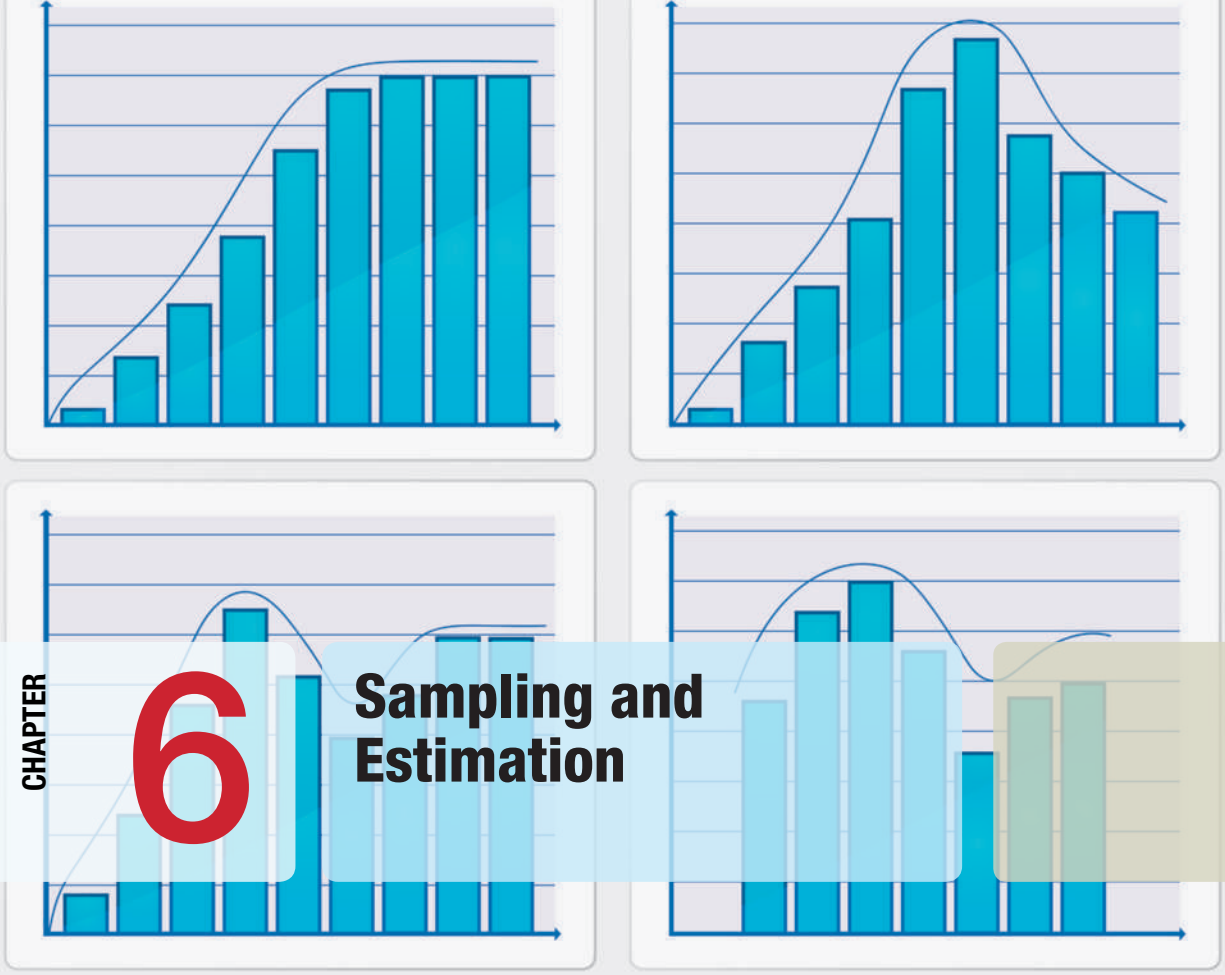


CHAPTER 6

Sampling and Estimation

KALABUKHAVA IRYNA/Shutterstock.com

Learning Objectives

After studying this chapter, you will be able to:

- Describe the elements of a sampling plan.
- Explain the difference between subjective and probabilistic sampling.
- State two types of subjective sampling.
- Explain how to conduct simple random sampling and use Excel to find a simple random sample from an Excel database.
- Explain systematic, stratified, and cluster sampling, and sampling from a continuous process.
- Explain the importance of unbiased estimators.
- Describe the difference between sampling error and nonsampling error.
- Explain how the average, standard deviation, and distribution of means of samples changes as the sample size increases.
- Define the sampling distribution of the mean.
- Calculate the standard error of the mean.
- Explain the practical importance of the central limit theorem.
- Use the standard error in probability calculations.
- Explain how an interval estimate differs from a point estimate.
- Define and give examples of confidence intervals.
- Calculate confidence intervals for population means and proportions using the formulas in the chapter and the appropriate Excel functions.
- Explain how confidence intervals change as the level of confidence increases or decreases.
- Describe the difference between the *t*-distribution and the normal distribution.
- Use confidence intervals to draw conclusions about population parameters.
- Compute a prediction interval and explain how it differs from a confidence interval.
- Compute sample sizes needed to ensure a confidence interval for means and proportions with a specified margin of error.

