chapter 2-1

Descriptive Statistics

Introduction

Any text discussing measurement and evaluation will have chapters like this with Statistics in the title. The word statistics strikes fear in the hearts of some and total cynicism in the minds of others. Fear because it is a topic filled with what they perceive to be hard to understand concepts; cynicism because statistics can apparently be made to say anything, so they say. There should be no reason to be fearful of the information presented here; the aim of this text is to provide an understanding of why a statistic is applied and a simple explanation of its interpretation. The cynical belief that statistics can be distorted to say whatever the presenter wants, is the best argument why a student should learn about statistical procedures. It is only when you have an understanding of various statistical procedures that you have the knowledge necessary to guard against being "lied" to. If you understand a given statistic and when it should be applied, you will be able to detect when statistics have been intentionally or unintentionally misapplied. The important issues for you are:

- Why is any given statistical test applied?
- What is the meaning of each statistic derived during the procedure, and
- How should the findings of this analysis be interpreted?

The purpose of this and subsequent chapters is to present answers to these questions as they pertain to the various statistics presented. In this chapter, the discussion will revolve around descriptive statistics. Descriptive statistics are generally univariate statistics, in that they describe the distributional characteristics of one variable. They include such statistics as means, standard deviations and percentiles. Having collected copious amounts of data on your sample selected from the target population, you must reduce the chaos of all that data into an easily comprehended form. This is the purpose of the so called, descriptive statistics, in that you want to describe your data.

There are two distinct categories of statistical procedures, namely, ***parametric*** and ***non-parametric***. A parametric test assumes that the data being analysed have a specific distribution; often it is assumed that the data are normally distributed. The normal probability distribution is often called the "bell-curve"; particularly by university students, when asking their course instructor how she or he intends to assign grades. In contrast to parametric tests, non-parametric tests make no assumptions about the distribution of the data. The majority of the statistics dealt with in this text are parametric; however, examples of non-parametric tests will be discussed in Chapter 2-9.

Normal Distribution

Although many theoretical distributions exist, the most commonly known distribution is the normal curve. If the distribution of a set of scores approximates the normal distribution, the known properties of the normal curve provide useful information about the distribution. Actually, few distributions of sampled data fit all of the requirements of normality. However, the normal curve can be closely approximated in most variables we study, if the distribution is based on a sufficiently large number of scores.

