CHAPTER 4

Probability Distributions

Compared to most mathematical sciences, statistics is young. Most methods discussed in this book were developed within the past century. By contrast, probability, the subject of this chapter, has a long history. For instance, mathematicians used probability in France in the seventeenth century to evaluate various gambling strategies. Probability is a highly developed subject, but this chapter limits attention to the basics that we'll need for statistical inference.

Following a brief introduction to probability in Section 4.1, Sections 4.2 and 4.3 introduce *probability distributions*, which provide probabilities for all the possible outcomes of a variable. The *normal distribution*, described by a bell-shaped curve, is the most important probability distribution for statistical analysis. Sections 4.4 and 4.5 introduce the *sampling distribution*, a fundamentally important type of probability distribution for statistical inference. It enables us to predict how close a sample mean falls to the population mean. We'll see that the main reason for the importance of the normal distribution is the remarkable result that sampling distributions are often bell shaped.

4.1 INTRODUCTION TO PROBABILITY

In Chapter 2 we learned that randomness is a key component of good ways to gather data. Consider a hypothetical random sample or randomized experiment. For each observation, the possible outcomes are known, but it's uncertain which will occur.

Probability as a Long-Run Relative Frequency

For a particular possible outcome for a random phenomenon, the *probability* of that outcome is the proportion of times that the outcome would occur in a very long sequence of observations.

Probability

With a random sample or randomized experiment, the ***probability*** an observation has a particular outcome is the proportion of times that outcome would occur in a very long sequence of observations.

Later in this chapter, we'll analyze data for the 2006 California gubernatorial election, for which the winner was the Republican candidate, Arnold Schwarzenegger.

Imagine the process of interviewing a random sample of voters in that election and asking whom they voted for. As we interview more and more people, the sample proportion who say they voted for Schwarzenegger gets closer and closer to the population proportion who voted for him. They are the same after we have interviewed everyone in the population of all voters. Suppose that population proportion is 0.56. Then, the probability that a randomly selected person voted for Schwarzenegger is 0.56.

Why does probability refer to the *long run*? Because you need a large number of observations to accurately assess a probability. If you sample only ten people and they are all right-handed, you can't conclude that the probability of being right-handed equals 1.0.

This book defines a probability as a proportion, so it is a number between 0 and 1. In practice, probabilities are often expressed also as percentages, then falling between 0 and 100. For example, if a weather forecaster says that the probability of rain today is 70%, this means that in a long series of days with atmospheric conditions like those today, rain occurs on 70% of the days.

This *long-run* approach to defining probability is not always helpful. If you decide to start a new business, you won't have a long run of trials with which to estimate the probability that the business is successful. You must then rely on *subjective* information rather than solely on *objective* data. In the subjective approach, the probability of an outcome is defined to be your degree of belief that the outcome will occur, based on the available information. A branch of statistics uses subjective probability as its foundation. It is called ***Bayesian statistics***, in honor of a British clergyman (Thomas Bayes) who discovered a probability rule on which it is based. This approach is beyond our scope in this text.

Basic Probability Rules

It's not the purpose of this text to go into detail about the many rules for finding probabilities. Here, we'll briefly mention four rules that are especially useful. We won't try to explain them with precise, mathematical reasoning, because for our purposes it suffices to have an intuitive feel for what each rule says.

Let $P(A)$ denote the probability of a possible outcome or set of outcomes denoted by the letter A. Then

1. **P(not A) = 1 − $P(A)$.**

   If you know the probability a particular outcome occurs, then the probability it does *not* occur is 1 minus that probability. Suppose A represents the outcome that a randomly selected voter cast his or her vote for Schwarzenegger. If $P(A) = 0.56$, then $1 - 0.56 = 0.44$ is the probability of *not* voting for Schwarzenegger, that is, voting instead for the Democratic candidate or some other candidate on the ballot.

2. **If A and B are distinct possible outcomes (with no overlap), then $P(A \text{ or } B) = P(A) + P(B)$.**

   Suppose you take a survey to estimate the population proportion of people who believe that embryonic stem cell research should be banned by the federal government. Let A represent your getting a sample proportion estimate that is much too low, being more than 0.10 *below* the population proportion. Let B represent your sample proportion estimate being much too high—at least 0.10 *above* the population proportion. Using methods from this chapter, perhaps you find that $P(A) = P(B) = 0.03$. Then the overall probability your sample proportion is in error by more than 0.10 (without specifying the direction of error) is

$$P(A \text{ or } B) = P(A) + P(B) = 0.03 + 0.03 = 0.06.$$

3. **If A and B are possible outcomes, then $P(A \text{ and } B) = P(A) \times P(B \text{ given } A)$.**
   From U.S. Census data, the probability that a randomly selected American adult is married equals 0.56. Of those who are married, General Social Surveys indicate that the probability a person reports being *very happy* when asked to choose among (very happy, pretty happy, not too happy) is about 0.40; that is, given you are married, the probability of being very happy is 0.40. So

$$P(\text{married and very happy}) =$$
$$P(\text{married}) \times P(\text{very happy given married}) = 0.56 \times 0.40 = 0.22.$$

   About 22% of the adult population is both married *and* very happy.

   In some cases, A and B are "independent," in the sense that whether one occurs does not depend on whether the other does. That is, $P(B \text{ given } A) = P(B)$, so the previous rule simplifies:

4. **If A and B are independent, then $P(A \text{ and } B) = P(A) \times P(B)$.**
   For example, an inference method presented in the next chapter often is used with the probability of a correct inference set at 0.95. Suppose A represents an inference about men in the population of interest (such as a prediction about the proportion of men who vote for Schwarzenegger) being correct. Let B represent a separate inference about women being correct. Then, since these are independent samples and inferences, the probability that *both* inferences are correct is

$$P(A \text{ and } B) = P(A) \times P(B) = 0.95 \times 0.95 = 0.90.$$

4.2 PROBABILITY DISTRIBUTIONS FOR DISCRETE AND CONTINUOUS VARIABLES

A variable can take at least two different values. For a random sample or randomized experiment, each possible outcome has a probability that it occurs. The variable itself is sometimes then referred to as a ***random variable***. This terminology emphasizes that the outcome varies from observation to observation according to random variation that can be summarized by probabilities. We'll continue to use the simpler "variable" terminology.

Recall from Section 2.1 that a variable is *discrete* if the possible outcomes are a set of separate values, for example, a variable expressed as "the number of..." with possible values 0, 1, 2, ... It is *continuous* if the possible outcomes are an infinite continuum. A ***probability distribution*** lists the possible outcomes and their probabilities. We'll next see how this is done for discrete and for continuous variables.

Probability Distributions for Discrete Variables

The probability distribution of a *discrete* variable assigns a probability to each possible value of the variable. Each probability is a number between 0 and 1. The sum of the probabilities of all possible values equals 1.

Let $P(y)$ denote the probability of a possible outcome for a variable y. Then

$$0 \leq P(y) \leq 1 \text{ and } \Sigma_{\text{all } y} P(y) = 1,$$

where the sum is over all the possible values of the variable.

EXAMPLE 4.1 Ideal Number of Children for a Family

Let y denote the response to the question, "What do you think is the ideal number of children for a family to have?" This is a discrete variable, taking the possible values 0, 1, 2, 3, and so forth. According to results from a GSS for a randomly chosen person in the U.S. the probability distribution of y is approximately as Table 4.1 shows. The

TABLE 4.1: Probability Distribution of y = Ideal Number of Children for a Family

y	$P(y)$
0	0.01
1	0.03
2	0.60
3	0.23
4	0.12
5	0.01
Total	1.0

table displays the recorded y-values and their probabilities. For instance, $P(4)$, the probability that $y = 4$ children is regarded as ideal, equals 0.12. Each probability in Table 4.1 is between 0 and 1, and the sum of the probabilities equals 1. ■

A *histogram* can portray the probability distribution. The rectangular bar over a possible value of the variable has height equal to the probability of that value. Figure 4.1 is a histogram for the probability distribution of the ideal number of children, from Table 4.1. The bar over the value 4 has height 0.12, the probability of the outcome 4.

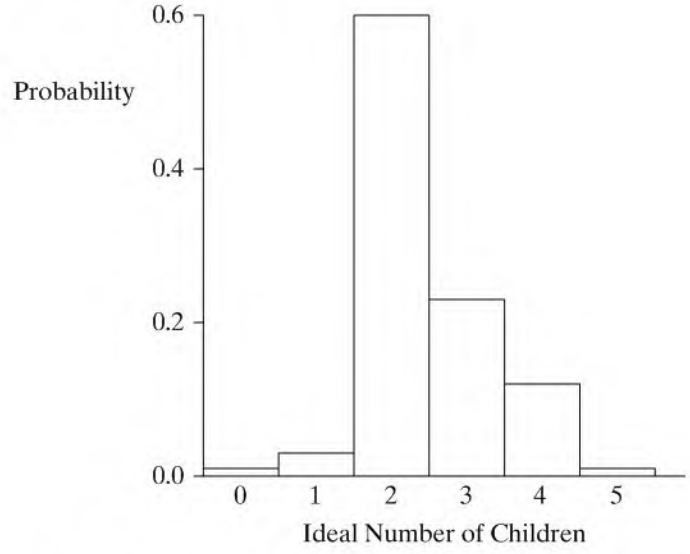

FIGURE 4.1: Histogram for the Probability Distribution of the Ideal Number of Children for a Family

Probability Distributions for Continuous Variables

Continuous variables have an infinite continuum of possible values. Probability distributions of continuous variables assign probabilities to *intervals* of numbers. The probability that a variable falls in any particular interval is between 0 and 1, and the probability of the interval containing all the possible values equals 1.

A graph of the probability distribution of a continuous variable is a smooth, continuous curve. The *area* under the curve for an interval of values represents the probability that the variable takes a value in that interval.

EXAMPLE 4.2 Commuting Time to Work

A recent study about commuting time for workers in the U.S. who commute to work[1] measured y = travel time (in minutes). The probability distribution of y provides probabilities such as $P(y < 10)$, the probability that travel time is less than 10 minutes, or $P(30 < y < 60)$, the probability that travel time is between 30 and 60 minutes.

Figure 4.2 portrays the approximate probability distribution of y. The shaded area in the figure refers to the region of values higher than 45. This area equals 15% of the total area under the curve, representing the probability of 0.15 that commuting time is more than 45 minutes. Those regions in which the curve has relatively high height have the values most likely to be observed. ■

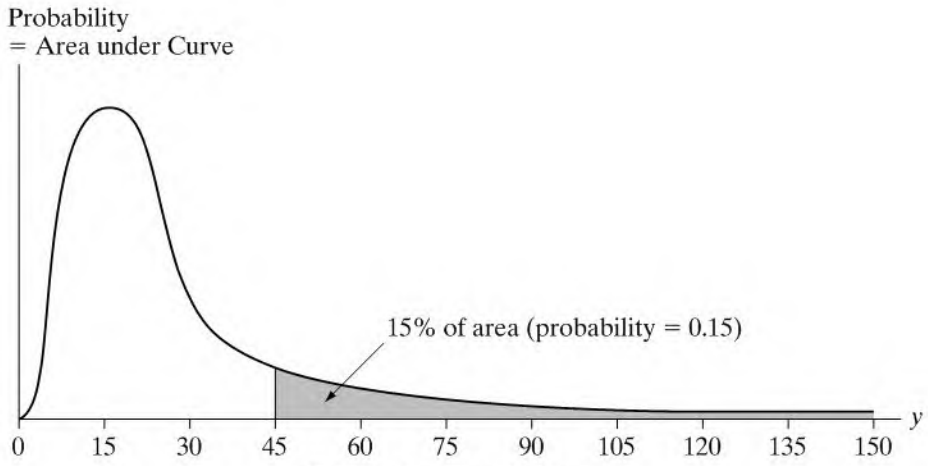

FIGURE 4.2: Probability Distribution of Commuting Time to Work. The area under the curve between two points represents the probability of that interval of values.

Parameters Describe Probability Distributions

Most probability distributions have formulas for calculating probabilities. For others, tables or graphs provide the probabilities. Section 4.3 shows how to calculate probabilities for the most important probability distribution.

Section 3.1 introduced the *population distribution* of a variable. This is, equivalently, the probability distribution of the variable for a subject selected randomly from the population. For example, if 0.12 is the population proportion of adults who believe the ideal number of children is 4, then the probability that an adult selected randomly from that population believes this is also 0.12.

Like a population distribution, a probability distribution has *parameters* describing center and variability. The *mean* describes center and the *standard deviation* describes variability. The parameter values are the values these measures would assume, in the long run, if the randomized experiment or random sample repeatedly took observations on the variable y having that probability distribution.

For example, suppose we take observations from the distribution in Table 4.1. Over the long run, we expect $y = 0$ to occur 1% of the time, $y = 1$ to occur 3% of the time, and so forth. In 100 observations, for instance, we expect about

one 0, 3 1's, 60 2's, 23 3's, 12 4's, and one 5.

[1] *Journey to Work*, issued 2004 by U.S. Census Bureau.

In that case, since the mean equals the total of the observations divided by the sample size, the mean equals

$$\frac{(1)0 + (3)1 + (60)2 + (23)3 + 12(4) + 1(5)}{100} = \frac{245}{100} = 2.45.$$

This calculation has the form

$$0(0.01) + 1(0.03) + 2(0.60) + 3(0.23) + 4(0.12) + 5(0.01),$$

the sum of the possible outcomes times their probabilities. In fact, for any discrete variable y, the mean of its probability distribution has this form.

Mean of a Probability Distribution (Expected Value)

The ***mean of the probability distribution*** for a discrete variable y is

$$\mu = \sum yP(y).$$

The sum is taken over all possible values of the variable. This parameter is also called the ***expected value of y*** and is denoted by $E(y)$.

