SASEG 3 Exercise: Fundamental Business Analytics -- Exploring the Data, Charts, and Creating Reports using SAS Enterprise Guide

(Fall 2015)

Sources (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

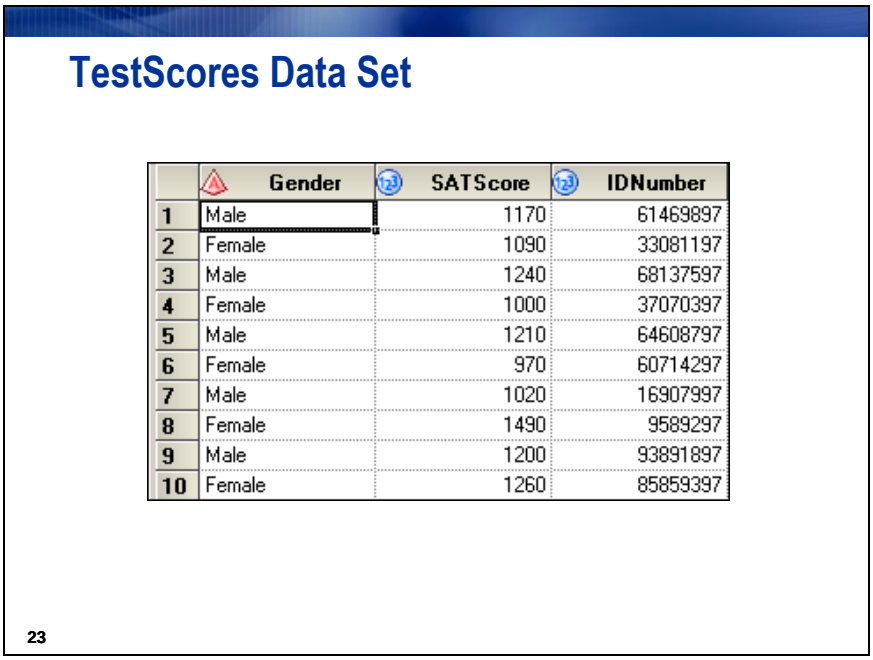
Remember the Test Scores Problem?

Defining the Problem

The purpose of the study is to determine whether or not the average combined Math and Verbal scores on the Scholastic Aptitude Test (SAT) at Carver County magnet high schools is 1200 – the goal set by the school board.

16

As a project, students in Ms. Chao's statistics course are to assess whether the students at magnet schools (schools with special curricula) in their district have accomplished the goal that the board of education set of having their graduating class scoring on average 1200 combined on the Math and Verbal portions of the SAT (Scholastic Aptitude Test), a college admissions exam. Each section of the SAT has a maximum score of 800. Eighty students are selected at random from among magnet school students in the district. The total scores are recorded and each sample member is assigned an identification number.



Example: The identification number of each student (**IDNumber**) and the total score on the SAT (**SATScore**) are recorded. The data is stored in the **TestScores** data set.

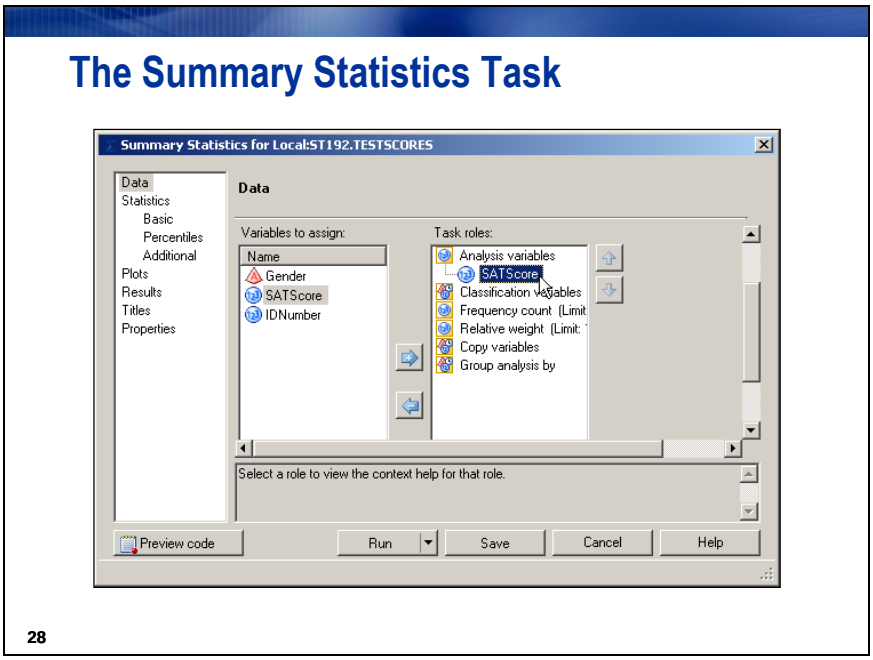
A *distribution* is a collection of data values that are arranged in order, along with the relative frequency. For any kind of data, it is important that you describe the location, spread, and shape of your distribution using graphical techniques and descriptive statistics.

For the example, these questions can be addressed using graphical techniques.

- Are the values of **SATScore** symmetrically distributed?
- Are any values of **SATScore** unusual?

You can answer these questions using descriptive statistics.

- What is the best estimate of the average of the values of **SATScore** for the population?
- What is the best estimate of the average spread or dispersion of the values of **SATScore** for the population?



The Summary Statistics task is used for generating descriptive statistics for your data.

Recall SASEG1 Exercise – Fundamental Summary Statistics

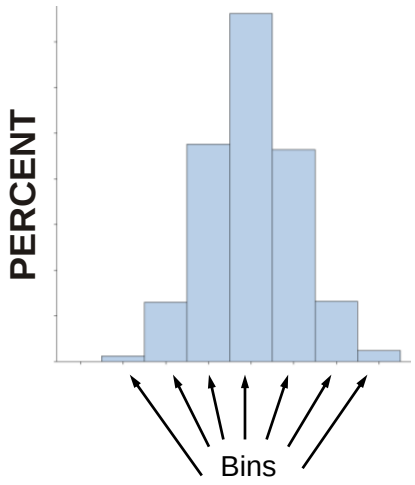
Distributions – Histograms - Plots

Objectives

- Look at distributions of continuous variables.
- Describe the normal distribution.
- Use the Distribution Analysis task to generate histograms and normal probability plots and to produce descriptive statistics.

33

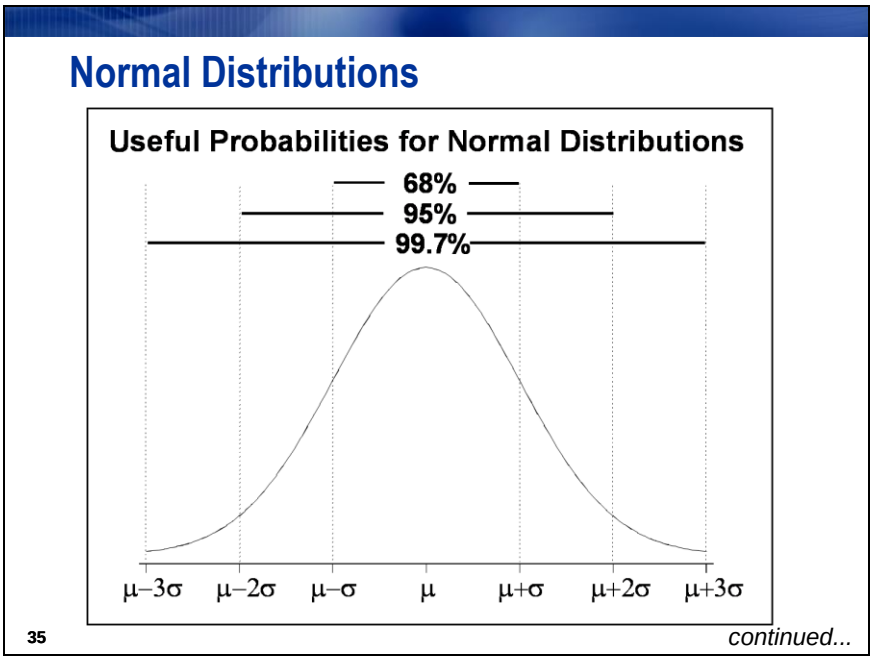
Picturing Distributions: Histogram



- Each bar in the histogram represents a group of values (a *bin*).
- The height of the bar is the percent of values in the bin.
- SAS determines the width and number of bins automatically, or you can specify them.

34

Most parametric statistical procedures (those in which parameters are to be estimated) assume an underlying distribution. It is a good idea to look at your data to see if the distribution of your sample data can reasonably be assumed to come from a population with the assumed distribution. A histogram is a good way to get an idea of what the population distribution is shaped like.



Quite often, although not always, a normal distribution is assumed.

The normal distribution is a mathematical function. The height of the function at any point on the horizontal axis is the "probability density" at that point. Normal distribution probabilities (which can be thought of as the proportion of the area under the curve) tend to be higher near the middle. The center of the distribution is the population mean (μ). The standard deviation (σ) describes how variable the distribution is about μ. A larger standard deviation implies a wider normal distribution. The mean locates the distribution (sets its center point) and the standard deviation scales it.

An observation value is considered unusual if it is far away from the mean. How far is far? You may use the mathematical properties of the normal probability distribution function (PDF) to determine that. If a population follows a normal distribution, then approximately:

- 68% of the data falls within 1 standard deviation of the mean
- 95% of the data falls within 2 standard deviations of the mean
- 99.7% of the data falls within 3 standard deviations of the mean.

Often, values that are more than 2 standard deviations from the mean are regarded as unusual. Now you can see why. Only about 5% of all values are as far away from the mean as that. (Sometimes, only values more than 3 standard deviations away from the mean are closely examined as unusual.)

You will also use this information later when talking about the concepts of confidence intervals and hypothesis tests.

