NORUŠIS, Marija J. 1998. *Guide to Data Analysis.* Upper Saddle River, Prentice Hall.

Standard Scores

The mean often serves as a convenient reference point to which individual observations are compared. Whenever you receive an examination back, the first question you ask is, How does my performance compare with the rest of the class? An initially dismal-looking score of 65% may turn stellar if that's the highest grade. Similarly, a usually respectable score of 80 loses its appeal if it places you in the bottom quarter of the class. If the instructor just tells you the mean score for the class, you can only tell if your score is less than, equal to, or greater than the mean. You can't say how far it is from the average unless you also know the standard deviation.

For example, if the average score is 70 and the standard deviation is 5, a score of 80 is quite a bit better than the rest. It is two standard deviations above the mean. If the standard deviation is 15, the same score is not very remarkable. It is less than one standard deviation above the mean. You can determine the position of a case in the distribution of observed values by calculating what's known as a standard score, or *z* score.

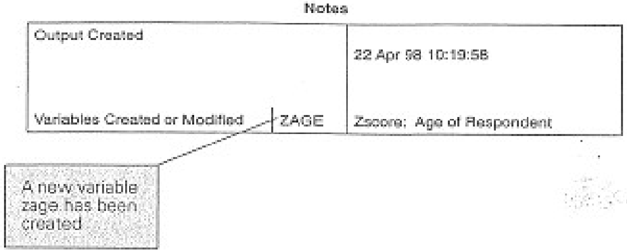
To calculate the standard score, first find the difference between the case's value and the mean and then divide this difference by the standard deviation.

$$\text{standard score} = \frac{\text{value} - \text{mean}}{\text{standard deviation}} \qquad \textbf{Equation 4.5}$$

A standard score tells you how many standard deviation units a case is above or below the mean. If a case's standard score is 0, the value for that case is equal to the mean. If the standard score is 1, the value for the case is one standard deviation above the mean. If the standard score is −1, the value for the case is one standard deviation below the mean. (For many types of distributions, including the normal distribution discussed in Chapter 10, most of the observed values fall within plus or minus two standard deviations of the mean.) The mean of the standard scores for a variable is always 0, and their standard deviation is 1.

You can use the Descriptives procedure in SPSS to obtain standard scores for your cases and to save them as a new variable. Figure 4.3 shows the notes from the Descriptives procedure that indicate that a new variable, the standard score for age, has been created. In addition, a new variable, *zage*, has been saved in the Data Editor, containing the standard scores for age (see Figure 4.4).

Figure 4.3 Descriptive statistics in the Viewer



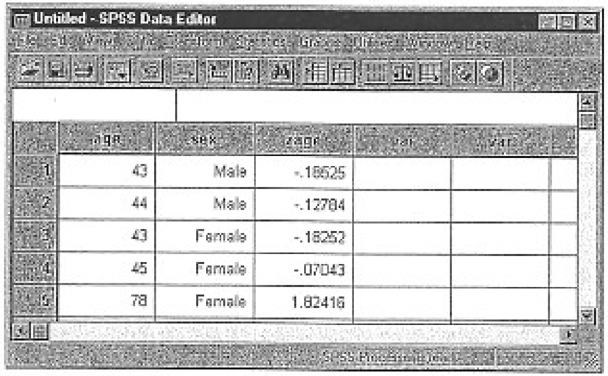
Notes

Output Created		22 Apr 98 10:19:58
Variables Created or Modified	ZAGE	Zscore: Age of Respondent

A new variable zage has been created

Figure 4.4 Data Editor with standard scores saved as a new variable

To save standardized scores, select Save standardized values as variables in the Descriptives dialog box, as shown in Figure 4.5.



	age	sex	zage	var	var
1	43	Male	-.18525		
2	44	Male	-.12784		
3	43	Female	-.18252		
4	45	Female	-.07043		
5	78	Female	1.82416		

You see that the first case has an age of 43. From the standard score, you know that the case has an age less than average, but not very much. The age for the case is less than a quarter of a standard deviation below the mean. The fifth case has an observed age of 78, which is almost two standard deviations above the mean.