For Table 4.1, for example,

$$\begin{aligned}\mu = \sum yP(y) &= 0P(0) + 1P(1) + 2P(2) + 3P(3) + 4P(4) + 5P(5) \\ &= 0(0.01) + 1(0.03) + 2(0.60) + 3(0.23) + 4(0.12) + 5(0.01) \\ &= 2.45.\end{aligned}$$

This is also the *expected value* of y, $E(y) = \mu = 2.45$. The terminology reflects that $E(y)$ represents what we expect for the average value of y in a long series of observations.

The ***standard deviation*** of a probability distribution, denoted by σ, measures its variability. The larger the value of σ, the more spread out the distribution. In a rough sense, σ describes how far the variable y tends to fall from the mean of its distribution. The Empirical Rule (Section 3.3) helps us to interpret σ. If a probability distribution is approximately bell shaped, about 68% of the probability falls between $\mu - \sigma$ and $\mu + \sigma$, about 95% falls between $\mu - 2\sigma$ and $\mu + 2\sigma$, and all or nearly all falls between $\mu - 3\sigma$ and $\mu + 3\sigma$. For example, suppose commuting time to work for a randomly selected worker in Toronto has a bell-shaped probability distribution with $\mu = 24$ minutes and $\sigma = 8$ minutes. Then there's about a 95% chance that commuting time falls between $24 - 2(8) = 8$ minutes and $24 + 2(8) = 40$ minutes.

The standard deviation is the square root of the ***variance*** of the probability distribution. The variance measures the average squared deviation of an observation from the mean. That is, it is the expected value of $(y - \mu)^2$. We shall not need to compute this measure, so we do not study its formula here (Exercise 4.55 shows the formula for σ for the discrete case).

4.3 THE NORMAL PROBABILITY DISTRIBUTION

Some probability distributions are important because they approximate well the distributions of variables in the real world. Some are important because of their uses in statistical inference. This section introduces the ***normal probability distribution***, which is important for both reasons. Its bell-shaped curve describes well many histograms of data for variables that are continuous or take a large number of

possible values. It is the most important distribution for statistical inference, because we'll see it is still useful even when the sample data are *not* bell shaped.

Normal Distribution

The ***normal distribution*** is symmetric, bell shaped, and characterized by its mean μ and standard deviation σ. The probability within any particular number of standard deviations of μ is the same for all normal distributions. This probability equals 0.68 within 1 standard deviation, 0.95 within 2 standard deviations, and 0.997 within 3 standard deviations.

Each normal distribution is specified by two parameters—its mean μ and standard deviation σ. For any real number for μ and any nonnegative number for σ, there is a normal distribution having that mean and standard deviation. Figure 4.3 illustrates. Essentially the entire distribution falls between $\mu - 3\sigma$ and $\mu + 3\sigma$.

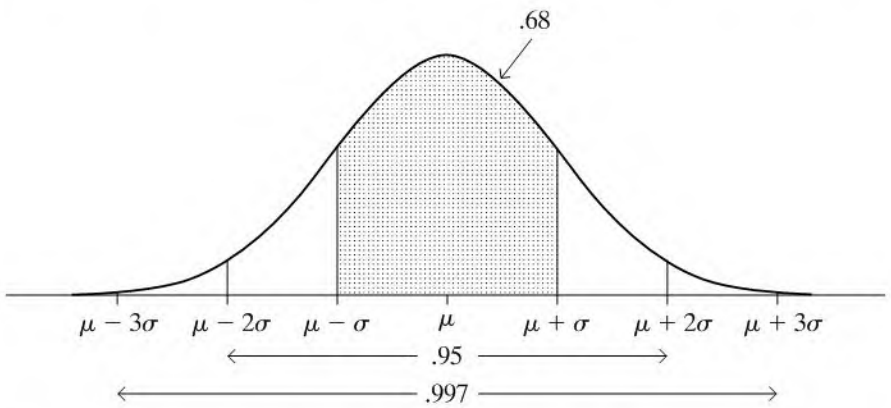

FIGURE 4.3: For Every Normal Distribution, the Probability Equals (Rounded) 0.68 within σ of μ, 0.95 within 2σ of μ, and 0.997 within 3σ of μ

For example, heights of adult females in North America have approximately a normal distribution with mean $\mu = 65.0$ inches and standard deviation $\sigma = 3.5$. The probability is nearly 1.0 that a randomly selected female has height between $\mu - 3\sigma = 65.0 - 3(3.5) = 54.5$ inches and $\mu + 3\sigma = 65.0 + 3(3.5) = 75.5$ inches. Adult male height has a normal distribution with $\mu = 70.0$ and $\sigma = 4.0$ inches. So the probability is nearly 1.0 that a randomly selected male has height between $\mu - 3\sigma = 70.0 - 3(4.0) = 58$ inches and $\mu + 3\sigma = 70.0 + 3(4.0) = 82$ inches. See Figure 4.4.

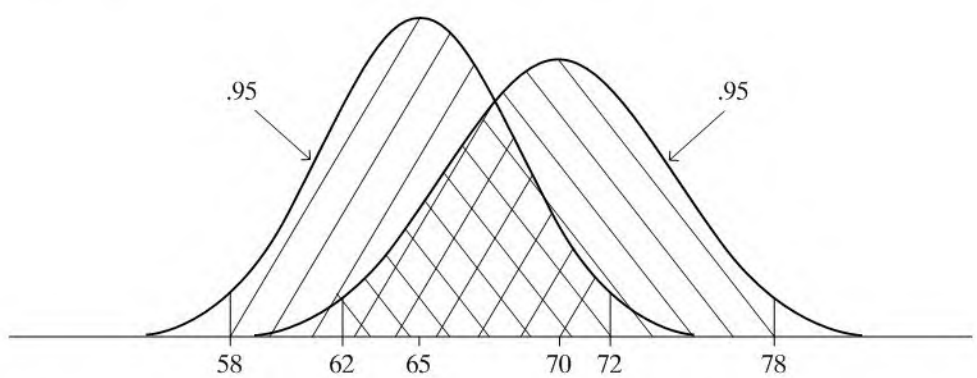

FIGURE 4.4: Normal Distributions for Women's Height ($\mu = 65, \sigma = 3.5$) and for Men's Height ($\mu = 70, \sigma = 4.0$)

Tabulated Normal Tail Probabilities

For the normal distribution, for each fixed number z, the probability falling within z standard deviations of the mean depends only on the value of z. This is the area under the bell-shaped normal curve between $\mu - z\sigma$ and $\mu + z\sigma$. For every normal distribution, this probability is 0.68 for $z = 1$, 0.95 for $z = 2$, and nearly 1.0 for $z = 3$.

For a normal distribution, the probability concentrated within $z\sigma$ of μ is the same for all normal curves even if z is not a whole number—for instance $z = 1.43$ instead of 1, 2, or 3. Table A in Appendix A, also shown next to the inside back cover, determines probabilities for any region of values. It tabulates the probability for the values falling in the right tail, at least z standard deviations above the mean. The left margin column of the table lists the values for z to one decimal point, with the second decimal place listed above the columns.

TABLE 4.2: Part of Table A Displaying Normal Right-Tail Probabilities

	Second Decimal Place of z									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
					....					
					....					
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
					....					
					....					

Table 4.2 displays a small excerpt from Table A. The probability for $z = 1.43$ falls in the row labeled 1.4 and in the column labeled .03. It equals 0.0764. This means that for every normal distribution, the right-tail probability above $\mu + 1.43\sigma$ (that is, more than 1.43 standard deviations above the mean) equals 0.0764.

Since the entries in Table A are probabilities for the right half of the normal distribution above $\mu + z\sigma$, they fall between 0 and 0.50. By the symmetry of the normal curve, these right-tail probabilities also apply to the left tail below $\mu - z\sigma$. For example, the probability below $\mu - 1.43\sigma$ also equals 0.0764. The left-tail probabilities, called *cumulative probabilities*, are given by many calculators and software.

Normal Probabilities and the Empirical Rule

The probabilities in Table A apply to the normal distribution and also apply approximately to other bell-shaped distributions. This table yields the probabilities for the Empirical Rule. That rule states that for bell-shaped histograms, about 68% of the data fall within 1 standard deviation of the mean, 95% within 2 standard deviations, and all or nearly all within 3 standard deviations.

For example, the value two standard deviations above the mean has a z-value of 2.00. The normal curve probability listed in Table A opposite $z = 2.00$ is 0.0228. The right-tail probability above $\mu + 2\sigma$ equals 0.0228 for every normal distribution. The left-tail probability below $\mu - 2\sigma$ also equals 0.0228, by symmetry (see Figure 4.5). The total probability more than two standard deviations from the mean is $2(0.0228) = 0.0456$. Since the probability more than two standard deviations from the mean equals 0.0456, the probability between $\mu - 2\sigma$ and $\mu + 2\sigma$ (i.e., within two standard deviations of the mean) equals $1 - 0.0456 = 0.9544$. (Here, we've used rule (1) of the probability rules at the end of Section 4.1, that $P(\text{not } A) = 1 - P(A)$.) When a

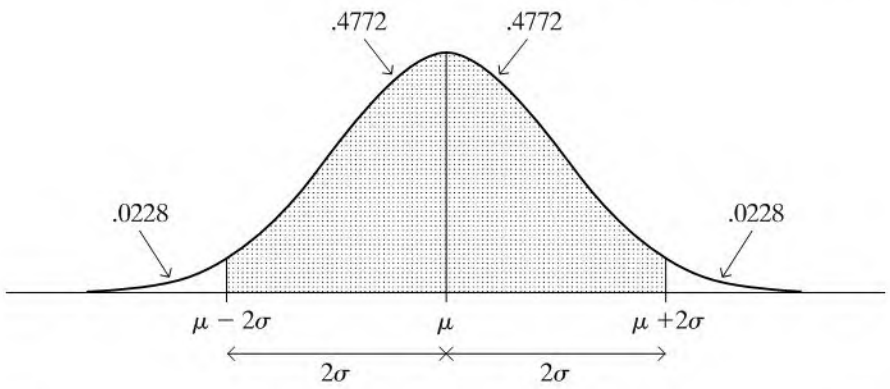

FIGURE 4.5: The Probability within Two Standard Deviations of the Mean for a Normal Distribution Is $1 - 2(0.0228)$, Which Is about 0.95

variable has a normal distribution, 95% of the observations fall within two standard deviations of the mean.

The probability equals 0.50 above the mean, since the normal distribution is symmetric about μ. So the probability between μ and $\mu + 2\sigma$ or between $\mu - 2\sigma$ and μ equals $0.50 - 0.0228 = 0.4772$, also shown in Figure 4.5. Again, we see that the total probability within two standard deviations of the mean equals $2(0.4772) = 0.9544$, or about 95%.

The approximate percentages in the Empirical Rule are the actual percentages for the normal distribution, rounded to two decimal places. For instance, with Table A you can verify that the probability within one standard deviation of the mean of a normal distribution equals 0.68. (*Hint:* Let $z = 1.00$.) The Empirical Rule stated the percentages as being *approximate* rather than *exact*. Why? Because that rule referred to *all approximately bell-shaped distributions*, not just the normal. Not all bell-shaped distributions are normal, only those described by the mathematical formula shown in Exercise 4.56 at the end of the chapter. We won't need that formula, but we will use the probabilities tabulated for it in Table A throughout the text.

Finding *z*-Values for Certain Tail Probabilities

Many inferential methods use z-values corresponding to certain normal curve probabilities. This entails the reverse use of Table A. Starting with a tail probability, which is listed in the body of Table A, we find the z-value that provides the number of standard deviations that that number falls from the mean.

To illustrate, let's find the z-value having a right-tail probability of 0.025. We look up 0.025 in the body of Table A. It corresponds to $z = 1.96$. This means that a probability of 0.025 falls above $\mu + 1.96\sigma$. Similarly, a probability of 0.025 falls below $\mu - 1.96\sigma$. So a total probability of $0.025 + 0.025 = 0.050$ falls more than 1.96σ from μ. We saw in the previous subsection that 95% of a normal distribution falls within two standard deviations of the mean. More precisely, 0.9544 falls within 2.00 standard deviations, and here we've seen that 0.950 falls within 1.96 standard deviations.

To check that you understand this reasoning, verify that the z-value for a right-tail probability of (1) 0.05 is $z = 1.64$, (2) 0.01 is $z = 2.33$, (3) 0.005 is $z = 2.58$. Show that 90% of a normal distribution falls between $\mu - 1.64\sigma$ and $\mu + 1.64\sigma$.

EXAMPLE 4.3 Finding the 99th Percentile of IQ Scores

Stanford-Binet IQ scores have approximately a normal distribution with mean = 100 and standard deviation = 16. What is the 99th percentile of IQ scores? In other words, what is the IQ score that falls above 99% of the scores?

To answer this, we need to find the value of z such that $\mu + z\sigma$ falls above 99% of a normal distribution. We use the normal curve probability in the right tail beyond the 99th percentile. Then we can use Table A to find the z-value corresponding to that probability. Now, for $\mu + z\sigma$ to represent the 99th percentile, the probability below $\mu + z\sigma$ must equal 0.99, by the definition of a percentile. So 1% of the distribution is above the 99th percentile. The right-tail probability equals 0.01, as Figure 4.6 shows.