Normal Distributions

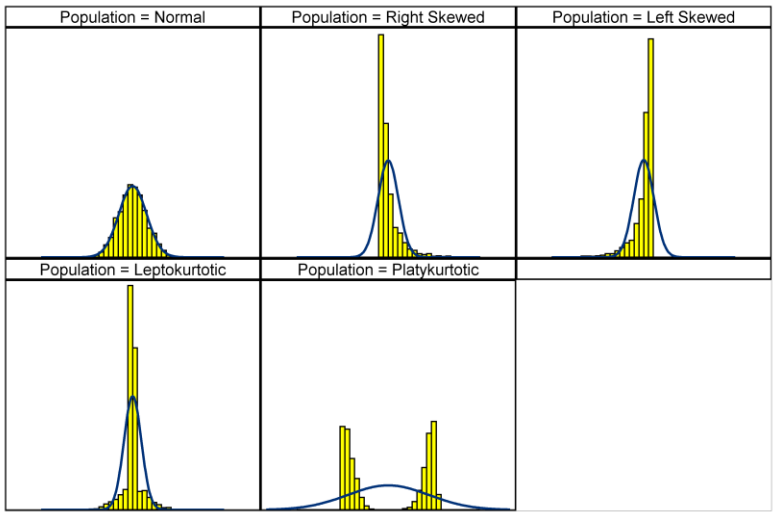
A normal distribution:

- is *symmetric*. If you draw a line down the center, you get the same shape on either side.
- is *fully characterized* by the mean and standard deviation. Given those two parameters, you know all there is to know about the distribution.
- is bell shaped.
- has mean = median = mode.

The blue line on each of the following graphs represents the shape of the normal distribution with the mean and variance estimated from the sample data.

36

Data Distributions Compared to Normal

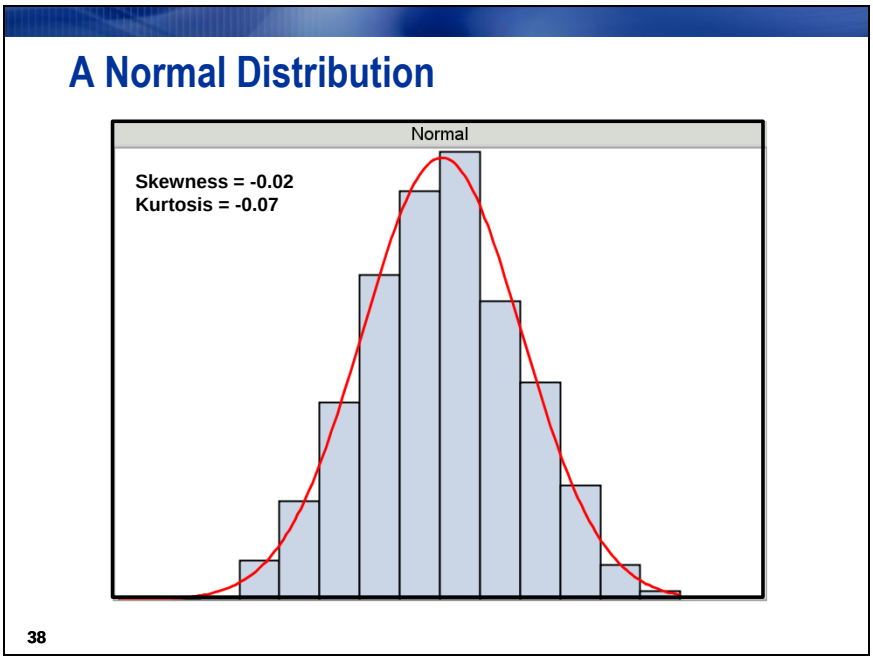


37

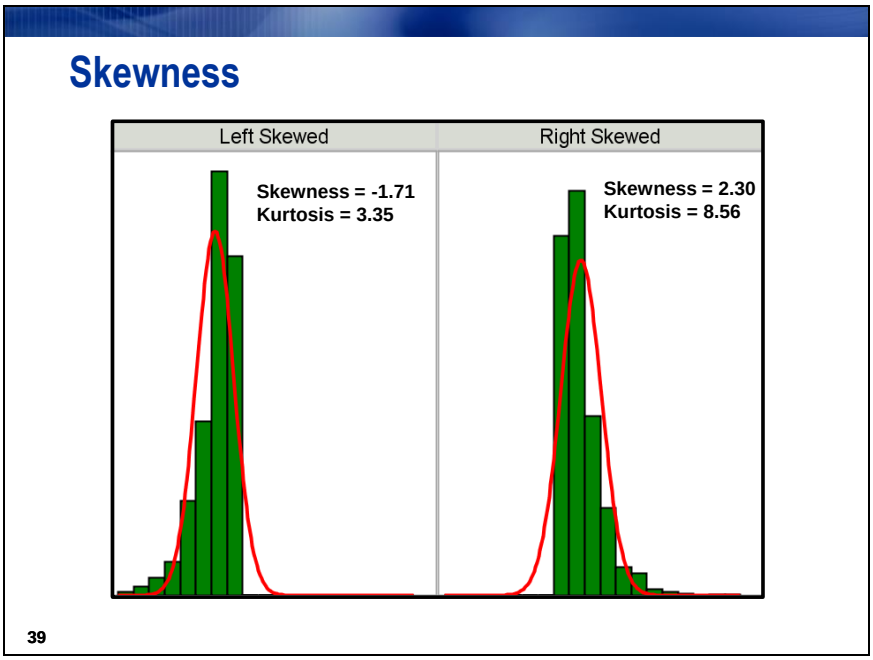
The distribution of your data might not look normal. There are infinitely different ways that a population can be distributed. When you look at your own data, you might take note of features of the distribution that indicate similarity or difference from the normal distribution.

In evaluating distributions, it is useful to look at statistical measures of the shape of the sample distribution compared to the normal.

Two such measures are skewness and kurtosis, which are defined over the next few pages.



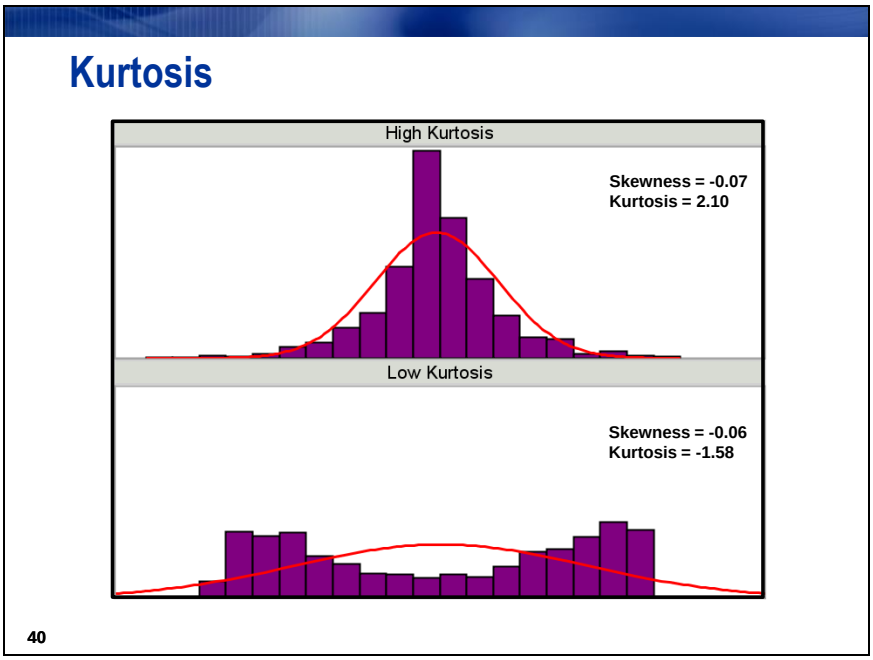
A histogram of data from a sample drawn from a normal population will generally show values of skewness and kurtosis near 0 in SAS Enterprise Guide output.



One measure of the shape of a distribution is skewness. The *skewness* statistic measures the tendency of your distribution to be more spread out on one side than the other. A distribution that is approximately symmetric has a skewness statistic close to 0.

If your distribution is more spread out on the

- left side, then the statistic is negative, and the mean is less than the median. This is sometimes referred to as a *left-skewed* or *negatively skewed* distribution.
- right side, then the statistic is positive, and the mean is greater than the median. This is sometimes referred to as a *right-skewed* or *positively skewed* distribution.



Kurtosis is often very difficult to assess visually. The kurtosis statistic measures the tendency of your data to be distributed toward the center or toward the tails of the distribution. A distribution that is approximately normal has a kurtosis statistic close to 0 in SAS.

If your kurtosis statistic is negative, the distribution is said to be *platykurtic* compared to the normal. If the distribution is symmetric, a platykurtic distribution tends to have a smaller-than-normal proportion of observations in the tails, and/or a somewhat flat peak. A platykurtic distribution is often referred to as *light-tailed*. Rectangular, bimodal, and multimodal distributions tend to have low values of kurtosis.

If your kurtosis statistic is positive, the distribution is said to be *leptokurtic* compared to the normal. If the distribution is symmetric, a leptokurtic distribution tends to have a larger-than-normal proportion of observations in the extreme tails and/or a taller peak than the normal. A leptokurtic distribution is often referred to as *heavy-tailed*. Leptokurtic distributions are also sometimes referred to as *outlier-prone distributions*.

Distributions that are asymmetric also tend to have nonzero kurtosis. In these cases, understanding kurtosis is considerably more complex than in situations where the distribution is approximately symmetric.

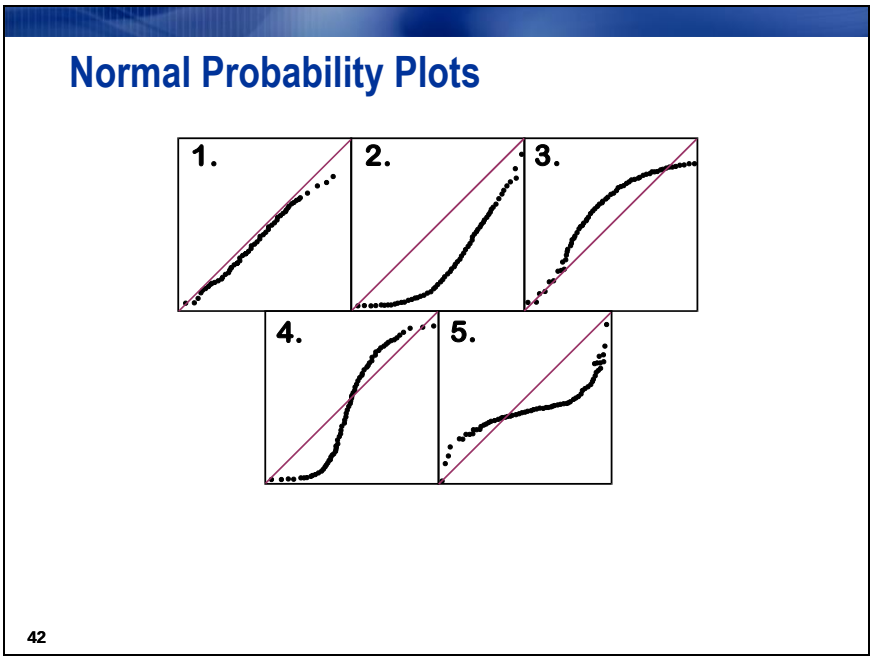
The normal distribution actually has a kurtosis value of 3, but SAS subtracts a constant of 3 from all reported values of kurtosis, making the constant-modified value for the normal distribution 0 in SAS output. That is the value against which to compare a sample kurtosis value in SAS when assessing normality.