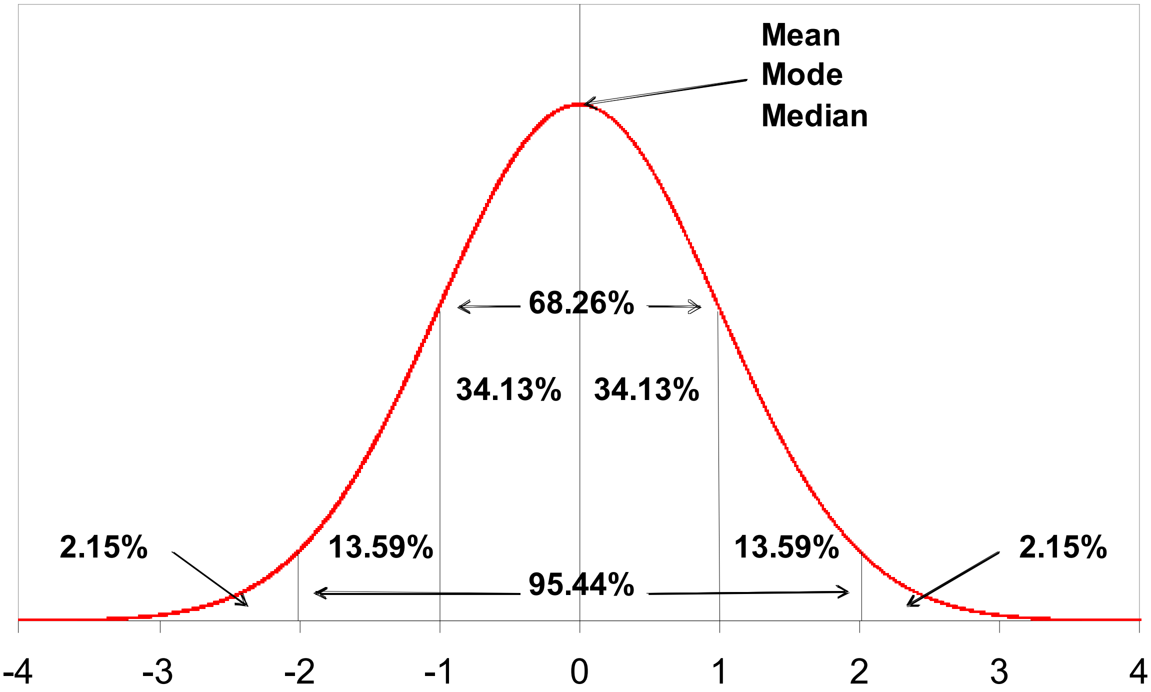


Figure 2-1.1: Normal Probability Distribution

The normal curve is represented by a bell-shaped curve, as shown in Figure 2-1.1. The defined characteristics of the normal curve are:

- The curve is symmetrical, that is, the left half of the curve is identical to the right.
- The ends of the curve never touch the base line. Theoretically, the distribution includes all possible scores to infinity. Thus, the curve has no upper or lower limits.
- The mean, mode and median are all at the centre of the distribution. These are referred to as measures of central tendency.

- One standard deviation above and below the mean encompasses 68.26% of the sample. The standard deviation is a measure of variability in the sample.
- The points of inflection of the curve (where it changes from convex to concave at specific points above and below the centre of the distribution), represent the location of one standard deviation above and one standard deviation below the mean.
- The area between any given points is determined by the percentage of scores that theoretically would fall between those points. More scores fall at the middle of the distribution, with fewer and fewer scores falling toward the extremes.

The general formula for the probability density function of the normal probability distribution (Figure 2-1.1) is:

$$f(X) = \frac{e^{-(X-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}$$

where μ is the mean and σ is the standard deviation. The case where μ = 0 and σ = 1 is called the standard normal distribution. By substituting μ = 0 and σ = 1 in the above equation, the equation for the standard normal distribution is produced:

$$f(X) = \frac{e^{-X^2/2}}{\sqrt{2\pi}}$$

Measures of Central Tendency

As discussed previously, many variables tend to be normally distributed, or close to normally distributed. As such, the observations are more frequent towards the centre of the distribution. Therefore, an important set of descriptive statistics indicate where the centre of the distribution tends to be – these are called *measures of central tendency*. The three most commonly used measures are the mean, the mode, and the median

Mean: The arithmetic average of the scores in the sample. The mean ($\overline{X}$) is the sum of all the scores divided by the sample size n.

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Mode: The mode is the most frequently observed value, which because of the symmetrical nature of the normal distribution will be at the middle.

Median: If scores are ranked from smallest to largest, the median is the middle one in the ranking. It is also called the 50th percentile as 50% of the scores will be at or below this value.

When the data are normally distributed, by definition the mean, mode and median have the same value. Usually you report the mean as the preferred measure of central tendency; however, when the data are not normally distributed a decision needs to be made upon choice of measure. When the data is positively skewed (see below) the mean will be distorted by the few large values. The relative orientation will tend to be that the mean will be to the right or higher than the mode with the median somewhere between the two. In this situation the median is often the preferred selection because it is less influenced by the few large scores. For example, the median is usually used when describing the typical selling price of a house in an area. Only a few multimillion dollar homes will have sold, and they will cause the house price distribution to be positively skewed. The median is therefore selected as the most representative of typical house selling price.

Measures of Variability

A measure of central tendency is not adequate to describe a distribution on its own. Some measure of variability is required. If you measure the height of all the men in a room and report the mean height is 177 cm we do not know if there were short and tall men in the room, or if they all happened to be 177 cm tall. All we know is that the average height was 177 cm. If the data can be assumed to be normally distributed then the standard deviation can be used as a convenient measure of variability. In figure 2-1.1 it can be seen that in the normal distribution by definition, 68.26% of the population lie between the mean ±1 **standard deviation**. If the mean and standard deviation are reported, we now have a much better picture of the distribution. In the case of the height of our men in the room, if the SD is reported as 10 cm, we now know that 68.26% of the men are between 167cm and 187cm (177cm ±10cm) in height.

A number that you may have seen in stats books is 1.96, particularly in reference to so-called 95% confidence intervals. 1.96 is the number of standard deviations above and below the mean that encompasses 95% of the scores in the normal distribution. The mean ±2 standard deviations actually encompasses 95.44% of the scores, although it is often stated in the Empirical Rule that the mean ±2 standard deviations encompasses 95%. This is merely an approximation in order to simplify.

