Chapter 3

Summarizing Data Numerically: Measures of Central Tendency

In addition to graphical summaries (Chapter 2), the primary features of a data set can be summarized through numerical indices. Measures of central tendency or location specify the "center" of a set of measurements. This chapter describes ways to use SPSS to obtain three common measures of location — the mode, the median, and the mean – of a sample. Measures of central tendency can be used to:

- find the most common college major for a group of students;
- find the midpoint of a set of ordered body weights that divides the set in half;
- calculate the average gross of the top movies from a given year;
- find the percentage of 13-year-old children who have a home computer.

3.1 THE MODE

The mode, especially useful in summarizing categorical or discrete numerical variables, is the category or value that occurs with the greatest frequency. One way to obtain the mode with SPSS for Windows is by using the Frequencies

procedure. This is the same procedure used to obtain frequency distributions, histograms, and bar charts discussed in Chapter 2.

We illustrate how to obtain the mode using the "movies.sav" data file, which contains information on top grossing movies of 2001. Data include the genre, opening week gross, total gross, number of theatres in which it was released, studio, and the number of weeks the movie was in the top 60. The "genre" variable is a categorical variable representing the type of movie. To obtain the mode of this variable:

1. Click on **Analyze** from the menu bar.
2. Click on **Descriptive Statistics** from the pull-down menu.
3. Click on **Frequencies** from the pull-down menu.
4. Click on the "genre" variable and then the **right arrow button** to move the variable into the Variable(s) box.
5. Click on the **Statistics button** at the bottom of the screen. This opens the Frequencies: Statistics dialog box, as shown in Figure 3.1.
6. Click on the **Mode** option in the Central Tendency section.
7. Click on **Continue** to close this dialog box.
8. Click on **OK** to close the Frequencies dialog box and execute the procedure.

The output is shown in Figure 3.2.

Notice that in addition to the frequency distribution, the output lists the mode of the variable in the statistics table; it is genre "4." Because genre is a categorical variable, a value of 4 is representative of a particular genre. In this case, it represents "comedy." So, there are more comedies in the top 100 movies than any other genre.

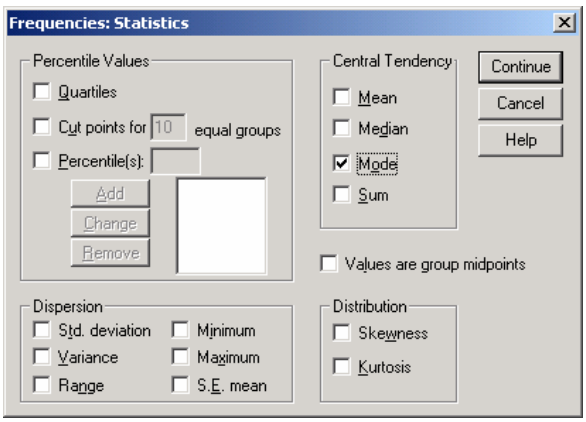

Figure 3.1 Frequencies: Statistics Dialog Box

Statistics

Movie type

N	Valid	100
	Missing	0
Mode		4

Movie type

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Thriller/horror	11	11.0	11.0	11.0
	family	7	7.0	7.0	18.0
	drama	22	22.0	22.0	40.0
	comedy	38	38.0	38.0	78.0
	adventure/fantasy	22	22.0	22.0	100.0
	Total	100	100.0	100.0	

Figure 3.2 Frequency Distribution with Mode for Movie Genre

It is also possible to determine the mode of a variable by examining the frequency distribution itself. As an example, refer back to Figure 3.2. Even without the Mode option, you could search through the Frequency column for the row with the highest number, here comedy. Or, you could look at the Valid Percent column for the largest percentage, here 38.0%. The value associated with these numbers is the most common value for the variable — the mode of the variable.

3.2 THE MEDIAN AND OTHER PERCENTILES

The Median

The median is a value that divides the set of ordered (from smallest to largest) observations in half. That is, one-half the observations are less than (or equal to) the median value, and one-half the observations are greater than (or equal to) the median value. The symbol for the median is *M*.

