Chapter 4

Summarizing Data Numerically: Measures of Variability

Chapter 3 described the use of SPSS to obtain measures of central tendency for a data set. Another important characteristic of numerical data is variability. Measures of variability indicate how spread out the observations are, that is, how much the values differ from individual to individual. For example, measures of variability let you:

- find the difference between the largest and smallest salary paid to people working at a particular company;
- determine how daily hours of sleep vary among different species of mammals;
- compare two or more distributions, such as whether basketball players or football players vary more in terms of height.

This chapter demonstrates ways to use SPSS to examine variability in a sample. We demonstrate how to calculate range, interquartile range, standard deviation, variance, and standard scores. A graphical display of measures of variability is demonstrated through box-and-whisker plots.

4.1 RANGES

The Range

The range is the difference between the maximum value in a distribution and the minimum value. There are several ways to find the range. One method is to create a frequency distribution of the scores as demonstrated in Section 2.2. To find the range from this distribution, subtract the minimum value from the maximum value. When you have a small data set with relatively few values, this method is adequate. However, if your data set has many different values, the frequency distribution will be very large and it is more efficient to compute the range using the Statistics option within the Frequencies procedure. First follow steps 1–5 given in Section 2.1 for creating a frequency distribution. Next:

1. Click on **Statistics** to open the Frequencies: Statistics dialog box (see Fig. 2.2).
2. Click on **Range** in the Dispersion box.
3. Click on **Continue**.
4. Click on **OK**.

A second method for obtaining the range and other measures of dispersion uses the Descriptives procedure. We shall illustrate with the "sleep.sav" data file, which contains information on physical, environmental, and sleep characteristics of 62 mammals. Let us find the range of hours of sleep per day. After opening the file:

1. Click on **Analyze** on the menu bar.
2. Click on **Descriptive Statistics** from the pull-down menu.
3. Click on **Descriptives** to open the Descriptives dialog box (see Chapter 3, Fig. 3.5).
4. Click on the variable name for which you wish to have the range ("totsleep"), and then click on the **right arrow button** to move the variable into the Variables box.
5. Click on **Options** to open the Descriptives: Options dialog box (Fig. 3.6).
6. In the Dispersion box, click on **Range**. (Notice that the mean, standard deviation, minimum, and maximum boxes are already checked. This is the default.)
7. Click on **Continue**.
8. Click on **OK**.

The summary statistics for the Descriptives procedure should appear as shown in Figure 4.1.

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation
TOTSLEEP	58	17.30	2.60	19.90	10.5328	4.60676
Valid N (listwise)	58					

Figure 4.1 Descriptives for Hours of Sleep Variable

The variable name appears in the left column of the table, followed by the descriptive statistics. The range, the difference between the maximum (19.90 hours) and the minimum (2.60 hours), is 17.30 hours. Notice that the N is 58. Thus, four of the mammals (62 mammals are in the data file) did not have data on total hours of sleep.

The Interquartile Range

Quartiles are one of the measures of location discussed in Chapter 3. The interquartile range is the difference between the first and third quartiles. The interquartile range is a more stable measure of variability than the range because it is less affected by extreme scores. The interquartile range can be obtained by using the Explore procedure, as described in Section 3.3. The interquartile range for total sleep is 5.575 hours.

4.2 THE STANDARD DEVIATION

The standard deviation (s) is a type of average of the distances of the values of individual observations from the mean. It is one of the most common and most useful measures of dispersion.

The variance (s^2) is the square of the standard deviation. If x_i ($i = 1,2,...,n$) is the value of observation i, and $\bar{x}$ is the sample mean, then the sample variance is:

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

The standard deviation is found by computing the square root of the variance, $s = \sqrt{s^2}$. The Descriptives procedure is one method for producing the standard deviation in SPSS.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
TOTSLEEP	58	2.60	19.90	10.5328	4.60676	21.222
Valid N (listwise)	58					

Figure 4.2 Descriptives for Hours of Sleep Variable Including the Variance

1. Click on **Analyze** on the menu bar.
2. Click on **Descriptive Statistics** from the pull-down menu.
3. Click on **Descriptives** to open the Descriptives dialog box.
4. Click on the variable name that you wish to examine and the **right arrow button** to move it to the Variable(s) box.

By default, the output will contain the mean, standard deviation, minimum, and maximum values. To obtain the variance, follow steps 1--4 above for the standard deviation, and then:

1. Click on **Options**.
2. Click on **Variance** in the Dispersion box.
3. Click on **Continue** and then click on **OK**.

The output is shown in Figure 4.2. Notice that the variance is the square of the standard deviation; the variance is $4.60676^2 = 21.222$.

