TESTING DATA FOR NORMALITY

OBJECTIVES

In this lecture, you will learn the following items:

- **How to find a data sample's kurtosis and skewness and determine if the sample meets acceptable levels of normality.**

- **How to use SPSS to find a data sample's kurtosis and skewness and determine if the sample meets acceptable levels of normality.**

INTRODUCTION

Parametric statistical tests, such as the t-test and one-way analysis of variance, are based on particular assumptions or parameters. The data samples meeting those parameters are randomly drawn from a normal population, based on independent observations, measured with an interval or ratio scale, possess an adequate sample size, and approximately resemble a normal distribution.

Moreover, comparisons of samples or variables should have approximately equal variances. If data samples violate one or more of these assumptions, you should consider using a nonparametric test.

Examining the data gathering method, scale type, and size of a sample are fairly straightforward. However, examining a data sample's resemblance to a normal distribution, or its normality, requires a more involved analysis. Visually inspecting a graphical representation of a sample, such as a stem and leaf plot or a box and whisker plot, might be the most simplistic examination of normality.

Statisticians advocate this technique in beginning statistics; however, this measure of normality does not suffice for strict levels of defensible analyses.

In this lecture, we present three quantitative measures of sample normality.
First, we discuss the properties of the normal distribution. Then, we describe how to examine a sample's kurtosis and skewness.

DESCRIBING DATA AND THE NORMAL DISTRIBUTION

We will attempt to summarize the concept and begin with a practical approach as it applies to data collection.

In research, we often identify some population we wish to study. Then, we strive to collect several independent, random measurements of a particular variable associated with our population. We call this set of measurements a *sample*. If we used good experimental technique and our sample adequately represents our population, we can study the sample to make inferences about our population.

For example, during a routine checkup, your physician draws a sample of your blood instead of all of your blood. This blood sample allows your physician to evaluate all of your blood even though he or she only tested the sample. Therefore, all of your body's blood cells represent the population about which your physician makes an inference using only the sample.

While a blood sample leads to the collection of a very large number of blood cells, other fields of study are limited to small sample sizes. It is not uncommon to collect less than 30 measurements for some studies in the behavioral and social sciences.

Moreover, the measurements lie on some scale over which the measurements vary about the mean value. This notion is called variance.

For example, a researcher uses some instrument to measure the intelligence of 25 children in a math class. It is highly unlikely that every child will have the same intelligence level. In fact, a good instrument for measuring intelligence should be sensitive enough to measure differences in the levels of the children.

The variance s^2 can be expressed quantitatively. It can be calculated using the next formula

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

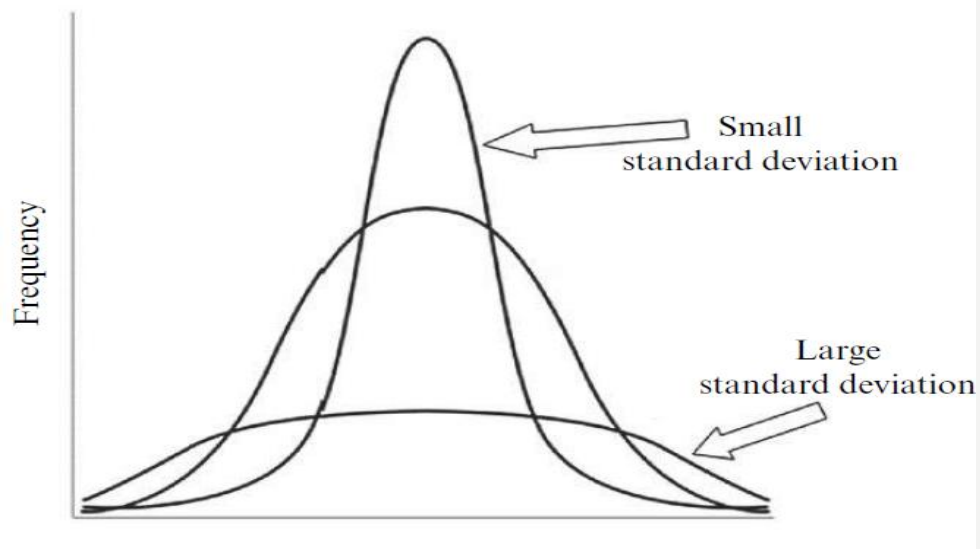
where xi is an individual value in the distribution, x is the distribution's mean, and n is the number of values in the distribution.

Parametric Tests assume that the variances of samples being compared are approximately the same. This idea is called homogeneity of variance. To compare sample variances, we obtain a variance ratio by taking the largest sample variance and dividing it by the smallest sample variance. The variance ratio should be less than 2. No sample's variance be twice as large as any other sample's variance. **If the homogeneity of variance assumption cannot be met, one would use a non-parametric test.**