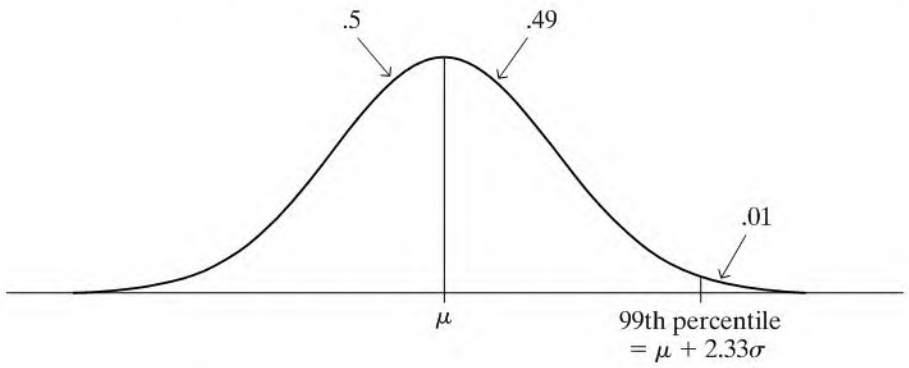

FIGURE 4.6: The 99th Percentile for a Normal Distribution Has 99% of the Distribution below that Point and 1% above It

The body of Table A does not contain the tail probability of exactly 0.0100. The tail probability = 0.0102 corresponds to $z = 2.32$, and tail probability = 0.0099 corresponds to $z = 2.33$. We could interpolate, but it is sufficient to use the z-value rounded to two decimal places. We select the one having probability closer to the desired value of 0.0100. Thus, the 99th percentile is 2.33 standard deviations above the mean. In summary, 99% of any normal distribution is located below $\mu + 2.33\sigma$, no more than 2.33 standard deviations above the mean.

For the IQ scores with mean = 100 and standard deviation = 16, the 99th percentile equals

$$\mu + 2.33\sigma = 100 + 2.33(16) = 137.$$

That is, about 99% of IQ scores fall below 137. ■

To check that you understand the reasoning above, show that the 95th percentile of a normal distribution is $\mu + 1.64\sigma$, and show that the 95th percentile for the IQ distribution equals 126.

z-Score Is Number of Standard Deviations from Mean

The z symbol in a normal table refers to the distance between a possible value y of a variable and the mean μ of its probability distribution, in terms of the *number of standard deviations* that y falls from μ.

EXAMPLE 4.4 *z*-Scores for College Board Test Scores

Scores on each portion of the Scholastic Aptitude Test (SAT), a college entrance examination, have traditionally been approximately normal with mean $\mu = 500$ and standard deviation $\sigma = 100$. The test score of $y = 650$ has a z-score of $z = 1.50$, because 650 is 1.50 standard deviations above the mean. In other words, $y = 650 = \mu + z\sigma = 500 + z(100)$, where $z = 1.50$. ■

For sample data, Section 3.4 introduced the z-score as a measure of position. Let's review how to find it. The distance between y and the mean μ equals $y - \mu$. The z-score expresses this difference in units of standard deviations.

z-Score

The ***z-score*** for a value y of a variable is the *number of standard deviations* that y falls from μ. It equals

$$z = \frac{\text{Observation} - \text{Mean}}{\text{Standard Deviation}} = \frac{y - \mu}{\sigma}.$$

To illustrate, when $\mu = 500$ and $\sigma = 100$, an observation of $y = 650$ has the z-score of

$$z = \frac{y - \mu}{\sigma} = \frac{650 - 500}{100} = 1.50.$$

Positive z-scores occur when the number y falls *above* the mean μ. *Negative* z-scores occur when the number y falls *below* the mean. For example, for SAT scores with $\mu = 500$ and $\sigma = 100$, a value of $y = 350$ has a z-score of

$$z = \frac{y - \mu}{\sigma} = \frac{350 - 500}{100} = -1.50.$$

The test score of 350 is 1.50 standard deviations below the mean. The value $y = 350$ falls below the mean, so the z-score is negative.

Table A contains only positive z-values. Since the normal distribution is symmetric about the mean, the left-tail probability below $-z$ equals the right-tail probability above $+z$. Looking up $z = 1.50$ in Table A, we see that the probability that a SAT score falls below 350 is 0.0668, as Figure 4.7 shows. Fewer than 7% of the scores are below 350, and fewer than 7% fall above 650.

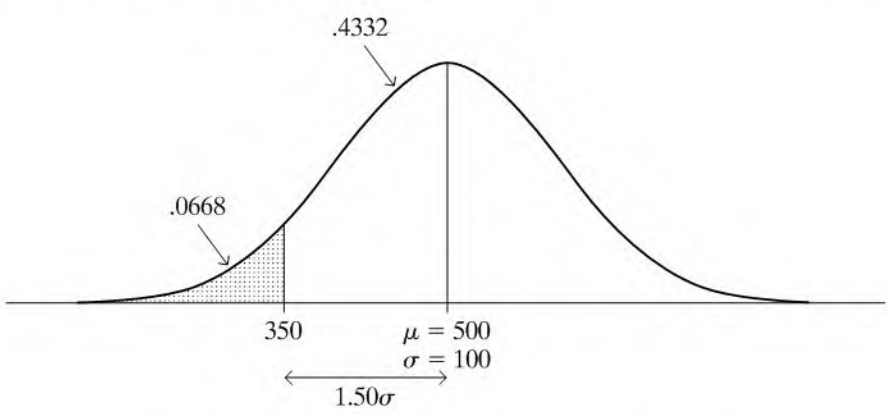

FIGURE 4.7: Normal Distribution for SAT Scores

The next example shows that z-scores provide a useful way to compare positions for different normal distributions.

EXAMPLE 4.5 Comparing SAT and ACT Test Scores

Suppose that when you applied to college, you took a SAT exam, scoring 550. Your friend took the ACT exam, scoring 30. Which score is better?

We cannot compare the test scores of 550 and 30 directly, because they have different scales. We convert them to z-scores, analyzing how many standard deviations each falls from the mean. If SAT has $\mu = 500$ and $\sigma = 100$, a SAT score of $y = 550$ converts to a z-score of

$$z = \frac{(y - \mu)}{\sigma} = \frac{(550 - 500)}{100} = 0.50.$$

The ACT has approximately $\mu = 18$ and $\sigma = 6$, so ACT $= 30$ converts to a z-score of $(30 - 18)/6 = 2.0$.

The ACT score of 30 is relatively higher than the SAT score of 650, because 30 is 2.0 standard deviations above its mean whereas 550 is only 0.5 standard deviations above its mean. The SAT and ACT scores both have approximate normal distributions. From Table A, $z = 2.0$ has a right-tail probability of 0.0228 and $z = 0.5$ has a right-tail probability of 0.3085. Of all students taking the ACT, only about 2% scored higher than 30, whereas of all students taking the SAT, about 31% scored higher than 550. In this relative sense, the ACT score is higher. ■

Here's a summary of how we've used z-scores:

Using *z*-Scores to Find Probabilities or *y*-Values

- If we have a value y and need to find a probability, convert y to a z-score using $z = (y - \mu)/\sigma$, and use a table of normal probabilities to convert it to the probability of interest.
- If we have a probability and need to find a value of y, convert the probability to a tail probability and find the z-score using a normal table, and then evaluate $y = \mu + z\sigma$.

For example, Example 4.6 used the equation $z = (y - \mu)/\sigma$ to determine how many standard deviations an SAT test score fell from the mean. Example 4.3 used the equation $y = \mu + z\sigma$ to find a percentile score for a normal distribution of IQ scores.

The Standard Normal Distribution

Many inferential statistical methods use a particular normal distribution, called the ***standard normal distribution***.

Standard Normal Distribution

The ***standard normal distribution*** is the normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

For the standard normal distribution, the number falling z standard deviations above the mean is $\mu + z\sigma = 0 + z(1) = z$. It is simply the z-score itself. For instance, the value of 2 is two standard deviations above the mean, and the value of -1.3 is 1.3 standard deviations below the mean. The original values are the same as the z-scores. See Figure 4.8.

When the values for an arbitrary normal distribution are converted to z-scores, those z-scores are centered around 0 and have a standard deviation of 1. The z-scores have the standard normal distribution.

***z*-Scores and the Standard Normal Distribution**

If a variable has a normal distribution, and if its values are converted to z-scores by subtracting the mean and dividing by the standard deviation, then the z-scores have the standard normal distribution.

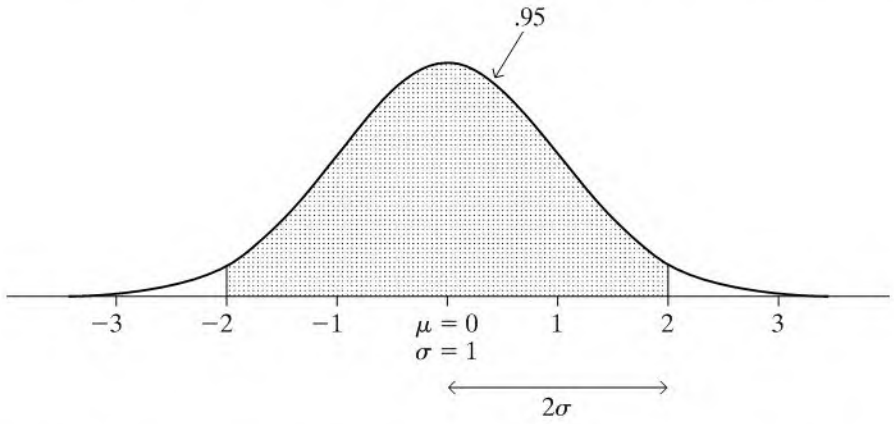

FIGURE 4.8: The Standard Normal Distribution Has Mean 0 and Standard Deviation 1. Its ordinary scores are the same as its ***z-scores***.

Suppose we convert each SAT score y to a z-score by using $z = (y - 500)/100$. For instance, $y = 650$ converts to $z = 1.50$, and $y = 350$ converts to $z = -1.50$. Then the entire set of z-scores has a normal distribution with a mean of 0 and a standard deviation of 1. This is the standard normal distribution.

Many inferential methods convert values of statistics to z-scores and then to normal curve probabilities. We use z-scores and normal probabilities often throughout the rest of the book.

4.4 SAMPLING DISTRIBUTIONS DESCRIBE HOW STATISTICS VARY

We've seen that probability distributions summarize probabilities of possible outcomes for a variable. So far, this chapter has treated these distributions as known. In practice, they are rarely known. We use sample data to make inferences about the parameters of those distributions. However, probability distributions with fixed parameter values are useful for many of those inferential methods. Let's look at an example that illustrates the connection between statistical inference and probability calculations with known parameter values.

EXAMPLE 4.6 Predicting an Election from an Exit Poll

Television networks sample voters on election day to help them predict the winners early. For the fall 2006 election for Governor of California, CNN[2] reported results of an exit poll of 2705 voters. CNN stated that 56.5% reported voting for the Republican candidate, Arnold Schwarzenegger. In this example, the probability distribution for a person's vote would state the probability that a randomly selected voter voted for Schwarzenegger. This equals the proportion of the population of voters who voted for him. When the exit poll was taken, this was an unknown population parameter.

To judge whether this is sufficient information to predict the outcome of the election, the network can ask, "Suppose only half the population voted for Schwarzenegger. Would it then be surprising that 56.5% of the sampled individuals voted for him?" If this would be very unlikely, the network infers that Schwarzenegger received more than half the population votes. The inference about the election outcome is based on finding the probability of the sample result under the supposition that the population parameter, the percentage of voters preferring Schwarzenegger, equals 50%. ■

[2]www.cnn.com/ELECTION/2006

Nearly 7 million people voted in this race. The exit poll sampled only 2705 voters, yet TV networks used it to predict that Schwarzenegger would win. How could there possibly have been enough information from this poll to make a prediction? We next see justification for making a prediction.

Simulating the Estimation Process

A ***simulation*** can tell us how well an exit poll result approximates the population proportion voting for a candidate. We simulate the vote of a voter randomly chosen from the population by selecting a two-digit random number from a random number table (such as Table 2.2) or software. Suppose exactly 50% of the population voted for Schwarzenegger and 50% voted for the Democratic candidate, Phil Angelides. Identify all 50 two-digit numbers between 00 and 49 as Republican votes and all 50 two-digit numbers between 50 and 99 as Democrat votes. Then each candidate has a 50% chance of selection on each choice of two-digit random number.

For instance, the first two digits of the first column of Table 2.2 provide the random numbers 10, 22, 24, 42, 37, 77, and so forth. So of the first 6 voters selected, 5 voted Republican (i.e., have numbers between 00 and 49). Selecting 2705 two-digit numbers simulates the process of observing the votes of a random sample of 2705 voters of the much larger population (which is actually treated as infinite in size).

Using a computer, we selected 2705 random two-digit numbers and got 1334 Republican votes and 1371 Democrat votes. (You can try this yourself with *applets* available on the Internet. See Exercise 4.41.) The sample proportion of Republican votes was $1334/2705 = 0.493$, quite close to the population proportion of 0.50. This particular estimate was good. Were we merely lucky? We repeated the process and selected 2705 more two-digit random numbers. This time the sample proportion of Republican votes was 0.511, also quite good. We next programmed the computer to perform this process of picking 2705 people a million times so we could search for a pattern in the results. Figure 4.9 shows a histogram of the million values of the sample proportion. Nearly all the simulated proportions fell between 0.47 and 0.53; that is, within 0.03 of the population proportion of 0.50. Apparently a sample of size 2705 provides quite a good estimate of a population proportion.