The standard deviation is calculated from the **variance** (s^2), which is calculated as:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

Where X_i is the value of the ith score; $\overline{X}$ is the mean and n is the size of the sample (n = 1 to i). The standard deviation (*SD*) is then the square root of the variance:

$$SD = \sqrt{s^2}$$

When the normal distribution does not apply, the range can be used as an indicator of variability. Analagous to the standard deviation is the interquartile distance. The quartiles are the scores that 25% of the other scores are at or below (1st quartile) and 75% of the other scores are at or below (3rd quartile). Also known as the 25th and 75th percentiles. The other quartile is the median or 50th percentile. These quartiles divide the distribution into four 25% chunks. The difference between the 25th and 75th percentile is often reported as the interquartile distance and quantifies the location of 50% of the sample.

Standard Error of the Mean (SEM)

A statistic that is useful to report is the standard error of the mean (SEM). The standard error of the mean represents your confidence that the mean of your sample truly reflects the mean of the population you are sampling from. Calculated as the standard deviation divided by the square root of the sample size, the SEM will obviously get smaller as the sample size increases. This makes sense since as the larger your sample is, the more confident you would be that your mean was a good estimate of the population mean.

$$SEM = \frac{SD}{\sqrt{n}}$$

where *SD* = Standard deviation, *n* = sample size

> **Central Limit Theorem:** If a sufficiently large number of random samples of the same size were drawn from an infinitely large population, and the mean (average) was computed for each sample, the distribution formed by these averages would be normal.

The standard error of the mean is actually the standard deviation of this distribution of averages. Obviously you would never carry out this enormous sampling procedure, so the SEM is estimated by the equation above. The SEM describes your confidence that the mean of the sample represents the mean of the population sampled from. It can be stated that you are 68.26% confident that the mean of the population is within the mean of the sample ±1 SEM, or

95% confident ±1.96 SEM. Although considered a descriptive statistic, in later chapters we will show how it can be used when describing the confidence in stating differences between means.

Calculating Descriptive Statistics with EXCEL

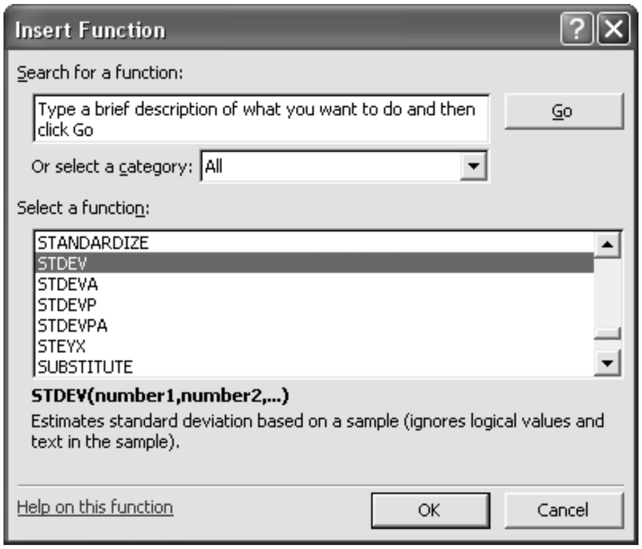
As discussed in Chapter 1-3, EXCEL provides many convenient prewritten functions that save you the trouble of having to insert all the equations necessary for some common statistical procedures. The descriptive statistics described earlier all exist as functions in EXCEL. As example calculating the mean is achieved using the AVERAGE() function. To enter an AVERAGE( ) function, begin by typing **=AVERAGE(.** Next, tell Excel which cells to average. Using the mouse, press and drag over the range of cells you wish to add. A dotted outline appears around the cells, and Excel displays the cell range in the formula bar. When you have the correct cells selected, release the mouse button, close the parenthesis, and press the Enter key. If you do not want to use the mouse, type in the cells you want Excel to average. For example, to average cells B6 through B8, type **=AVERAGE(B6:B8).** Excel interprets B6:B8 as the range of cells from B6 to B8.