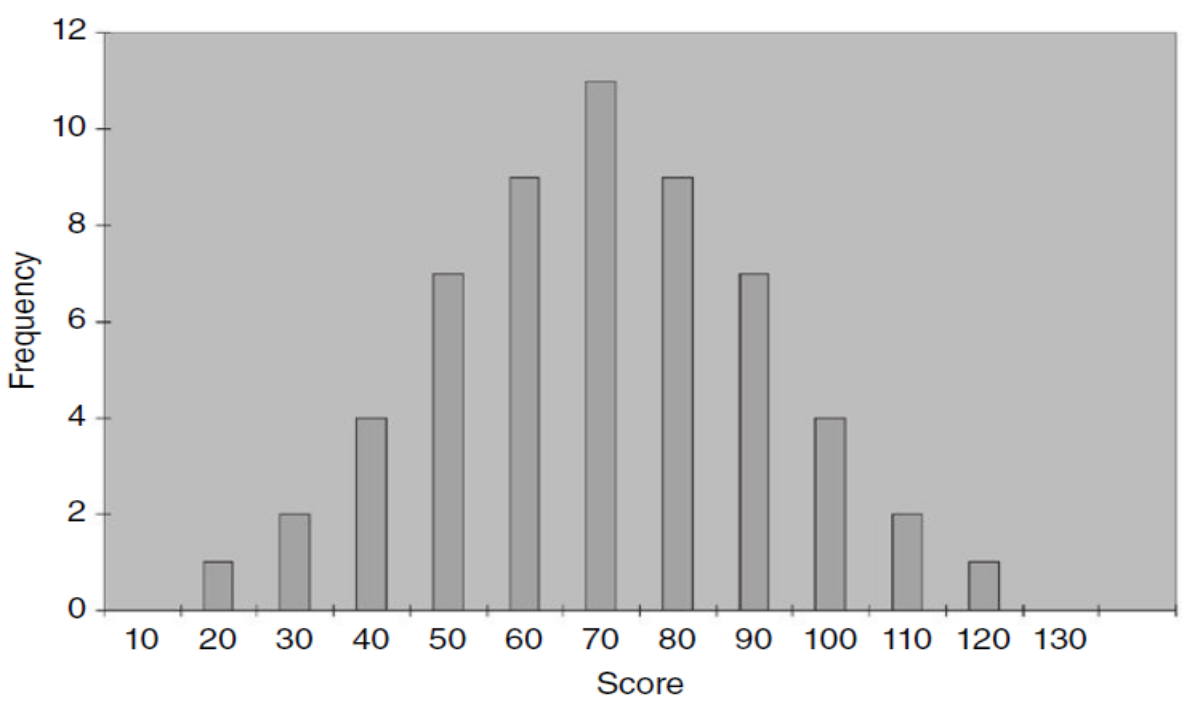
A more common way of expressing a sample's variability is with its standard deviation, s. Standard deviation is the square root of variance where s. In other words, standard deviation is calculated using formula:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

As illustrated in the next figure, a small standard deviation indicates that a sample's values are fairly concentrated about its mean, whereas a large standard deviation indicates that a sample's values are fairly spread out.



A histogram is a useful tool for graphically illustrating a sample's frequency distribution and variability. This graph plots the value of the measurements horizontally and the frequency of each particular value vertically. The middle value is called the median and the greatest frequency is called the mode.

