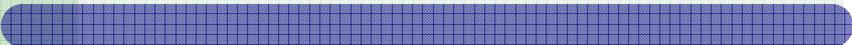


# **Exploratory Data Analysis**

**Brawijaya Professional Statistical Analysis  
BPSA MALANG  
Jl. Kertoasri 66 Malang  
(0341) 580342**



## Exploratory Data Analysis

Exploring data can help to determine whether the statistical techniques that you are considering for data analysis are appropriate. The Explore procedure provides a variety of visual and numerical summaries of the data, either for all cases or separately for groups of cases. The dependent variable must be a **scale** variable, while the grouping variables may be **ordinal** or **nominal**.

A variable can be treated as scale when its values represent ordered categories with a meaningful metric, so that distance comparisons between values are appropriate. Examples of scale variable include age in years and income in thousands of dollars.

A variable can be treated as ordinal when its values represent categories with some intrinsic ranking; for example, levels of service satisfaction from highly dissatisfied to highly satisfied. Examples of ordinal variables include attitude scores representing degree of satisfaction or confidence and preference rating scores.

A variable can be treated as nominal when its values represent categories with no intrinsic ranking; for example, the department of the company in which an employee works. Examples of nominal variables include region, zip code, or religious affiliation.

With the Explore procedure, you can:

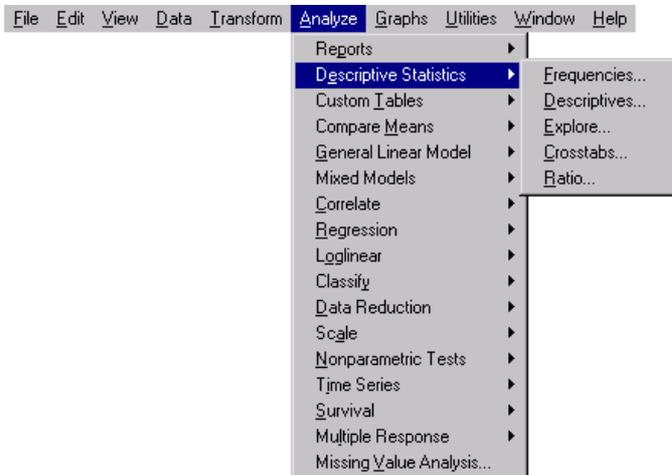
- Screen data
- Identify **outliers**
- Check assumptions
- Characterize differences among groups of cases

### *Descriptive Statistics across Groups*

Corn crops must be tested for aflatoxin, a poison whose concentration varies widely between and within crop yields. A grain processor has received eight crop yields, but the distribution of aflatoxin in parts per billion (PPB) must be assessed before they can be accepted.

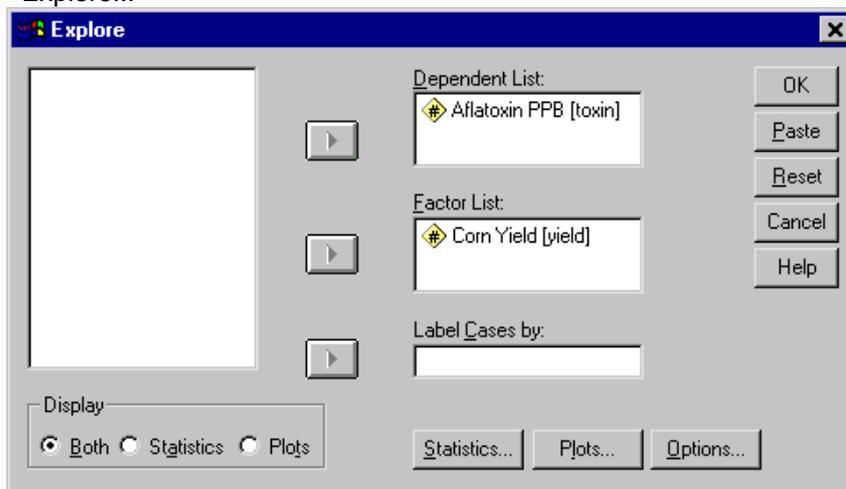
This example uses the file [aflatoxin.sav](#) . The data consist of 16 samples from each of the 8 crop yields.

## Running the Analysis



To begin the analysis, from the menus choose:

Analyze  
Descriptive Statistics  
Explore...