	A	B	C	D	E
4					
5	**Subject #**	**Age**	**Weight**	**Height**	**BMI**
6	**141**	23.3	72.3	176.6	**=C6/(D6^2)**
7	**142**	26.5	67.8	172.7	
8	**143**	22.8	86.9	184.9	
9					
10	**Mean**	**=AVERAGE(B6:B8)**			
11	**S.D.**	**=STDEV(B6:B8)**			

Figure 2-1.2: EXCEL worksheet illustrating inclusion of descriptive statistical functions

Excel has many more functions besides the AVERAGE() function described above. For example, you might want to calculate the sum of a column of numbers, or count how many entries are in a row, or calculate a standard deviation.

Select FUNCTION from the INSERT menu and you will bring up the dialog box shown in Figure 2-1.3. You have a choice of hundreds of functions in categories such as Statistical, Math & Trig, Financial etc. When selected, a line of explanation will be presented about the function, and when OK is clicked the function will be entered into the cell. The function can now be completed by entering the cell addresses of the parameters required. This is particularly useful if you require the result of one of these functions for direct inclusion in further calculations.


Insert Function

Search for a function:

Type a brief description of what you want to do and then click Go

Or select a category: All

Select a function:

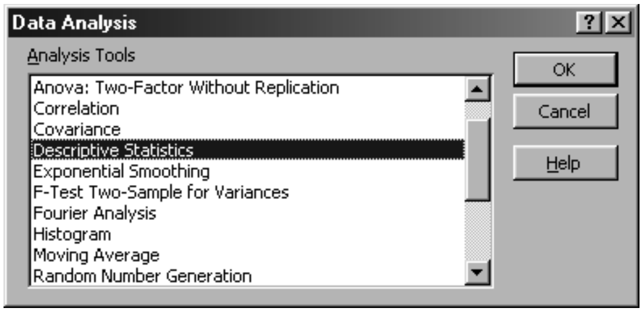
STANDARDIZE
STDEV
STDEVA
STDEVP
STDEVPA
STEYX
SUBSTITUTE

STDEV(number1,number2,...)

Estimates standard deviation based on a sample (ignores logical values and text in the sample).

Figure 2-1.3: INSERT FUNCTION dialog box

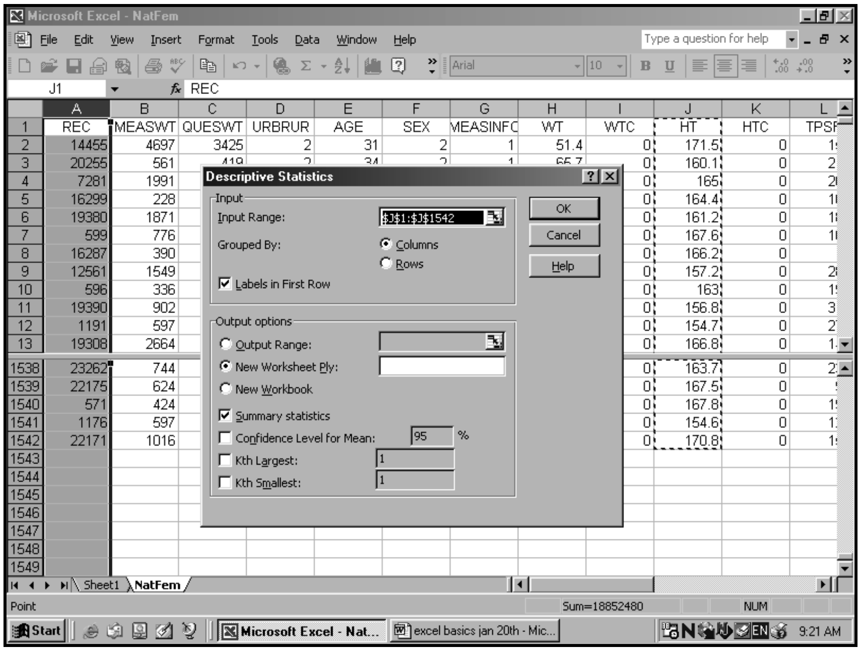
Although the statistical functions in EXCEL are useful, it is more convenient if all of the results of a statistical analysis are reported in a table. EXCEL has this facility but with a limited statistical package available. EXCEL has many statistical tests that can be carried out. It should be pointed out that EXCEL is not intended to be a sophisticated statistical analysis package. It is a spreadsheeting program that offers some statistical analysis features. We will show its facility in several chapters in this text. Sometimes EXCEL might be the only software available to you and the truth is it does provide a lot of utility in this area. Given the option we would like to have a sophisticated package like SPSS, so in many illustrations in this text we will show both applications being used for statistical analysis.