We discussed the difference between population and samples in Chapter 4. Sampling is the foundation of statistical analysis. We use sample data in business analytics applications for many purposes. For example, we might wish to estimate the mean, variance, or proportion of a very large or unknown population; provide values for inputs in decision models; understand customer satisfaction; reach a conclusion as to which of several sales strategies is more effective; or understand if a change in a process resulted in an improvement. In this chapter, we discuss sampling methods, how they are used to estimate population parameters, and how we can assess the error inherent in sampling.

Statistical Sampling

The first step in sampling is to design an effective sampling plan that will yield representative samples of the populations under study. A **sampling plan** is a description of the approach that is used to obtain samples from a population prior to any data collection activity. A sampling plan states

- the objectives of the sampling activity,
- the target population,
- the **population frame** (the list from which the sample is selected),
- the method of sampling,
- the operational procedures for collecting the data, and
- the statistical tools that will be used to analyze the data.

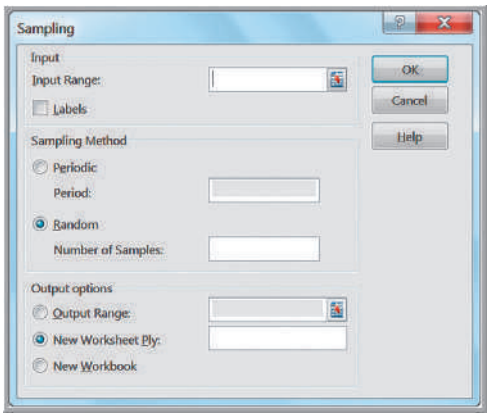
EXAMPLE 6.1 A Sampling Plan for a Market Research Study

Suppose that a company wants to understand how golfers might respond to a membership program that provides discounts at golf courses in the golfers' locality as well as across the country. The *objective* of a sampling study might be to estimate the proportion of golfers who would likely subscribe to this program. The *target population* might be all golfers over 25 years old. However, identifying all golfers in America might be impossible. A practical *population frame* might be a list of golfers who have purchased equipment from national golf or sporting goods companies through which the discount card will be sold. The *operational procedures* for collecting the data might be an e-mail link to a survey site or direct-mail questionnaire. The data might be stored in an Excel database; *statistical tools* such as PivotTables and simple descriptive statistics would be used to segment the respondents into different demographic groups and estimate their likelihood of responding positively.

Sampling Methods

Many types of sampling methods exist. Sampling methods can be *subjective* or *probabilistic*. Subjective methods include **judgment sampling**, in which expert judgment is used to select the sample (survey the "best" customers), and **convenience sampling**, in which samples are selected based on the ease with which the data can be collected (survey all customers who happen to visit this month). Probabilistic sampling involves selecting the

Figure 6.1
Excel *Sampling* Tool Dialog



items in the sample using some random procedure. Probabilistic sampling is necessary to draw valid statistical conclusions.

The most common probabilistic sampling approach is simple random sampling. **Simple random sampling** involves selecting items from a population so that every subset of a given size has an equal chance of being selected. If the population data are stored in a database, simple random samples can generally be easily obtained.

EXAMPLE 6.2 Simple Random Sampling with Excel

Suppose that we wish to sample from the Excel database *Sales Transactions.* Excel provides a tool to generate a random set of values from a given population size. Click on *Data Analysis* in the *Analysis* group of the *Data* tab and select *Sampling*. This brings up the dialog shown in Figure 6.1. In the *Input Range* box, we specify the data range from which the sample will be taken. This tool requires that the data sampled be numeric, so in this example we sample from the first column of the data set, which corresponds to the customer ID number. There are two options for sampling:

1. Sampling can be *periodic*, and we will be prompted for the *Period*, which is the interval between sample observations from the beginning of the data set. For instance, if a period of 5 is used, observations 5, 10, 15, and so on, will be selected as samples.
2. Sampling can also be *random*, and we will be prompted for the *Number of Samples*. Excel will then randomly select this number of samples from the specified data set. However, this tool generates random samples *with replacement*, so we must be careful to check for duplicate observations in the sample created.

Figure 6.2 shows 20 samples generated by the tool. We sorted them in ascending order to make it easier to identify duplicates. As you can see, two of the customers were duplicated by the tool.