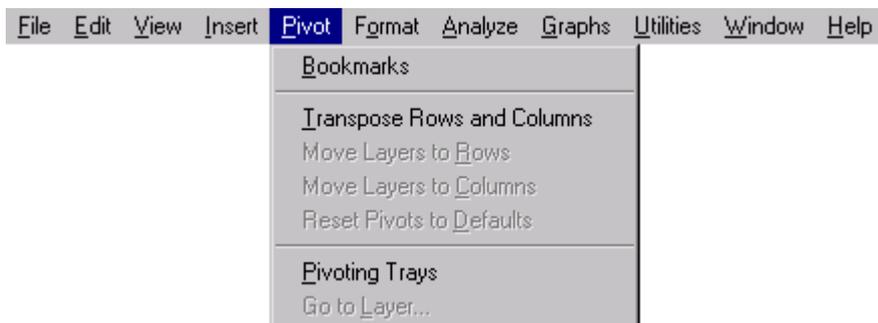
- ▶ Select *Aflatoxin PPB* as the dependent variable.
- ▶ Select *Corn Yield* as the factor variable.
- ▶ Click **OK**.

## Pivoting the Descriptives Table

Descriptives			
Corn Yield			Statistic
Aflatoxin PPB	1	Mean	20.2500
		95% Confidence Interval for Mean	Lower Bound 17.9519 Upper Bound 22.5481
		5% Trimmed Mean	20.4444
		Median	21.5000
		Variance	18.6000
		Std. Deviation	4.31277

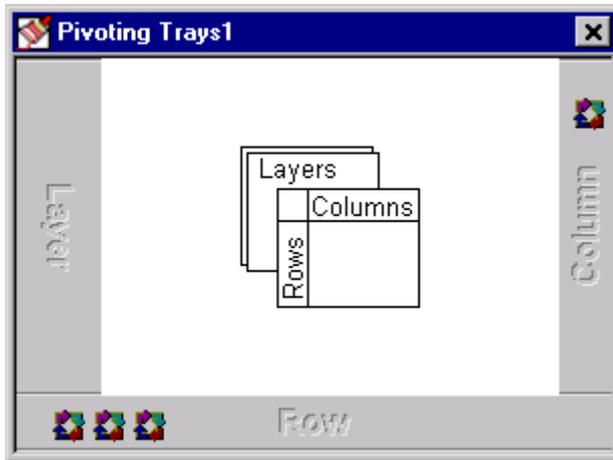
To assess how the mean of *Aflatoxin PPB* varies by *Corn Yield*, you can pivot the Descriptives table to display the statistics that you want.

- ▶ In the output window, double-click the Descriptives table to activate it.

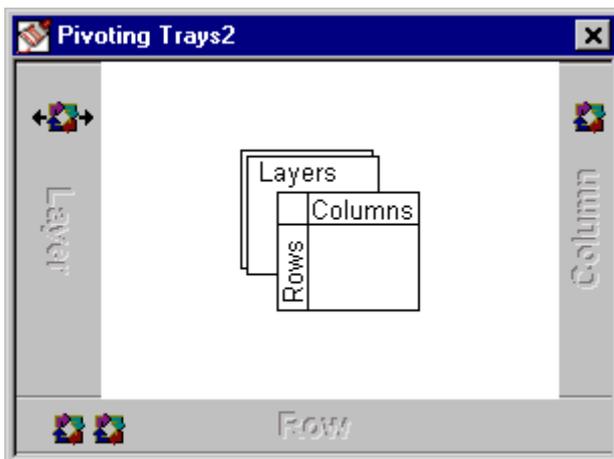


- ▶ From the Viewer menus choose:

Pivot  
Pivoting Trays...



- ▶ Select the Statistics icon (the third element) in the Row tray.