4.3 SOME USES OF LOCATION AND DISPERSION MEASURES TOGETHER

Box-and-Whisker Plots

A box-and-whisker plot is a useful graphical display that shows the median, interquartile range, and extremes of a data set. SPSS for Windows creates these plots using the Explore procedure. To illustrate this procedure, we continue with the "sleep.sav" data file, but this time with the "lifespan" variable, which indicates the maximum life span (in years) for the mammals. To create the box-and-whisker plot:

1. Click on **Analyze** on the menu bar.
2. Click on **Descriptive Statistics** from the pull-down menu.
3. Click on **Explore** to open the Explore dialog box.
4. Click on the "lifespan" variable then click the **top right arrow button**.

5. Click on the **Plots button** to open the Explore: Plots dialog box.
6. Click off the **stem-and-leaf** option in the descriptive section and leave clicked the default **Factor levels together** in the boxplots section.
7. Click on **Continue**.
8. Click on **OK**.

The output is shown in Figure 4.3; we are primarily concerned with the graph. The red box in the box-and-whisker plot represents the interquartile range. The top of the red box demarcates the third quartile (28 years), and the bottom denotes the first quartile (6.4 years). (You may wish to check these figures by using the Frequencies procedure to report the quartiles.) The horizontal line in the middle of the red box represents the median of the distribution. Here, it is 15.1 years.

The horizontal line (whisker) at 2 years indicates the minimum life span. The whisker at 50 years denotes the maximum lifespan, excluding outliers. The circle and asterisk at 69 and 100 years, respectively, represent outliers.

Standard Scores

Standard scores, also called z-scores, indicate the relative position of a single observation in the sample, that is, the number of standard deviations the observation is above or below the mean. For values above the mean, the z-score is positive. Values below the mean have negative z-scores, and values equal to the mean have a z-score of 0. The z-scores are calculated using the Descriptives procedure.

We shall illustrate using the "sleep.sav" data file. To create standard scores for the variable "lifespan":

1. Click on **Analyze** on the menu bar.
2. Click on **Descriptive Statistics** and then **Descriptives** to open the Descriptives dialog box.
3. Click on the "lifespan" variable and move it into the Variable(s) box with the **right arrow button**.
4. Click on the **Save standardized values as variables** box.
5. Click on **OK**.

This will cause SPSS to create a new z-score variable. By default, the new variable is named by prefixing the letter Z to the original variable. For example, "lifespan" becomes "Zlifespan." You can examine the standard scores using the Frequencies procedure (see Section 3.1). The histogram is displayed in Figure 4.4.

Descriptives

			Statistic	Std. Error
LIFESPAN	Mean		19.8776	2.39060
	95% Confidence Interval for Mean	Lower Bound	15.0905	
		Upper Bound	24.6647	
	5% Trimmed Mean		17.8594	
	Median		15.1000	
	Variance		331.468	
	Std. Deviation		18.20626	
	Minimum		2.00	
	Maximum		100.00	
	Range		98.00	
	Interquartile Range		21.6250	
	Skewness		2.014	.314
	Kurtosis		5.885	.618

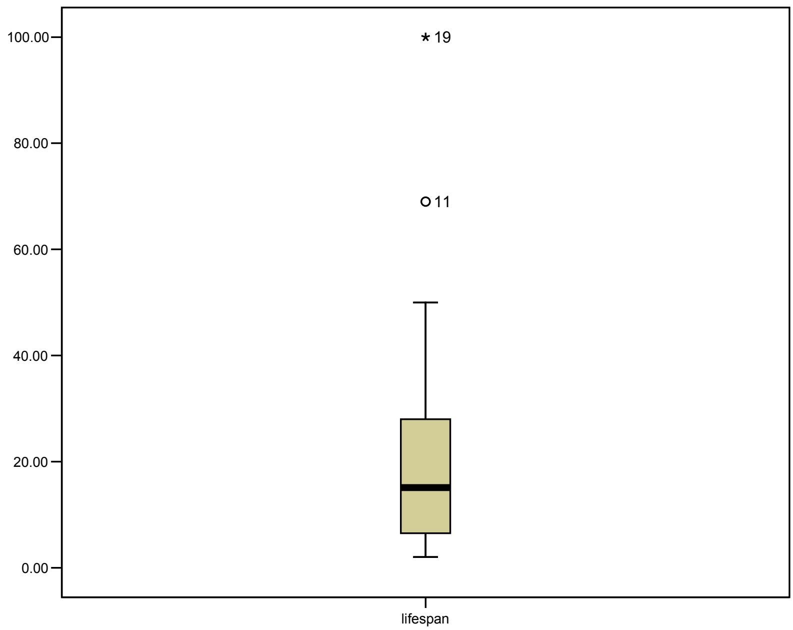

Figure 4.3 Box-and-Whisker Plot of Maximum Lifespan of Mammals

Statistics

Zscore(LIFESPAN)