Standard scores allow you to compare relative values of several different variables for a case. For example, if a person has a standard score of 2 for income, and a standard score of −1 for education, you know that the person has a larger income than most and somewhat fewer years of education. You couldn't meaningfully compare the original values, since the variables all have different units of measurement, different means, and different standard deviations.

NORUŠIS, Marija J. 1998. *Guide to Data Analysis.* Upper Saddle River, Prentice Hall.

10

The Normal Distribution

What is the normal distribution, and why is it important for data analysis?

- What does a normal distribution look like?
- What is a standard normal distribution?
- What is the Central Limit Theorem, and why is it important?

In Chapter 9, you learned how to evaluate a claim about the mean of a variable that has two possible values. Using the binomial test, you calculated the probabilities of getting various sample results when the probability of a success was assumed to be known. In this chapter, you'll learn how to test claims about the mean of a variable that has more than two values. You'll also learn about the normal distribution and the important role it plays in statistics.

▶ This chapter examines data on serum cholesterol levels from the *electric.sav* data file. In addition, some figures use simulated data sets included in the file *simul.sav*. The histograms and output shown can be obtained using the SPSS Graphs menu (see Appendix A) and the Descriptives procedure (see Chapter 4).

The Normal Distribution

You may have noticed that the shapes of the two stem-and-leaf plots in Chapter 9 are similar. They look like bells (on their sides). The same data are displayed as histograms in Figure 10.1 and Figure 10.2, where a bell-shaped distribution with the same mean and variance as the data is superimposed. You can see that most of the values are bunched in the center. The farther you move from the center, in either direction, the fewer the number of observations. The distributions are also more or less symmetric. That is, if you divide the distribution into two pieces at the peak, the two halves of the distribution are very similar in shape, but mirror images of each other. (The theoretical bell distribution is perfectly symmetric.)

You can obtain histograms using the Graphs menu, as described in Appendix A.

In the Histograms dialog box, select the variables cured10 and cured40.

Figure 10.1 Simulated experiments: sample size 10

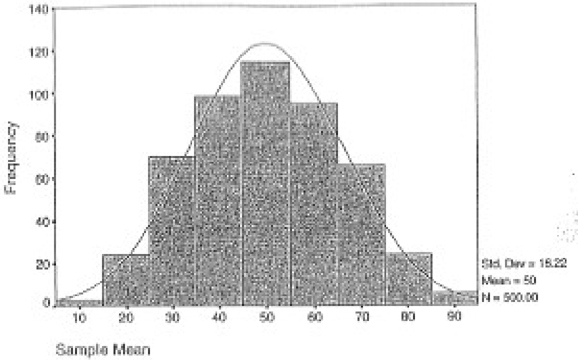
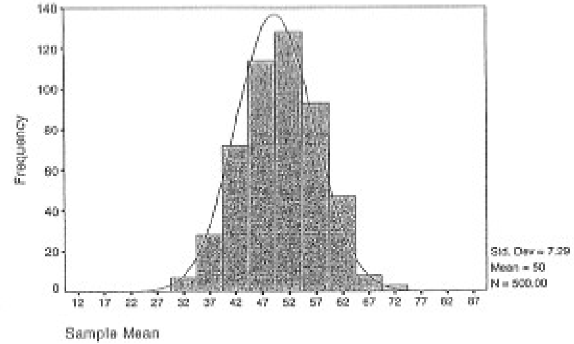


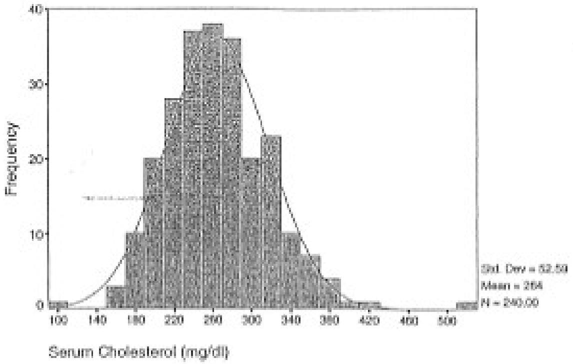
Figure 10.2 Simulated experiments: sample size 40