Other methods of sampling include the following:

- ***Systematic (Periodic) Sampling.*** **Systematic, or periodic, sampling** is a sampling plan (one of the options in the Excel *Sampling* tool) that selects every *n*th item from the population. For example, to sample 250 names from a list of 400,000, the first name could be selected at random from the first 1,600, and then every 1,600th name could be selected. This approach can be used for telephone sampling when supported by an automatic dialer that is programmed to dial numbers in a systematic manner. However, systematic sampling is not the same

Figure 6.2

Samples Generated Using the Excel *Sampling* Tool

	A
1	Sample of Customer IDs
2	10009
3	10092
4	10102
5	10118
6	10167
7	10176
8	10256
9	10261
10	10266
11	10293
12	10320
13	10336
14	10355
15	10355
16	10377
17	10393
18	10413
19	10438
20	10438
21	10455

as simple random sampling because for any sample, every possible sample of a given size in the population does not have an equal chance of being selected. In some situations, this approach can induce significant bias if the population has some underlying pattern. For instance, sampling orders received every 7 days may not yield a representative sample if customers tend to send orders on certain days every week.

- ***Stratified Sampling.*** **Stratified sampling** applies to populations that are divided into natural subsets (called *strata*) and allocates the appropriate proportion of samples to each stratum. For example, a large city may be divided into political districts called wards. Each ward has a different number of citizens. A stratified sample would choose a sample of individuals in each ward proportionate to its size. This approach ensures that each stratum is weighted by its size relative to the population and can provide better results than simple random sampling if the items in each stratum are not homogeneous. However, issues of cost or significance of certain strata might make a disproportionate sample more useful. For example, the ethnic or racial mix of each ward might be significantly different, making it difficult for a stratified sample to obtain the desired information.
- ***Cluster Sampling.*** **Cluster sampling** is based on dividing a population into subgroups (clusters), sampling a set of clusters, and (usually) conducting a complete census within the clusters sampled. For instance, a company might segment its customers into small geographical regions. A cluster sample would consist of a random sample of the geographical regions, and all customers within these regions would be surveyed (which might be easier because regional lists might be easier to produce and mail).
- ***Sampling from a Continuous Process.*** Selecting a sample from a continuous manufacturing process can be accomplished in two main ways. First, select a time at random; then select the next n items produced after that time. Second, select n times at random; then select the next item produced after each of these times. The first approach generally ensures that the observations will come from a homogeneous population; however, the second approach might include items from different populations if the characteristics of the process should change over time, so caution should be used.

Analytics in Practice: Using Sampling Techniques to Improve Distribution[1]

U.S. breweries rely on a three-tier distribution system to deliver product to retail outlets, such as supermarkets and convenience stores, and on-premise accounts, such as bars and restaurants. The three tiers are the manufacturer, wholesaler (distributor), and retailer. A distribution network must be as efficient and cost effective as possible to deliver to the market a fresh product that is damage free and is delivered at the right place at the right time.

To understand distributor performance related to overall effectiveness, MillerCoors brewery defined seven attributes of proper distribution and collected data from 500 of its distributors. A field quality specialist (FQS) audits distributors within an assigned region of the country and collects data on these attributes. The FQS uses a handheld device to scan the universal product code on each package to identify the product type and amount. When audits are complete, data are summarized and uploaded from the handheld device into a master database.

This distributor auditing uses stratified random sampling with proportional allocation of samples based on the distributor's market share. In addition to providing a more representative sample and better logistical control of sampling, stratified random sampling enhances statistical precision when data are aggregated by market area served by the distributor. This enhanced precision is a consequence of smaller and typically homogeneous market regions, which are able to provide realistic estimates of variability, especially when compared to another market region that is markedly different.

Stephen Finn/Shutterstock.com

Randomization of retail accounts is achieved through a specially designed program based on the GPS location of the distributor and serviced retail accounts. The sampling strategy ultimately addresses a specific distributor's performance related to out-of-code product, damaged product, and out-of-rotation product at the retail level. All in all, more than 6,000 of the brewery's national retail accounts are audited during a sampling year. Data collected by the FQSs during the year are used to develop a performance ranking of distributors and identify opportunities for improvement.

Estimating Population Parameters

Sample data provide the basis for many useful analyses to support decision making. **Estimation** involves assessing the value of an unknown population parameter—such as a population mean, population proportion, or population variance—using sample data. **Estimators** are the measures used to estimate population parameters; for example, we use the sample mean $\bar{x}$ to estimate a population mean μ. The sample variance s^2 estimates a population variance σ^2, and the sample proportion p estimates a population proportion π. A **point estimate** is a single number derived from sample data that is used to estimate the value of a population parameter.