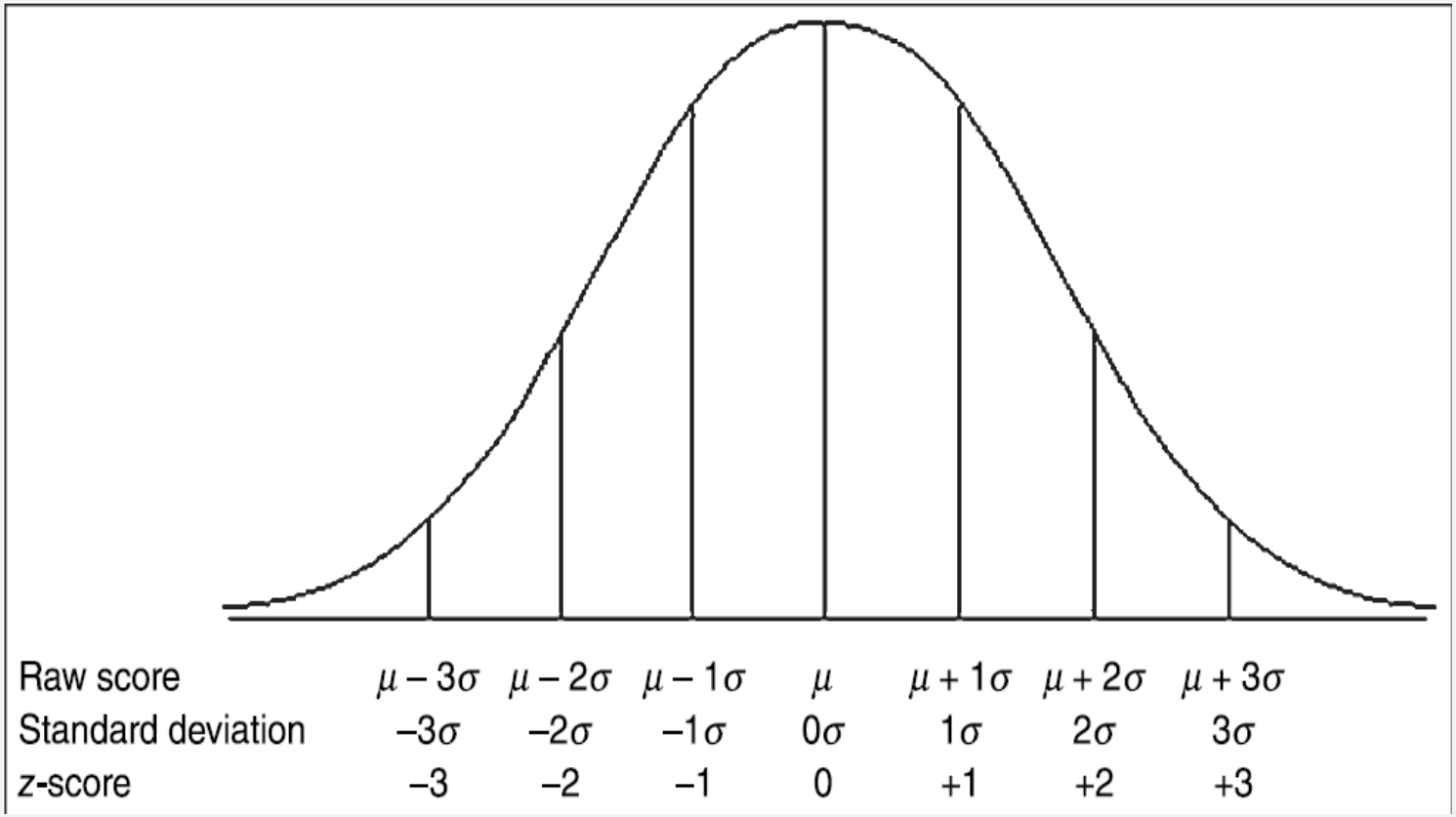


The mean and standard deviation of one distribution differ from the next. If we want to compare two or more samples, then we need some type of standard. A standard score is a way we can compare multiple distributions. The standard score that we use is called a z-score, and it can be calculated using the next formula:

$$z = \frac{x_i - \overline{x}}{s}$$

where x_i is an individual value in the distribution, x is the distribution's mean, and s is the distribution's standard deviation.

The following figure shows the relationship among the raw values, standard deviation, and z-scores of a population. Since we are describing a population, we use sigma, , to represent standard deviation and mu, μ, to represent the mean.



COMPUTING AND TESTING KURTOSIS AND SKEWNESS FOR SAMPLE NORMALITY

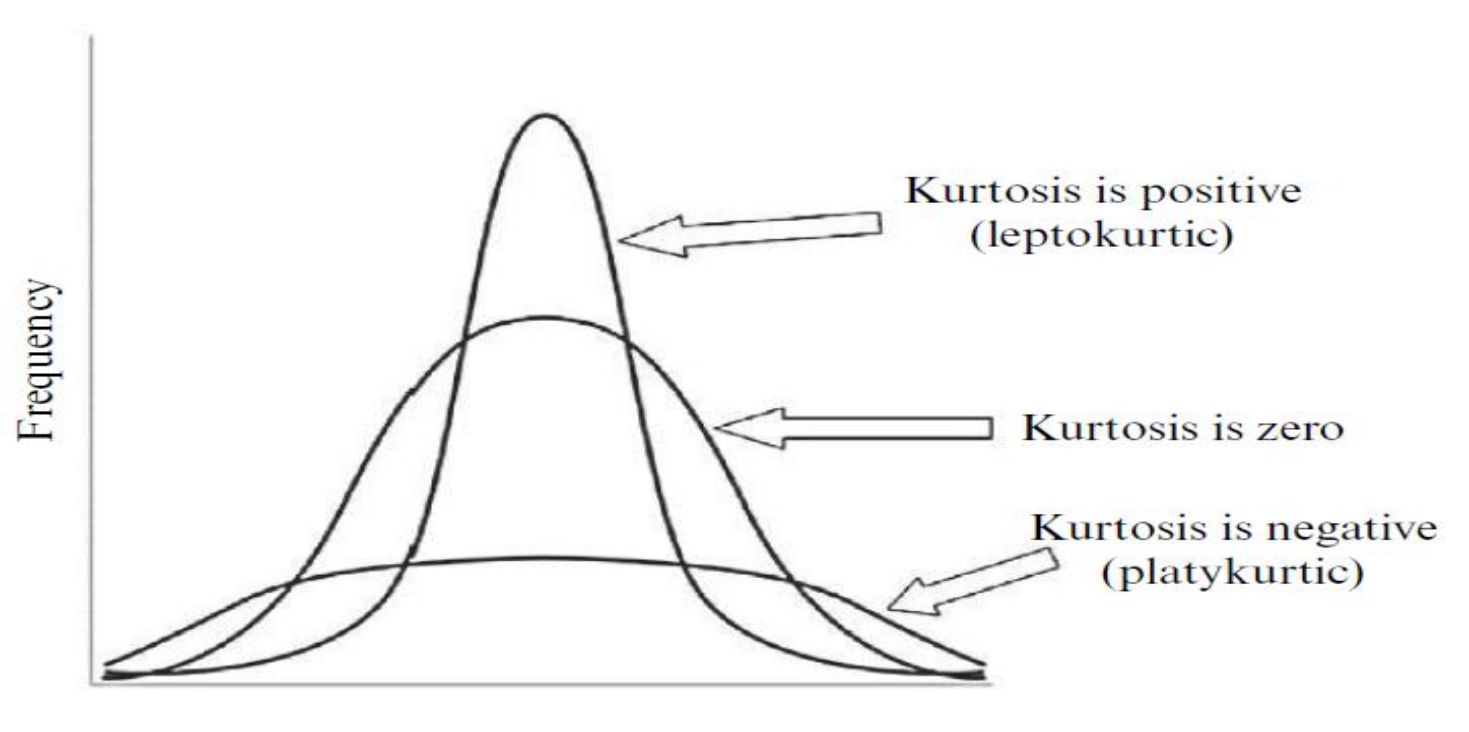
A frequency distribution that resembles a normal curve is approximately normal. However, not all frequency distributions have the approximate shape of a normal curve. The values might be densely concentrated in the center or substantially spread out. The shape of the curve may lack symmetry with many values concentrated on one side of the distribution. We use the terms kurtosis and skewness to describe these conditions, respectively.