The procedure for determining the median of a variable is similar to that for obtaining the mode. You simply need to click on the **Median** option instead of the Mode option in the Central Tendency box of the Frequencies: Statistics dialog box. (See Fig. 3.1.)

Continuing with the “movies.sav” data file we used in Section 3.1, we will

examine the number of weeks the movies were in the Top 60. Your output should look like that in Figure 3.3. The median of the distribution is 14 weeks, as indicated on the first table in Figure 3.3.

Statistics

Weeks in top 60

N	Valid	100
	Missing	0
Median		14.00

Weeks in top 60

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	7	8	8.0	8.0	8.0
	8	6	6.0	6.0	14.0
	9	4	4.0	4.0	18.0
	10	4	4.0	4.0	22.0
	11	10	10.0	10.0	32.0
	12	6	6.0	6.0	38.0
	13	10	10.0	10.0	48.0
	14	5	5.0	5.0	53.0
	15	9	9.0	9.0	62.0
	16	5	5.0	5.0	67.0
	17	2	2.0	2.0	69.0
	18	6	6.0	6.0	75.0
	19	3	3.0	3.0	78.0
	20	2	2.0	2.0	80.0
	21	1	1.0	1.0	81.0
	22	2	2.0	2.0	83.0
	23	2	2.0	2.0	85.0
	24	6	6.0	6.0	91.0
	25	2	2.0	2.0	93.0
	26	1	1.0	1.0	94.0
	27	3	3.0	3.0	97.0
	29	1	1.0	1.0	98.0
	31	1	1.0	1.0	99.0
	38	1	1.0	1.0	100.0
	Total	100	100.0	100.0	

Figure 3.3 Frequency Distribution with Median for Total Gross of Movies

As with the mode, you can also determine the median from the frequency distribution. To do so, recall that the cumulative percentage column represents the percentage of cases at or below a given value. Because the median is the value of the ordered distribution at which 50% of the values are below it, to find the median, locate the first row in the distribution that has a cumulative percentage of 50 or greater. If it is not exactly 50%, the value associated with the percentage is the median. In this example, it is 53%. The value (number of weeks) associated with 53% is 14 weeks.

If the percentage in the cumulative percent column is exactly 50%, the median is halfway between that value and the subsequent value on the frequency distribution. This can occur when the distribution has an even number of observation. Then, median is halfway between: the observation at the $n/2$ position, and the observation at the $n/2 + 1$ position. For example, when there are 100 observations in the data set, the median is halfway between the 50^{th} and 51^{st} positions.

There are some additional features of the Frequencies Procedure that may be useful in some cases. For instance, if you would like to obtain both the mode and the median of a variable, you can select more than one option from the Frequencies: Statistics dialog box and obtain several statistics at once. There may also be times that you wish to only obtain the statistics, but not the frequency distribution. This is particularly useful for examining continuous variables from very large data sets. Suppose, for instance, you have a data file containing heights of 500 people. If the heights were measured to the nearest 100^{th} of an inch, there would be very few data points with more than one observation. Thus, the frequency distribution would be a long ordered listing of the data points. There is an option on the Frequencies dialog box called "Display frequency tables" that governs whether or not the frequency distribution is displayed. The default for this option is "yes," but you may manually turn off the option by clicking on the box to the left of the phrase.

Quartiles, Deciles, Percentiles, and Other Quantiles

Just as the median divides the set of ordered observations into halves, quartiles divide the set into quarters, deciles divide the set of ordered observations into tenths, and percentiles divide the set of observations into hundredths. Quantile is a general term that includes quartiles, deciles, and percentiles.

You can obtain quartiles or percentiles from the Frequencies procedure by selecting the appropriate option in the upper left box of the Frequencies: Statistics dialog box. For instance, click on **quartiles** to generate a list of the quartiles. Deciles can be obtained by clicking on the **"Cut points"** option and selecting 10 equal groups. Other percentiles can be obtained by clicking on the **percentiles** option and entering the desired figures.