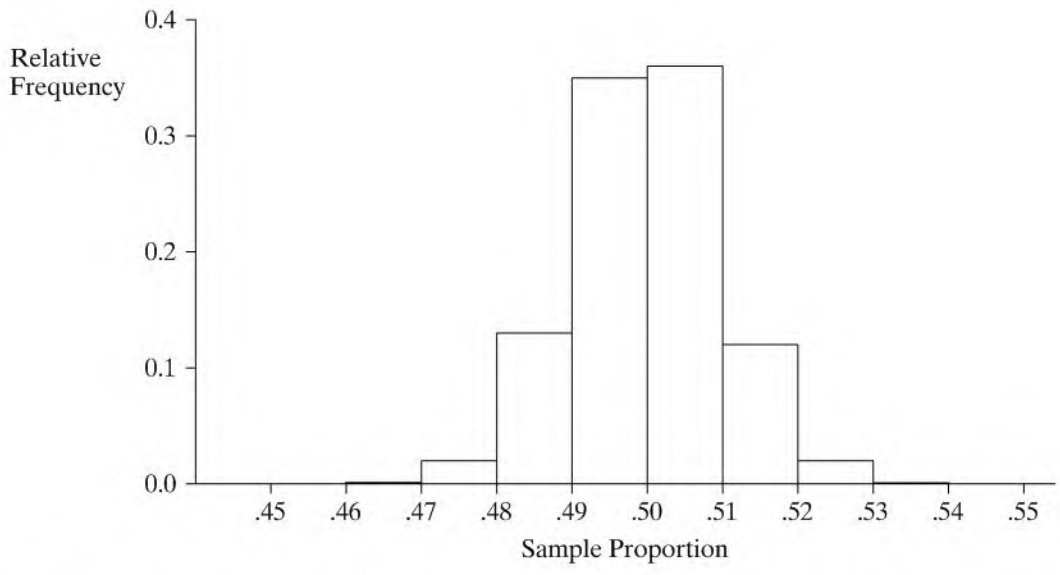

FIGURE 4.9: Results of Simulating the Sample Proportion Favoring the Republican Candidate, for Random Samples of 2705 Subjects from a Population in which Half Voted for Each Candidate. In nearly all cases, the sample proportion fell within 0.03 of the population proportion of 0.50.

In summary, if half the population of voters had voted for Schwarzenegger, we would have expected between about 47% and 53% of voters in an exit poll of size 2705 to have voted for him. So it would have been very unusual to observe 56.5% voting for him, as happened in the actual exit poll. If *less than half* the population voted for Schwarzenegger, it would have been even more unusual to observe this. This is the basis of the network's prediction, from its exit poll, that Schwarzenegger won the election.

It is possible to perform this simulation using any population proportion value. For instance, we could simulate sampling when the population proportion voting for the Republican is 0.45 by letting the 45 random numbers between 00 and 44 represent Republican votes and the 55 between 45 and 99 represent Democrat votes. Likewise, we could change the size of each random sample in the simulation to study the impact of the sample size. From results of the next section, for a random sample of size 2705, the sample proportion is very likely to fall within 0.03 of the population proportion, regardless of its value.

Representing Sampling Variability by a Sampling Distribution

Voter preference is a variable, varying among voters. Likewise, so is the sample proportion voting for a given candidate a variable: Before the sample is obtained, its value is unknown, and that value varies from sample to sample. If several random samples of size $n = 2705$ each were selected, a certain predictable amount of variation would occur in the sample proportion values. A probability distribution with appearance similar to Figure 4.9 describes the variation that occurs from repeatedly selecting samples of a certain size n and forming a particular statistic. This distribution is called a *sampling distribution*. It also provides probabilities of the possible values of the statistic for a *single* sample of size n.

Sampling Distribution

A ***sampling distribution*** of a statistic is the probability distribution that specifies probabilities for the possible values the statistic can take.

Each sample statistic has a sampling distribution. There is a sampling distribution of a sample mean, a sampling distribution of a sample proportion, a sampling distribution of a sample median, and so forth. A sampling distribution is merely a type of probability distribution. Unlike the distributions studied so far, a sampling distribution specifies probabilities not for individual observations but for possible values of a statistic computed from the observations. A sampling distribution allows us to calculate, for example, probabilities about the sample proportion of individuals who voted for the Republican in an exit poll. Before the voters are selected for the exit poll, this is a variable. It has a sampling distribution that describes the probabilities of the possible values.

The sampling distribution is important in inferential statistics because it helps us predict how close a statistic falls to the parameter it estimates. From Figure 4.9, for instance, with a sample of size 2705 the probability is apparently high that a sample proportion falls within 0.03 of the population proportion.

EXAMPLE 4.7 Constructing a Sampling Distribution

It is sometimes possible to construct the sampling distribution without resorting to simulation or complex mathematical derivations. To illustrate, we construct the sampling distribution of the sample proportion, for an exit poll of $n = 4$ voters from

a population in which half voted for each candidate. For each voter, define the y variable representing the vote as follows:

$y = 1$, vote for the Republican
$y = 0$, vote for the Democrat.

We use a symbol with four entries to represent the y-values for a potential sample of size 4. For instance, (1, 0, 0, 1) represents a sample in which the first and fourth subjects voted for the Republican and the second and third subjects voted for the Democrat. The 16 possible samples are

$$(1,1,1,1)\ (1,1,1,0)\ (1,1,0,1)\ (1,0,1,1)$$
$$(0,1,1,1)\ (1,1,0,0)\ (1,0,1,0)\ (1,0,0,1)$$
$$(0,1,1,0)\ (0,1,0,1)\ (0,0,1,1)\ (1,0,0,0)$$
$$(0,1,0,0)\ (0,0,1,0)\ (0,0,0,1)\ (0,0,0,0).$$

Since half the population voted for each candidate, the 16 samples are equally likely.

Now let's construct the sampling distribution of the proportion of the sample that voted for the Republican candidate. For a sample of size 4, that proportion can be 0, 0.25, 0.50, 0.75, or 1.0. The proportion 0 occurs with only one of the 16 possible samples, (0, 0, 0, 0), so its probability equals $1/16 = 0.0625$. The proportion 0.25 occurs for four samples, (1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0), and (0, 0, 0, 1), so its probability equals $4/16 = 0.25$. Based on this reasoning, Table 4.3 shows the probability for each possible sample proportion value.

TABLE 4.3: Sampling Distribution of Sample Proportion, for Random Sample of Size $n = 4$ when Population Proportion Is 0.50. For example, a sample proportion of 0.0 occurs for only 1 of 16 possible samples, namely (0, 0, 0, 0), so its probability is $1/16 = 0.0625$.

Sample Proportion	Probability
0.0	0.0625
0.25	0.2500
0.50	0.3750
0.75	0.2500
1.0	0.0625

Figure 4.10 portrays the sampling distribution of the sample proportion for $n = 4$. It is much more spread out than the one in Figure 4.9 for samples of size $n = 2705$, which falls nearly entirely between 0.47 and 0.53. With such a small sample ($n = 4$), the sample proportion need not be near the population proportion. This is not surprising. In practice, samples are usually much larger than $n = 4$. We used a small value in this example so it was simpler to write down all the potential samples and find probabilities for the sampling distribution. ■

With the two possible outcomes denoted by 0 and 1, Section 3.2 observed that the proportion of times that 1 occurs is the sample mean of the data. For instance, for the sample (0, 1, 0, 0) in which only the second subject voted for the Republican, the sample mean equals $(0 + 1 + 0 + 0)/4 = 1/4 = 0.25$, the sample proportion voting

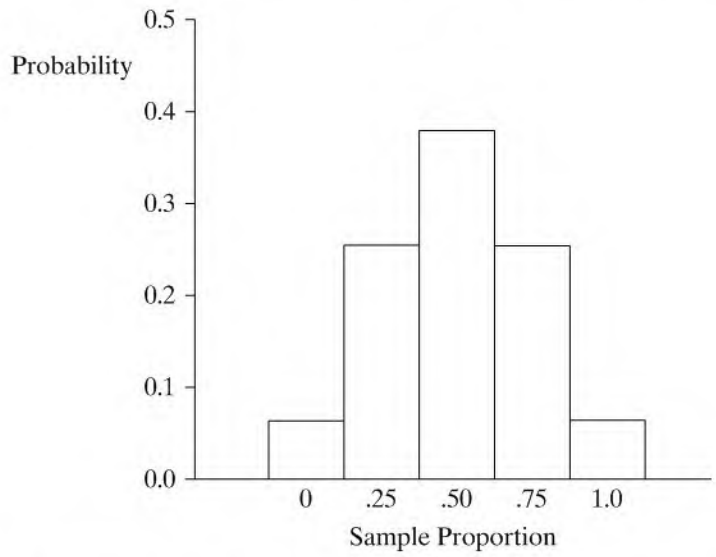

FIGURE 4.10: Sampling Distribution of Sample Proportion, for Random Sample of Size $n = 4$ when Population Proportion Is 0.50

for the Republican. So Figure 4.10 is also an example of a sampling distribution of a sample mean. Section 4.5 provides some general results about the sampling distribution of the sample mean.

Repeated Sampling Interpretation of Sampling Distributions

Sampling distributions portray the sampling variability that occurs in collecting data and using sample statistics to estimate parameters. If different polling organizations each take their own exit poll and estimate the population proportion voting for the Republican candidate, they will get different estimates, because the samples have different people. Likewise, Figure 4.9 describes the variability in sample proportion values that occurs in selecting a huge number of samples of size $n = 2705$ and constructing a histogram of the sample proportions. By contrast, Figure 4.10 describes the variability for a huge number of samples of size $n = 4$.

A sampling distribution of a statistic based on n observations is the relative frequency distribution for that statistic resulting from repeatedly taking samples of size n, each time calculating the statistic value. It's possible to form such a distribution empirically, as in Figure 4.9, by repeated sampling or through simulation. In practice, this is not necessary. The form of sampling distributions is often known theoretically, as shown in the previous example and in the next section. We can then find probabilities about the value of the sample statistic for one sample of the given size n.

4.5 SAMPLING DISTRIBUTIONS OF SAMPLE MEANS

Because the sample mean $\bar{y}$ is used so much, its sampling distribution merits special attention. In practice, when we analyze data and find $\bar{y}$, we don't know how close it falls to the population mean μ, because we do not know the value of μ. Using information about the spread of the sampling distribution, though, we can predict how close it falls. For example, the sampling distribution might tell us that with high probability, $\bar{y}$ falls within 10 units of μ.

In this section we'll learn two main results about the sampling distribution of the sample mean. One provides formulas for the center and spread of the sampling distribution. The other describes its shape.

Mean and Standard Error of Sampling Distribution of $\bar{y}$

The sample mean $\bar{y}$ is a variable, because its value varies from sample to sample. For random samples, it fluctuates around the population mean μ, sometimes being smaller and sometimes being larger. In fact, the mean of the sampling distribution of $\bar{y}$ equals μ. If we repeatedly took samples, then in the long run, the mean of the sample means would equal the population mean μ.

The spread of the sampling distribution of $\bar{y}$ is described by its standard deviation, which is called the *standard error* of $\bar{y}$.

Standard Error

The standard deviation of the sampling distribution of $\bar{y}$ is called the ***standard error*** of $\bar{y}$. The standard error of $\bar{y}$ is denoted by $\sigma_{\bar{y}}$.

The standard error describes how $\bar{y}$ varies from sample to sample. Suppose we repeatedly selected samples of size n from the population, finding $\bar{y}$ for each set of n observations. Then, in the long run, the standard deviation of the $\bar{y}$-values would equal the standard error. The symbol $\sigma_{\bar{y}}$ (instead of σ) and the terminology *standard error* (instead of *standard deviation*) distinguish this measure from the standard deviation σ of the population distribution.

In practice, we don't need to take samples repeatedly to find the standard error of $\bar{y}$, because a formula is available. For a random sample of size n, the standard error of $\bar{y}$ depends on n and the population standard deviation σ by

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}.$$

Figure 4.11 displays a population distribution having $\sigma = 10$ and shows the sampling distribution of $\bar{y}$ for $n = 100$, for which the standard error is $\sigma_{\bar{y}} = \sigma/\sqrt{n} = 10/\sqrt{100} = 1.0$. The sampling distribution has only a tenth of the spread of the population distribution. This means that individual observations tend to vary much more than sample means vary from sample to sample.

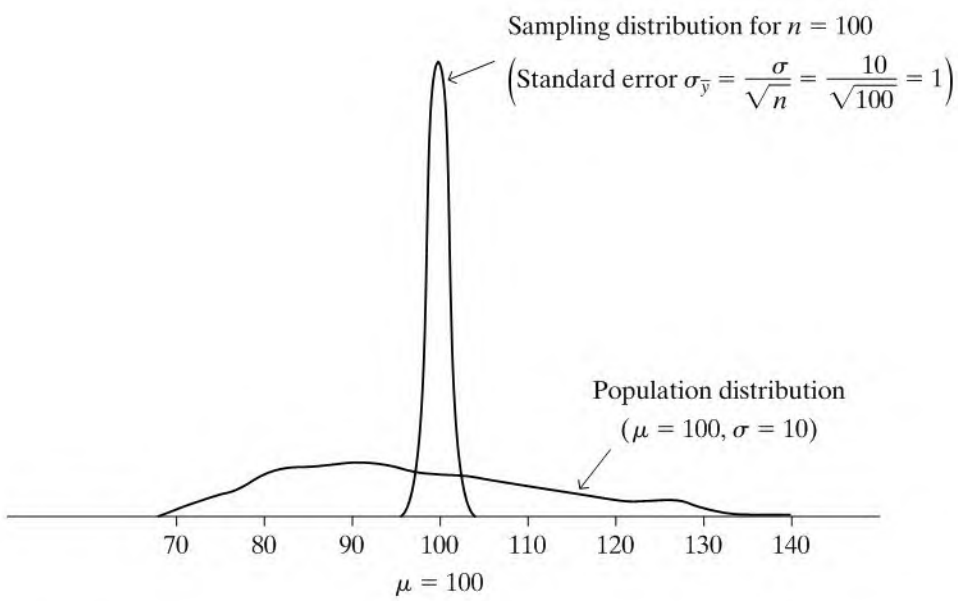

FIGURE 4.11: A Population Distribution and the Sampling Distribution of $\bar{y}$ for $n = 100$

In summary, the following result describes the center and spread of the sampling distribution of $\bar{y}$:

Mean and Standard Error of $\bar{y}$

Consider a random sample of size n from a population having mean μ and standard deviation σ. The sampling distribution of $\bar{y}$ gives the probabilities for the possible values of $\bar{y}$. It has mean μ and standard error $\sigma_y = \sigma/\sqrt{n}$.