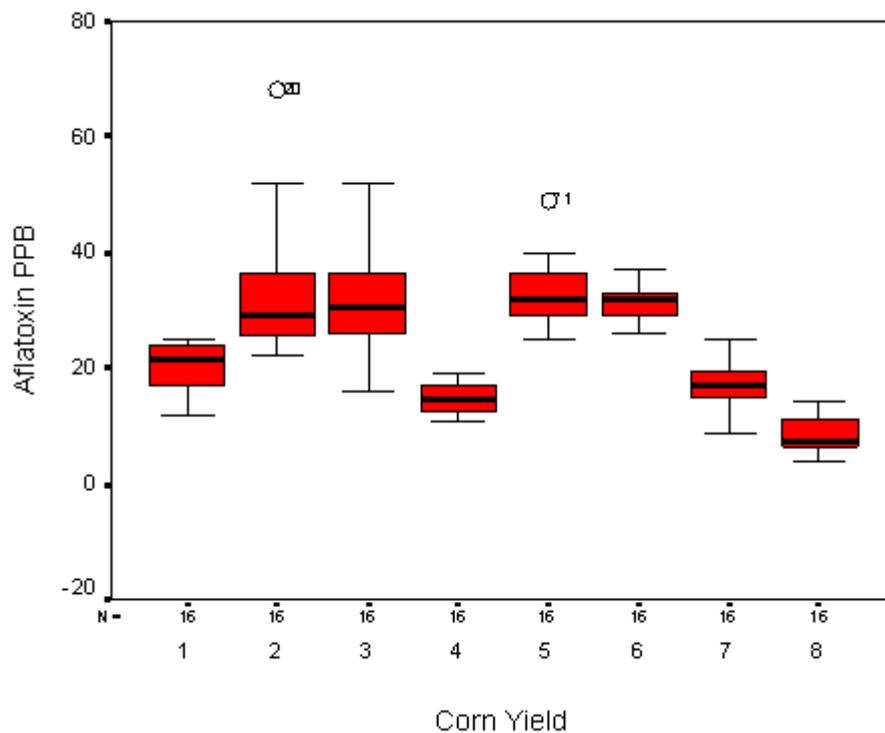


- ▶ Drag the Statistics icon to the Layer tray.
- ▶ Close the Pivoting Trays window.
- ▶ Deactivate the table by clicking outside of its boundaries.

Mean		Statistic	Std. Error
Aflatoxin PPB	Corn Yield 1	20.2500	1.07819
	2	33.0625	3.04339
	3	32.6875	2.57669
	4	14.6875	.66281
	5	33.0000	1.55724
	6	31.3750	.71224
	7	17.0625	1.04670
	8	8.4375	.76903

According to U.S. law, a yield is unfit for human consumption if aflatoxin exceeds 20 PPB. The pivoted table makes it clear that in these data, only yields 4, 7, and 8 fall below the 20 PPB cutoff.

### *Using Boxplots to Compare Groups*



**Boxplots** allow you to compare each group using a five-number summary: the median, the 25th and 75th percentiles, and the minimum and maximum observed values that are not statistically outlying. Outliers and extreme values are given special attention.

- The heavy black line inside each box marks the 50th percentile, or **median**, of that distribution. For example, the median aflatoxin level of yield 1 is 21.50 PPB. Notice that the medians vary quite a bit across the boxplots.

- The lower and upper **hinges**, or box boundaries, mark the 25th and 75th percentiles of each distribution, respectively. For yield 2, the lower hinge value is 24.75, and the upper hinge value is 36.75.
- **Whiskers** appear above and below the hinges. Whiskers are vertical lines ending in horizontal lines at the largest and smallest observed values that are not statistical outliers. For yield 2, the smallest value is 22, and the largest value that is not an outlier is 52.

**Outliers** are identified with an **O**. Yield 2 has an outlying value of 68, labeled **2O**, and Yield 5 has an outlier of 49, labeled **71**. The label refers to the row number in the Data Editor where that observation is found. **Extreme** values are marked with an asterisk (\*). There are no extreme values in these data.

## *Summary*

Using the Explore procedure, you created a summary table that showed the mean alfatoxin level to be unsafe for five of eight corn yields. You also created a boxplot for visual confirmation of these results.