Kurtosis is a measure of a sample or population that identifies how flat or peaked it is with respect to a normal distribution. Stated another way, kurtosis refers to how concentrated the values are in the center of the distribution.

As shown in the following figure, a peaked distribution is said to be leptokurtic. A leptokurtic distribution has a positive kurtosis.

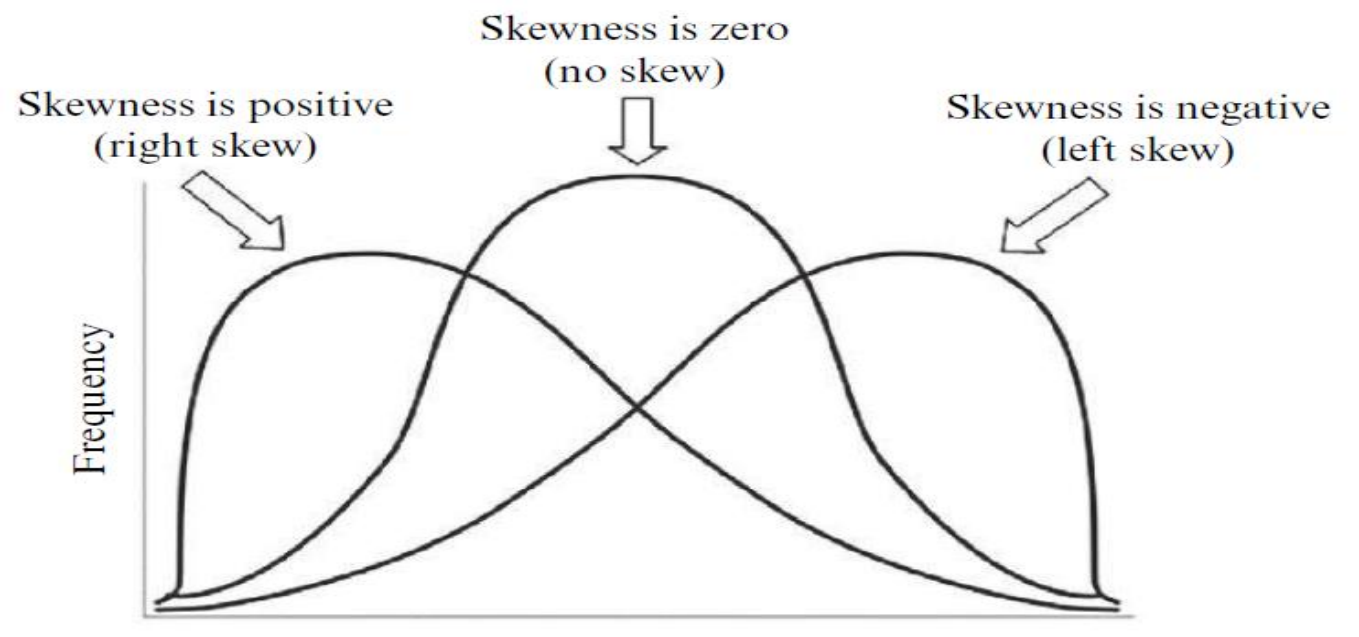
If a distribution is flat, it is said to be platykurtic.

A platykurtic distribution has a negative kurtosis.



The skewness of a sample can be described as a measure of horizontal symmetry with respect to a normal distribution. As shown in the following figure, if a distribution's scores are concentrated on the right side of the curve, it is said to be left skewed.

A left skewed distribution has a negative skewness. If a distribution's scores are concentrated on the left side of the curve, it is said to be right skewed. A right skewed distribution has a positive skewness.



The kurtosis and skewness can be used to determine if a sample approximately resembles a normal distribution. There are five steps for examining sample normality in terms of kurtosis and skewness.

1. Determine the sample's mean and standard deviation.

2. Determine the sample's kurtosis and skewness.

3. Calculate the standard error of the kurtosis and the standard error of the skewness.

4. Calculate the z-score for the kurtosis and the z-score for the skewness.

5. Compare the z-scores with the critical region obtained from the normal distribution.

The kurtosis K is found using the following formula:

$$K = \left[\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^4 \right] - \frac{3(n-1)^2}{(n-2)(n-3)}$$

The standard error of the kurtosis is found using the following formula:

$$SE_K = \sqrt{\frac{24n(n-1)^2}{(n-2)(n-3)(n+5)(n+3)}}$$

The skewness S_k is found using the following formula:

$$S_k = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

The standard error of the skewness, SE_{Sk}, is found using the following formula:

$$SE_{S_k} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}$$

Normality can be evaluated using the z-score for the kurtosis, z_K, and the z-score for the skewness, z_{Sk}. Use the following two formula to find those z-scores:

$$z_K = \frac{K - 0}{SE_K}$$

$$z_{S_k} = \frac{S_k - 0}{SE_{S_k}}$$

Compare these z-scores with the values of the standard normal distribution for a desired level of confidence .

