error (the error of concluding that the mean SAT sum score is not 1200 when it really is 1200). If making a Type I error is especially egregious, you might consider choosing a lower significance level when planning your analysis.

Performing a Hypothesis Test

To test the null hypothesis H_0: $\mu = \mu_0$, SAS software calculates the *t* statistic:

$$t = \frac{(\bar{x} - \mu_0)}{s_{\bar{x}}}$$

For the test score example:

$$t = \frac{(1190.625 - 1200)}{16.4416} = -0.5702$$

p-value = 0.5702

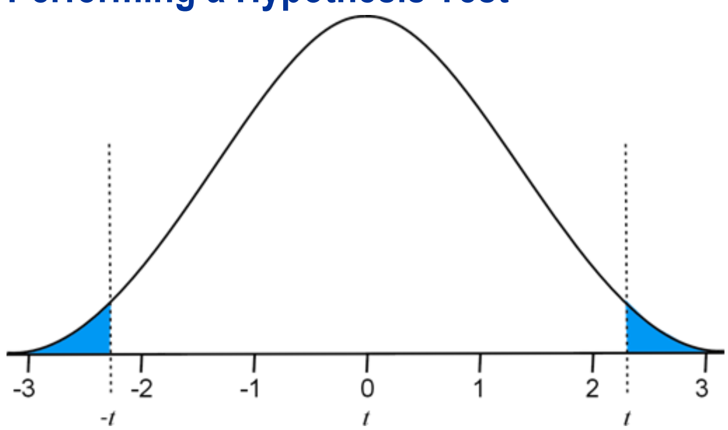
Therefore, the null hypothesis is not rejected.

85

For the test score example, μ_0 is the hypothesized value of 1200, $\bar{x}$ is the sample mean SAT score of students selected from the school district, and $s_{\bar{x}}$ is the standard error of the mean.

- This statistic measures how far $\bar{x}$ is from the hypothesized mean.
- To reject a test with this statistic, the *t* statistic should be much higher or lower than 0 and have a small corresponding *p*-value.
- The results of this test are valid if the distribution of sample means is normally distributed.

Performing a Hypothesis Test



The *t* statistic can be positive or negative.

86

For a two-sided test of a hypothesis, the rejection region is contained in both tails of the *t* distribution. If the *t* statistic falls in the rejection region (in the shaded region in the graph above), then you reject the null hypothesis. Otherwise, you fail to reject the null hypothesis.

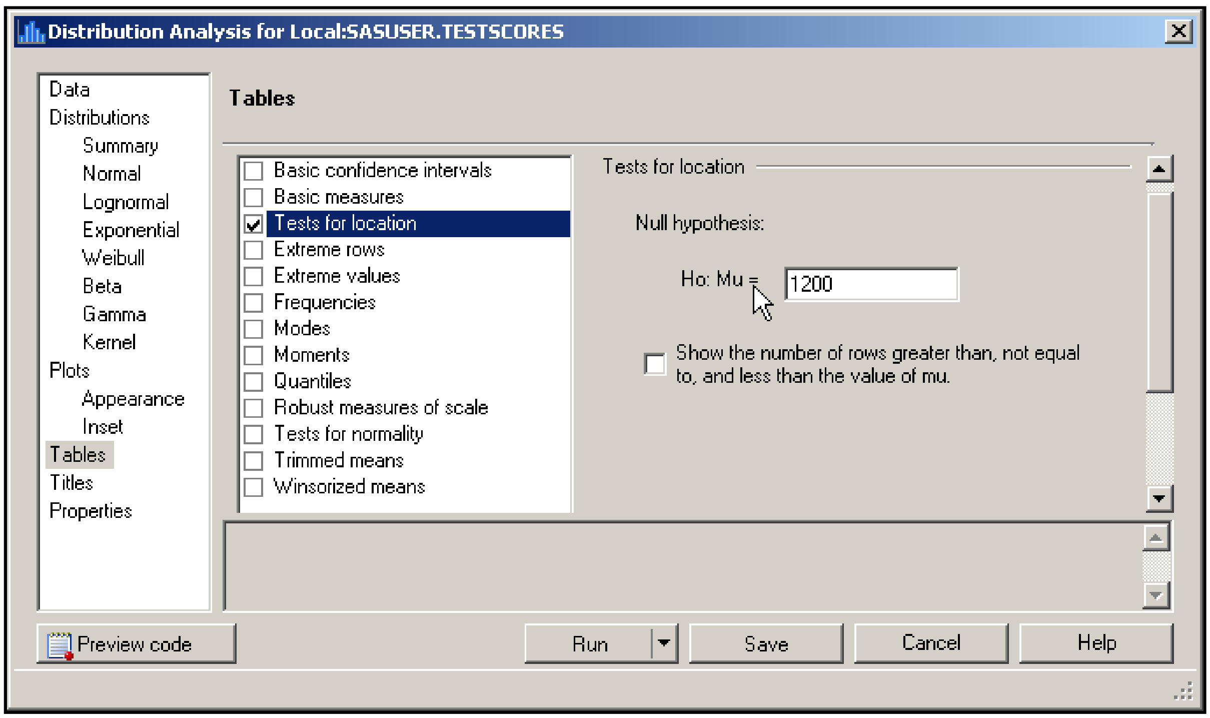
The area in each of the tails corresponds to $\alpha/2$ or 2.5%. The sum of the areas under the tails is 5%, which is alpha.

The alpha and *t*-distribution mentioned here are the same as those in the section on confidence intervals. In fact, there is a direct relationship. The rejection region based on α begins at the point where the $(1.00-\alpha)$ confidence interval will no longer include the true value of μ_0.

Exercise - Hypothesis Testing

With the **TESTSCORES** SAS dataset, use the Distribution Analysis task to test the hypothesis that the mean of SAT Math+Verbal score is equal to 1200.

1. Open the **TESTSCORES** dataset.
2. Use **Describe** > **Distribution Analysis**.
3. Use the **SATscore** variable as the analysis variable.
4. Click **Tables** and uncheck all checked boxes.
5. Check the box for **Tests for location** and then type the value `1200` in the field next to `Ho: Mu=`.



6. Run this task, but do not replace the previous results.

Tests for Location: Mu0=1200				
Test	Statistic		p Value	
Student's t	t	-0.5702	Pr > |t|	0.5702
Sign	M	-5	Pr >= |M|	0.3019
Signed Rank	S	-207	Pr >= |S|	0.2866

The *t* statistic and *p*-value are labeled Student's t and Pr > |t|, respectively.