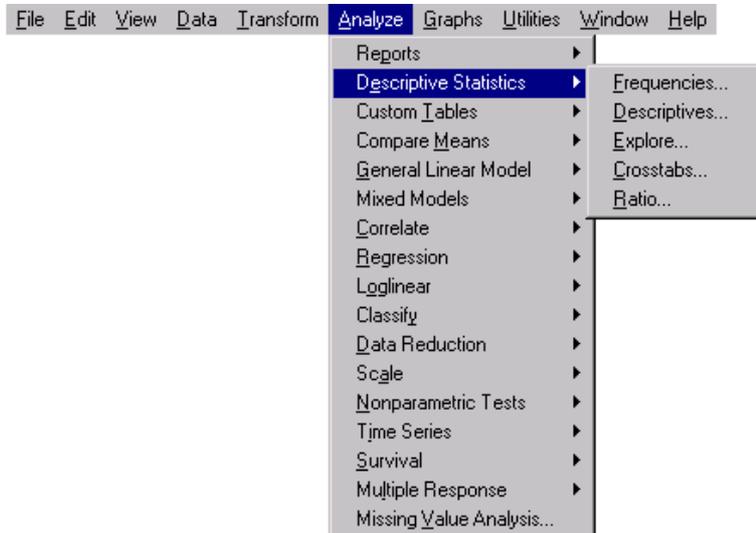
Boxplots provide a quick, visual summary of any number of groups. Further, all the groups within a single factor are arrayed on the same axes, making comparisons easier. While boxplots provide some evidence about shape of the distributions, the Explore procedure offers many options that allow a more detailed look at how groups may differ from each other or from expectation.

## *Exploring Distributions*

A manufacturing firm uses silver nitride to create ceramic bearings, which must resist temperatures of 1500 degrees centigrade or higher. The heat resistance of a standard alloy is known to be normally distributed. However, a new premium alloy is under test, and its distribution is unknown.

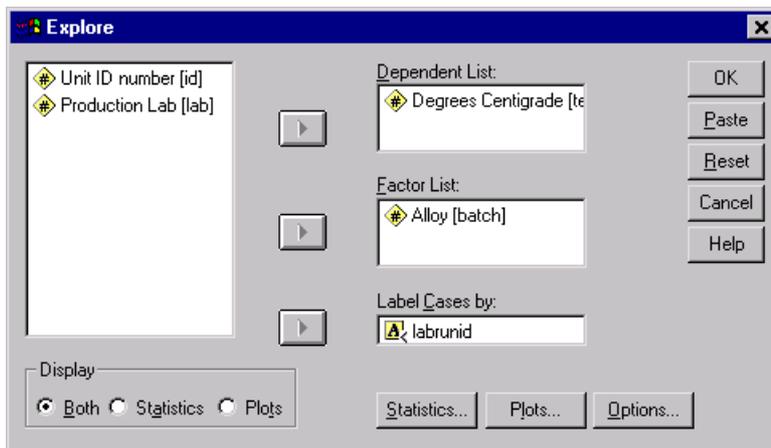
This example uses the file [ceramics.sav](#) .

## Running the Analysis

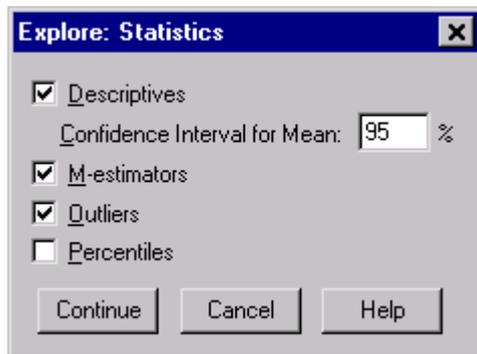


To begin the analysis, from the menus choose:

Analyze  
Descriptive Statistics  
Explore...

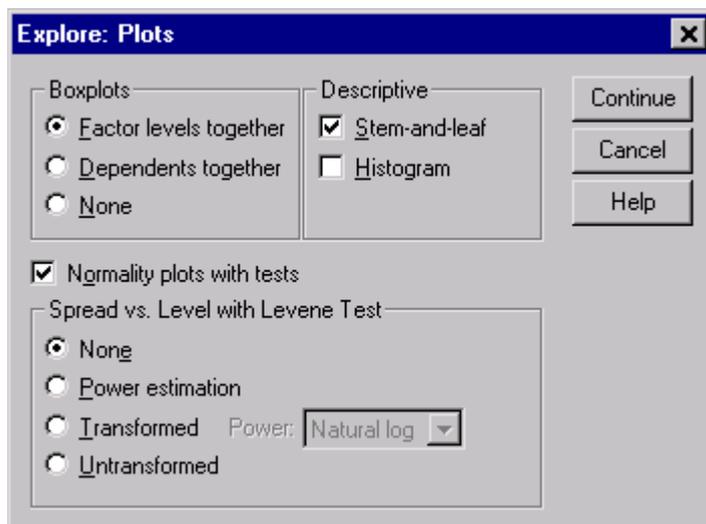


- ▶ Select *Degrees Centigrade* as the dependent variable.
- ▶ Select *Alloy* as the factor variable.
- ▶ Label cases by *labrunid*.
- ▶ Click **Statistics**.