[1]Based on Tony Gojanovic and Ernie Jimenez, "Brewed Awakening: Beer Maker Uses Statistical Methods to Improve How Its Products Are Distributed," *Quality Progress* (April 2010).

Unbiased Estimators

It seems quite intuitive that the sample mean should provide a good point estimate for the population mean. However, it may not be clear why the formula for the sample variance that we introduced in Chapter 4 has a denominator of $n - 1$, particularly because it is different from the formula for the population variance (see formulas (4.4) and (4.5) in Chapter 4). In these formulas, the population variance is computed by

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$$

whereas the sample variance is computed by the formula

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

Why is this so? Statisticians develop many types of estimators, and from a theoretical as well as a practical perspective, it is important that they "truly estimate" the population parameters they are supposed to estimate. Suppose that we perform an experiment in which we repeatedly sampled from a population and computed a point estimate for a population parameter. Each individual point estimate will vary from the population parameter; however, we would hope that the long-term average (expected value) of all possible point estimates would equal the population parameter. If the expected value of an estimator equals the population parameter it is intended to estimate, the estimator is said to be *unbiased*. If this is not true, the estimator is called *biased* and will not provide correct results.

Fortunately, all the estimators we have introduced are unbiased and, therefore, are meaningful for making decisions involving the population parameter. In particular, statisticians have shown that the denominator $n - 1$ used in computing s^2 is necessary to provide an unbiased estimator of σ^2. If we simply divided by the number of observations, the estimator would tend to underestimate the true variance.

Errors in Point Estimation

One of the drawbacks of using point estimates is that they do not provide any indication of the magnitude of the potential error in the estimate. A major metropolitan newspaper reported that, based on a Bureau of Labor Statistics survey, college professors were the highest-paid workers in the region, with an average salary of $150,004. Actual averages for two local universities were less than $70,000. What happened? As reported in a follow-up story, the sample size was very small and included a large number of highly paid medical school faculty; as a result, there was a significant error in the point estimate that was used.

When we sample, the estimators we use—such as a sample mean, sample proportion, or sample variance—are actually random variables that are characterized by some distribution. By knowing what this distribution is, we can use probability theory to quantify the uncertainty associated with the estimator. To understand this, we first need to discuss sampling error and sampling distributions.

Sampling Error

In Chapter 4, we observed that different samples from the same population have different characteristics—for example, variations in the mean, standard deviation, frequency distribution, and so on. **Sampling (statistical) error** occurs because samples are only a subset of the total population. Sampling error is inherent in any sampling process, and although it can be minimized, it cannot be totally avoided. Another type of error, called **nonsampling error**, occurs when the sample does not represent the target population adequately. This is generally a result of poor sample design, such as using a convenience sample when a simple random sample would have been more appropriate or choosing the wrong population frame. It may also result from inadequate data reliability, which we discussed in Chapter 1. To draw good conclusions from samples, analysts need to eliminate nonsampling error and understand the nature of sampling error.

Sampling error depends on the size of the sample relative to the population. Thus, determining the number of samples to take is essentially a statistical issue that is based on the accuracy of the estimates needed to draw a useful conclusion. We discuss this later in this chapter. However, from a practical standpoint, one must also consider the cost of sampling and sometimes make a trade-off between cost and the information that is obtained.

Understanding Sampling Error

Suppose that we estimate the mean of a population using the sample mean. How can we determine how accurate we are? In other words, can we make an informed statement about how far the sample mean might be from the true population mean? We could gain some insight into this question by performing a sampling experiment.

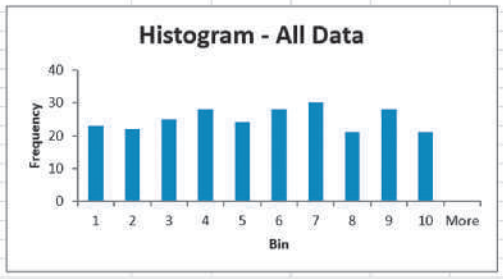
EXAMPLE 6.3 A Sampling Experiment

Let us choose a population that is uniformly distributed between $a = 0$ and $b = 10$. Formulas (5.17) and (5.18) state that the expected value is $(0 + 10)/2 = 5$, and the variance is $(10 - 0)^2/12 = 8.333$. We use the Excel *Random Number Generation* tool described in Chapter 5 to generate 25 samples, each of size 10 from this population. Figure 6.3 shows a portion of a spreadsheet for this experiment, along with a histogram of the data (on the left side) that shows that the 250 observations are approximately uniformly distributed. (This is available in the Excel file *Sampling Experiment*.)