N	Valid	58
	Missing	4
Mean		.0000000
Std. Deviation		1.000000
Minimum		-.98195
Maximum		4.40082

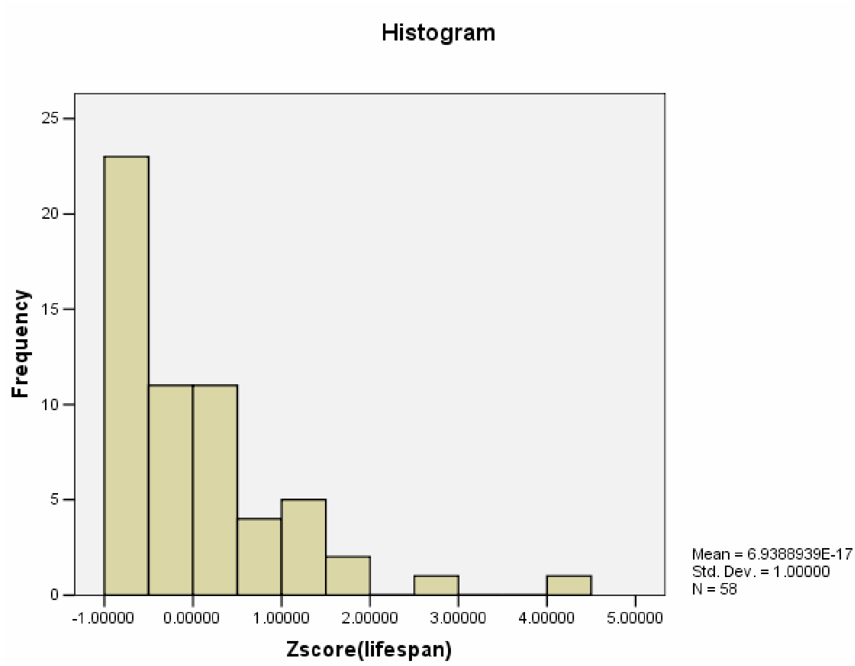

Figure 4.4 Output from Frequencies Procedure Summarizing "Zlifespan"

By definition, the mean and standard deviation of the z-scores are 0 and 1, respectively. If the distribution of "lifespan" were normal, than the range of "Zlifespan" would be generally between –2.0 and 2.0. But, as we discovered with the box-and-whisker plot (and can see again in the histogram in Figure 4.4), this variable has a positively skewed distribution. The two outliers have standardized scores greater than 2.0.

Chapter Exercises

4.1 The “library.sav” data file contains measurements staff and size of 22 college libraries. Use this data file to perform the following analyses with SPSS for Windows:

- **a.** Find the range of the number of volumes held by the libraries using either the Frequencies and Descriptives procedures.
- **b.** Find the interquartile range. In what situation would the interquartile range be a better measure of dispersion than the range?
- **c.** Compute the standard deviation and variance of number of volumes.
- **d.** Are most of the scores within 2 standard deviations of the mean?

4.2 Using the “movies.sav” data file, compute z-scores for the “opening” variable, represents the opening week gross of the movies.

- **a.** What is the z-score for the movie *Cats and Dogs*?
- **b.** What is the z-score for the movie that grossed the least amount of money in the first week? The most?
- **c.** Verify that the mean of the z-scores is 0.

4.3 Using SPSS and the data on 28 firefighter applicants in “fire.sav” file, do the following:

- **a.** Find the standard deviation of the candidates’ body drag time.
- **b.** Suppose that each firefighter received agility training and decreased his or her time by 1 second. Compute the standard deviation of the new times (Hint: use the Compute command). Did the standard deviation change? Why or why not?
- **c.** Suppose that each firefighter received more training and halved his or her original time. Compute the new scores and find the standard deviation of them. How did dividing by a constant affect the standard deviation?

4.4 Use the “hotdog.sav” data set containing information on number of calories and milligrams of sodium in three types of hot dogs (beef, meat, and poultry) to do the following:

- **a.** Use the Explore procedure to produce summary statistics and a box-and-whisker plot for calories each of the three types of hotdog. (HINT: the “type” variable will be used in the Factor List box and “calories” in the Dependent List box.)
- **b.** What is the median number of calories for each type of hot dog?

c. What is the minimum and maximum number of calories for beef hot dogs?

d. Are there any outliers for the poultry hot dogs?

e. Based on the standard deviations, which type of hot dog has the most variability in calories?