- The *t* statistic value is -0.5702 and the *p*-value is .5702.
- *Therefore, you cannot reject the null hypothesis at the 0.05 level. Thus, even though the mean of the student scores in this sample (1190.625) is slightly lower than the magnet school goal of 1200, there is not enough evidence to reject the hypothesis that the population mean of all magnet school students in the district is1200.*

7. Save the project as **`SASEG5A`**.

<u>Note:</u>

SAS EG performs a *two tailed* test of hypothesis to test the hypothesis that H_o: $\mu = \mu_0$. To perform a one tailed hypothesis, a small calculation is needed as follows:

H_o: $\mu <= \mu_0$

H_a: $\mu > \mu_0$

----

if t > 0, p–value is p/2

if t < 0, p–value is (1.0 - p/2)

H_o: $\mu => \mu_0$

H_a: $\mu < \mu_0$

----

if t > 0, p–value is (1.0 – p/2)

if t < 0, p–value is p/2

Exercises – One Sample t-Test

1. Performing a One-Sample *t*-Test

- The data set **`NormTemp`** comes from a paper in the *Journal of Statistics Education (*Shoemaker 1996). The data was simulated based on distributions shown in an article in the *Journal of the American Medical Association* that examined whether true mean body temperature is 98.6 degrees Fahrenheit. The data is used with permission from Dr. Allen L. Shoemaker of Calvin College.

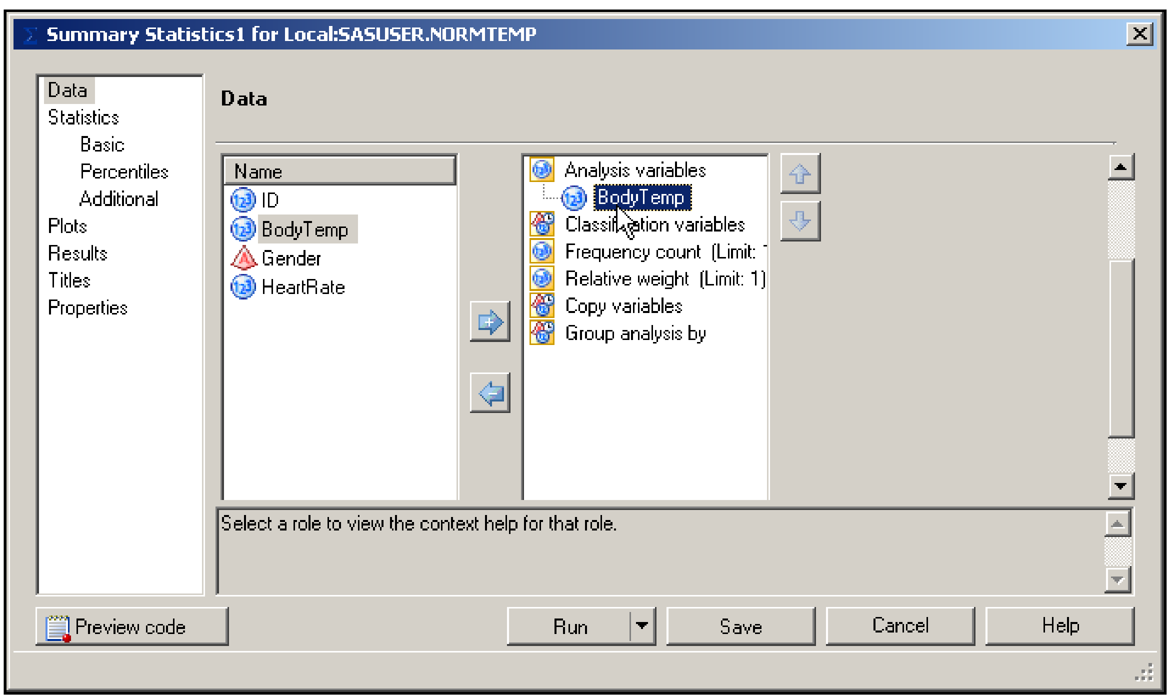
  Perform a one-sample *t*-test to determine whether the mean of body temperatures (the variable `BodyTemp` in `NormTemp`) is truly the value 98.6 that everyone assumes it to be.

Using the **ISYS 5503 Shared Datasets** folder, open **`NORMTEMP`** SAS dataset by double-clicking it or by highlighting it and selecting Open.

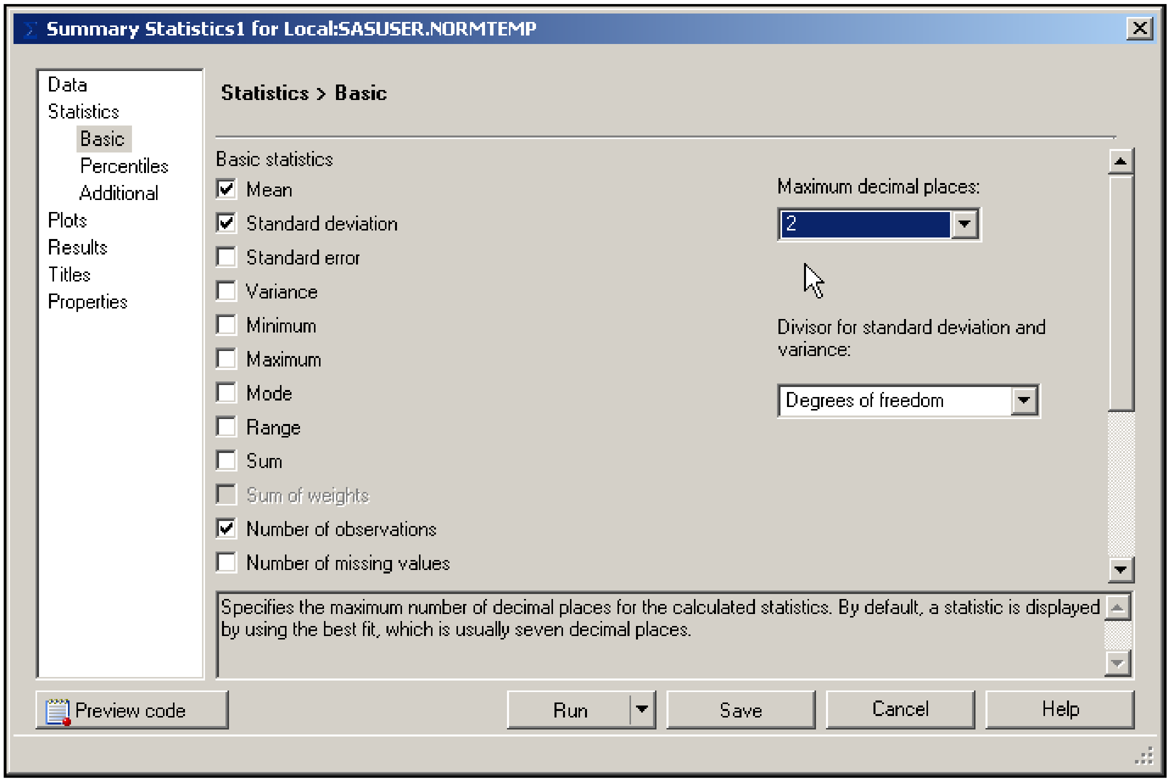
1. Calculating Basic Statistics Using the Summary Statistics Task

- With the **`NORMTEMP`** data table open, click **<u>Describe</u>** ⇨ **<u>Summary Statistics…</u>**.