Graphical Displays of Distributions

You can produce three kinds of plots for examining the distribution of your data values:

- histograms
- normal probability plots
- box-and-whisker plots

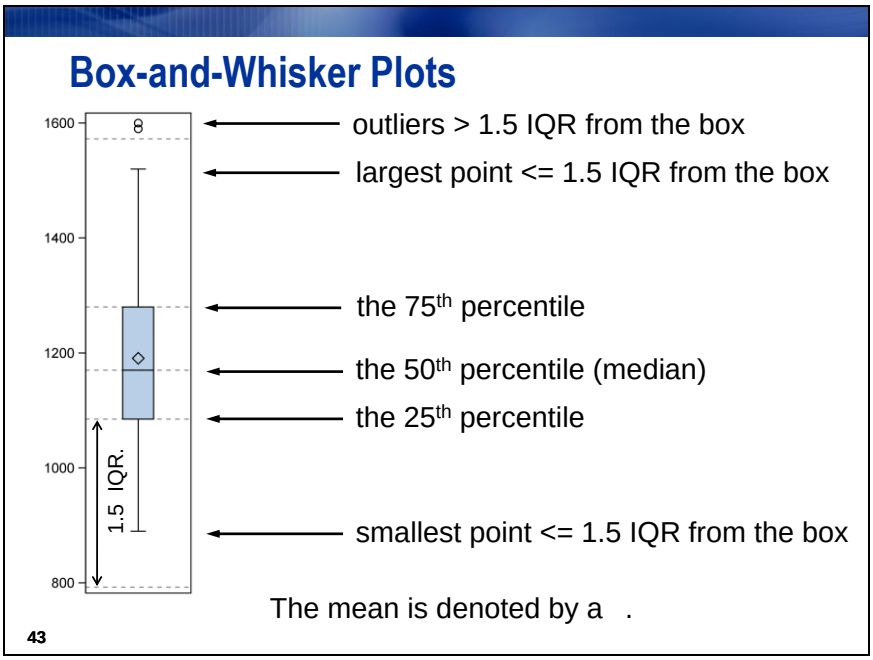
41



A *normal probability plot* is a visual method for determining whether or not your data comes from a distribution that is approximately normal. The vertical axis represents the actual data values, and the horizontal axis displays the expected percentiles from a standard normal distribution.

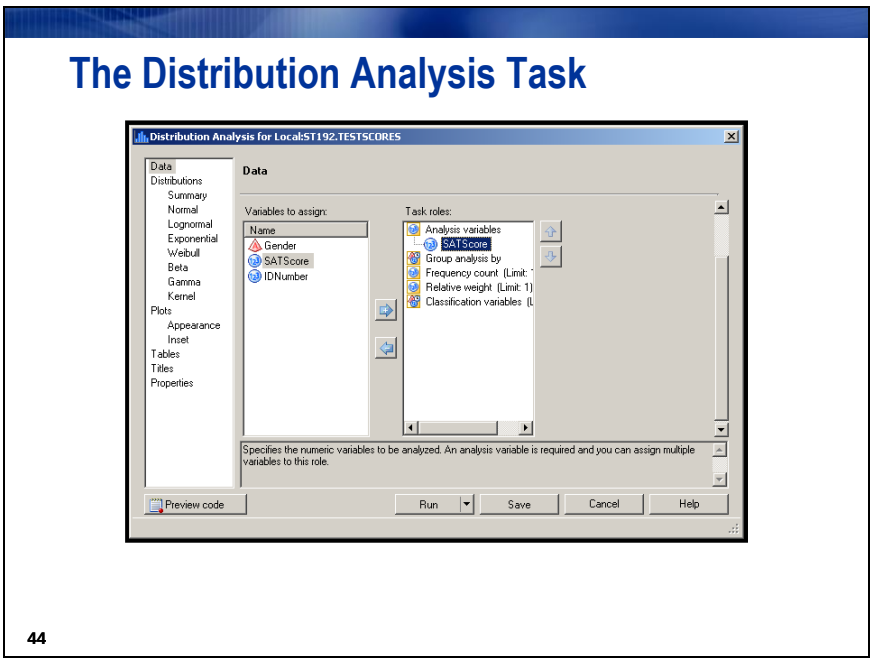
The above diagrams illustrate some possible normal probability plots for data from a

1. normal distribution (the observed data follow the reference line)
2. skewed-to-the-right distribution
3. skewed-to-the-left distribution
4. light-tailed distribution
5. heavy-tailed distribution.



Box-and-Whisker plots (sometimes referred to simply as *Box Plots*) provide information about the variability of data and the extreme data values. The box represents the middle 50% of your data (between the 25th and 75th percentile values). You get a rough impression of the symmetry of your distribution by comparing the mean and median, as well as assessing the symmetry of the box and whiskers around the median line. The whiskers extend from the box as far as the data extends, to a distance of, at most, 1.5 interquartile range (IQR) units. If any values lie more than 1.5 IQR from either end of the box, they are represented in SAS by individual plot symbols.

The plot above shows that the data is approximately symmetric.

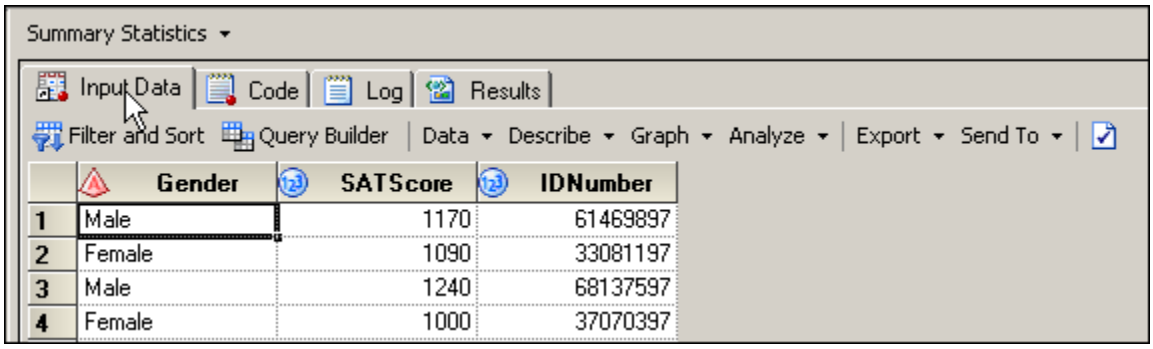


Exercise - Examining Distributions

This demonstration illustrates how to create statistical tables, histograms, normal probability plots, and box plots using the Distribution Analysis task.

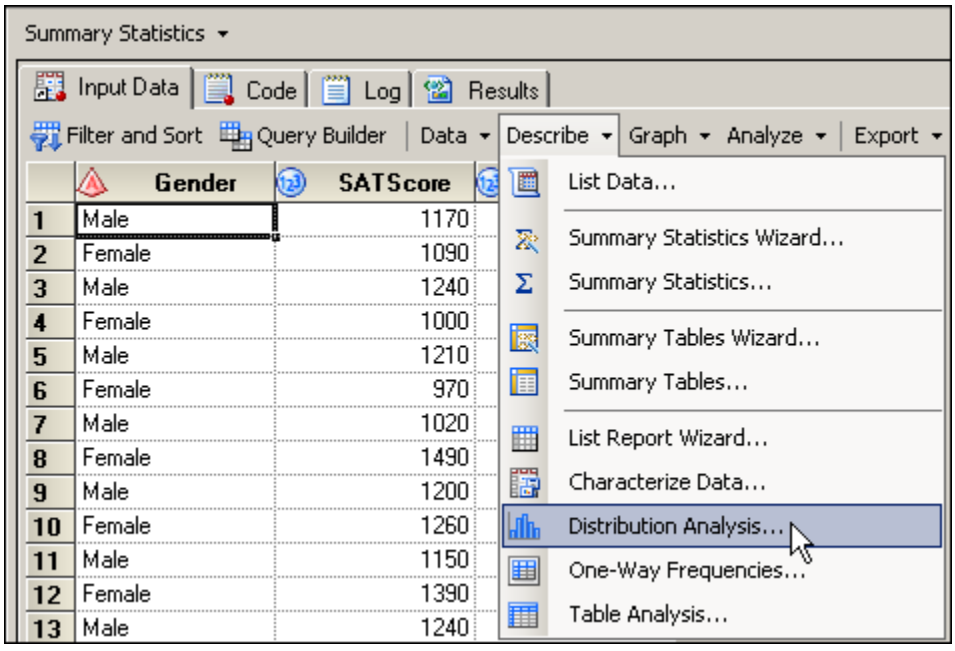
1. Click the **Input Data** tab under Summary Statistics from the previous demonstration to show the **`TestScores`** data table.