In row 12 we compute the mean of each sample. These statistics vary a lot from the population values because of sampling error. The histogram on the right shows the distribution of the 25 sample means, which vary from less than 4 to more than 6. Now let's compute the average and standard deviation of the sample means in row 12 (cells AB12 and AB13). Note that the average of all the sample means is quite close to the true population mean of 5.0.

Now let us repeat this experiment for larger sample sizes. Table 6.1 shows some results. Notice that as the sample size gets larger, the averages of the 25 sample means are all still close to the expected value of 5; however, the standard deviation of the 25 sample means becomes smaller for increasing sample sizes, meaning that the means of samples are clustered closer together around the true expected value. Figure 6.4 shows comparative histograms of the sample means for each of these cases. These illustrate the conclusions we just made and, also, perhaps even more surprisingly, the distribution of the sample means appears to assume the shape of a normal distribution for larger sample sizes. In our experiment, we used only 25 sample means. If we had used a much-larger number, the distributions would have been more well defined.

	A	B	C	D	E	F	W	X	Y	Z	AA	AB
1	Observation	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 22	Sample 23	Sample 24	Sample 25		
2	1	5.3935	0.8756	9.9338	4.3294	7.1908	8.2244	8.4655	1.9404	9.9133		
3	2	2.8282	2.5047	6.4480	5.9877	7.3946	3.3000	0.3632	9.8871	5.1079		
4	3	5.2715	0.6949	1.5015	8.3935	3.1559	7.1023	9.3628	1.7844	7.3937		
5	4	5.4912	0.7739	8.1466	5.5205	2.4586	8.7262	9.1598	7.5820	1.8513		
6	5	9.3158	4.4591	6.3573	3.8679	1.1493	3.3854	1.2482	1.9391	5.4405		
7	6	7.9745	6.9784	7.9962	2.3157	7.8564	8.9032	3.8716	5.8525	5.4164		
8	7	6.7043	8.4039	5.1088	9.1098	1.1802	2.7732	2.4815	9.0817	4.3889		
9	8	1.3041	2.5678	6.1794	7.8396	6.2709	0.5692	2.5800	1.1911	7.2430		
10	9	0.9870	6.3964	8.2269	9.6112	6.6814	2.8306	4.6004	9.0274	6.1232		
11	10	9.9493	9.3936	4.5015	5.2385	0.6970	3.7074	5.9062	0.6592	7.5021		
12	Sample mean	5.5219	4.3048	6.4400	6.2214	4.4035	4.9522	4.8039	4.8945	6.0380	Average	5.0108
13											Standard Dev.	0.816673



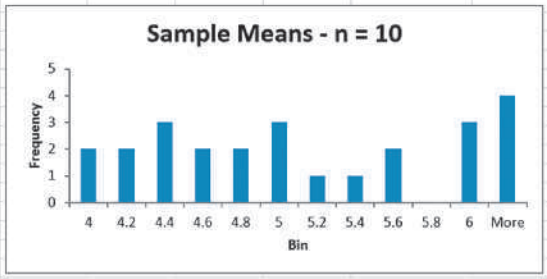


Figure 6.3

Portion of Spreadsheet for *Sampling Experiment*

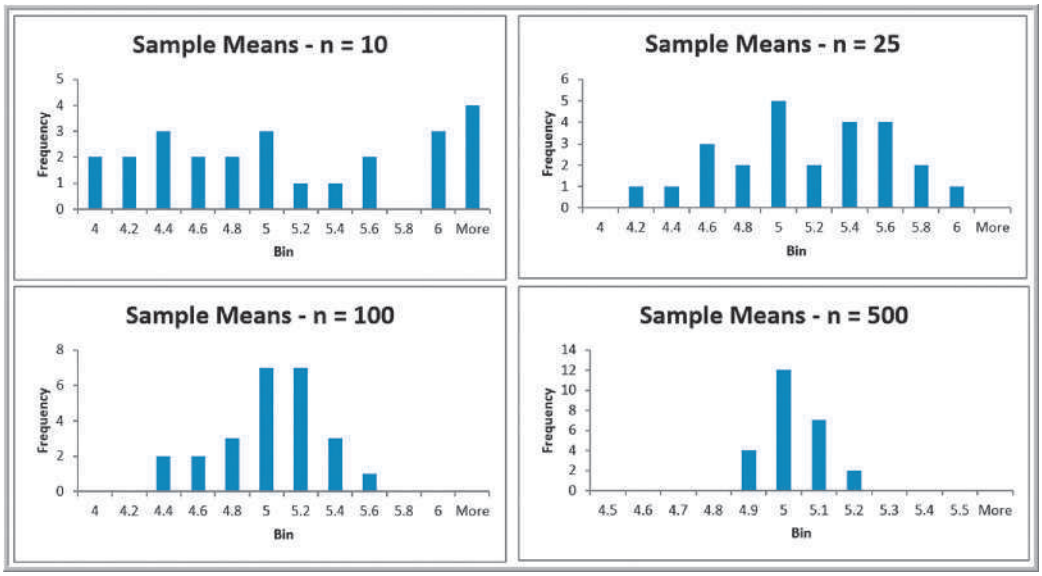
Table 6.1

Results from *Sampling Experiment*

Sample Size	Average of 25 Sample Means	Standard Deviation of 25 Sample Means
10	5.0108	0.816673
25	5.0779	0.451351
100	4.9173	0.301941
500	4.9754	0.078993