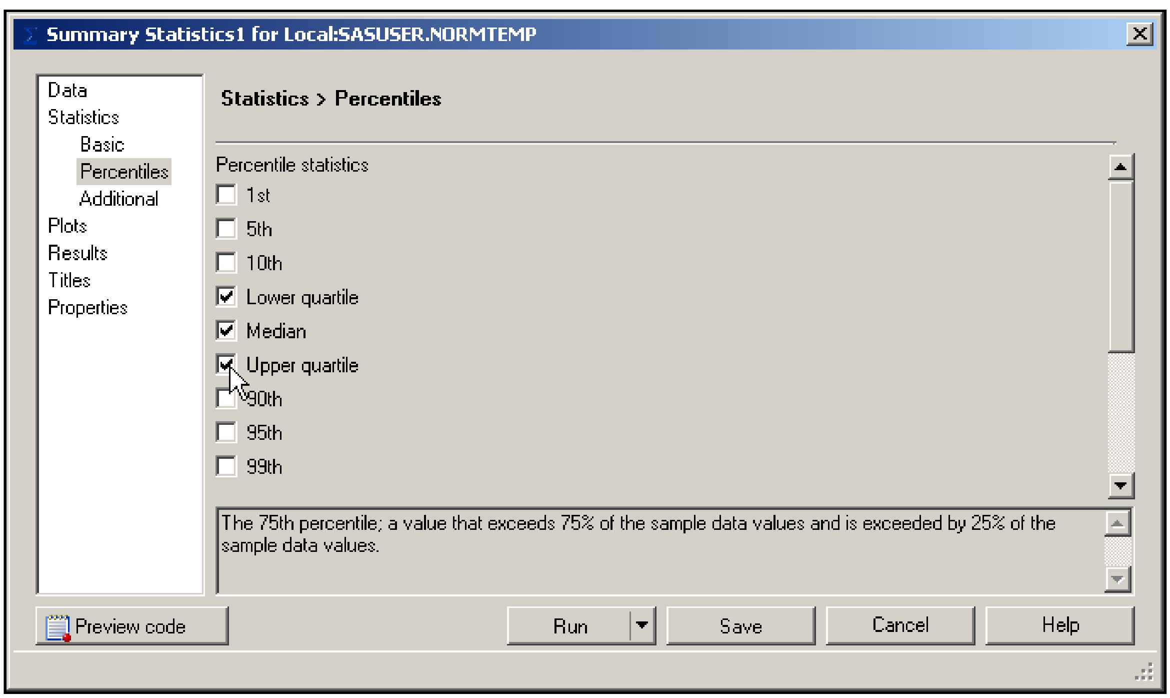
- Add **`BodyTemp`** to the analysis variables task role.



- Click **Basic** under Statistics and check and uncheck boxes until the only ones left checked are for the number of observations, sample mean, and standard deviation. For `Maximum decimal places`, select **2** from the drop-down menu.



- Click **Percentiles** under **Statistics** and check the boxes for the lower and upper quartiles, as well as the median.



- Run the task.

Analysis Variable : BodyTemp					
Mean	Std Dev	N	Lower Quartile	Median	Upper Quartile
98.25	0.73	130	97.80	98.30	98.70

a. What is the overall mean and standard deviation of body temperature in the sample?

The overall mean is 98.25 and the standard deviation is 0.73.

b. What is the interquartile range of body temperature?

The interquartile range is 0.90 (98.70 – 97.80).

2. Producing Confidence Intervals

Generate the 95% confidence interval for the mean of **`BodyTemp`** in the **`NormTemp`** data set.

- Reopen the Summary Statistics task by right-clicking the task icon in the process flow and clicking **Modify Summary Statistics**.

- Click **Additional** under **Statistics** at the left and then check the box for **Confidence limits of the mean**.

- Select **Yes** to replace the previous output.

Analysis Variable : BodyTemp							
Mean	Std Dev	N	Lower Quartile	Median	Upper Quartile	Lower 95% CL for Mean	Upper 95% CL for Mean
98.25	0.73	130	97.80	98.30	98.70	98.12	98.38

a. What is the confidence interval?

The 95% confidence interval is 98.12 to 98.38 degrees Fahrenheit.

b. How do you interpret this interval with regards to the true population mean for body temperature?

You are 95% confident that the true mean body temperature for the population of all people in the world is somewhere between 98.12 and 98.38 degrees.

3. Performing a One-Sample *t*-Test

a. Perform a one-sample *t*-test to determine whether the mean of body temperatures (the variable **`BodyTemp`** in **`NormTemp`**) is truly the value 98.6 that everyone assumes it to be.

- Use **Describe** > **Distribution Analysis** and use BodyTemp as the analysis variable

- Click **Tables** and deselect all currently selected tables. Check the box for **Tests for location** and then type the number **`98.6`** in the box next to `Ho: Mu0=`.

- Click **Run** and do not replace the results from the previous run.

Tests for Location: Mu0=98.6				
Test	Statistic		p Value	
Student's t	t	-5.45482	Pr > |t|	<.0001
Sign	M	-21	Pr >= |M|	0.0002
Signed Rank	S	-1963	Pr >= |S|	<.0001

1) What is the value of the *t* statistic and the corresponding *p*-value?

   They are -5.45482 and <.0001, respectively.

2) Do you reject or fail to reject the null hypothesis at the .05 level that the average temperature is 98.6 degrees?

   Because the p-value is less than the stated alpha level of .05, you do reject the null hypothesis.

3) Above, we tested the null hypothesis that H_o: Mu_0= 98.6.

   What if we tested whether the average temperature is greater than or equal to 98.6 degrees?