   File > Open >Data--> Servers > SASApp-->Files > D: > ISYS 5503--> ISYS 5503 Shared Datasets



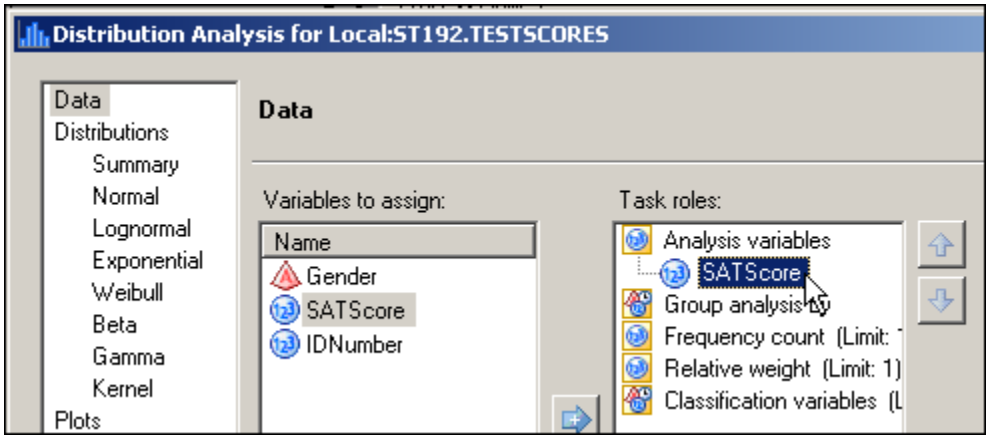
	Gender	SATScore	IDNumber
1	Male	1170	61469897
2	Female	1090	33081197
3	Male	1240	68137597
4	Female	1000	37070397

2. Select **Describe** ⇨ **Distribution Analysis…**.

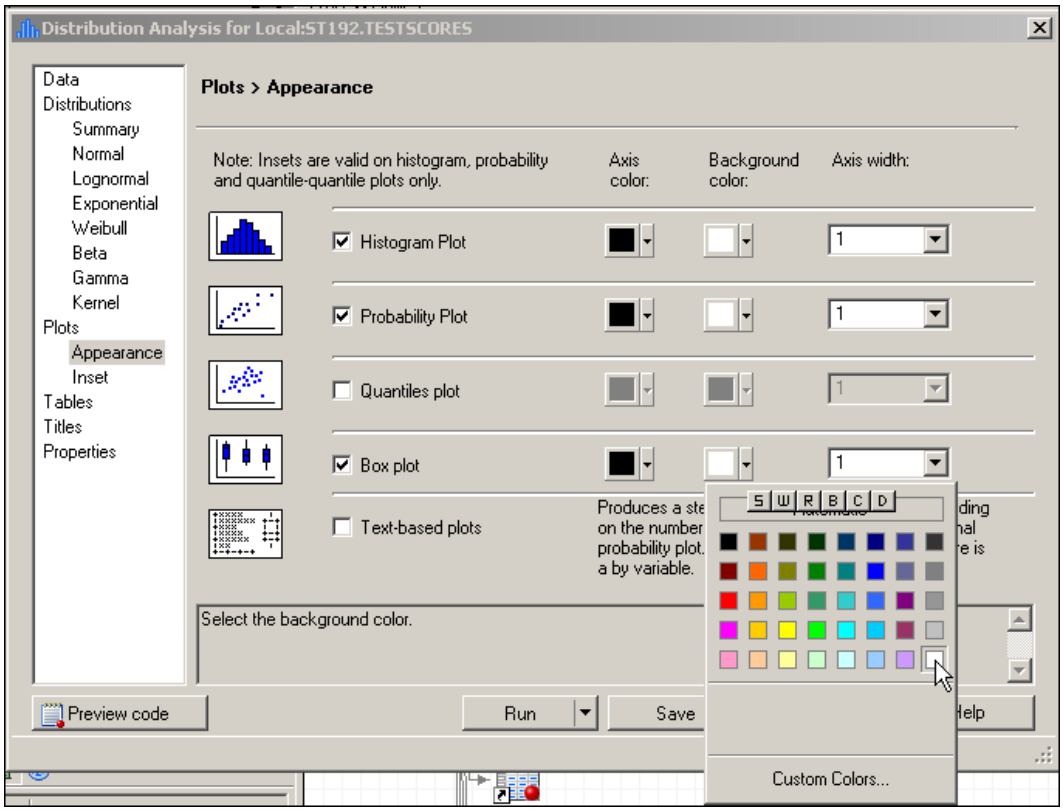


	Gender	SATScore
1	Male	1170
2	Female	1090
3	Male	1240
4	Female	1000
5	Male	1210
6	Female	970
7	Male	1020
8	Female	1490
9	Male	1200
10	Female	1260
11	Male	1150
12	Female	1390
13	Male	1240

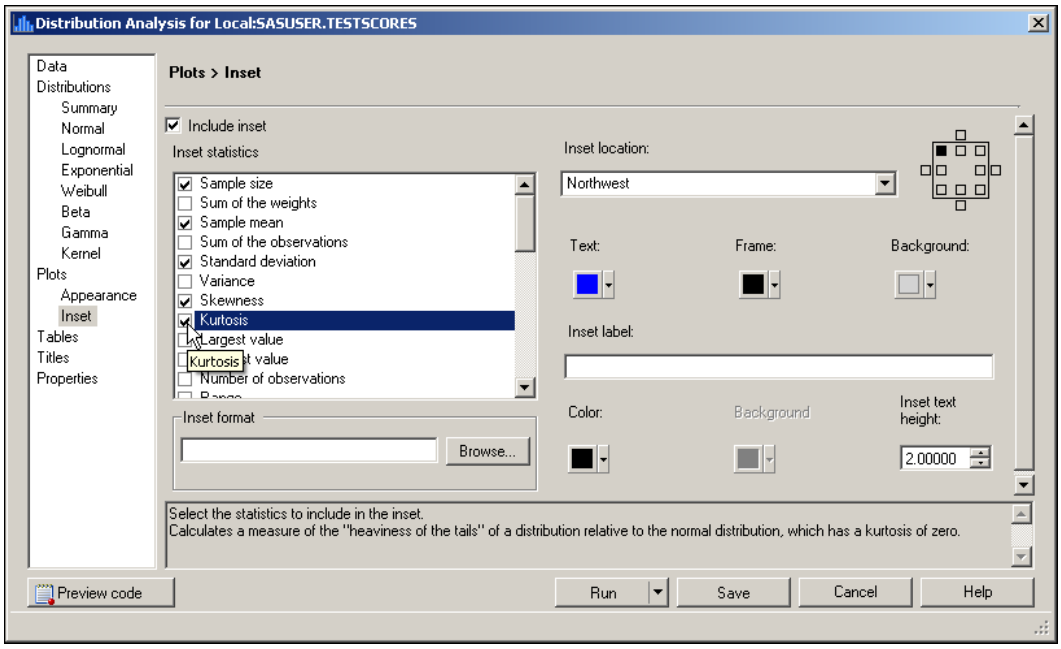
3. With **Task Roles** selected on the left, drag and drop **`SATScore`** to the analysis variables role.



4. Select **Appearance** under Plots on the left, and check the box next to **Histogram Plot**, **Probability Plot**, and **Box Plot**. Change the background color in each case to white.

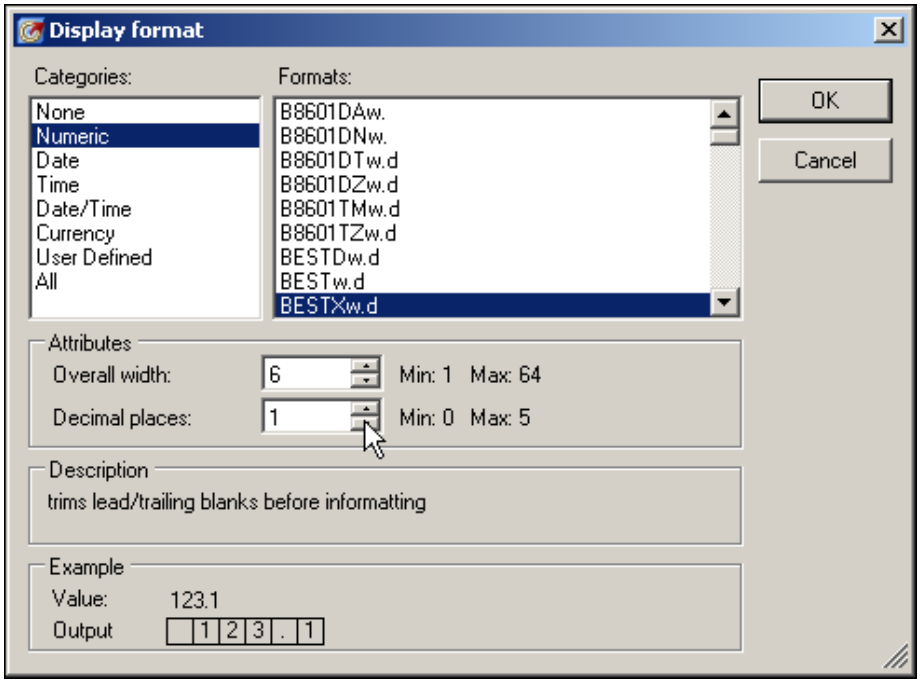


5. Select **Inset** under Plots on the left, check the **Include inset** box, and check the boxes for **Sample Size**, **Sample Mean**, **Standard Deviation**, **Skewness**, and **Kurtosis**.



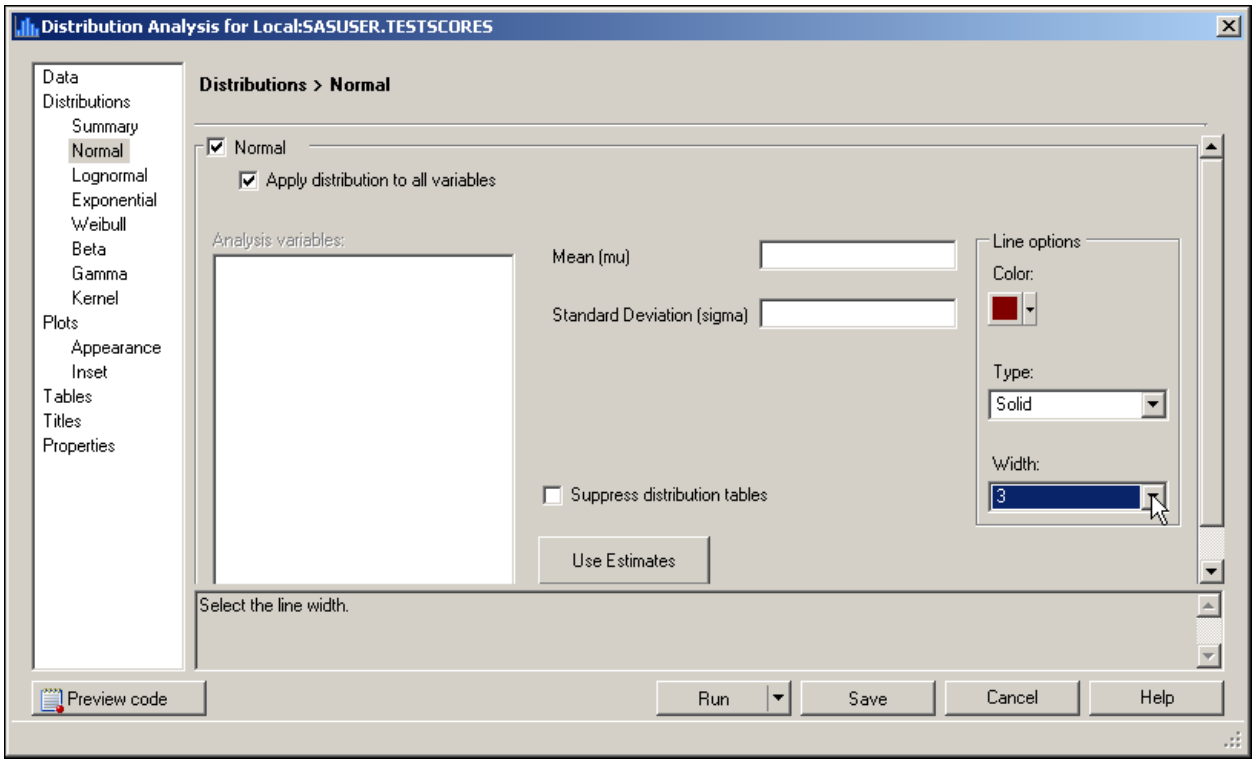
6. Click Browse... to change the format for the inset statistics.