For example: If you set alpha = 0.05, then the calculated z-scores for an approximately normal distribution must fall between −1.96 and +1.96.

Sample Problem for Examining Kurtosis

The scores in the next table represent students' quiz performance during the first week of class. Use alpha= 0.05 for your desired level of confidence. Determine if the samples of week 1 quiz scores are approximately normal in terms of its kurtosis.

Week 1 quiz scores		
90	72	90
64	95	89
74	88	100
77	57	35
100	64	95
65	80	84
90	100	76

First, find the mean and standard deviation of the sample:

$$\overline{x} = 80.24 \qquad s = 16.62$$

Compute the kurtosis and standard error of the kurtosis:

$$K = 1.153 \qquad SE_K = 0.972$$

Finally, use the kurtosis and the standard error of the kurtosis to find a z-score:

$$z_K = \frac{K - 0}{SE_K} = \frac{1.153 - 0}{0.972}$$
$$z_K = 1.186$$

Use the z-score to examine the sample's approximation to a normal distribution. This value must fall between −1.96 and +1.96 to pass the normality assumption for = 0.05. Since this z-score value does fall within that range, the sample has passed our normality assumption for kurtosis.

Sample Problem for Examining Skewness

Based on the same values from the example listed earlier, determine if the samples of week 1 quiz scores are approximately normal in terms of its skewness. Use the mean and standard deviation from the previous example to find the skewness.

The skewness and the standard error of the skewness are:

$$S_k = -1.018 \qquad SE_{S_k} = 0.501$$

Finally, use the skewness and the standard error of the skewness to Find a z-score:

$$z_{S_k} = -2.032$$

Use the z-score to examine the sample's approximation to a normal distribution. This value must fall between −1.96 and +1.96 to pass the normality assumption for alpha = 0.05.

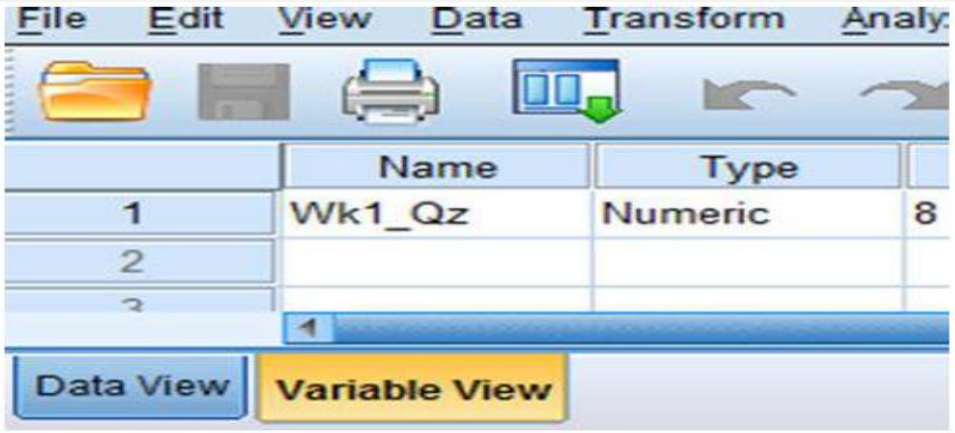
Since this z-score value does not fall within that range, the sample has failed our normality assumption for skewness.

Therefore, either the sample must be modified and rechecked or you must use a **Non-parametric statistical test**.

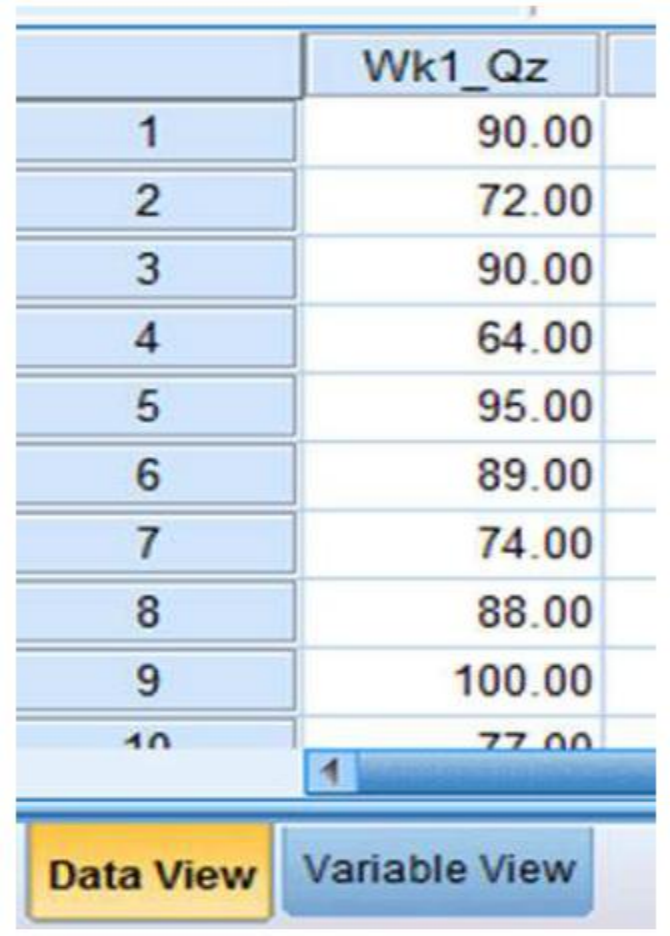
Examining Skewness and Kurtosis for Normality Using SPSS

We will analyze the examples earlier using SPSS.