EXAMPLE 4.8 Standard Error of Sample Proportion in Election Exit Poll

Following Example 4.7 (page 85), we conducted a simulation to investigate how much variability to expect from sample to sample in an exit poll of 2705 voters. Instead of conducting a simulation, we can get similar information directly by finding a standard error. Knowing the standard error helps us answer the following question: If half the population voted for each candidate, how much would a sample proportion for an exit poll of 2705 voters tend to vary from sample to sample?

As in Example 4.8, let the variable y equal 1 for a vote for the Republican and 0 for a vote for the Democrat. Figure 4.12 shows the distribution for which half the population voted for the Republican, so that $P(1) = 0.50$ and $P(0) = 0.50$. The mean of the distribution equals 0.50, which is the population proportion voting for the Republican. (Or, from the formula, $\mu = \Sigma y P(y) = 0(0.50) + 1(0.50) = 0.50$.) The squared deviation of y from the mean, $(y - \mu)^2$, equals $(0 - 0.50)^2 = 0.25$ when $y = 0$ and it equals $(1 - 0.50)^2 = 0.25$ when $y = 1$. The variance is the expected value of this squared deviation. Thus, it equals $\sigma^2 = 0.25$. So the standard deviation of the population distribution of y equals $\sigma = \sqrt{0.25} = 0.50$.

For a sample, the mean of the 0 and 1 values is the sample proportion of voters who voted for the Republican. Its sampling distribution has mean that is the mean of the population distribution of y, namely, $\mu = 0.50$. For repeated samples of a fixed size n, the sample proportions fluctuate around 0.50, being larger about half the time and smaller half the time. The standard deviation of the sampling distribution is the standard error. For a sample of size 2705, this is

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{0.50}{\sqrt{2705}} = 0.01.$$

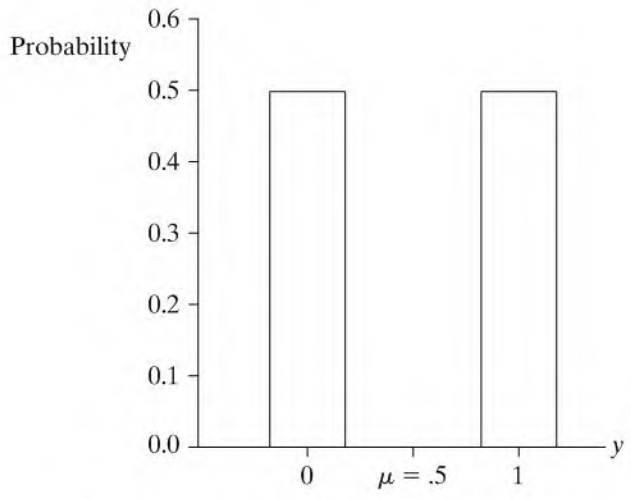

FIGURE 4.12: The Population Distribution when $y = 0$ or 1, with Probability 0.50 Each. This is the distribution for a vote, with 1 = vote for Republican candidate and 0 = vote for Democratic candidate.

A result from later in this section says that this sampling distribution is bell shaped. Thus, with probability close to 1.0 the sample proportion falls within three standard errors of μ, that is, within $3(0.01) = 0.03$ of 0.50, or between 0.47 and 0.53. For a random sample of size 2705 from a population in which 50% voted for the Republican, it would be extremely surprising if fewer than 47% or more than 53% voted for the Republican. We've now seen how to get this result either using simulation, as shown in Figure 4.9, or using the information about the mean and standard error of the sampling distribution. ■

Effect of Sample Size on Sampling Distribution and Precision of Estimates

The standard error gets smaller as the sample size n gets larger. The reason for this is that the denominator ($\sqrt{n}$) of the standard error formula $\sigma_{\bar{y}} = \sigma/\sqrt{n}$ increases as n increases. For instance, when the population standard deviation is $\sigma = 0.50$, we've just seen that the standard error is 0.01 when $n = 2705$. When $n = 100$, a less typical size for a poll, the standard error equals

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{0.50}{\sqrt{n}} = \frac{0.50}{\sqrt{100}} = 0.05.$$

With $n = 100$, since three standard errors equals $3(0.05) = 0.15$, the probability is very high that the sample proportion falls within 0.15 of 0.50, or between 0.35 and 0.65.

Figure 4.13 shows the sampling distributions of the sample proportion when $n = 100$ and when $n = 2705$. As n increases, the standard error decreases and the sampling distribution gets narrower. This means that the sample proportion tends to fall closer to the population proportion. It's more likely that the sample proportion closely approximates an unknown population proportion when $n = 2705$ than when $n = 100$. This agrees with our intuition that larger samples provide more precise estimates of population characteristics.

In summary, error results from estimating μ by $\bar{y}$, because we sampled only part of the population. This error, which is the ***sampling error***, tends to decrease as the sample size n increases. The standard error is fundamental to inferential procedures that predict the sampling error in using $\bar{y}$ to estimate μ.

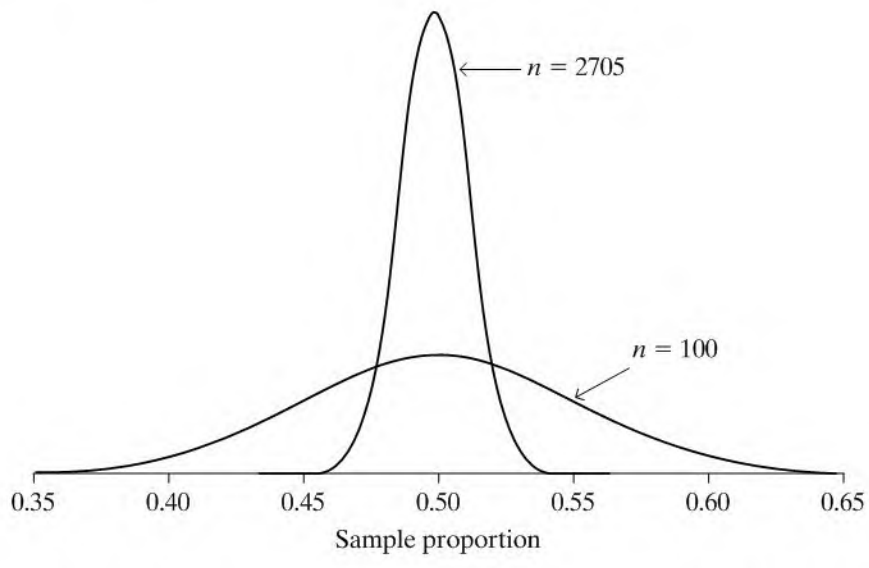

FIGURE 4.13: The Sampling Distributions of the Sample Proportion, when $n = 100$ and when $n = 2705$. These refer to sampling from the population distribution in Figure 4.12.

Sampling Distribution of Sample Mean Is Approximately Normal

For the population distribution for the vote in an election, shown in Figure 4.12, the outcome has only two possible values. It is highly discrete. Nevertheless, the two sampling distributions shown in Figure 4.13 have bell shapes. This is a consequence of the second main result of this section, which describes the *shape* of the sampling distribution of $\bar{y}$. This result can be proven mathematically, and it is often called the *Central Limit Theorem*.

Central Limit Theorem

For random sampling with a large sample size n, the sampling distribution of the sample mean $\bar{y}$ is approximately a normal distribution.

Here are some implications and interpretations of this result:

- The approximate normality of the sampling distribution applies *no matter what the shape* of the population distribution. This is quite remarkable. For large random samples, the sampling distribution of $\bar{y}$ is approximately normal even if the population distribution is highly skewed, U shaped, or highly discrete such as the binary distribution in Figure 4.12. We'll see that this enables us to make inferences even when the population distribution is highly irregular. This is helpful, because many social science variables are very skewed or highly discrete.

  Figure 4.14 displays sampling distributions of $\bar{y}$ for four different shapes for the population distribution, shown at the top of the figure. Below them are

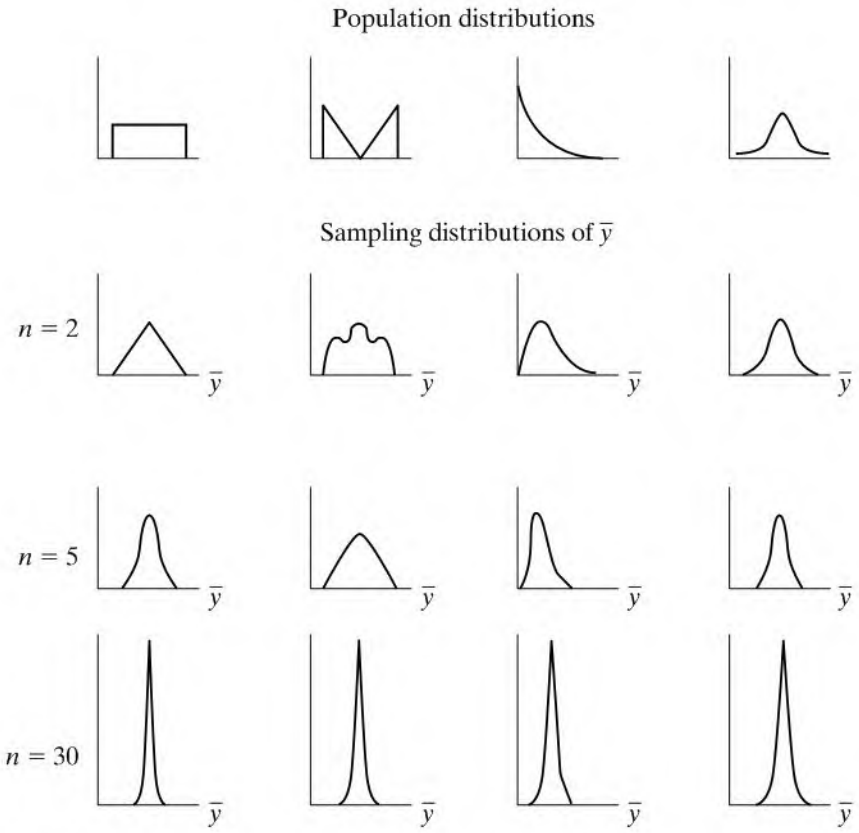

FIGURE 4.14: Four Different Population Distributions and the Corresponding Sampling Distributions of $\bar{y}$. As n increases, the sampling distributions get narrower and have more of a bell shape.

portrayed the sampling distributions for random samples of sizes $n = 2, 5$, and 30. As n increases, the sampling distribution has more of a bell shape.

- How large n must be before the sampling distribution is bell shaped largely depends on the skewness of the population distribution. If the *population* distribution is bell shaped, then the sampling distribution is bell shaped for *all* sample sizes. The rightmost panel of Figure 4.14 illustrates. More skewed distributions require larger sample sizes. For most cases, a sample size of about 30 is sufficient (although it may not be large enough for precise inference). So, in practice for random sampling, the sampling distribution is nearly always approximately bell shaped.
- We could verify the Central Limit Theorem empirically by repeatedly selecting random samples, calculating $\bar{y}$ for each sample of n observations. Then the histogram of the $\bar{y}$-values would be approximately a normal curve about μ with standard error equal to $\sigma/\sqrt{n}$, the population standard deviation divided by the square root of the sample size of each sample.
- Knowing that the sampling distribution of $\bar{y}$ is approximately normal helps us to find probabilities for possible values of $\bar{y}$. For instance, $\bar{y}$ almost certainly falls within $3\sigma_{\bar{y}} = 3\sigma/\sqrt{n}$ of μ. We'll see that reasoning of this nature is vital to inferential statistical methods.

EXAMPLE 4.9 Is Sample Mean Income of Migrant Workers Near Population Mean?

For the population of migrant workers in California, suppose that weekly income has a distribution that is skewed to the right with a mean of $\mu = \$380$ and a standard deviation of $\sigma = \$80$. A researcher, unaware of these values, plans to randomly sample 100 migrant workers and use the sample mean income $\bar{y}$ to estimate μ. What is the sampling distribution of the sample mean? What is the probability that $\bar{y}$ falls above \$400?

By the Central Limit Theorem, the sampling distribution of the sample mean $\bar{y}$ is approximately normal, even though the population distribution is skewed. The sampling distribution has the same mean as the population distribution, namely, $\mu = \$380$. Its standard error is

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{80}{\sqrt{100}} = 8.0 \text{ dollars.}$$

Thus, it is highly likely that $\bar{y}$ falls within about \$24 (three standard errors) of μ.

For the normal sampling distribution with mean 380 and standard error 8, the possible $\bar{y}$ value of 400 has a z-score of

$$z = (400 - 380)/8 = 2.5.$$

From a table of normal probabilities (such as Table A), the corresponding right-tail probability above 400 is 0.0062. It is very unlikely that the sample mean would fall above \$400. ■

This last example is unrealistic, because it used the value of the population mean μ. In practice, this would be unknown. However, the sampling distribution of $\bar{y}$ provides the probability that the sample mean falls within a certain distance of the population mean μ, even when μ is unknown. We illustrate for the study of income of California migrant workers. Let's calculate the probability that the sample mean weekly income $\bar{y}$ falls within \$10 of the true mean income μ for all such workers.