7. Select **Numeric** under Categories. Find the BESTXw.d format and assign an overall width of **6** with **1** decimal place. This process will limit the reported output in the inset to 6 columns total, with one assigned after the decimal point.

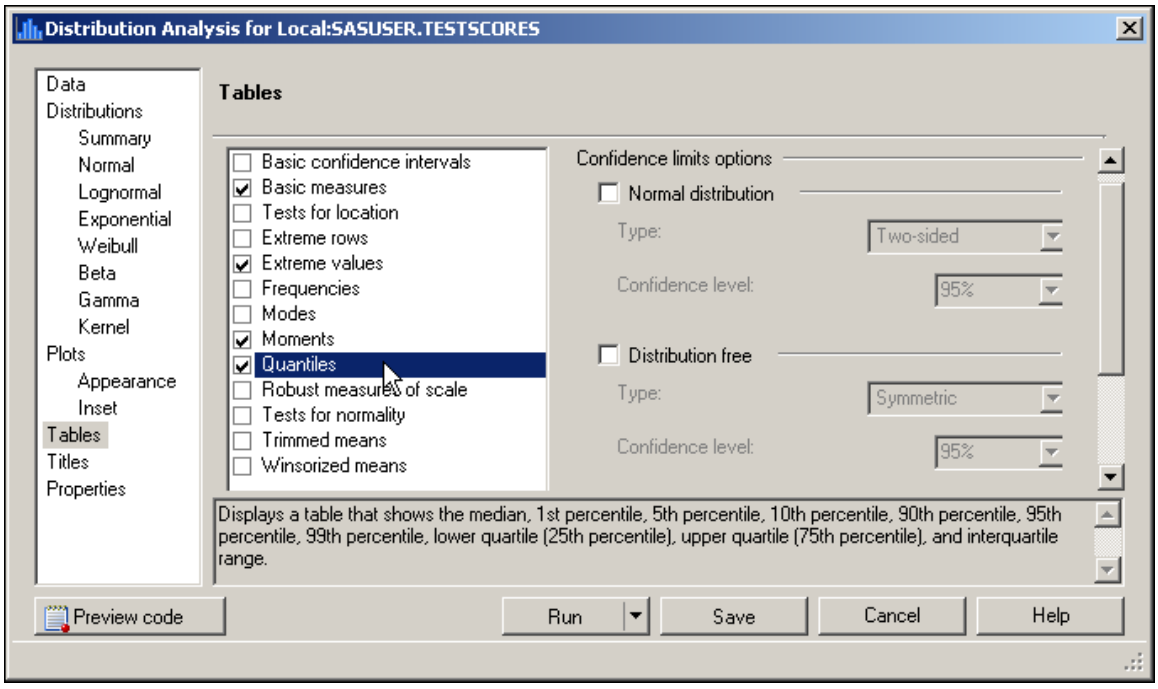


8. Click OK.

9. In order to draw a diagonal reference line for the normal probability plot and a normal curve for the histogram, select **Normal** under Distribution at left. Then check the box for **Normal** and change the color to dark red and the width to **3**.



10. Select **Tables** on the left. Deselect the boxes for **Basic confidence intervals** and **Tests for location**. Select the boxes for **Extreme values**, **Moments** and **Quantiles**. (You might need to click twice to select and deselect the tables.)



11. Change the titles and footnotes if desired.
12. Click **Run**.

Moments			
N	80	Sum Weights	80
Mean	1190.625	Sum Observations	95250
Std Deviation	147.058447	Variance	21626.1867
Skewness	0.64202018	Kurtosis	0.42409987
Uncorrected SS	115115500	Corrected SS	1708468.75
Coeff Variation	12.3513656	Std Error Mean	16.4416342

Basic Statistical Measures			
Location		Variability	
Mean	1190.625	Std Deviation	147.05845
Median	1170.000	Variance	21626
Mode	1050.000	Range	710.00000
		Interquartile Range	195.00000

Quantiles (Definition 5)	
Level	Quantile
100% Max	1600
99%	1600
95%	1505
90%	1375
75% Q3	1280
50% Median	1170
25% Q1	1085
10%	1020
5%	995
1%	890
0% Min	890

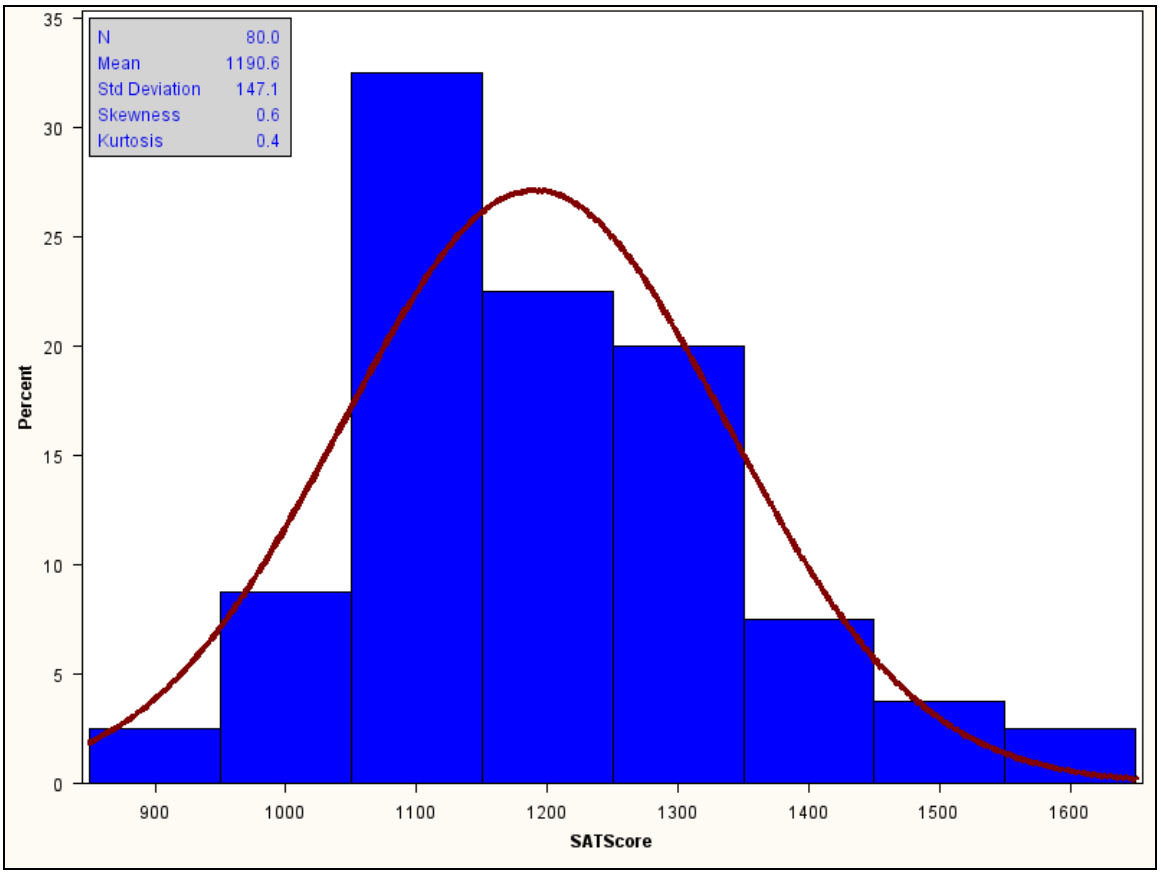
Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
890	69	1490	8
910	74	1520	42
970	6	1520	54
990	51	1590	70
1000	4	1600	25

Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	890	1	39	1390	1
2	910	1	40	1490	1
3	970	1	41	1520	2
4	990	1	42	1590	1
5	1000	1	43	1600	1