Statistics

Weeks in top 60

N	Valid	100
	Missing	0
Percentiles	25	11.00
	50	14.00
	75	18.75

Figure 3.4 Quartiles for Weeks in Top 60

Quartiles for the variable "weekstop" in the "movies.sav" data file are displayed in Figure 3.4. The 25th percentile is equivalent to the first quartile. Thus, one-quarter (25%) of the movies were in the top 60 for 11 or fewer weeks. The 50th percentile is the same as the median — 14 weeks. The 75th percentile, or third quartile is 18.75 weeks.

3.3 THE MEAN

The mean of a set of numbers is the arithmetic average of those numbers. The mean summarizes all of the units in every observed value, and is the most frequently used measure of central tendency for numerical variables. (When data are skewed, however, the median is generally a more appropriate measure of central tendency.) The symbol for the mean in a sample is $\bar{x}$, which is often referred to as "x bar."

There are several methods for obtaining the mean of a distribution with SPSS for Windows. You can use the Frequencies procedure by clicking on **Mean** in the Central Tendency box in the Frequencies: Statistics dialog box. Try this with the movies data and the "weekstop" variable. The mean should be 15.26 weeks. In this example, the mean is somewhat larger than the median (14 weeks), suggesting that the distribution may be positively skewed. (In normally distributed distributions, the mean and median are similar in value.)

The mean can also be calculated with SPSS using the Descriptives or the Explore procedures. To obtain the mean using the Descriptives procedure:

1. Click on **Analyze** from the menu bar.
2. Click on **Descriptive Statistics** from the pull-down menu.
3. Click on **Descriptives** from the pull-down menu. This opens the Descriptives dialog box, as shown in Figure 3.5.
4. Move the "weekstop" variable to the Variable(s) box by clicking on the variable and then on the **right arrow button**.

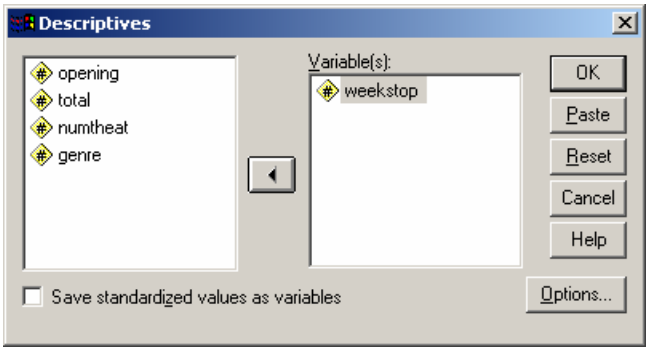

Figure 3.5 Descriptives Dialog Box

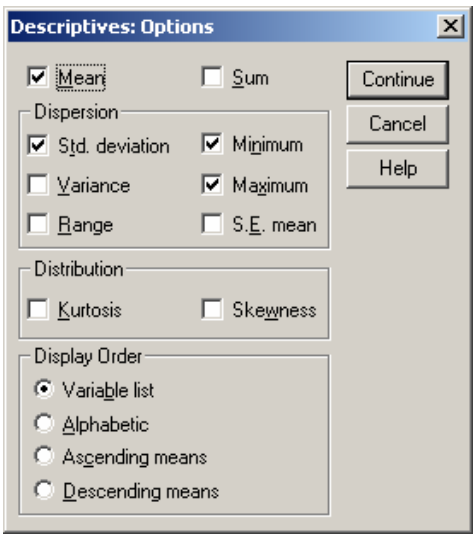

Figure 3.6 Descriptives: Options Dialog Box

5. Click on the **Options** button to open the Descriptives: Options dialog box (Fig. 3.6). The default options selected are: Mean, Std. deviation, Minimum and Maximum. We will accept the default in this example.
6. Click on **Continue** to close this dialog box.
7. Click on **OK** to run the procedure.

Did you obtain the same mean as you did when you used the Frequencies procedure?