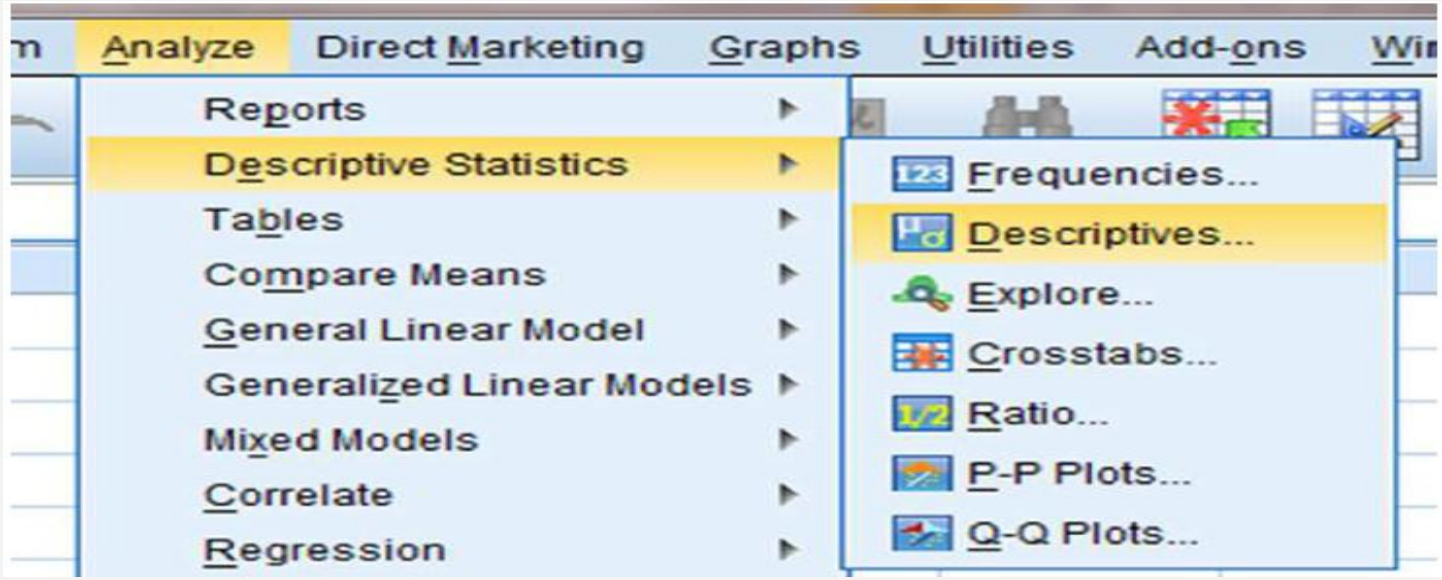
Define Your Variables First, click the "Variable View" tab at the bottom of your screen. Then, type the name of your variable(s) in the "Name" column. As shown in the following figur, we have named our variable "Wk1_Qz."