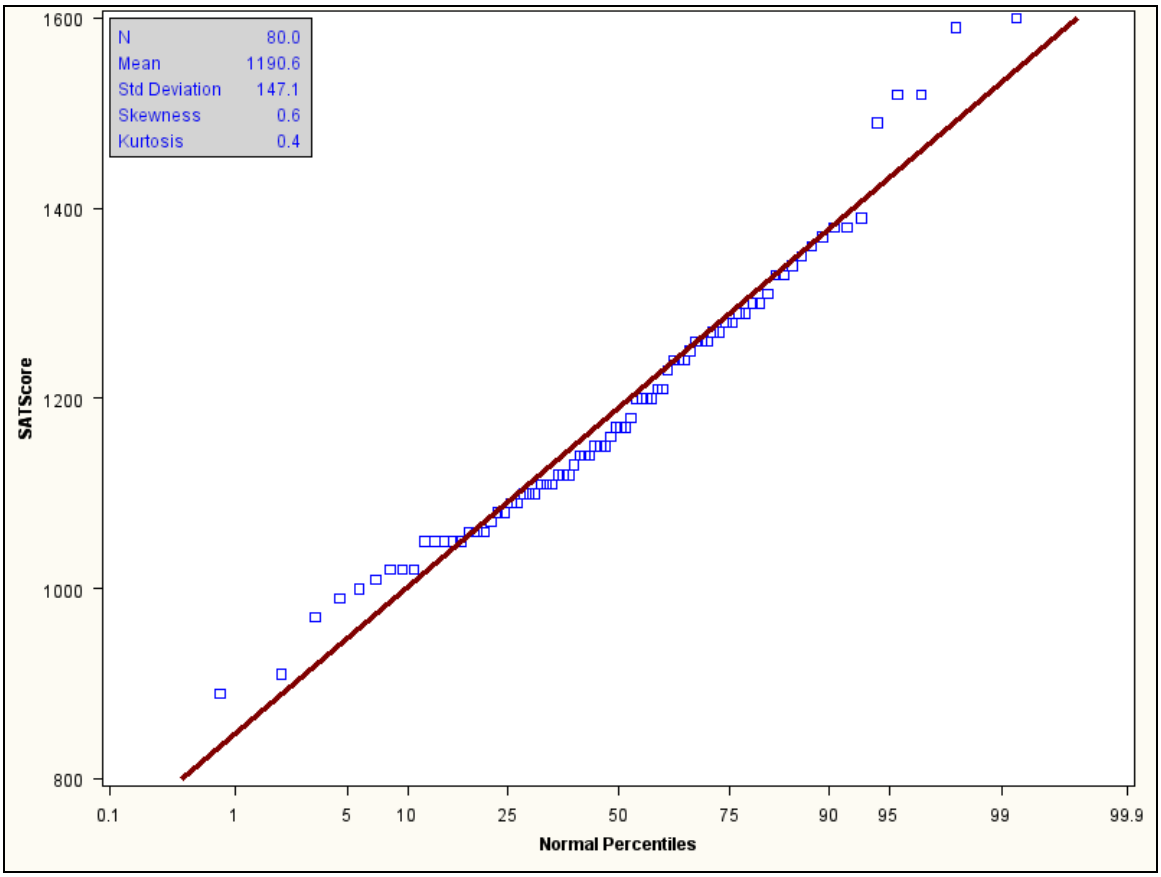
The tabular output indicates that

- the mean of the data is 1190.625. This is approximately equal to the median (1170), which indicates the distribution is fairly symmetric.
- the standard deviation is 147.058447, which means that the average variability around the mean is approximately 147 points.
- the distribution is slightly skewed to the right (Skewness = +0.64).
- the distribution has slightly heavier tails than the normal distribution (Kurtosis = +0.42).
- the student with the lowest score is observation (row number) 69, with a score of 890. The student with the highest score is row number 25, with a score of 1600 (highest possible score for the SAT.

In the Quantiles table, Definition 5 indicates that PROC UNIVARIATE is using the default definition for calculating percentile values. You can use the PCTLDEF= option in the PROC UNIVARIATE statement to specify one of five methods. These methods are listed in the "Percentile Definitions" appendix.



The bin identified with the midpoint of 1100 has approximately 33% of the values. The skewness and kurtosis values are reported in the inset.

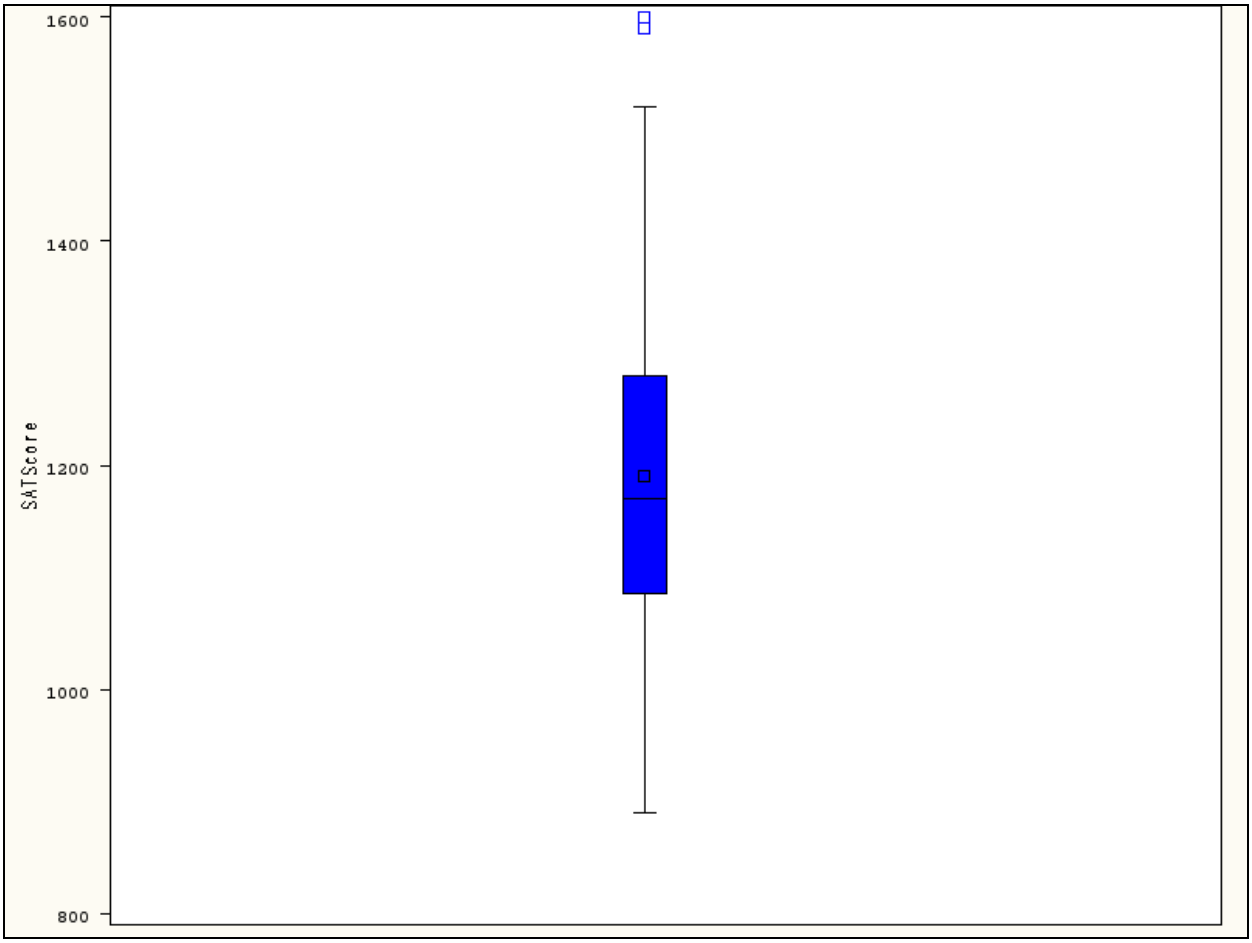


The normal probability plot is shown above. The 45-degree line represents where the data values would fall if they came from a normal distribution. The squares represent the observed data values. Because the squares follow the 45-degree line in the graph, you can conclude that there does not appear to be any severe departure from the normality.

Asking for the normal reference curve for the histogram also produces a set of tables relating to assessing whether the distribution is normal or not. There is a table with three tests presented: the Kolmogorov-Smirnov; Anderson-Darling; and Cramer-von Mises. In each case, the null hypothesis is that the distribution is normal. Therefore, high *p*-values are desirable.

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.08382224	**Pr > D**	>0.150
Cramer-von Mises	**W-Sq**	0.09964577	**Pr > W-Sq**	0.114
Anderson-Darling	**A-Sq**	0.70124822	**Pr > A-Sq**	0.068

All three tests are not significant, implying that the distribution of **`SATScore`** is approximately normal.



There are two outliers (values beyond 1.5 interquartile units from the box).

Sporting Goods: Exploring and Creating a Report Using SAS Enterprise Guide

Exploring the Data and Creating a Report

Investigate the products sold by country.

- Create a report of product category revenues by country.

Which product categories sell the greatest quantity of products in each country? Which account for the highest revenue in each country?

- Create a graph of total quantity for each country.
- Create a graph of total revenue for each country.

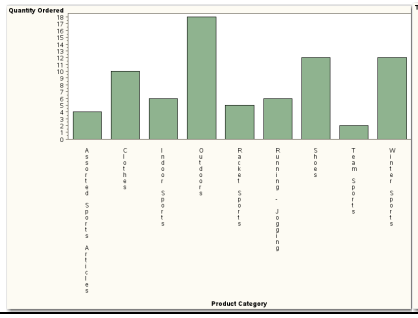
228

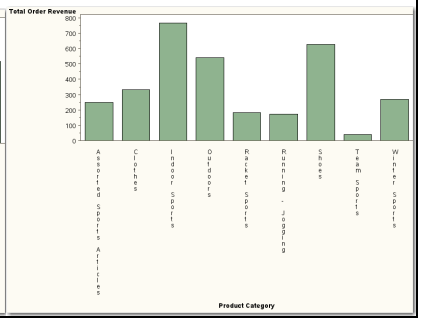
Exploratory Analysis

Summary Statistics

Results

Analysis Variable : Total Order Revenue		Total Order Revenue	
Country	N Obs	Sum	N
Germany	43	2640.55	43
Italy	12	536.8500000	12





229

Exercise -- Sporting Goods Case Study: Exploring the Data and Creating a Basic Report