You may also obtain the mean and several other descriptive statistics using the Explore procedure as follows:

Descriptives

			Statistic	Std. Error
Weeks in top 60	Mean		15.26	.632
	95% Confidence Interval for Mean	Lower Bound	14.01	
		Upper Bound	16.51	
	5% Trimmed Mean		14.88	
	Median		14.00	
	Variance		39.891	
	Std. Deviation		6.316	
	Minimum		7	
	Maximum		38	
	Range		31	
	Interquartile Range		7.75	
	Skewness		.939	.241
	Kurtosis		.747	.478

Figure 3.7 Output from the Explore Procedure

1. Click on **Analyze** from the menu bar.
2. Click on **Descriptive Statistics** from the pull-down menu.
3. Click on **Explore** from the pull-down menu.
4. Click on and move the "weekstop" variable to the Dependent List box using the **right arrow button**.
5. In the display box, click on the **Statistics** option.
6. Click on **OK**.

In addition to the mean, this procedure lists the median and several other descriptive statistics (see Fig. 3.7). For instance, the median is 14 weeks. We will discuss some of the other statistics, such as the variance, standard deviation, and range in the Chapter 4.

Proportion as a Mean

There is an exception to the rule that requires a variable to have numerical properties in order to calculate the mean — it is the dichotomous categorical variable. A dichotomous variable is a variable with only two possible values. If such a variable is coded with values 0 and 1, the mean will be the proportion of the cases with a value of 1. We will illustrate this using the "titanic.sav" data file as described in Chapter 2. The variable "survived' is coded so 0 = no, and 1 = yes. Thus, the mean represents proportion of passengers who survived the ship's sinking.

Statistics

SURVIVED

N	Valid	2201
	Missing	0
Mean		.32

SURVIVED

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	no	1490	67.7	67.7	67.7
	yes	711	32.3	32.3	100.0
	Total	2201	100.0	100.0	

Figure 3.8 Frequency Distribution and Mean of a Dichotomous Variable

Compute the mean using the Frequencies procedure. From the frequency distribution (Fig. 3.8), note that the percentage of survivors (coded 1) in the sample is about 32.3%. Within rounding, this is the same as the mean of .32.

Chapter Exercises

3.1 The file "library.sav" contains information on the size (number of books in the collection) and staff in 22 college libraries. Use this file to perform the following analyses with SPSS:

a. Determine the mean and median of the distribution of staff using the Frequencies procedure. How do these measures compare? What do you think accounts for the differences?

b. What value is at the 10th percentile? The 90th percentile?

c. Create a histogram of the variable. Describe the distribution. Do any of the observations seem to be outliers? If so, how do these observations affect your findings in parts (a) and (b)?

3.2 Use the "fire.sav" data file, which contains demographic and test performance data on 28 firefighter applicants, to do the following:

a. Determine the mean obstacle course time of the sample using either the Frequencies or the Descriptives procedure.

b. Suppose you discovered that the clock device that kept time was set to start at 2 seconds, rather than 0 seconds. In order to obtain more accurate

time, subtract 2 seconds from each candidate's time, and calculate the mean of the revised times. (Hint: use the Compute procedure.)

c. How does your result in (a) compare to that in part (b)? What principle does this illustrate?

d. Repeat the procedure, this time dividing each time by 2. How does the mean of the revised times compare to the mean of the original times?

3.3 The "IQ.sav" data file contains information on language and non-language IQ scores for a sample of children. Using these data use SPSS to complete the following:

a. Compute the mean, median, and mode of each type of IQ using the Frequencies procedure.

b. Which is the "best" measure of central tendency for the language IQ scores? Why? For the nonlanguage scores? Why?

3.4 The "movies.sav" data file contains information on top grossing movies of 2001. Following the steps below, use SPSS to illustrate the principle that the sum of all deviations from the mean is zero.

a. In Section 3.3 we found that the mean number of weeks these movies spent in the top 60 was 15.26 weeks. Compute a new variable, called "dev," representing deviations from this mean. Details for computing variables are contained in Chapter 1. The algebraic expression, which you will enter in the Numeric Expression box of the Compute Variable dialog box, is: "weekstop-15.26."

b. Now, compute the sum and mean of this new variable, "dev," using the Descriptives procedure.

c. Did you find that both the sum and mean are 0? Why is this?