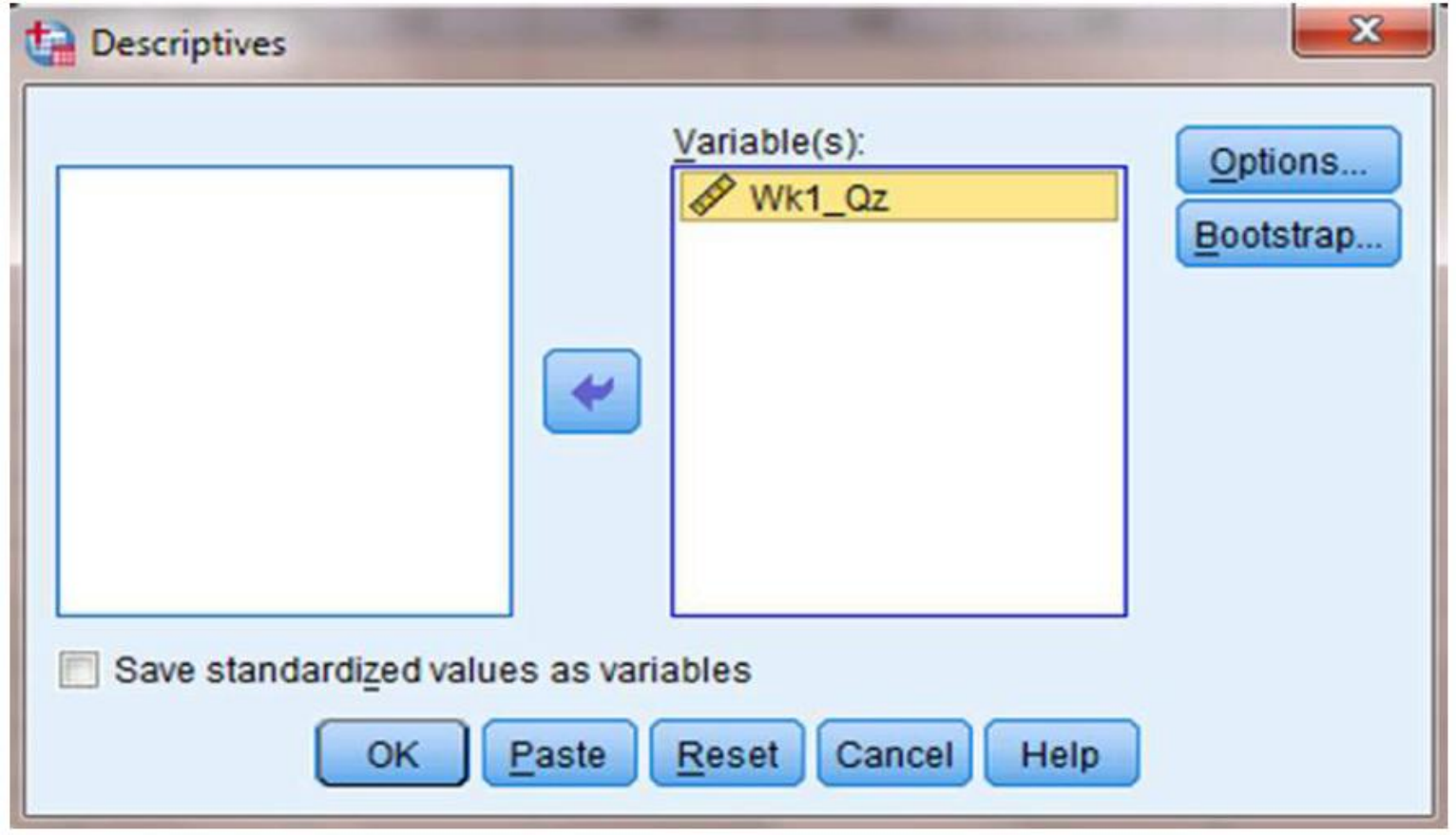
Type in Your Values Click the "Data View" tab at the bottom of your screen and type your data under the variable names. As shown in the following figure, we have typed the values for the "Wk1_Qz" sample.



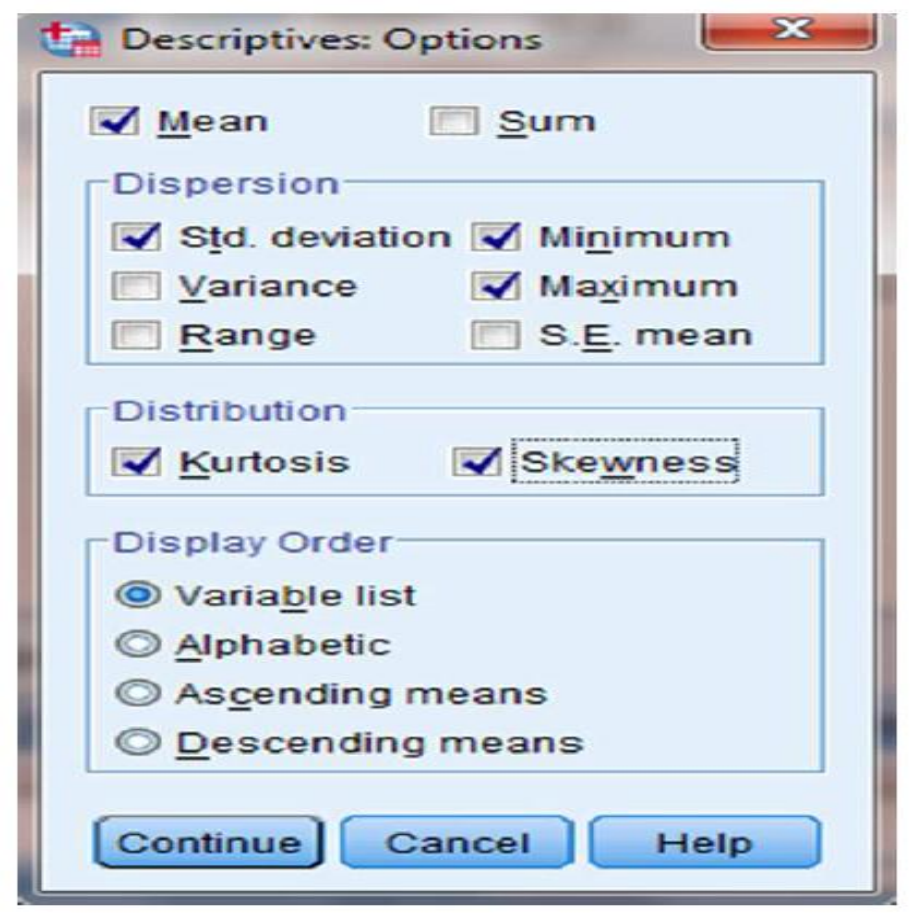
Analyze Your Data As shown in the following figure, use the pull-down menus to choose “Analyze,” “Descriptive Statistics,” and “Descriptives . . .”