   That is, H_o: Mu_0= > 98.6 (a one tailed test)

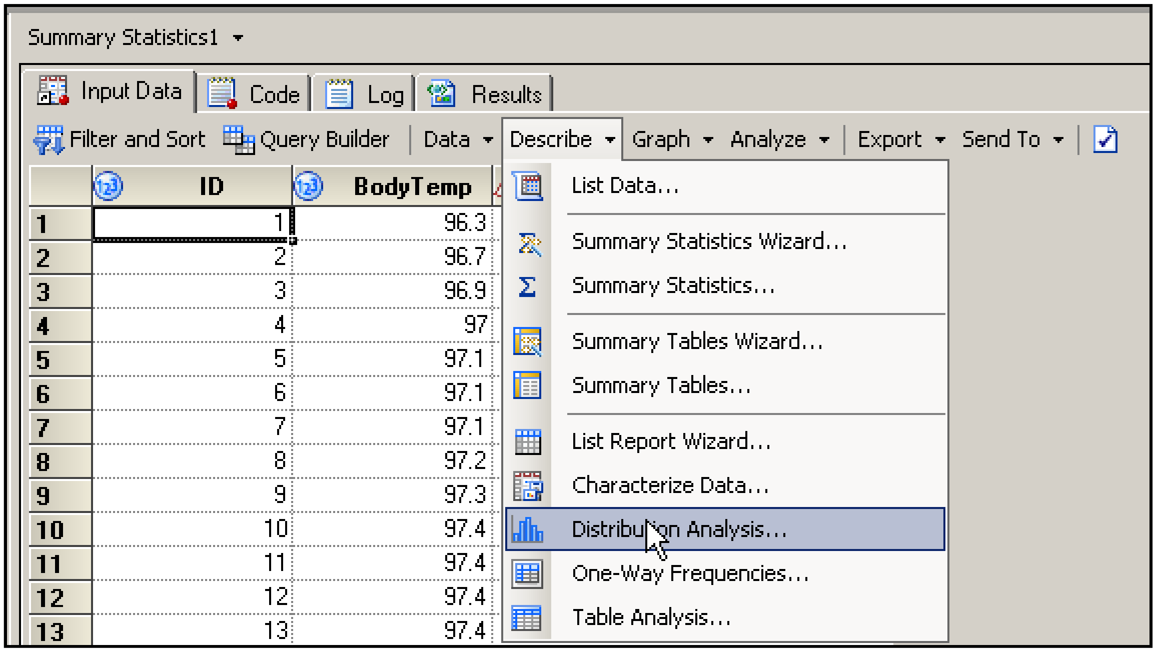
   H_a: Mu_0 < 98.6

Using the previous note on page 11, t < 0, therefore, the p–value is p/2 (.0001/2). In this case, <u>we reject the null hypothesis</u> at the .05 level that the average temperature is greater than or equal to 98.6 degrees *because the p-value is less than the stated alpha level of .05.*

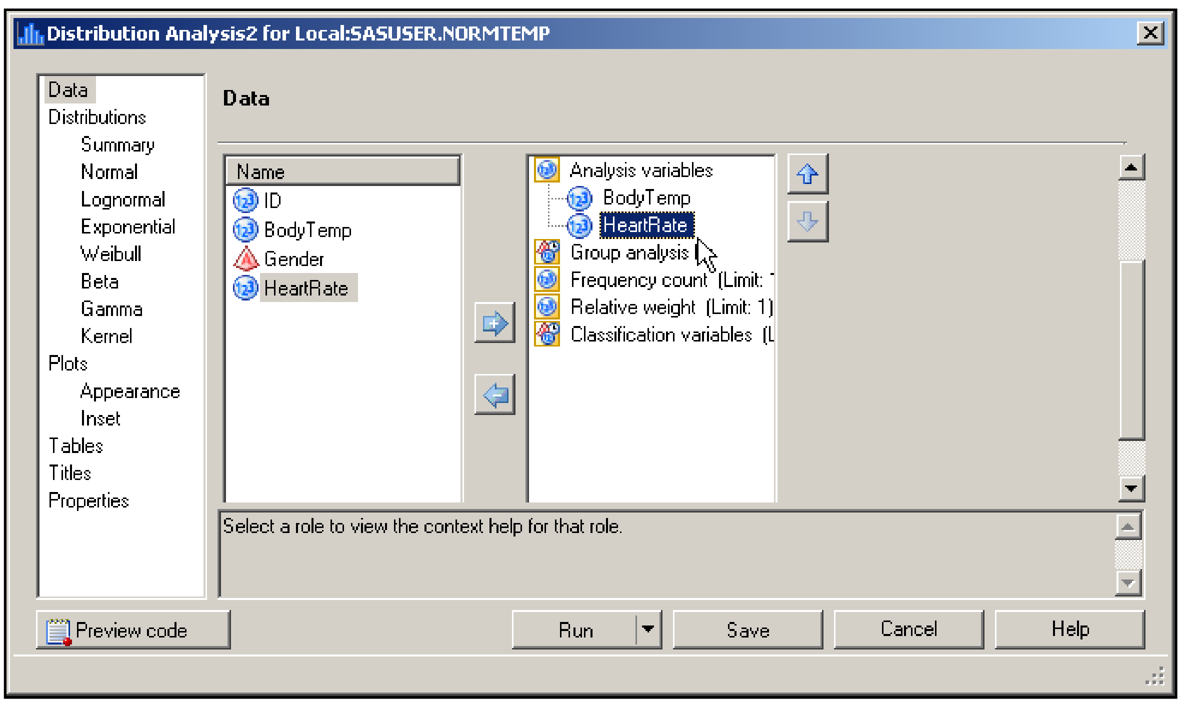
4. (Going above and beyond) - Producing Distributions and Descriptive Statistics

Use the **NormTemp** data set to answer the following:

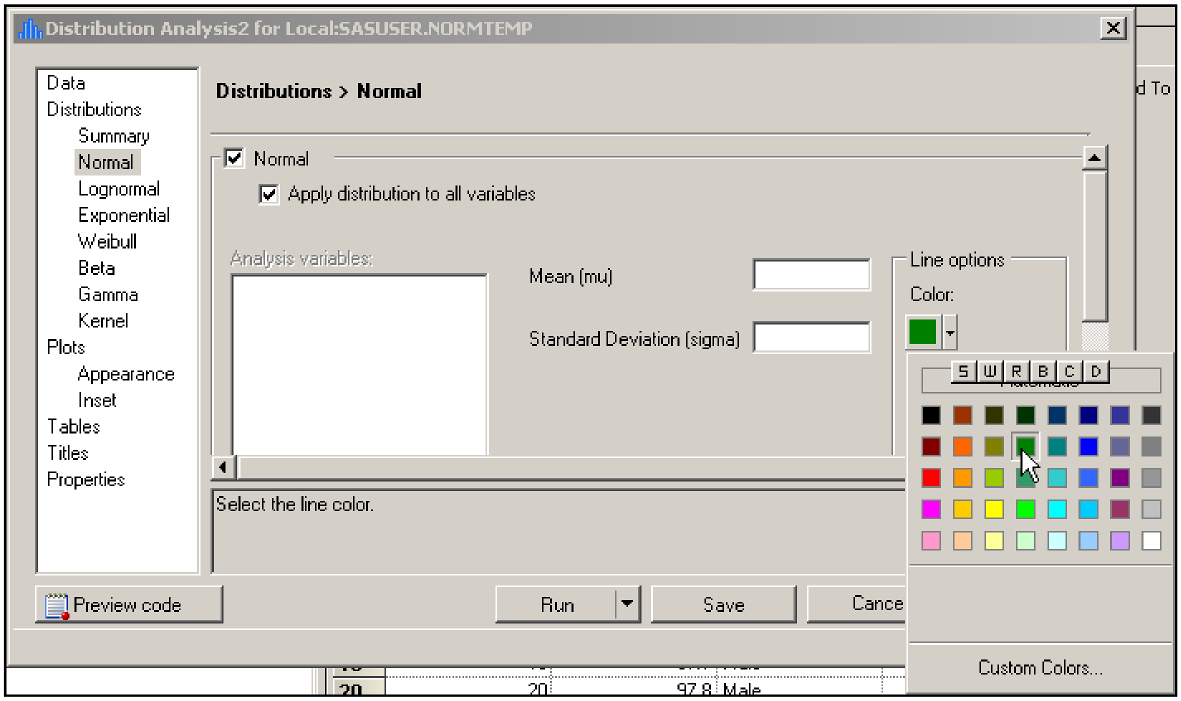
- With the **NORMTEMP** data set selected, click **Describe** ⇨ **Distribution Analysis…**.



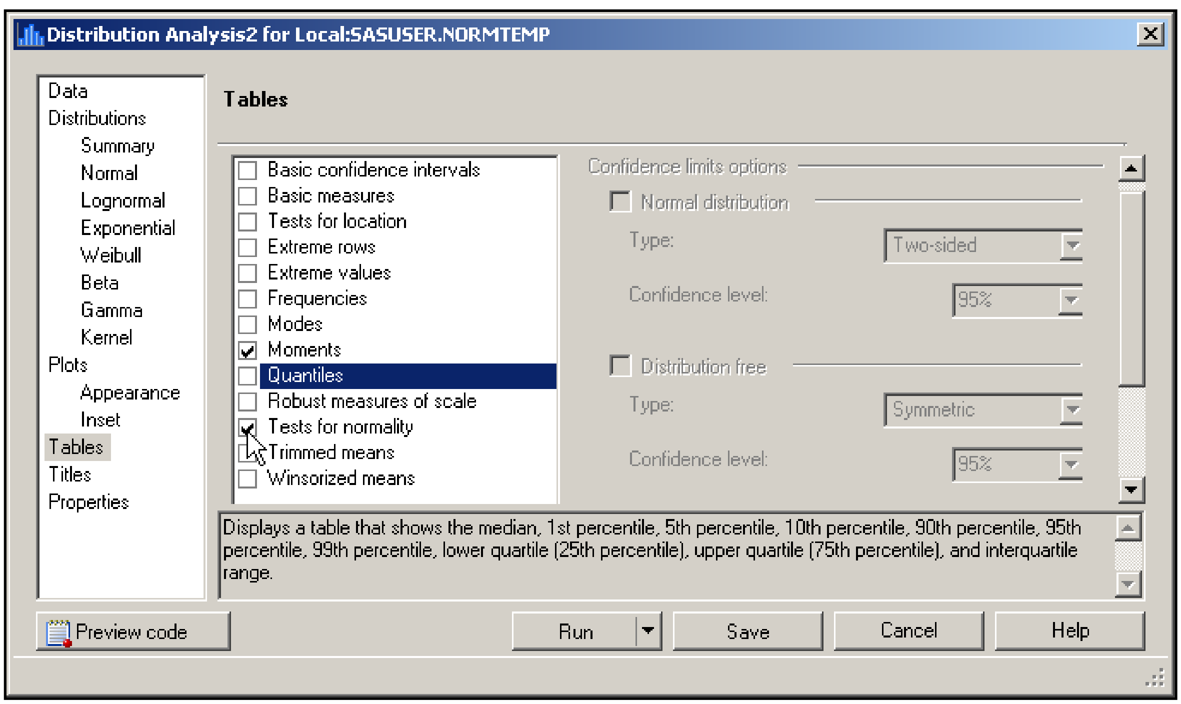
- Add **BodyTemp** and **HeartRate** to the analysis variables task role.



- Click **Normal** under Distributions and then check the box for **Normal**. Change the line options color to any color that you want.



- Click **Appearance** under **Plots** and select **Histogram**, **Probability Plot**, and **Box Plot**. Choose any color scheme.
- Click **Tables** and then check the boxes for **Moments**, and **Tests for Normality**. Deselect every other box.



- Click 

.

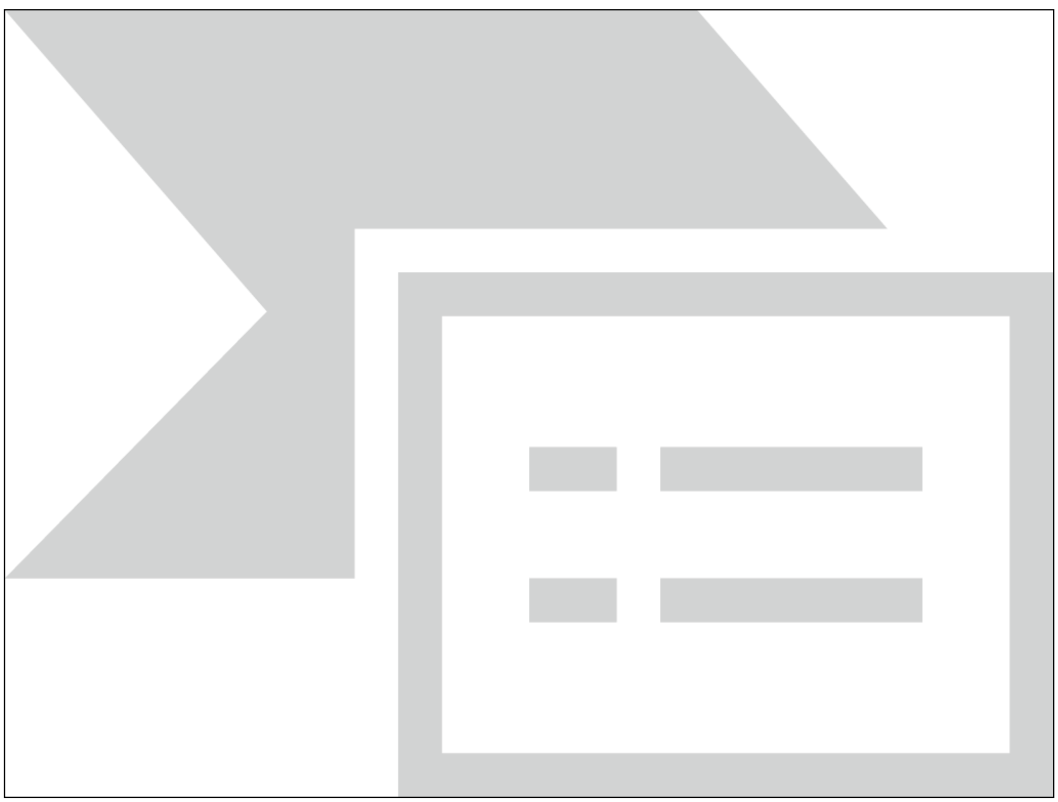
a. Complete the descriptive statistics table below. Do the variables appear to be normally distributed?