Data Analysis

Analysis Tools

Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram
Moving Average
Random Number Generation

Figure 2-1.4: DATA ANALYSIS dialog box



Descriptive Statistics

Input

Input Range: $J$1:$J$1542

Grouped By: ● Columns ○ Rows

☑ Labels in First Row

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

☑ Summary statistics

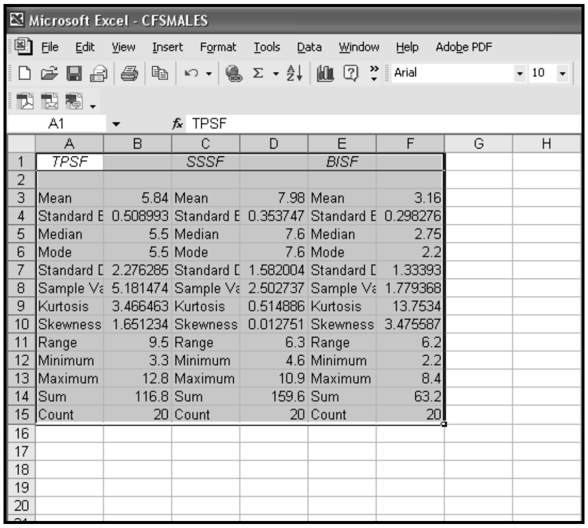
☐ Confidence Level for Mean: 95 %

☐ Kth Largest: 1

☐ Kth Smallest: 1

Figure 2-1.5: DESCRIPTIVE STATISTICS dialog box

EXCEL statistical analysis can be found under the TOOLS menu in the DATA ANALYSIS option. When you click on DATA ANALYSIS, the dialogue box shown in Figure 2-1.4 will appear. An extensive menu of statistical procedures is available. Click on the procedure you want and you will be presented with the appropriate dialog box. Figure 2-1.5 shows the result of selecting DESCRIPTIVE STATISTICS. The dialog boxes are similar for all the procedures. IINPUT RANGE refers to the cells where the data is for analysis. Note that you must check LABELS IN FIRST ROW if you wish the output to have variable names included. OUTPUT RANGE is where you want the output table to be printed. You can select a new location on the same sheet or send it to a new worksheet. Dependent upon the particular procedure you may be offered options such as confidence levels which tend to default to 95%.



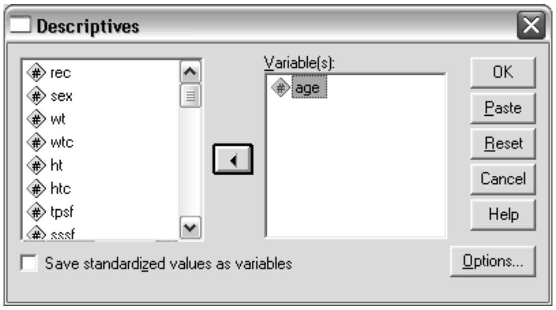
	A	B	C	D	E	F
1	*TPSF*		*SSSF*		*BISF*	
2						
3	Mean	5.84	Mean	7.98	Mean	3.16
4	Standard E	0.508993	Standard E	0.353747	Standard E	0.298276
5	Median	5.5	Median	7.6	Median	2.75
6	Mode	5.5	Mode	7.6	Mode	2.2
7	Standard D	2.276285	Standard D	1.582004	Standard D	1.33393
8	Sample Va	5.181474	Sample Va	2.502737	Sample Va	1.779368
9	Kurtosis	3.466463	Kurtosis	0.514886	Kurtosis	13.7534
10	Skewness	1.651234	Skewness	0.012751	Skewness	3.475587
11	Range	9.5	Range	6.3	Range	6.2
12	Minimum	3.3	Minimum	4.6	Minimum	2.2
13	Maximum	12.8	Maximum	10.9	Maximum	8.4
14	Sum	116.8	Sum	159.6	Sum	63.2
15	Count	20	Count	20	Count	20

Figure 2-1.6: DESCRIPTIVE STATISTICS dialog box

Figure 2-1.6 shows the output for DESCRIPTIVE STATISTICSS for three variables labelled TPSF, SSSF and BISF. An irritating feature of the outputs from the statistical procedures is that no optimizing of column width occurs. This means that labels often are not completed within the cell width. This can be cured by manually changing column widths as discussed earlier. In later chapters as specific statistical procedures are referred to, the details of EXCEL operations will be discussed.