1. Use the **CUSTPRODORDERS** data set.

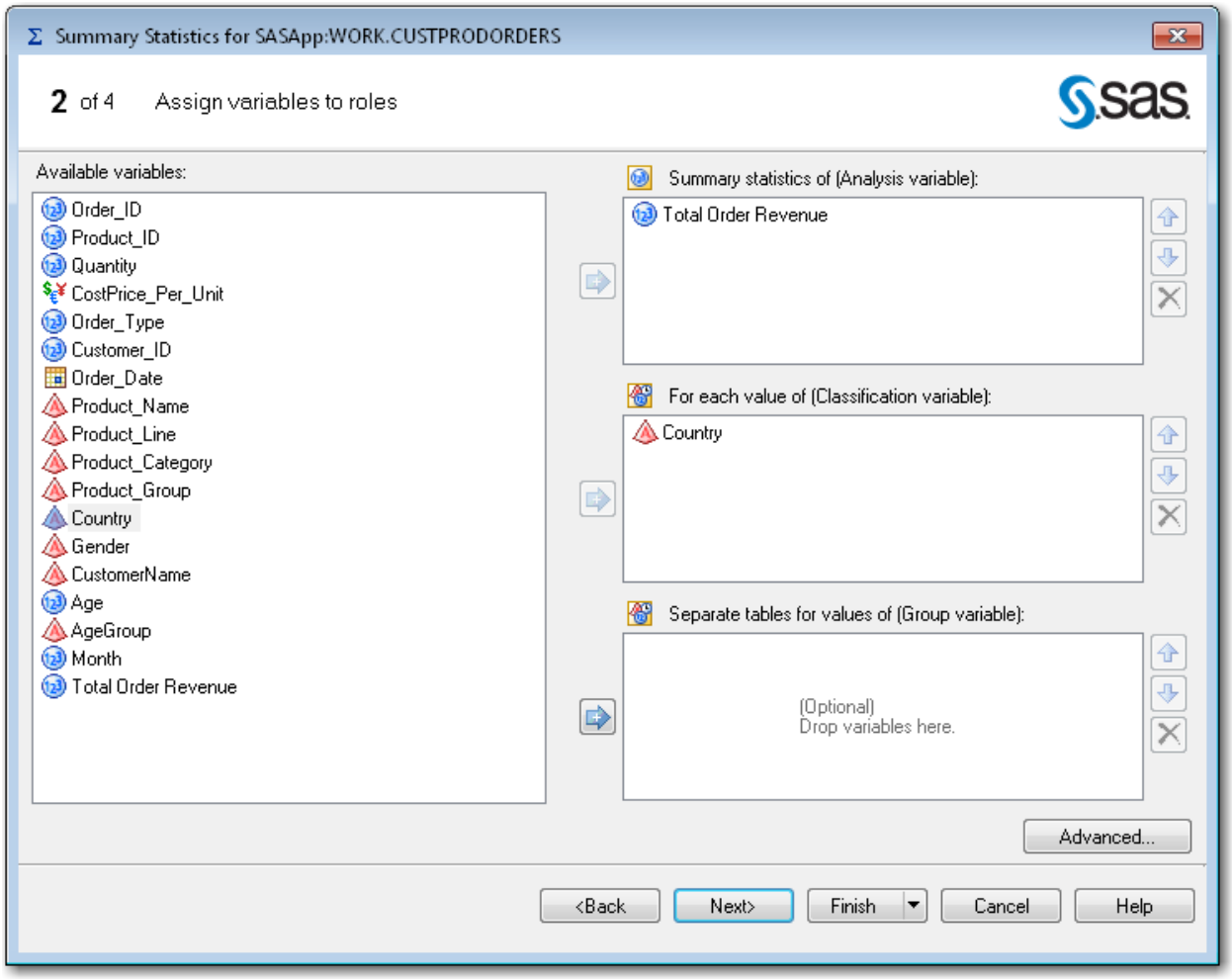
   Recall where this is --- File > Open >Data--> Servers > SASApp-->Files > D: > ISYS 5503--> ISYS 5503 Shared Datasets

2. Select **Describe** ⇨ **Summary Statistics Wizard**.

3. Click Next> .

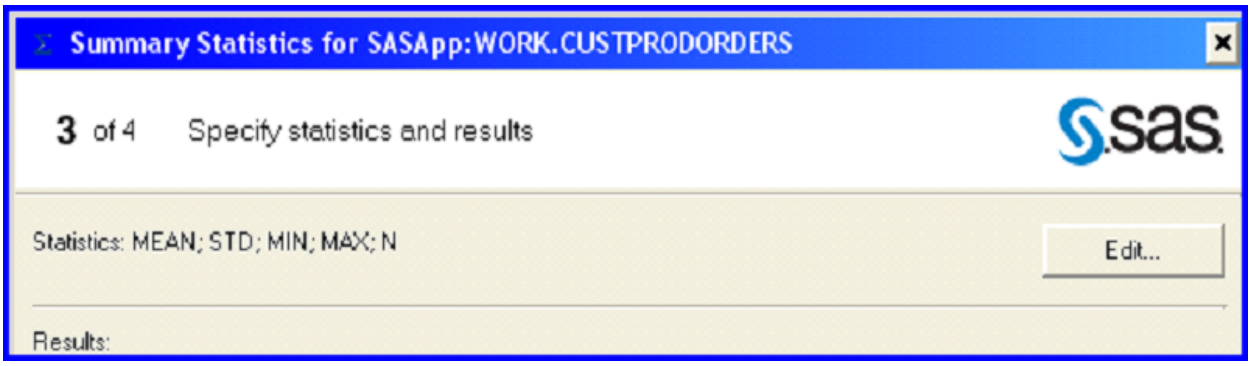
4. Assign **Total Order Revenue** to the Summary statistics of (Analysis variable) role.

5. Assign **Country** to the For each value of (Classification variables) role.

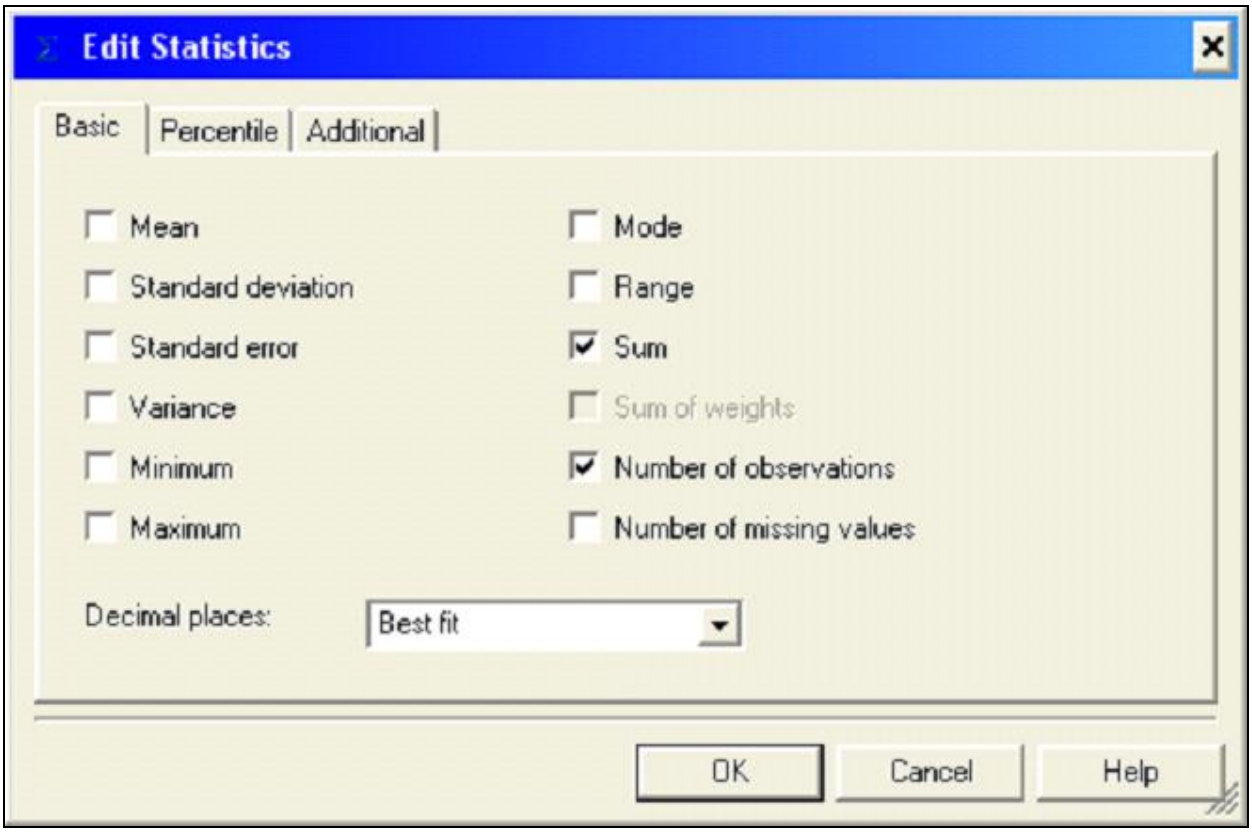


6. Click Next> .

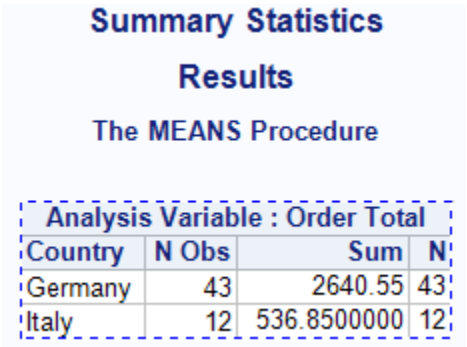
The default statistics are MEAN, STD, MIN, MAX, and N. To get the revenue totals per country, you need the sum to be computed.



7. Click Edit... . Select **Sum**. Clear all statistics except for **Sum** and **Number of observations**.



8. Click OK . Then click Finish .



Summary Statistics

Results

The MEANS Procedure

Analysis Variable : Order Total			
Country	N Obs	Sum	N
Germany	43	2640.55	43
Italy	12	536.8500000	12

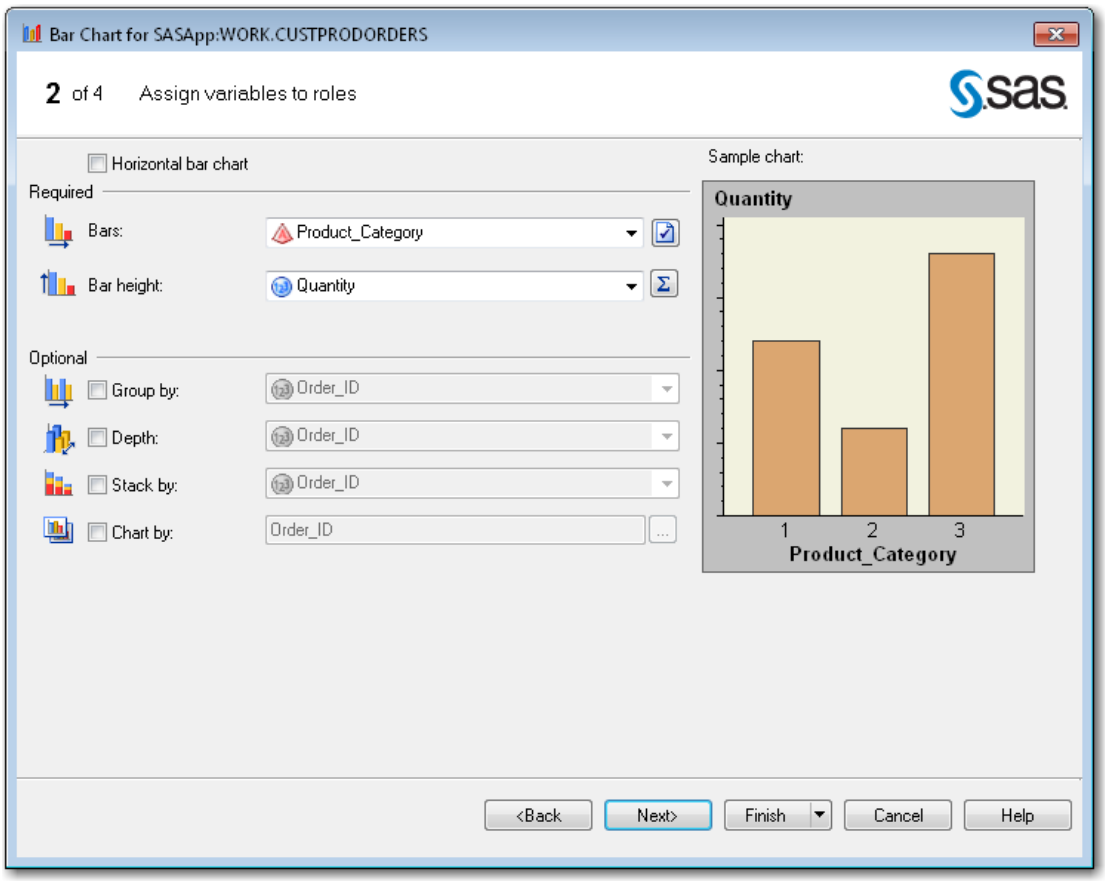
Two countries are represented in the data: Italy and Germany. Germany accounts for greater revenue (2640.55) than Italy (536.85).