Figure 6.4

Histograms of Sample Means for Increasing Sample Sizes



If we apply the empirical rules to these results, we can estimate the sampling error associated with one of the sample sizes we have chosen.

EXAMPLE 6.4 Estimating Sampling Error Using the Empirical Rules

Using the results in Table 6.1 and the empirical rule for three standard deviations around the mean, we could state, for example, that using a sample size of 10, the distribution of sample means should fall approximately from $5.0 - 3(0.816673) = 2.55$ to $5.0 + 3(0.816673) = 7.45$. Thus, there is considerable error in estimating the mean using a sample of only 10. For a sample of size 25, we would expect the sample means to fall between $5.0 - 3(0.451351) = 3.65$ to $5.0 + 3(0.451351) = 6.35$. Note that as the sample size increased, the error decreased. For sample sizes of 100 and 500, the intervals are [4.09, 5.91] and [4.76, 5.24].

Sampling Distributions

We can quantify the sampling error in estimating the mean for any unknown population. To do this, we need to characterize the sampling distribution of the mean.

Sampling Distribution of the Mean

The means of *all possible* samples of a fixed size n from some population will form a distribution that we call the **sampling distribution of the mean**. The histograms in Figure 6.4 are approximations to the sampling distributions of the mean based on 25 samples. Statisticians have shown two key results about the sampling distribution of the mean. First, the standard deviation of the sampling distribution of the mean, called the **standard error of the mean**, is computed as

$$\text{Standard Error of the Mean} = \sigma/\sqrt{n} \tag{6.1}$$

where σ is the standard deviation of the population from which the individual observations are drawn and n is the sample size. From this formula, we see that as n increases, the standard error decreases, just as our experiment demonstrated. This suggests that the estimates of the mean that we obtain from larger sample sizes provide greater accuracy in estimating the true population mean. In other words, *larger sample sizes have less sampling error*.

EXAMPLE 6.5 Computing the Standard Error of the Mean

For our experiment, we know that the variance of the population is 8.33 (because the values were uniformly distributed). Therefore, the standard deviation of the population is $\sigma = 2.89$. We may compute the standard error of the mean for each of the sample sizes in our experiment using formula (6.1). For example, with $n = 10$, we have

$$\text{Standard Error of the Mean} = \sigma/\sqrt{n} = 2.89/\sqrt{10} = 0.914$$

For the remaining data in Table 6.1 we have the following:

Sample Size, n	Standard Error of the Mean
10	0.914
25	0.577
100	0.289
500	0.129

The standard deviations shown in Table 6.1 are simply estimates of the standard error of the mean based on the limited number of 25 samples. If we compare these estimates with the theoretical values in the previous example, we see that they are close but not exactly the same. This is because the true standard error is based on *all possible* sample means in the sampling

distribution, whereas we used only 25. If you repeat the experiment with a larger number of samples, the observed values of the standard error would be closer to these theoretical values.

In practice, we will never know the true population standard deviation and generally take only a limited sample of n observations. However, we may estimate the standard error of the mean using the sample data by simply dividing the sample standard deviation by the square root of n.

The second result that statisticians have shown is called the **central limit theorem**, one of the most important practical results in statistics that makes systematic inference possible. The central limit theorem states that if the sample size is large enough, the sampling distribution of the mean is approximately normally distributed, *regardless* of the distribution of the population and that the mean of the sampling distribution will be the same as that of the population. This is exactly what we observed in our experiment. The distribution of the population was uniform, yet the sampling distribution of the mean converges to the shape of a normal distribution as the sample size increases. The central limit theorem also states that if the population is normally distributed, then the sampling distribution of the mean will also be normal for *any* sample size. The central limit theorem allows us to use the theory we learned about calculating probabilities for normal distributions to draw conclusions about sample means.

Applying the Sampling Distribution of the Mean

The key to applying sampling distribution of the mean correctly is to understand whether the probability that you wish to compute relates to an individual observation or to the mean of a sample. If it relates to the mean of a sample, then you must use the sampling distribution of the mean, whose standard deviation is the standard error, $\sigma/\sqrt{n}$.