Because the heat-resistant properties of premium bearings are unknown, robust estimates of central tendency and a table of outliers should be requested.

- ▶ Select **M-estimators** and **Outliers**.
- ▶ Click **Continue**.
- ▶ Click **Plots** in the Explore dialog box.



You should also request tests of normality for these data. These tests will be calculated individually for each alloy.

- ▶ Select **Normality plots with tests**.
- ▶ Click **Continue**.
- ▶ Click **OK** in the Explore dialog box.

## Numerical Descriptions of Shape

Alloy: Standard

			Statistic	Std. Error
Degrees	Mean		1514.6564	.62004
Centigrade	95% Confidence Interval for Mean	Lower Bound	1513.4350	
		Upper Bound	1515.8779	
	5% Trimmed Mean		1514.7302	
	Median		1514.5317	
	Variance		92.269	
	Std. Deviation		9.60566	
	Minimum		1488.30	
	Maximum		1537.99	
	Range		49.69	
	Interquartile Range		13.5098	
	Skewness		-.078	.157
	Kurtosis		-.343	.313

The Descriptives table is pivoted so that *Alloy* is in the **layers** of the pivot table, with standard bearings displayed first. The **mean**, **trimmed mean**, and **median** are nearly equal, and the **skewness** and **kurtosis** statistics are close to 0. This is strong evidence that heat resistance in standard bearings is **normally distributed**.

Alloy: Premium

			Statistic	Std. Error
Degrees	Mean		1542.0787	.61165
Centigrade	95% Confidence Interval for Mean	Lower Bound	1540.8738	
		Upper Bound	1543.2836	
	5% Trimmed Mean		1541.2805	
	Median		1539.7181	
	Variance		89.789	
	Std. Deviation		9.47569	
	Minimum		1530.44	
	Maximum		1591.04	
	Range		60.61	
	Interquartile Range		11.5051	
	Skewness		1.439	.157
	Kurtosis		3.036	.313

The premium bearings tell a different story. The mean is higher than either the trimmed mean or the median; outliers or extreme values are pulling it upward. The skewness and kurtosis statistics also provide evidence of disproportionate values at the upper tail of the distribution.

### *Robustness and Influential Values*

Alloy		Huber's M-Estimator <sup>a</sup>	Tukey's Biweight <sup>b</sup>	Hampel's M-Estimator <sup>c</sup>	Andrews' Wave <sup>d</sup>
Degrees	Premium	1540.0953	1539.5658	1540.2052	1539.5506
Centigrade	Standard	1514.6413	1514.6925	1514.6828	1514.6955

- a. The weighting constant is 1.339.
- b. The weighting constant is 4.685.
- c. The weighting constants are 1.700, 3.400, and 8.500
- d. The weighting constant is  $1.340 \cdot \pi$ .

In this case, the robust estimates for premium bearings are quite close to the median (1539.72). Since none of these measures are near the mean, this may be an indication that the distribution is not reasonably normal.

Alloy			Case Number	LABRUNID	Value	
Degrees Centigrade	Premium	Highest	1	211	d421	1591.04
			2	417	g837	1574.62
			3	17	a 17	1571.77
			4	437	h917	1568.10
			5	357	f657	1567.07
	Lowest		1	139	c289	1530.44
			2	475	h955	1530.73
			3	199	d379	1530.75
			4	373	g733	1530.76
			5	207	d387	1530.79
Standard	Highest		1	408	g828	1537.99
			2	198	d378	1534.29
			3	20	a 20	1534.06
			4	168	c318	1533.43
			5	184	d364	1533.35
	Lowest		1	396	g816	1488.30
			2	100	b190	1488.36
			3	80	b170	1494.09
			4	154	c304	1494.64
			5	240	d450	1495.15

The table of extreme values lists the five highest and lowest values for each *Alloy*. The premium bearings range from five standard deviations above to one standard deviation below the mean. At times, these can withstand much higher heat than standard bearings and never fail below 1530 degrees centigrade.