Distribution analysis of: BodyTemp, HeartRate

The UNIVARIATE Procedure
Variable: BodyTemp

Moments			
N	130	Sum Weights	130
Mean	98.2492308	Sum Observations	12772.4
Std Deviation	0.73318316	Variance	0.53755754
Skewness	-0.0044191	Kurtosis	0.7804574
Uncorrected SS	1254947.82	Corrected SS	69.3449231
Coeff Variation	0.74624824	Std Error Mean	0.06430442

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.986577	Pr < W	0.2332
Kolmogorov-Smirnov	D	0.064727	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.081952	Pr > W-Sq	0.2014
Anderson-Darling	A-Sq	0.520104	Pr > A-Sq	0.1916



Distribution analysis of: BodyTemp, HeartRate

The UNIVARIATE Procedure
Fitted Normal Distribution for BodyTemp

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	98.24923
Std Dev	Sigma	0.733183

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.06472685	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.08195196	Pr > W-Sq	0.201
Anderson-Darling	A-Sq	0.52010388	Pr > A-Sq	0.192

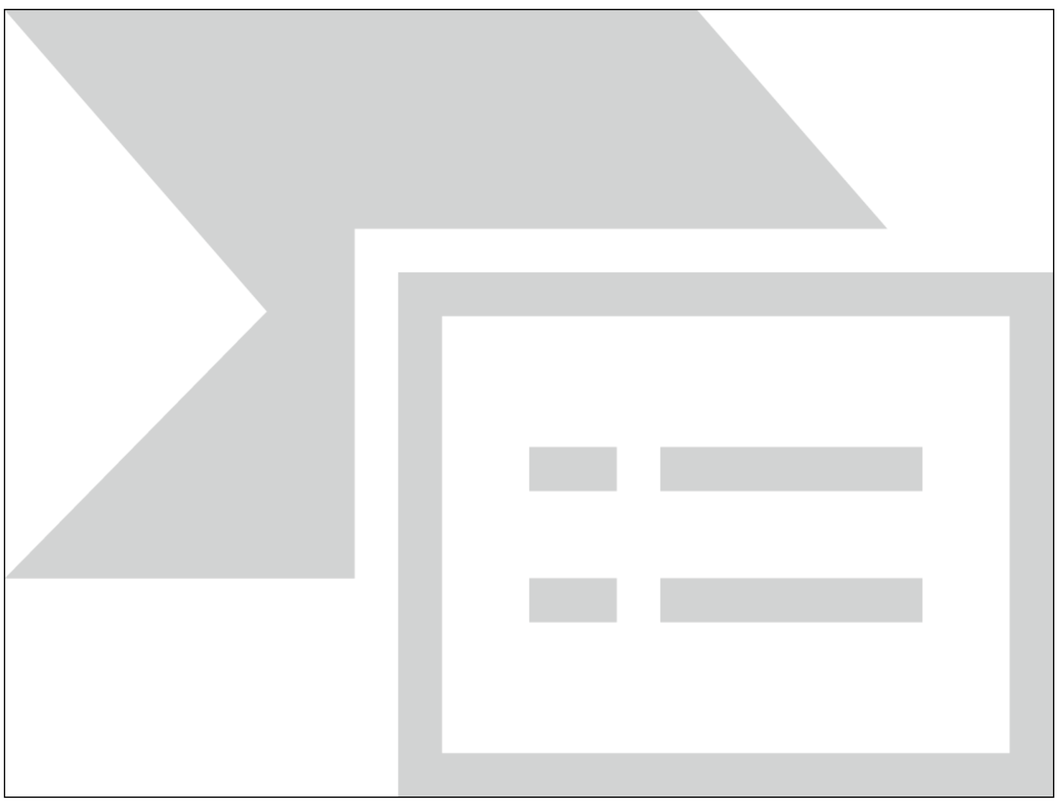
Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	96.4000	96.5436
5.0	97.0000	97.0433
10.0	97.2500	97.3096
25.0	97.8000	97.7547
50.0	98.3000	98.2492
75.0	98.7000	98.7438
90.0	99.1000	99.1888
95.0	99.3000	99.4552
99.0	100.0000	99.9549

Generated by the SAS System ('SASApp', X64_ES08R2) on September 12, 2015 at 3:13:32 PM