Now the sampling distribution of $\bar{y}$ is approximately normal in shape and is centered about μ. We saw in the previous example that when $n = 100$, the standard error is $\sigma_{\bar{y}} = \$8.0$. Hence, the probability that $\bar{y}$ falls within \$10 of μ is the probability that a normally distributed variable falls within $10/8 = 1.25$ standard deviations of its mean. That is, the number of standard errors that $\mu + 10$ (or $\mu - 10$) falls from μ is

$$z = \frac{(\mu + 10) - \mu}{8} = \frac{10}{8} = 1.25,$$

as Figure 4.15 shows. From a normal table, the probability that $\bar{y}$ falls *more than* 1.25 standard errors from μ (in either direction) is $2(0.1056) = 0.21$. Thus, the probability that $\bar{y}$ falls no more than \$10 from μ equals $1 - 0.21 = 0.79$.

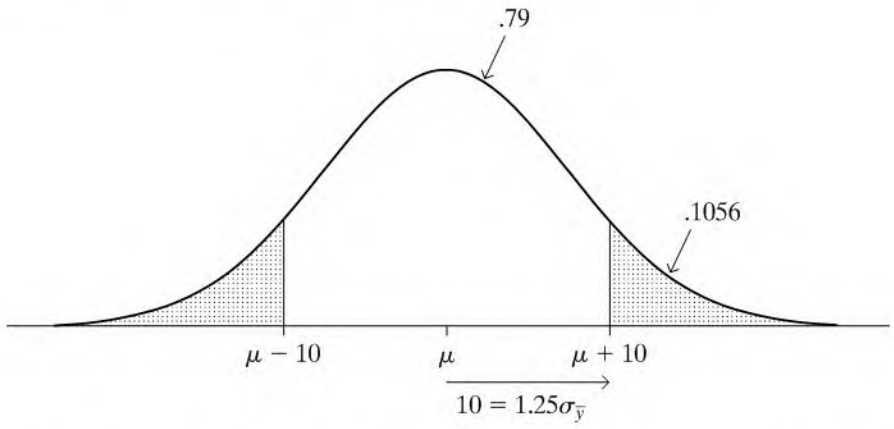

FIGURE 4.15: Sampling Distribution of $\bar{y}$ for Unknown μ and Standard Error $\sigma_{\bar{y}} = 8$

This example is still unrealistic, because it used the population standard deviation σ. In practice, we'd estimate this value. The next chapter shows that, to conduct inference, we can estimate σ by the sample standard deviation s.

To get a feel for the Central Limit Theorem and how sampling distributions become more bell shaped as n increases, it can be very helpful to use an applet on the Internet. We strongly recommend that you try Exercises 4.41 and 4.42.

4.6 REVIEW: POPULATION, SAMPLE DATA, AND SAMPLING DISTRIBUTIONS

Sampling distributions are fundamental to statistical inference and methodology presented in the rest of this text. Because of this, we now review and elaborate on the distinction between the sampling distribution and the two types of distributions presented in Section 3.1—the ***population*** distribution and the ***sample data*** distribution.

Here is a capsule description of the three types of distribution:

- ***Population distribution***: This is the distribution from which we select the sample. It is usually unknown. We make inferences about its characteristics, such as the parameters μ and σ that describe its center and spread. Denote the population size by N.
- ***Sample data distribution***: This is the distribution of data that we actually observe; that is, the sample observations $y_1, y_2, \ldots, y_n$. We can describe it by statistics such as the sample mean $\bar{y}$ and sample standard deviation s. The larger the sample size n, the closer the sample data distribution resembles the population distribution, and the closer the sample statistics such as $\bar{y}$ fall to the population parameters such as μ.

- ***Sampling distribution*** of a statistic: This is the probability distribution for the possible values of a sample statistic, such as $\bar{y}$. A sampling distribution describes the variability that occurs in the statistic's value among samples of a certain size. This distribution determines the probability that the statistic falls within a certain distance of the population parameter it estimates.

EXAMPLE 4.10 Three Distributions for a General Social Survey Item

In 2006, the GSS asked about the number of hours a week spent on the World Wide Web, excluding e-mail (variable denoted WWWHR). The *sample data distribution* for the $n = 2778$ subjects in the sample was very highly skewed to the right. It is described by the sample mean $\bar{y} = 5.7$ and sample standard deviation $s = 10.5$.

Because the GSS cannot sample the entire population of adult Americans (N of about 200 million), we don't know the *population distribution*. Because the sample data distribution had a large sample size, probably the population distribution looks like it. Most likely the population distribution would also be highly skewed to the right. Its mean and standard deviation would be similar to the sample values. Values such as $\mu = 6.0$ and $\sigma = 10.3$ would be realistic.

If the GSS repeatedly took random samples of 2778 adult Americans, the sample mean time $\bar{y}$ spent on the WWW would vary from survey to survey. The *sampling distribution* describes how $\bar{y}$ would vary. For example, if the population has mean $\mu = 6.0$ and standard deviation $\sigma = 10.3$, then the sampling distribution of $\bar{y}$ also has mean 6.0, and it has a standard error of

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{10.3}{\sqrt{2778}} = 0.20.$$

Unlike the population and sample data distributions, the sampling distribution would be bell shaped and narrow. Nearly all of that distribution would fall within $3(0.20) = 0.6$ of the mean of 6.0. So it's very likely that any sample of size 2778 would have a sample mean within 0.6 of 6.0. In summary, the sample data and population distributions are highly skewed and spread out, whereas the sampling distribution of $\bar{y}$ is approximately normal and has nearly all its probability in a narrow range. ■

In reality, the GSS uses a multistage cluster sample rather than a simple random sample. Because of this, the true standard error is actually a bit larger than given by this formula. (This is discussed in Appendix A of the codebook at sda.berkeley.edu/GSS.) It's beyond the scope of this text to adjust standard errors for clustering effects. For purposes of illustration, we'll treat GSS data as if they come from a simple random sample, keeping in mind that in practice some adjustment may be necessary as explained at the GSS website.

EXAMPLE 4.11 Three Distributions for Exit Poll Example

We consider, once again, the variable y = vote in the 2006 California gubernatorial election for a randomly selected voter. Let $y = 1$ for Republican and $y = 0$ for another candidate. In fact, of the 6,921,442 adult residents of California who voted, 55.9% voted for Schwarzenegger. So the probability distribution for y has probability 0.559 at $y = 1$ and probability 0.441 at $y = 0$. The mean of this distribution is $\mu = 0.559$, which is the population proportion of votes for Schwarzenegger. From a formula we'll study in the next chapter, the standard deviation of this distribution equals $\sigma = 0.497$.

The population distribution of candidate preference consists of $N = 6{,}921{,}442$ values of y, 44.1% of which are 0 and 55.9% of which are 1. This distribution is described by the parameters $\mu = 0.559$ and $\sigma = 0.497$. Figure 4.16 portrays this distribution, which is highly discrete (binary). It is not at all bell shaped.

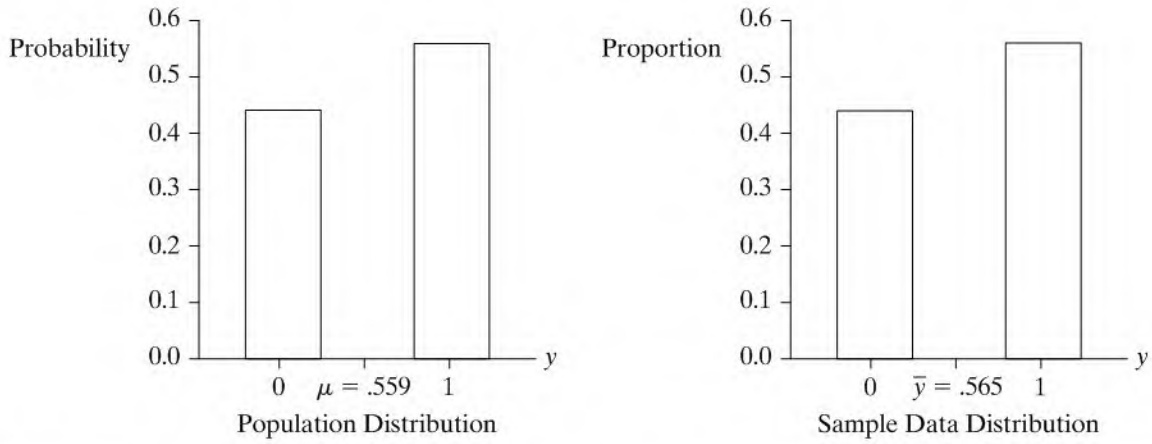

FIGURE 4.16: The Population ($N = 6{,}921{,}442$) and Sample Data ($n = 2705$) Distributions of Vote in 2006 California Gubernatorial Election, where 1 = Schwarzenegger and 0 = Other Candidates

Before all the votes were counted, the population distribution was unknown. When polls closed, CNN reported results of an exit poll of size $n = 2705$ to predict the outcome. A histogram of the 2705 votes in the sample describes the sample data distribution. Of the 2705 voters, 56.5% said they voted for Schwarzenegger (i.e., have $y = 1$) and 43.5% said they voted for another candidate ($y = 0$). Figure 4.16 also displays the histogram of these sample data values. Like the population distribution, the sample data distribution concentrates at $y = 0$ and $y = 1$. It is described by sample statistics such as $\bar{y} = 0.565$, which is the sample proportion voting for Schwarzenegger. The larger the sample size, the more this sample data distribution tends to resemble the population distribution, since the sample observations are a subset of the population values. If the entire population is sampled, as when all the votes are counted, then the two distributions are identical.

For a random sample of size $n = 2705$, the sampling distribution of $\bar{y}$ is approximately normal. Its mean is $\mu = 0.559$ and its standard error is

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{0.497}{\sqrt{2705}} = 0.01.$$

Figure 4.17 portrays this sampling distribution, relative to the population distribution of votes.

By contrast, the population distribution and sample data distribution of votes are concentrated at the values 0 and 1. The sampling distribution looks completely different from them, being much less spread out and bell shaped. The population and sample data distributions of the vote are not bell shaped. They are highly discrete, concentrated at 0 and 1. With $n = 2705$ the sample proportion can take a large number of values between 0 and 1, and its sampling distribution is essentially continuous, being approximately normal by the Central Limit Theorem. ■

Effect of Sample Size on Sample Data and Sampling Distributions

We've seen that the sampling distribution is more nearly normal in shape for larger values of n. For sampling only one observation ($n = 1$), $\bar{y} = y_1$, and the sampling

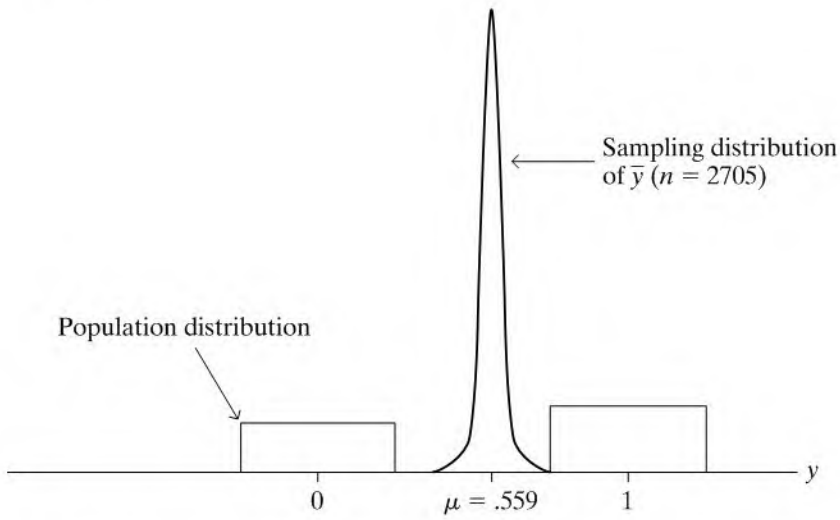

FIGURE 4.17: The Population Distribution (where $y = 1$ Is Vote for Schwarzenegger and $y = 0$ Is Vote for Other Candidate) and the Sampling Distribution of $\bar{y}$ for $n = 2705$

distribution of $\bar{y}$ is the same as the probability distribution for one observation on y. This is simply the population distribution of y, which need not be the least bit normal. As n increases, the sampling distribution of $\bar{y}$ assumes more of a bell shape. For $n \geq 30$, the approximation is usually good. As the sample size n approaches the population size N, the normal sampling distribution of $\bar{y}$ gets narrower and eventually converges to a spike at the single number μ. When the entire population is sampled, $\bar{y} = \mu$ with probability 1 (i.e., the two measures are the same), and the sampling distribution concentrates at the point μ.

Figure 4.17 showed a great difference between the population distribution and the sampling distribution of $\bar{y}$ (for $n = 2705$). Note also the great difference between the sample data distribution (Figure 4.16) and the sampling distribution (Figure 4.17). The sample data distribution looks much like the population distribution, more so as the sample size increases. The sampling distribution, on the other hand, has a bell-shaped appearance and gets narrower as n increases. As Figure 4.16 shows, the sample values of y can be only 0 or 1. On the other hand, sample mean values (which are sample proportions) can fall between 0 and 1. According to the sampling distribution of $\bar{y}$ for $n = 2705$, it is practically impossible that a random sample of that size has a sample mean anywhere near 0 or 1; nearly all the probability falls between 0.53 and 0.59 (within about three standard errors of the mean of the sampling distribution).