EXAMPLE 6.6 Using the Standard Error in Probability Calculations

Suppose that the size of individual customer orders (in dollars), X, from a major discount book publisher Web site is normally distributed with a mean of \$36 and standard deviation of \$8. The probability that the next individual who places an order at the Web site will make a purchase of more than \$40 can be found by calculating

1 − NORM.DIST(40,36,8,TRUE) = 1 − 0.6915 = 0.3085

Now suppose that a sample of 16 customers is chosen. What is the probability that the *mean purchase* for these 16 customers will exceed \$40? To find this, we must realize that we must use the sampling distribution of the mean to carry out the appropriate calculations. The sampling distribution of the mean will have a mean of \$36 but a standard error of $\$8/\sqrt{16} = \$2$. Then the probability that the mean purchase exceeds \$40 for a sample size of $n = 16$ is

1 − NORM.DIST(40,36,2,TRUE) = 1 − 0.9772 = 0.0228

Although about 30% of individuals will make purchases exceeding \$40, the chance that 16 customers will collectively average more than \$40 is much smaller. It would be very unlikely for all 16 customers to make high-volume purchases, because some individual purchases would as likely be less than \$36 as more, making the variability of the mean purchase amount for the sample of 16 much smaller than for individuals.

Interval Estimates

An **interval estimate** provides a range for a population characteristic based on a sample. Intervals are quite useful in statistics because they provide more information than a point estimate. Intervals specify a range of plausible values for the characteristic of interest and a way of assessing "how plausible" they are. In general, a $100(1-\alpha)\%$ **probability interval** is any interval $[A, B]$ such that the probability of falling between A and B is $1-\alpha$. Probability intervals are often centered on the mean or median. For instance,

in a normal distribution, the mean plus or minus 1 standard deviation describes an approximate 68% probability interval around the mean. As another example, the 5th and 95th percentiles in a data set constitute a 90% probability interval.

EXAMPLE 6.7 Interval Estimates in the News

We see interval estimates in the news all the time when trying to estimate the mean or proportion of a population. Interval estimates are often constructed by taking a point estimate and adding and subtracting a margin of error that is based on the sample size. For example, a Gallup poll might report that 56% of voters support a certain candidate with a margin of error of ±3%. We would conclude that the true percentage of voters that support the candidate is most likely between 53% and 59%. Therefore, we would have a lot of confidence in predicting that the candidate would win a forthcoming election. If, however, the poll showed a 52% level of support with a margin of error of ±4%, we might not be as confident in predicting a win because the true percentage of supportive voters is likely to be somewhere between 48% and 56%.

The question you might be asking at this point is how to calculate the error associated with a point estimate. In national surveys and political polls, such margins of error are usually stated, but they are never properly explained. To understand them, we need to introduce the concept of confidence intervals.

Confidence Intervals

Confidence interval estimates provide a way of assessing the accuracy of a point estimate. A **confidence interval** is a range of values between which the value of the population parameter is believed to be, along with a probability that the interval correctly estimates the true (unknown) population parameter. This probability is called the **level of confidence**, denoted by $1 - \alpha$, where α is a number between 0 and 1. The level of confidence is usually expressed as a percent; common values are 90%, 95%, or 99%. (Note that if the level of confidence is 90%, then $\alpha = 0.1$.) The margin of error depends on the level of confidence and the sample size. For example, suppose that the margin of error for some sample size and a level of confidence of 95% is calculated to be 2.0. One sample might yield a point estimate of 10. Then, a 95% confidence interval would be [8, 12]. However, this interval may or may not include the true population mean. If we take a different sample, we will most likely have a different point estimate, say, 10.4, which, given the same margin of error, would yield the interval estimate [8.4, 12.4]. Again, this may or may not include the true population mean. If we chose 100 different samples, leading to 100 different interval estimates, we would expect that 95% of them—the level of confidence—would contain the true population mean. We would say we are "95% confident" that the interval we obtain from sample data contains the true population mean. The higher the confidence level, the more assurance we have that the interval contains the true population parameter. As the confidence level increases, the confidence interval becomes wider to provide higher levels of assurance. You can view α as the risk of incorrectly concluding that the confidence interval contains the true mean.

When national surveys or political polls report an interval estimate, they are actually confidence intervals. However, the level of confidence is generally not stated because the average person would probably not understand the concept or terminology. While not stated, you can probably assume that the level of confidence is 95%, as this is the most common value used in practice (however, the Bureau of Labor Statistics tends to use 90% quite often).

Many different types of confidence intervals may be developed. The formulas used depend on the population parameter we are trying to estimate and possibly other characteristics or assumptions about the population. We illustrate a few types of confidence intervals.