Page Break

Distribution analysis of: BodyTemp, HeartRate

The UNIVARIATE Procedure

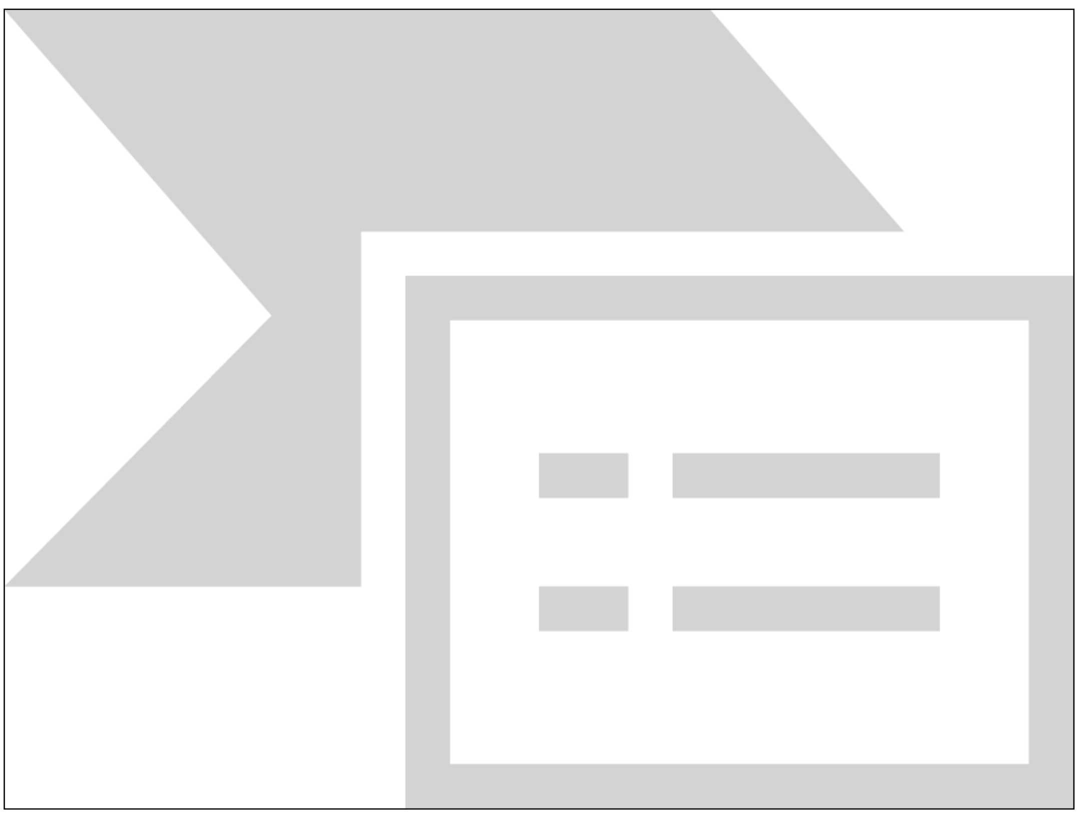


Distribution analysis of: BodyTemp, HeartRate

The UNIVARIATE Procedure
Variable: HeartRate

Moments			
N	130	Sum Weights	130
Mean	73.7615385	Sum Observations	9589
Std Deviation	7.06207674	Variance	49.8729278
Skewness	-0.178353	Kurtosis	-0.463021
Uncorrected SS	713733	Corrected SS	6433.60769
Coeff Variation	9.57419935	Std Error Mean	0.6193851

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.988544	Pr < W	0.3550
Kolmogorov-Smirnov	D	0.076729	Pr > D	0.0600
Cramer-von Mises	W-Sq	0.065767	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.393271	Pr > A-Sq	>0.2500



Distribution analysis of: BodyTemp, HeartRate

The UNIVARIATE Procedure
Fitted Normal Distribution for HeartRate

Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	73.76154
Std Dev	Sigma	7.062077

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	D	0.07672876	Pr > D	0.060
Cramer-von Mises	W-Sq	0.06576727	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.39327143	Pr > A-Sq	>0.250

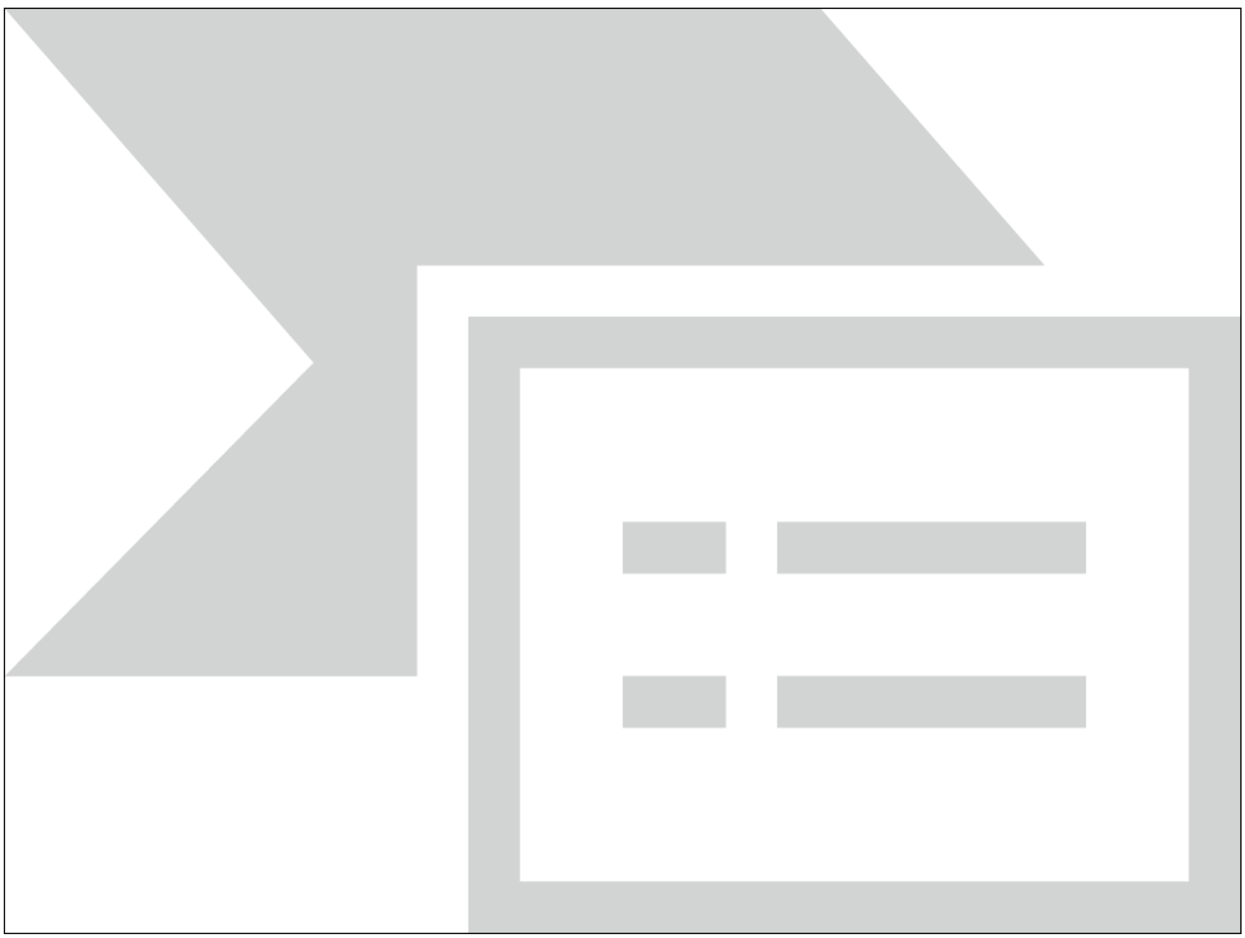
Quantiles for Normal Distribution		
	Quantile	
Percent	**Observed**	**Estimated**
1.0	57.0000	57.3327
5.0	62.0000	62.1455
10.0	64.0000	64.7111
25.0	69.0000	68.9982
50.0	74.0000	73.7615
75.0	79.0000	78.5248
90.0	83.0000	82.8120
95.0	84.0000	85.3776
99.0	89.0000	90.1904