## Are the Distributions Normal?

Alloy		Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Degrees	Premium	.123	240	.000	.888	240	.000
Centigrade	Standard	.027	240	.200*	.995	240	.602

\*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

The tests of normality overlay a normal curve on actual data, to assess the fit. A significant test means the fit is poor. For the standard alloy, the test is not significant; they fit the normal curve well. However, for the premium alloy, the test is significant; they fit the normal curve poorly.

Degrees Centigrade Stem-and-Leaf Plot for  
 BATCH= Premium

```

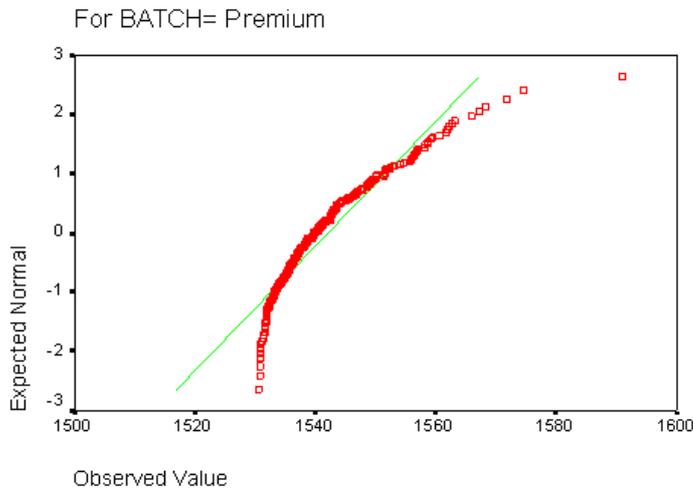
Frequency      Stem & Leaf

 24.00      153 . 00000001111111111111111111
 22.00      153 . 222222222233333333333333
 26.00      153 . 44444444445555555555555555
 26.00      153 . 6666666666666666777777777777
 24.00      153 . 8888888888888888999999999999
 19.00      154 . 00000000001111111111
 25.00      154 . 2222222222222222333333333333
 10.00      154 . 4444455555
 12.00      154 . 666666667777
 10.00      154 . 8888999999
  8.00      155 . 00111111
  4.00      155 . 2223
  6.00      155 . 445555
  6.00      155 . 666667
  6.00      155 . 888899
  3.00      156 . 011
  3.00      156 . 223
  6.00 Extremes      (>=1566)
  
```

```

Stem width:      10.00
Each leaf:       1 case(s)
  
```

**Stem-and-leaf plots** use the original data values to display the distribution's shape. The plot for premium bearings visualizes the positive skew statistic seen in the descriptives table; the values cluster uniformly in a range of 1530 to 1543 degrees, then disperse gradually at the higher temperatures.



Finally, a **Q-Q** plot is displayed. The straight line in the plot represents expected values when the data are normally distributed. The observed premium bearing values deviate markedly from that line, especially as temperature increases.

### ***Summary***

Using the Explore procedure, you found the premium alloy to have a different distribution from the standard alloy. On a positive note, the mean heat resistance for the new alloy is considerably higher than that for the standard alloy. Unfortunately, there is evidence that the mean may not be a good measure of central tendency for the premium alloy. However, robust estimates of central tendency reconfirm the superiority of the premium alloy.

### ***Related Procedures***

The Explore procedure is a very useful procedure for visually and numerically comparing groups, summarizing distributions, examining the assumption of normality, and looking for outlying observations. It is easy to assume without looking that your data have no outliers, extreme values, or distributional problems. Fortunately, the Explore procedure makes it just as easy to see how well the data validate those assumptions.

- If your dependent variable is categorical, try the [Crosstabs](#) procedure.
- Other procedures allow you to **layer** grouping variables so that you can examine summary statistics for cross-classifications of factors. See [The Summarize Procedure](#) for more information. See [The Means Procedure](#) for more information. See [The OLAP Cubes Procedure](#) for more information.
- You can alternately use the [One-Sample Kolmogorov-Smirnov Test](#) to test your dependent variable for normality. That procedure also allows you to check your dependent variable against the Poisson, Uniform, or Exponential distributions.

See the following text for more information on summarizing data (for complete bibliographic information, hover over the reference):

Hays, W. L. 1981. *Statistics*. New York: Holt, Rinehart, and Winston.