The Key Role of Sampling Distributions in Statistical Inference

We've seen that, by the Central Limit Theorem, we can often use the normal distribution to find probabilities about $\bar{y}$. The next two chapters will show that statistical inferences rely on this theorem.

The result about sample means having approximately normal sampling distributions, for large random samples, is important also because similar results hold for many other statistics. For instance, most sample statistics used to estimate population parameters have approximately normal sampling distributions, for large random samples. The primary reason for the key role of the normal distribution is that so many statistics have approximately normal sampling distributions.

4.7 CHAPTER SUMMARY

For an observation in a random sample or a randomized experiment, the ***probability*** of a particular outcome is the proportion of times that the outcome would occur in a long sequence of observations.

- A ***probability distribution*** specifies probabilities for the possible values of a variable. We let $P(y)$ denote the probability of the value y. The probabilities are nonnegative and sum to 1.0.
- Probability distributions have summary parameters, such as the mean μ and standard deviation σ. The mean for a probability distribution of a discrete variable is

$$\mu = \sum yP(y).$$

  This is also called the ***expected value*** of y.
- The ***normal distribution*** has a graph that is a symmetric bell-shaped curve specified by the mean μ and standard deviation σ. For any z, the probability falling within z standard deviations of the mean is the same for every normal distribution.
- The ***z-score*** for an observation y equals

$$z = (y - \mu)/\sigma.$$

  It measures the number of standard deviations that y falls from the mean μ. For a normal distribution, the z-scores have the *standard normal distribution*, which has mean = 0 and standard deviation = 1.
- A ***sampling distribution*** is a probability distribution of a sample statistic, such as the sample mean or sample proportion. It specifies probabilities for the possible values of the statistic for all the possible samples.
- The sampling distribution of the sample mean $\bar{y}$ centers at the population mean μ. Its standard deviation, called the ***standard error***, relates to the standard deviation σ of the population by $\sigma_{\bar{y}} = \sigma/\sqrt{n}$. As the sample size n increases, the standard error decreases, so the sample mean tends to be closer to the population mean.
- The ***Central Limit Theorem*** states that for large random samples, the sampling distribution of the sample mean is approximately normal. This holds no matter what the shape of the population distribution. The result applies also to proportions, since the sample proportion is a special case of the sample mean for observations coded as 0 and 1 (such as for two candidates in an election).

The bell shape for the sampling distribution of many statistics is the main reason for the importance of the normal distribution. The next two chapters show how the Central Limit Theorem is the basis of methods of statistical inference.

PROBLEMS

Practicing the Basics

4.1. In a GSS, in response to the question, "Do you believe in life after death?" 907 people answered yes and 220 answered no. Based on this survey, estimate the probability that a randomly selected adult in the United States believes in life after death.

4.2. A GSS estimates the probability that an American adult believes in heaven is 0.85.

(a) Estimate the probability that an American adult does not believe in heaven.

(b) Of those who believe in heaven, about 84% believe in hell. Estimate the probability a randomly chosen American adult believes in both heaven and hell.

4.3. In 2000, the GSS asked subjects whether they are a member of an environmental group (variable GRNGROUP) and whether they would be very willing to pay much higher prices to protect the environment (variable GRNPRICE). Table 4.4 shows results.

(a) Explain why $96/1117 = 0.086$ estimates the probability that a randomly selected American adult is a member of an environmental group.

(b) Show that the estimated probability of being very willing to pay much higher prices to protect the environment is (i) 0.312, given that the person is a member of an environmental group, (ii) 0.086, given not a member of an environmental group.

(c) Show that the estimated probability a person is both a member of an environmental group *and* very willing to pay much higher prices to protect the environment is 0.027 (i) directly using the counts in the table, (ii) using the probability estimates from (a) and (b).

(d) Show that the estimated probability that a person answers yes to both questions or no to both questions is 0.862.

TABLE 4.4

		Pay Higher Prices		
		Yes	No	Total
Member of Environmental Group	Yes	30	66	96
	No	88	933	1021
Total		118	999	1117

4.4. Let y = number of languages in which a person is fluent. According to Statistics Canada, for residents of Canada this has probability distribution $P(0) = 0.02$, $P(1) = 0.81$, $P(2) = 0.17$, with negligible probability for higher values of y.

(a) Is y a discrete or a continuous variable? Why?

(b) Construct a table showing the probability distribution of y.

(c) Find the probability a Canadian is *not* multilingual.

(d) Find the mean of this probability distribution.

4.5. Let y denote the number of people known personally who were victims of homicide within the past 12 months. According to results from recent General Social Surveys, for a randomly chosen person in the U.S. the probability distribution of y is approximately: $P(0) = 0.91, P(1) = 0.06, P(2) = 0.02, P(3) = 0.01$.

(a) Explain why it is not valid to find the mean of this probability distribution as $(0 + 1 + 2 + 3)/4 = 1.5$.

(b) Find the correct mean of the probability distribution.

4.6. A ticket for a statewide lottery costs \$1. With probability 0.0000001, you win a million dollars (\$1,000,000), and with probability 0.9999999 you win nothing. Let y denote the winnings from buying one ticket. Construct the probability distribution for y. Show that the mean of the distribution equals 0.10, corresponding to an expected return of 10 cents for the dollar paid.

4.7. Let y be the outcome of selecting a single digit from a random number table.

(a) Construct the probability distribution for y. (This type of distribution is called a *uniform* distribution because of the uniform spread of probabilities across the possible outcomes.)

(b) Find the mean of this probability distribution.

(c) The standard deviation σ of this distribution is one of the following: 0.4, 2.9, 7.0, 12.0. Which do you think is correct? Why?

4.8. For a normal distribution, find the probability that an observation falls

(a) At least one standard deviation above the mean

(b) At least one standard deviation below the mean

(c) At least 0.67 standard deviations above the mean

4.9. For a normally distributed variable, verify that the probability between:

(a) $\mu - \sigma$ and $\mu + \sigma$ equals 0.68

(b) $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ equals 0.95

(c) $\mu - 3\sigma$ and $\mu + 3\sigma$ equals 0.997

(d) $\mu - 0.67\sigma$ and $\mu + 0.67\sigma$ equals 0.50

4.10. Find the z-value for which the probability that a normal variable exceeds $\mu + z\sigma$ equals

(a) 0.01

(b) 0.025

(c) 0.05

(d) 0.10

(e) 0.25

(f) 0.50

4.11. Find the z-value such that for a normal distribution the interval from $\mu - z\sigma$ to $\mu + z\sigma$ contains

(a) 50%

(b) 90%

(c) 95%

(d) 98%

(e) 99% of the probability.

4.12. Find the z-values corresponding to the

(a) 90th,

(b) 95th,

(c) 98th, and

(d) 99th percentiles of a normal distribution.

4.13. Show that if z is the number such that the interval from $\mu - z\sigma$ to $\mu + z\sigma$ contains 90% of a normal distribution, then $\mu + z\sigma$ equals the 95th percentile.

4.14. If z is the positive number such that the interval from $\mu - z\sigma$ to $\mu + z\sigma$ contains 50% of a normal distribution, then

(a) Which percentile is (i) $\mu + z\sigma$? (ii) $\mu - z\sigma$?

(b) Find this value of z.

(c) Using this result, explain why the upper quartile and lower quartile of a normal distribution are $\mu + 0.67\sigma$ and $\mu - 0.67\sigma$.

4.15. What proportion of a normal distribution falls in the following ranges?

(a) Above a z-score of 2.10

(b) Below a z-score of -2.10

(c) Above a z-score of -2.10

(d) Between z-scores of -2.10 and 2.10

4.16. Find the z-score for the number that is less than only 1% of the values of a normal distribution.

4.17. Mensa is a society of high-IQ people whose members have a score on an IQ test at the 98th percentile or higher.

(a) How many standard deviations above the mean is the 98th percentile?

(b) For the normal IQ distribution with mean 100 and standard deviation 16, what is the IQ score for the 98th percentile?

4.18. According to a recent *Current Population Reports*, self-employed individuals in the United States work an average of 45 hours per week, with a standard deviation of 15. If this variable is approximately normally distributed, what proportion averaged more than 40 hours per week?

4.19. The Mental Development Index (MDI) of the Bayley Scales of Infant Development is a standardized measure used in studies with high-risk infants. It has approximately a normal distribution with a mean of 100 and a standard deviation of 16.

(a) What proportion of children have a MDI of at least 120?

(b) Find the MDI score that is the 90th percentile.

(c) Find and interpret the lower quartile, median, and upper quartile of MDI.

4.20. For 5459 pregnant women using Aarhus University Hospital in Denmark in a two-year period who reported information on length of gestation until birth, the mean was 281.9 days, with standard deviation 11.4 days.[3] A baby is classified as premature if the gestation time is 258 days or less.

(a) If gestation times are normally distributed, what proportion of babies would be born prematurely?

(b) The actual proportion born prematurely during this period was 0.036. Based on this information, how would you expect the distribution of gestation time to differ from normal?

4.21. Suppose that the weekly use of gasoline for motor travel by adults in North America is approximately normally distributed, with a mean of 16 gallons and a standard deviation of 5 gallons.

(a) What proportion of adults use more than 20 gallons per week?

(b) Assuming that the standard deviation and the normal form would remain constant, to what level must the mean reduce so that only 5% use more than 20 gallons per week?

(c) If the distribution of gasoline use is not actually normal, how would you expect it to deviate from normal?

4.22. On the midterm exam in introductory statistics, an instructor always gives a grade of B to students who score between 80 and 90. One year, the scores have approximately a normal distribution with mean 83 and standard deviation 5. About what proportion of the students get a B?

4.23. For an SAT distribution ($\mu = 500, \sigma = 100$) and an ACT distribution ($\mu = 21, \sigma = 4.7$), which score is relatively higher, SAT = 600 or ACT = 29? Explain.

4.24. Suppose that property taxes on homes in Iowa City, Iowa have an approximately normal distribution with a mean of \$2500 and a standard deviation of \$1500. The property tax for one particular home is \$5000.

(a) Find the z-score corresponding to that value.

(b) What proportion of the property taxes exceed \$5000?

(c) If the true distribution is not normal, how do you think it deviates from normal? Why?

4.25. An energy study in Gainesville, Florida, found that in March 2006, household use of electricity had a mean of 673 and a standard deviation of 556 kWh (kilowatt-hours).

(a) If the distribution were normal, what percentage of the households had use above 1000 kWh?

(b) Do you think the distribution is truly normal? Why or why not?

4.26. Five students, Ann, Betty, Clint, Douglas, and Edward, are rated equally qualified for admission to law school, ahead of other applicants. However, all but two positions have been filled for the entering class. Since the admissions committee can admit only two more students, it decides to randomly select two of these five candidates. For

[3] *British Medical Journal*, Vol. 307, 1993, p. 234.

this strategy, let y = number of females admitted. Using the first letter of the name to denote a student, the different combinations that could be admitted are (A, B), (A, C), (A, D), (A, E), (B, C), (B, D), (B, E), (C, D), (C, E), and (D, E).

(a) Construct the probability distribution for y.

(b) Construct the sampling distribution of the sample *proportion* of the students selected who are female.

4.27. Construct the sampling distribution of the sample proportion of heads, for flipping a balanced coin:

(a) Once.

(b) Twice. (*Hint*: The possible samples are (H, H), (H, T), (T, H), (T, T).)

(c) Three times. (*Hint*: There are 8 possible samples.)

(d) Four times. (*Hint*: There are 16 possible samples.)

(e) Describe how the shape of the sampling distribution seems to be changing as the number of flips increases.

4.28. The probability distribution associated with the outcome of rolling a balanced die has probability 1/6 attached to each integer, $\{1,2,3,4,5,6\}$. Let (y_1, y_2) denote the outcomes for rolling the die twice.

(a) Enumerate the 36 possible (y_1, y_2) pairs (e.g., (2, 1) represents a 2 followed by a 1).

(b) Treating the 36 pairs as equally likely, construct the sampling distribution for the sample mean $\bar{y}$ of the two numbers rolled.

(c) Construct a histogram of the (i) probability distribution for each roll, (ii) sampling distribution of $\bar{y}$ in (b). Describe their shapes.

(d) What are the means of the two distributions in (c)? Why are they the same?

(e) Explain why the sampling distribution of $\bar{y}$ has relatively more probability near the middle than at the minimum and maximum values. (*Hint*: Note there are many more (y_1, y_2) pairs that have a sample mean near the middle than near the minimum or maximum.)

4.29. An exit poll of 2293 voters in the 2006 Ohio Senatorial election indicated that 44% voted for the Republican candidate, Mike DeWine, and 56% voted for the Democratic candidate, Sherrod Brown.

(a) If actually 50% of the population voted for DeWine, find the standard error of the sample proportion voting for him, for this exit poll. (Recall from Example 4.8 on page 91 that the population standard deviation is 0.50.)

(b) If actually 50% of the population voted for DeWine, would it have been surprising to obtain the results in this exit poll? Why?

(c) Based on your answer in (b), would you be willing to predict the outcome of this election? Explain.

4.30. According to *Current Population Reports*, the population distribution of number of years of education for self-employed individuals in the United States has a mean of 13.6 and a standard deviation of 3.0. Find the mean and standard error of the sampling distribution of $\bar{y}$ for a random sample of

(a) 9 residents,

(b) 36 residents,

(c) 100 residents. Describe the pattern as n increases.

4.31. Refer to Exercise 4.6. The mean and standard deviation of the probability distribution for the lottery winnings y are $\mu = 0.10$ and $\sigma = 316.23$. Suppose you play the lottery 1 million times. Let $\bar{y}$ denote your average winnings.

(a) Find the mean and standard error of the sampling distribution of $\bar{y}$.