Generated by the SAS System ('SASApp', X64_ES08R2) on September 12, 2015 at 3:13:32 PM

Page Break

Distribution analysis of: BodyTemp, HeartRate

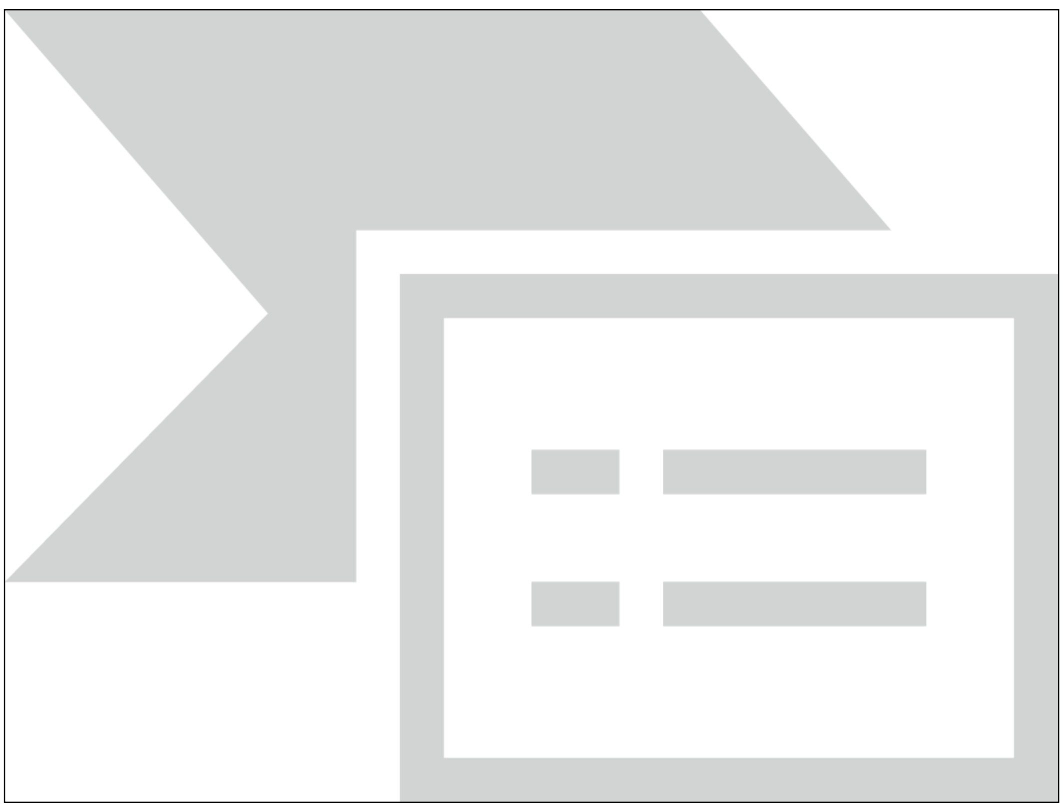
The UNIVARIATE Procedure

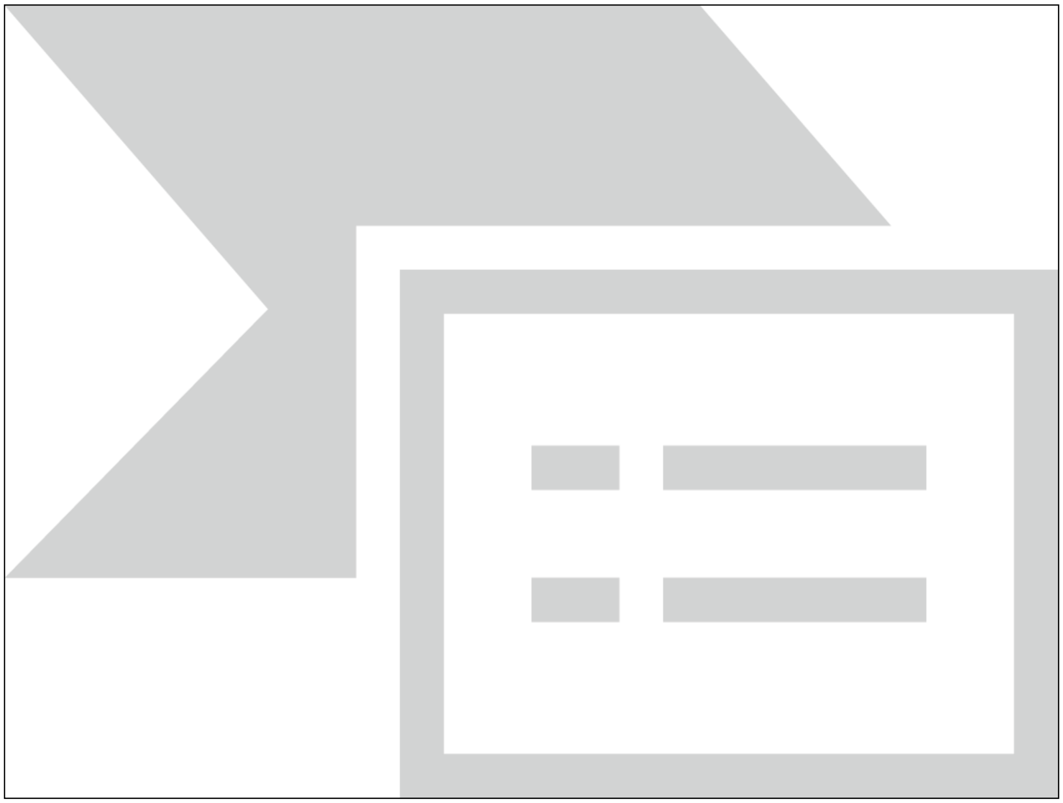


	BodyTemp	HeartRate
Minimum	96.30	57.00
Maximum	100.80	89.00
Mean	98.25	73.76
Standard Deviation	0.73	7.06
Skewness	-0.00	-0.02
Kurtosis	0.89	-0.46
Distribution: Normal	Yes/No	Yes/No

The distributions for both variables look approximately normal. None of the tests for normality are statistically significant.

b. Create box-and-whisker plots for the **BodyTemp** and **HeartRate** variables. Do there appear to be any outliers?





There appear to be three outliers for ***`BodyTemp`*** *and none for* ***`HeartRate`****.*