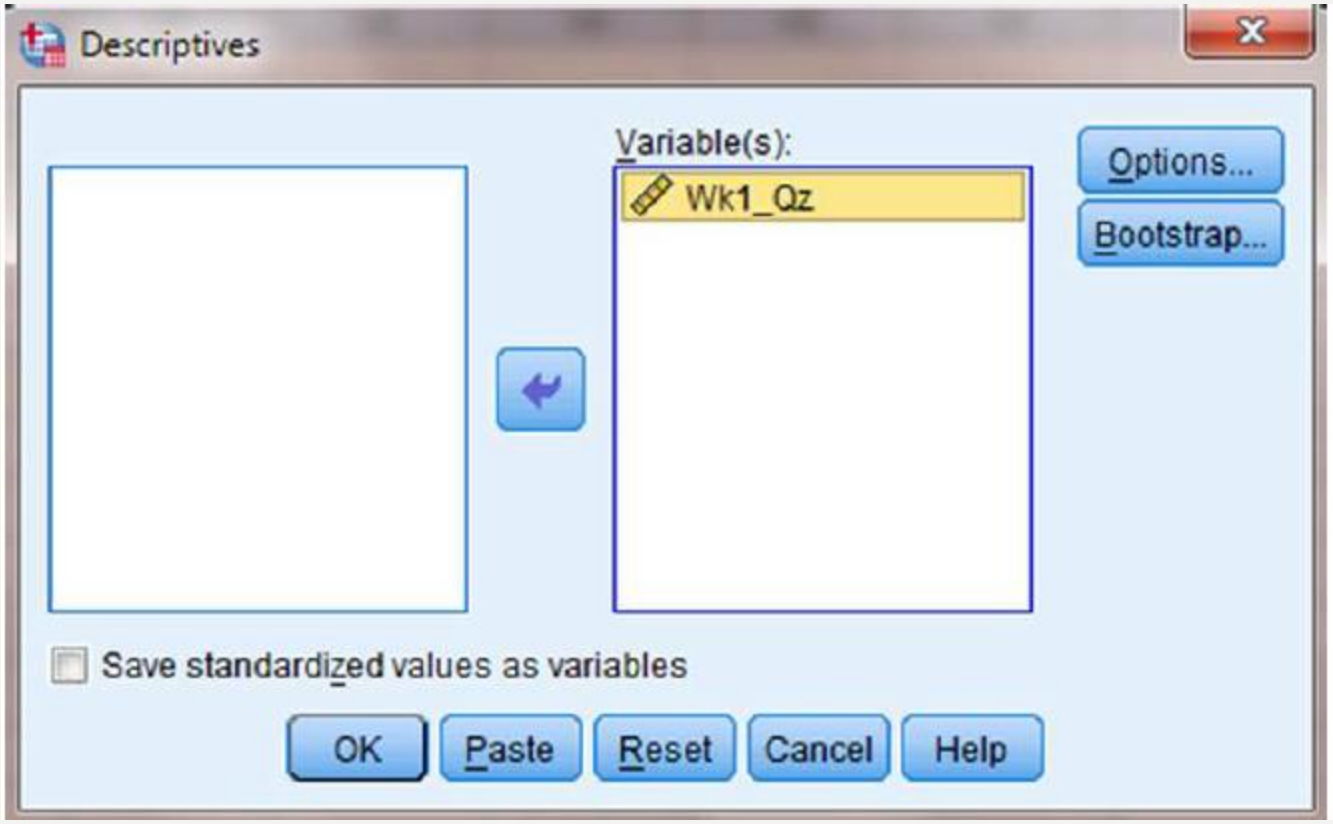
Choose the variable(s) that you want to examine. Then, click the button in the middle to move the variable to the "Variable(s)" box, as shown in the following figure:



Next, click the "Options . . ." button to open the "Descriptives: Options" window shown in Figure 2.11. In the "Distribution" section, check the boxes next to "Kurtosis" and "Skewness." Then, click "Continue."



Finally, once you have returned to the "Descriptives" window, as shown in the following figure, click "OK" to perform the analysis.



Interpret the Results from the SPSS Output Window

The output provides the kurtosis and the skewness, along with their associated standard errors. In our example, the skewness is −1.018 and its standard error is 0.501. The kurtosis is 1.153 and its standard error is 0.972.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Wk1_Qz	21	35.00	100.00	80.2381	16.62199	-1.018	.501	1.153	.972
Valid N (listwise)	21								

At this stage, we need to manually compute the z-scores for the skewness and kurtosis as we did in the previous examples. First, compute the z-score for kurtosis and skewness:

$$z_K = \frac{K - 0}{SE_K} = \frac{1.153 - 0}{0.972}$$

$$z_K = 1.186$$

$$z_{S_k} = \frac{S_k - 0}{SE_{S_k}} = \frac{-1.018}{0.501}$$

$$z_{S_k} = -2.032$$

Both of these values must fall between −1.96 and +1.96 to pass the normality assumption for = 0.05.

The z-score for kurtosis falls within the desired range, but the z-score for skewness does not. Using alpha = 0.05, the sample has passed the normality assumption for kurtosis, yet failed the normality assumption for skewness.

Therefore, either the sample must be modified and rechecked or you must use a **Non-parametric statistical test.**