Many variables—such as blood pressure, weight, and scores on standardized tests—turn out to have distributions that are bell-shaped. For example, look at Figure 10.3, which is a histogram of cholesterol levels for a sample of 239 men enrolled in the Western Electric study (Paul et al., 1963). Note that the shape of the distribution is very similar to that in Figure 10.2. That's a pretty remarkable coincidence, since Figure 10.2 is a plot of many sample means from a distribution that has only two values (1=cured, 0=not cured), while Figure 10.3 is a plot of actual cholesterol values.

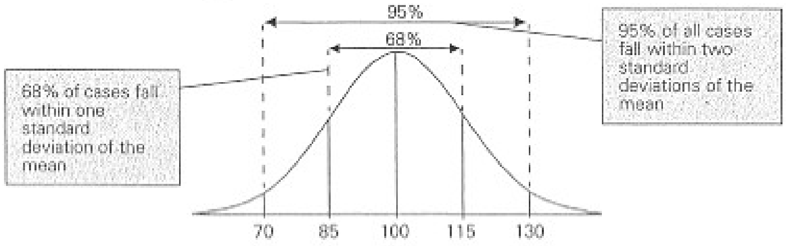
To obtain this histogram, open the electric.sav data file and select chol58 in the Histograms dialog box.

Figure 10.3 Histogram of cholesterol values



The bell distribution that is superimposed on Figure 10.1, Figure 10.2, and Figure 10.3 is called the **normal distribution.** A mathematical equation specifies exactly the distribution of values for a variable that has a normal distribution. Consider Figure 10.4, which is a picture of a normal distribution that has a mean of 100 and a standard deviation of 15. The center of the distribution is at the mean. The mean of a normal distribution has the same value as the most frequently occurring value (the mode), and as the median, the value that splits the distribution into two equal parts.

Figure 10.4 A normal distribution



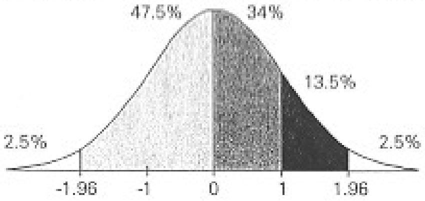
68% of cases fall within one standard deviation of the mean

95% of all cases fall within two standard deviations of the mean

If a variable has exactly a normal distribution, you can calculate the percentage of cases falling within any interval. All you have to know are the mean and the standard deviation. Suppose that scores on IQ tests are normally distributed, with a mean of 100 and a standard deviation of 15, as was once thought to be true. In a normal distribution, 68% of all values fall within one standard deviation of the mean, so you would expect 68% of the population to have IQ scores between 85 (one standard deviation below the mean) and 115 (one standard deviation above the mean). Similarly, 95% of the values in a normal distribution fall within two standard deviations of the mean, so you would expect 95% of the population to have IQ scores between 70 and 130.

Since a normal distribution can have any mean and standard deviation, the location of a case within the distribution is usually given by the number of standard deviations it is above or below the mean. (Recall from Chapter 4 that this is called a standard score, or z score.) A normal distribution in which all values are given as standard scores is called a **standard normal distribution.** A standard normal distribution has a mean of 0, and a standard deviation of 1. For example, a person with an IQ of 100 would have a standard score of 0, since 100 is the mean of the distribution. Similarly a person with an IQ of 115 would have a standard score of +1, since the score is one standard deviation (15 points) above the mean, while a person with an IQ of 70 would have a standard score of −2, since the score is two standard deviation units (30 points) below the mean.