Calculating Descriptive Statistics with SPSS

Figure 2-1.7 shows the DESCRIPTIVE STATISTICS dialog box. The list of variables can be selected from by highlighting the variable and moving into Variable(s) window by clicking on the right arrow button. When a variable is selected in the left hand list the arrow will blacken and point to the right. If the arrow is clicked the variable selected will be moved into the variable list on the right. In the example, the variable Age has been put into the list.



Descriptives

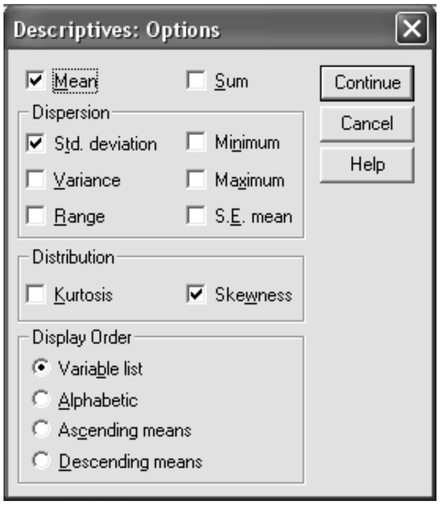
rec
sex
wt
wtc
ht
htc
tpsf
sssf

Variable(s):
age

□ Save standardized values as variables

Figure 2-1.7: COMPUTE VARIABLE Dialog Box (left) and COMPUTE VARIABLE: IF CASES dialog box (right)

Almost all of the statistical analysis dialog boxes have an OPTIONS button. This will bring up a dialog box that will allow selections about test specific configurations. In the case of our DESCRIPTIVES example the OPTIONS dialog box is shown in Figure 2-1.8. Here you can select which statistics will be calculated and the order of presentation of results. What is contained in the OPTIONS dialog box is specific to the test being used.



Descriptives: Options

☑ Mean □ Sum

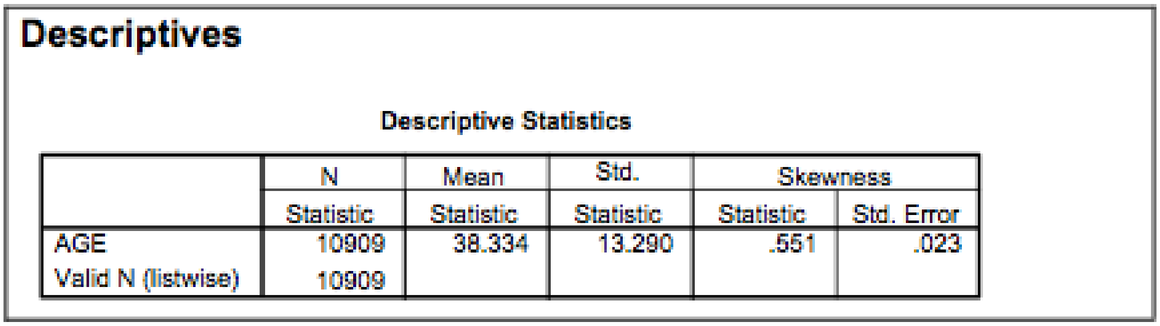
Dispersion
☑ Std. deviation □ Minimum
□ Variance □ Maximum
□ Range □ S.E. mean

Distribution
□ Kurtosis ☑ Skewness

Display Order
● Variable list
○ Alphabetic
○ Ascending means
○ Descending means

Figure 2-1.8: DESCRIPTIVES: OPTIONS Dialog Box

When the test is run, the results are sent to the OUTPUT window. Figure 2-1.9 shows the output for the DESCRIPTIVES dialog box shown in Figure 2-1.7. It simply gives the selected descriptive statistics of Age for all subjects (N = 10,909) in the CFS (Canada Fitness Survey) adult data set. The output can be printed or component parts can be copied and pasted into WORD files as desired. This output was customized with the options dialog box to report mean, standard deviation and the coefficient of skewness.



Descriptives

Descriptive Statistics

	N	Mean	Std.	Skewness	
	Statistic	Statistic	Statistic	Statistic	Std. Error
AGE	10909	38.334	13.290	.551	.023
Valid N (listwise)	10909				

Figure 2-1.9: DESCRIPTIVES output of Descriptive statistics for Age in the whole Canada Fitness Survey adults data set (N = 10909)