Graphical Exploration

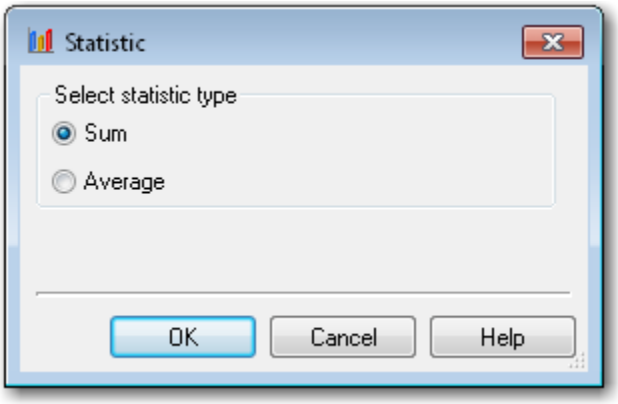
Create graphs of the categories of sales summarized by quantity and by revenue. Which categories sell the most products? Which categories bring in the most total revenue?

1. Select the Input Data tab – revealing the dataset once again
2. Select **Graph** ⇨ **Bar Chart Wizard**.
3. Click 

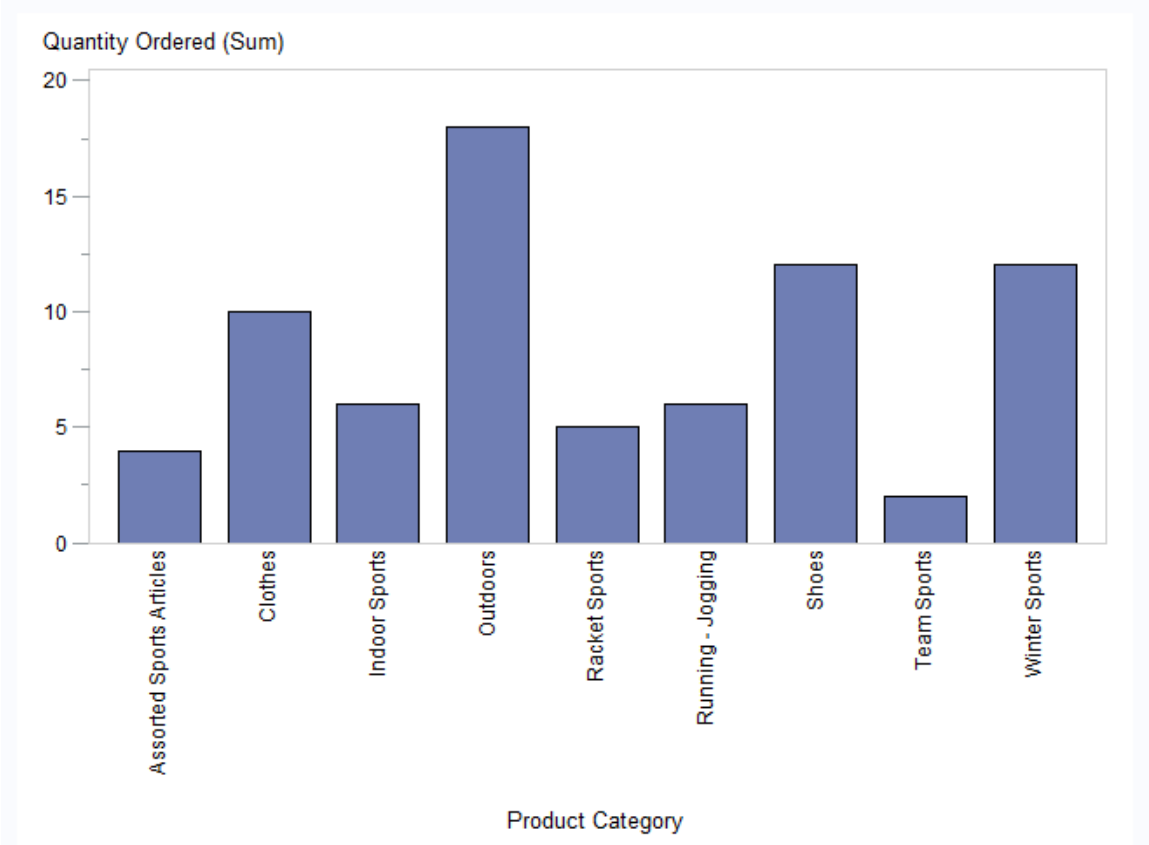
.
4. Assign **Product_Category** to the Bars role. Assign **Quantity** to the Bar height role.



5. Verify that the Sum statistic is being used by clicking the Σ button to the right of the Bar height role.



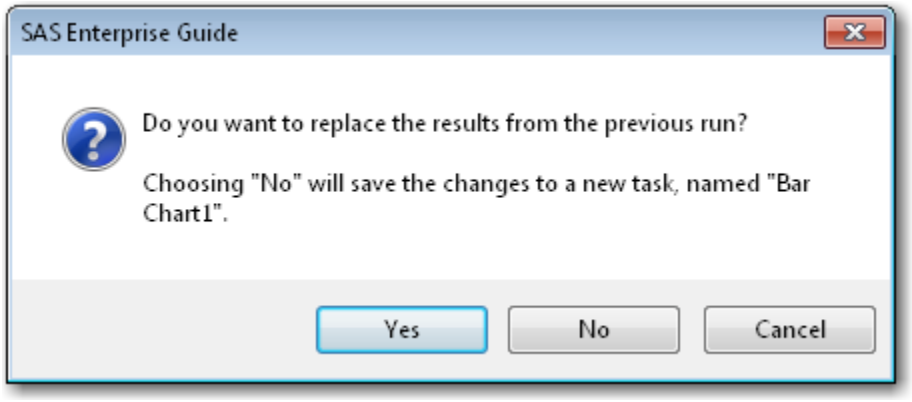
6. Click OK.
7. Click Finish.



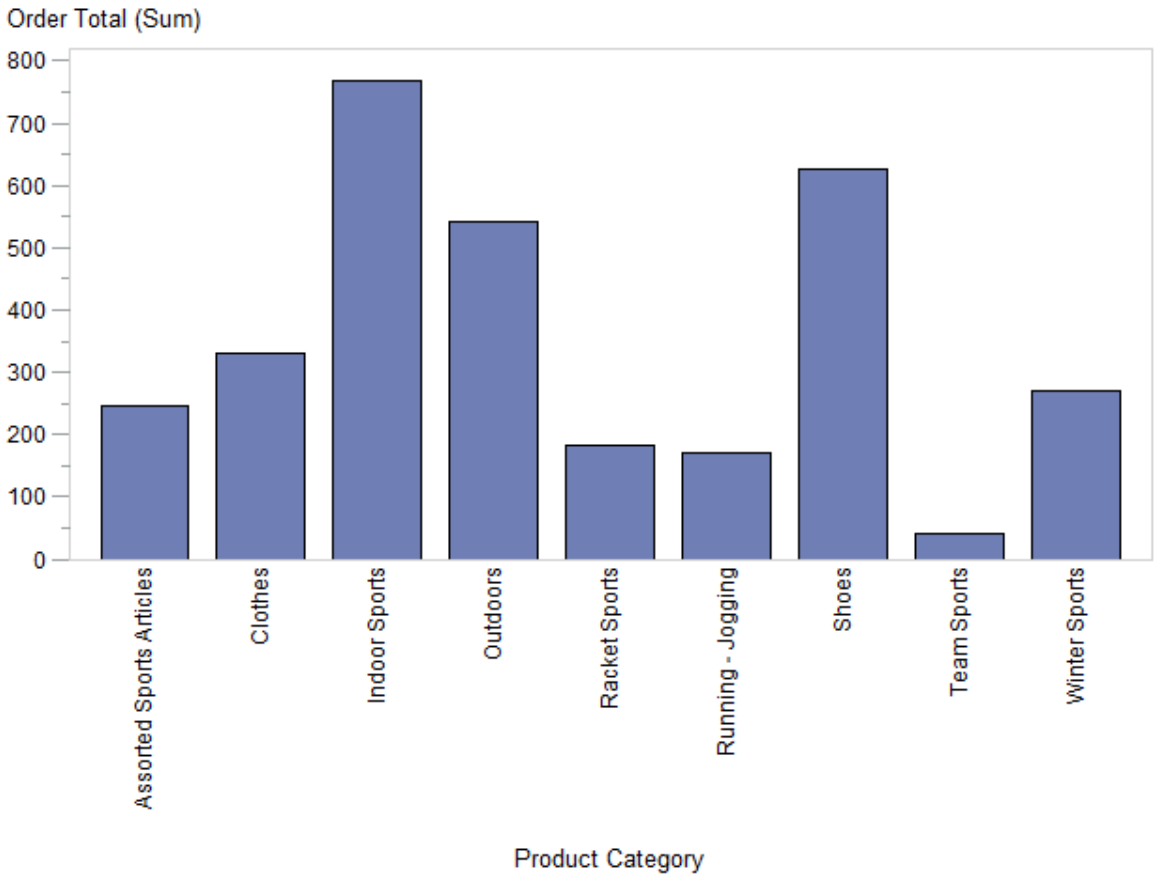
Outdoors accounts for the highest volume of sales. **Team Sports** is the lowest volume category.

8. Create a plot of the total order revenue for each category. Select **Modify Task**.

9. Change the bar height to **Total Order Revenue**. Click Finish.



10. Click No so that you create a new graph.



While **Outdoors** was the highest volume category, it is the third highest revenue category, after **Indoor Sports** and **Shoes**. **Team Sports** is still the lowest revenue category.