Figure 10.5 The standard normal distribution



Some of the areas in a standard normal distribution are shown in Figure 10.5. Since the distribution is symmetric, half of the values are greater than 0, and half are less. Also, the area to the right of any given positive score is the same as the area to the left of the same negative score. For example, 16% of cases have standardized scores greater than +1, and 16% of cases have standardized scores less than −1. Appendix D gives areas of

NORUŠIS, Marija J. 1998. *Guide to Data Analysis*. Upper Saddle River, Prentice Hall.

the normal distribution for various standard scores. The exercises show you how to use SPSS to calculate areas in a normal distribution.

? *If you're more than two standard deviations from the mean on some characteristic, does that mean you're abnormal?* Not necessarily. For example, pediatricians often evaluate a child's size by finding percentile values. They may tell the parents that their child is at the 2.5th percentile, or 97.5th percentile for height. (For a normal distribution, these percentiles correspond to standardized scores of –2 and +2.) The small or large percentile values don't necessarily indicate that something is wrong. Even if you took a group of healthy children and looked at their height distribution, some of them would be more than two standard deviations from the mean. Somebody has to fall into the tails of the normal distribution. This also leads to a convincing argument against grading on the curve. Even in a brilliant, hard-working class, some students will receive scores more than 2 standard deviations below the mean. Does that make their performance unacceptable? Not necessarily. ■■■

Samples from a Normal Distribution

If you look again at Figure 10.3, you'll see that the normal distribution that is superimposed on the cholesterol data doesn't fit the data values exactly. The observed data are not perfectly normal. Instead, the distribution of the data values can be described as approximately normal. That's not surprising. Even if you assume that cholesterol values have a perfect normal distribution in the population, you wouldn't expect a sample from this distribution to be exactly normal. You know that a sample is not a perfect picture of the population. You expect that samples from a normal population would appear to be more or less bell shaped, but it would be unrealistic to expect that every sample is exactly normal. In fact, even the population distribution of most variables is not exactly normal. Instead, it's usually the case that the normal distribution is a good approximation. Slight departures from the normal distribution have little effect on statistical analyses that assume that the distribution of data values is normal.

Means from a Normal Population

Since we've established that the normal distribution is a reasonable representation of the distribution of data values for many variables, we can use this information in testing statistical hypotheses about such variables. For example, suppose you want to test whether highly paid CEO's have average cholesterol levels which are different from the population as a whole. In 1991, *Forbes* sent out a survey to the 200 most highly compensated CEO's requesting their cholesterol levels. The 21 CEO's who responded had an average cholesterol of 193 mg/dL. Assume that, in the population, cholesterol levels are normally distributed with a mean of 205 and a standard deviation of 35. Based on this information, how would you determine if the CEO's differ from the rest of us not only in their net worth but in average cholesterol as well?

To answer this question, you need to know whether 193 is an unlikely sample value for the mean, when the true population value is 205. To arrive at this information, you'll follow the same procedure as you did in Chapter 9. However, instead of taking samples from a population in which only two values can occur, you'll take repeated samples from a normal population.

Figure 10.6 Distribution of 500 sample means

To obtain this histogram, open the simul.sav file and select the variable normal21 in the Histograms dialog box.

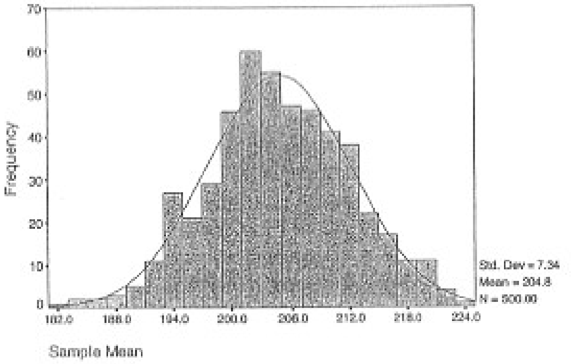


Figure 10.6 shows the distribution of 500 sample means from a normal distribution with a mean of 205 and a standard deviation of 35. Each mean is based on 21 cases. As you can see, the distribution of sample means is also approximately normal. That's always the case when you