(b) About how likely is it that you would "come out ahead," with your average winnings exceeding \$1, the amount you paid to play each time?

4.32. According to recent General Social Surveys (variable PARTNERS), in the United States the distribution of y = number of sex partners you have had in the past 12 months has a mean of about 1.1 and a standard deviation of about 1.1. Suppose these are the population mean and standard deviation.

(a) Does y have a normal distribution? Explain.

(b) For a random sample of 2400 adults (the size of the 2006 GSS for this variable), describe the sampling distribution of $\bar{y}$ by giving its shape, mean, and standard error.

(c) Refer to (b). Report an interval within which the sample mean would almost surely fall.

4.33. The scores on the Psychomotor Development Index (PDI), a scale of infant development, are approximately normal with mean 100 and standard deviation 15.

(a) An infant is selected at random. Find the probability that PDI is below 90.

(b) A study uses a random sample of 25 infants. Specify the sampling distribution of the sample mean PDI, and find the probability that the sample mean is below 90.

(c) Would you be surprised to observe a PDI score of 90? Would you be surprised to observe a sample mean PDI of 90? Why?

(d) Sketch the population distribution for PDI. Superimpose a sketch of the sampling distribution for $n = 25$.

4.34. A study plans to sample randomly 100 government records of farms in Ontario to estimate the mean

acreage of farms in that province. Results from an earlier study suggest that 200 acres is a reasonable guess for the population standard deviation of farm size.

(a) Find the probability that the sample mean acreage falls within 10 acres of the population mean acreage.

(b) If in reality the population standard deviation is larger than 200, would the probability be larger or smaller than you found in (a)?

4.35. According to the U.S. Census Bureau, in 2000 the number of people in a household had a mean of 2.6 and a standard deviation of 1.5. Suppose the Census Bureau instead had estimated this mean using a random sample of 225 homes, and that sample had a mean of 2.4 and standard deviation of 1.4.

(a) Identify the variable y.

(b) Describe the center and spread of the population distribution.

(c) Describe the center and spread of the sample data distribution.

(d) Describe the center and spread of the sampling distribution of the sample mean for 225 homes. What does that distribution describe?

4.36. The distribution of family size in a particular tribal society is skewed to the right, with $\mu = 5.2$ and $\sigma = 3.0$. These values are unknown to an anthropologist, who samples families to estimate mean family size. For a random sample of 36 families, she gets a mean of 4.6 and a standard deviation of 3.2.

(a) Identify the population distribution. State its mean and standard deviation.

(b) Identify the sample data distribution. State its mean and standard deviation.

(c) Identify the sampling distribution of $\bar{y}$. State its mean and standard error and explain what it describes.

4.37. Refer to the previous exercise.

(a) Find the probability that her sample mean falls within 0.5 of the population mean.

(b) Suppose she takes a random sample of size 100. Find the probability that the sample mean falls within 0.5 of the true mean, and compare the answer to that in (a).

(c) Refer to (b). If the sample were truly random, would you be surprised if the anthropologist obtained $\bar{y} = 4.0$? Why? (This could well happen if the sample were not random.)

4.38. At a university, 60% of the 7400 students are female. The student newspaper reports results of a survey of a random sample of 50 students about various topics involving alcohol abuse, such as participation in binge drinking. They report that their sample contained 26 females.

(a) Explain how you can set up a variable y to represent gender.

(b) Identify the population distribution of gender at this university.

(c) Identify the sample data distribution of gender for this sample.

(d) The sampling distribution of the sample proportion of females in the sample is approximately a normal distribution with mean 0.60 and standard error 0.07. Explain what this means.

4.39. Sunshine City was designed to attract retired people. Its current population of 50,000 residents has a mean age of 60 years and a standard deviation of 16 years. The distribution of ages is skewed to the left, reflecting the predominance of older individuals. A random sample of 100 residents of Sunshine City has $\bar{y} = 58.3$ and $s = 15.0$.

(a) Describe the center and spread of the population distribution.

(b) Describe the center and spread of the sample data distribution. What shape does it probably have?

(c) Find the center and spread of the sampling distribution of $\bar{y}$ for $n = 100$. What shape does it have and what does it describe?

(d) Explain why it would not be unusual to observe a person of age 40 in Sunshine City, but it would be highly unusual to observe a sample mean of 40, for a random sample size of 100.

4.40. Refer to the previous exercise.

(a) Describe the sampling distribution of $\bar{y}$ for a random sample of size $n = 1$.

(b) Describe the sampling distribution of $\bar{y}$ if you sample all 50,000 residents.

Concepts and Applications

4.41. You can use an ***applet*** on a computer or on the Internet to repeatedly generate random samples from artificial populations and analyze them to study the properties of statistical methods. To try this, go to www.prenhall.com/agresti and use the *sampling distribution* applet. Select binary for the parent population, setting the population proportion as 0.50. Select for the sample size $n = 100$.

(a) Simulate once (setting the number of simulations $N = 1$ and clicking on *Sample*) and report the counts and the proportions for the two categories. Did you get a sample proportion close to 0.50? Perform this simulation of a random sample of size 100 ten times, each time observing from the graphs the counts and the corresponding sample proportion of yes votes. Summarize.

(b) Now plot the results of simulating a random sample of size 100 and finding the sample

proportion 1000 times, by setting $N = 1000$ on the menu. How does this plot reflect the Central Limit Theorem?

4.42. Refer to the previous exercise.

(a) For this applet, select the skewed population distribution. Take 1000 samples of size 30 each. How does the empirical distribution of sample means compare to the population distribution? What does this reflect?

(b) Repeat, this time choosing a sample size of only 2 for each sample. Why is the sampling distribution not symmetric and bell-shaped?

4.43. (*Class Exercise*) Refer to Exercises 1.11 and 1.12 (page 8). Using the population defined by your class or using the student survey, the instructor will select a variable, such as weekly time watching television.

(a) Construct a histogram or stem-and-leaf plot of the population distribution of the variable for the class.

(b) Using a random number table, each student should select nine students at random and compute the sample mean response for those students. (Each student should use different random numbers.) Plot a histogram of the sample means obtained by all the students. How do the spread and shape compare to the histogram in (a)? What does this illustrate?

4.44. (*Class Exercise*) Table 4.5 provides the ages of all 50 heads of households in a small Nova Scotian fishing village. The distribution of these ages is characterized by $\mu = 47.18$ and $\sigma = 14.74$.

(a) Construct a stem-and-leaf plot of the population distribution.

(b) Using a random number table, each student should select nine random numbers between 01 and 50. Using these numbers, each student should sample nine heads of households and compute their sample mean age. Plot the empirical sampling distribution of the $\bar{y}$-values. Compare it to the distribution in (a).

(c) What do you expect for the mean of the $\bar{y}$-values in a long run of repeated samples of size 9?

(d) What do you expect for the standard deviation of the $\bar{y}$-values in a long run of repeated samples of size 9?

4.45. (*Class Exercise*) For a single toss of a coin, let $y = 1$ for a head and $y = 0$ for a tail. This simulates the vote in an election with two equally-preferred candidates.

(a) Construct the probability distribution for y, and find its mean.

(b) The coin is flipped ten times, yielding six heads and four tails. Construct the sample data distribution.

(c) Each student in the class should flip a coin 10 times and calculate the proportion of heads in the sample. Summarize the empirical sampling distribution by plotting the proportions for all students. Describe the shape and spread of the sampling distribution compared to the distributions in (a) and (b).

(d) If we performed the experiment of flipping the coin 10 times a very large number of times, what would we get for the (i) mean and (ii) standard deviation of the sample proportion values? You can use 0.50 as the standard deviation of the distribution in (a).

4.46. (a) Which distribution does the sample data distribution tend to resemble more closely—the sampling distribution or the population distribution? Explain.

(b) Explain carefully the difference between a *sample data distribution* and the *sampling*

TABLE 4.5

Name	Age	Name	Age	Name	Age	Name	Age
Alexander	50	Griffith	66	McTell	49	Staines	33
Bell	45	Grosvenor	51	MacLeod	30	Stewart	36
Bell	23	Ian	57	McNeil	28	Stewart	25
Bok	28	Jansch	40	McNeil	31	Thames	29
Clancy	67	Keelaghan	36	McNeil	45	Thomas	57
Cochran	62	Lavin	38	McNeil	43	Todd	39
Fairchild	41	Lunny	81	Mitchell	43	Trickett	50
Finney	68	MacColl	27	Muir	54	Trickett	64
Fisher	37	McCusker	37	Oban	62	Tyson	76
Francey	60	McCusker	56	Reid	67	Watson	63
Fricker	41	McDonald	71	Renbourn	48	Young	29
Gaughan	70	McDonald	39	Rogers	32		
Graham	47	McDonald	46	Rush	42		

distribution of $\bar{y}$. Illustrate your answer for a variable y that can take only values of 0 and 1.

4.47. The Palestinian Central Bureau of Statistics (www.pcbs.gov.ps) asked mothers of age 20–24 about the ideal number of children. For those living on the Gaza Strip, the probability distribution is approximately $P(1) = 0.01$, $P(2) = 0.10$, $P(3) = 0.09$, $P(4) = 0.31$, $P(5) = 0.19$, and $P(6 \text{ or more}) = 0.29$.

(a) Because the last category is open-ended, it is not possible to calculate the mean exactly. Find a lower bound for the mean.

(b) Explain why you can find the median of the distribution, and find it.

4.48. For a normal distribution, show that

(a) The upper quartile equals $\mu + 0.67\sigma$.

(b) According to the 1.5(IQR) criterion, an outlier is an observation falling more than 2.7 standard deviations below or above the mean, and this happens for only 0.7% of the data.

4.49. In an exit poll of 1336 voters in the 2006 Senatorial election in New York State, 67% said they voted for Hillary Clinton. Based on this information, would you be willing to predict the winner of the election? Explain your reasoning.

4.50. For an election exit poll for a Senatorial election, find the standard error of the sample proportion voting for a candidate for whom the population proportion is 0.50, when $n = 100, 1000$, and 10,000. In each case, predict an interval within which the sample proportion is almost certain to fall. Notice that the interval shrinks in width as the sample size increases. This is a consequence of the ***law of large numbers***, which states that the sample proportion tends to get closer and closer to the population proportion as n increases indefinitely.

Select the correct response(s) in multiple-choice questions 4.51–4.52. (There may be more than one correct answer.)

4.51. The standard error of a statistic describes

(a) The standard deviation of the sampling distribution of that statistic.

(b) The standard deviation of the sample data.

(c) How close that statistic is likely to fall to the parameter that it estimates.

(d) The variability in the values of the statistic for repeated random samples of size n.

(e) The error that occurs due to nonresponse and measurement errors.

4.52. The Central Limit Theorem implies that

(a) All variables have bell-shaped sample data distributions if a random sample contains at least about 30 observations.

(b) Population distributions are normal whenever the population size is large.

(c) For large random samples, the sampling distribution of $\bar{y}$ is approximately normal, regardless of the shape of the population distribution.

(d) The sampling distribution looks more like the population distribution as the sample size increases.

(e) All of the above

4.53. True or False: As the sample size increases, the standard error of the sampling distribution of $\bar{y}$ increases. Explain your answer.

***4.54.** Lake Wobegon Junior College admits students only if they score above 400 on a standardized achievement test. Applicants from group A have a mean of 500 and a standard deviation of 100 on this test, and applicants from group B have a mean of 450 and a standard deviation of 100. Both distributions are approximately normal, and both groups have the same size.

(a) Find the proportion not admitted for each group.

(b) Of the students who are not admitted, what proportion are from group B?

(c) A state legislator proposes that the college lower the cutoff point for admission to 300, thinking that the proportion of the students who are not admitted who are from group B would decrease. If this policy is implemented, determine the effect on the answer to (b), and comment.

***4.55.** The standard deviation of a discrete probability distribution is

$$\sigma = \sqrt{\sum (y - \mu)^2 P(y)}.$$

(a) Suppose $y = 1$ with probability 0.50 and $y = 0$ with probability 0.50, such as in Example 4.8 (page 91). Show that $\sigma = 0.50$.

(b) Suppose $y = 1$ with probability π and $y = 0$ with probability $1 - \pi$, where π represents a number between 0 and 1. Show that $\mu = \pi$ and that $\sigma = \sqrt{\pi(1 - \pi)}$.

(c) Show that the standard error of a sample proportion for a random sample of size n equals $\sqrt{\pi(1 - \pi)/n}$.

***4.56.** The curve for a normal distribution with mean μ and standard deviation σ has mathematical formula

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/(2\sigma^2)}.$$

Show that this curve is symmetric, by showing that for any constant c, the curve has the same value at $y = \mu + c$ as at $y = \mu - c$. (The integral of $f(y)$ for y between $\mu + z\sigma$ and ∞ equals the tail probability tabulated in Table A.)

***4.57.** The standard error formula $\sigma_{\bar{y}} = \sigma/\sqrt{n}$ treats the population size N as *infinitely* large relative to the sample size n. The formula for σ_y for a *finite* population size N is

$$\sigma_{\bar{y}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right).$$

The term $\sqrt{(N-n)/(N-1)}$ is called the ***finite population correction***.

(a) When $n = 300$ students are selected from a college student body of size $N = 30{,}000$, show that $\sigma_{\bar{y}} = 0.995\sigma/\sqrt{n}$. (In practice, n is usually small relative to N, so the correction has little influence.)

(b) If $n = N$ (i.e., we sample the entire population), show that $\sigma_{\bar{y}} = 0$. In other words, no sampling error occurs, because $\bar{y} = \mu$.

(c) For $n = 1$, explain why the sampling distribution of $\bar{y}$ and its standard error are identical to the population distribution and its standard deviation.