Confidence Interval for the Mean with Known Population Standard Deviation

The simplest type of confidence interval is for the mean of a population where the standard deviation is assumed to be known. You should realize, however, that in nearly all practical sampling applications, the population standard deviation will *not* be known. However, in some applications, such as measurements of parts from an automated machine, a process might have a very stable variance that has been established over a long history, and it can reasonably be assumed that the standard deviation is known.

A $100(1 - \alpha)\%$ confidence interval for the population mean μ based on a sample of size n with a sample mean $\bar{x}$ and a known population standard deviation σ is given by

$$\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n}) \tag{6.2}$$

Note that this formula is simply the sample mean (point estimate) plus or minus a margin of error.

The margin of error is a number $z_{\alpha/2}$ multiplied by the standard error of the sampling distribution of the mean, $\sigma/\sqrt{n}$. The value $z_{\alpha/2}$ represents the value of a standard normal random variable that has an upper tail probability of $\alpha/2$ or, equivalently, a cumulative probability of $1 - \alpha/2$. It may be found from the standard normal table (see Table A.1 in Appendix A at the end of the book) or may be computed in Excel using the value of the function NORM.S.INV$(1 - \alpha/2)$. For example, if $\alpha = 0.05$ (for a 95% confidence interval), then NORM.S.INV$(0.975) = 1.96$; if $\alpha = 0.10$ (for a 90% confidence interval), then NORM.S.INV$(0.95) = 1.645$, and so on.

Although formula (6.2) can easily be implemented in a spreadsheet, the Excel function CONFIDENCE.NORM(*alpha*, *standard_deviation, size*) can be used to compute the margin of error term, $z_{\alpha/2}\,\sigma/\sqrt{n}$; thus, the confidence interval is the sample mean $\pm$ CONFIDENCE.NORM(*alpha*, *standard_deviation*, *size*).

EXAMPLE 6.8 Computing a Confidence Interval with a Known Standard Deviation

In a production process for filling bottles of liquid detergent, historical data have shown that the variance in the volume is constant; however, clogs in the filling machine often affect the average volume. The historical standard deviation is 15 milliliters. In filling 800-milliliter bottles, a sample of 25 found an average volume of 796 milliliters. Using formula (6.2), a 95% confidence interval for the population mean is

$$\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$$
$$= 796 \pm 1.96(15/\sqrt{25}) = 796 \pm 5.88, \text{ or } [790.12, 801.88]$$

The worksheet *Population Mean Sigma Known* in the Excel workbook *Confidence Intervals* computes this interval using the CONFIDENCE.NORM function to compute the margin of error in cell B9, as shown in Figure 6.5.

As the level of confidence, $1 - \alpha$, decreases, $z_{\alpha/2}$ decreases, and the confidence interval becomes narrower. For example, a 90% confidence interval will be narrower than a 95% confidence interval. Similarly, a 99% confidence interval will be wider than a 95% confidence interval. Essentially, you must trade off a higher level of accuracy with the risk that the confidence interval does not contain the true mean. Smaller risk will result in a

Figure 6.5

Confidence Interval for Mean Liquid Detergent Filling Volume

	A	B	C	D	E	F
1	**Confidence Interval for Population Mean, Standard Deviation Known**					
2						
3	**Alpha**	0.05				
4	**Standard deviation**	15				
5	**Sample size**	25				
6	**Sample average**	796				
7						
8	**Confidence Interval**	95%				
9	**Error**	5.879892				
10	**Lower**	790.1201				
11	**Upper**	801.8799				

wider confidence interval. However, you can also see that as the sample size increases, the standard error decreases, making the confidence interval narrower and providing a more accurate interval estimate for the same level of risk. So if you wish to reduce the risk, you should consider increasing the sample size.

The *t*-Distribution

In most practical applications, the standard deviation of the population is unknown, and we need to calculate the confidence interval differently. Before we can discuss how to compute this type of confidence interval, we need to introduce a new probability distribution called the ***t*-distribution**. The *t*-distribution is actually a family of probability distributions with a shape similar to the standard normal distribution. Different *t*-distributions are distinguished by an additional parameter, **degrees of freedom (*df*)**. The *t*-distribution has a larger variance than the standard normal, thus making confidence intervals wider than those obtained from the standard normal distribution, in essence correcting for the uncertainty about the true standard deviation, which is not known. As the number of degrees of freedom increases, the *t*-distribution converges to the standard normal distribution (Figure 6.6). When sample sizes get to be as large as 120, the distributions are virtually identical; even for sample sizes as low as 30 to 35, it becomes difficult to distinguish between the two. Thus, for large sample sizes, many people use *z*-values to establish confidence intervals even when the standard deviation is unknown. We must point out, however, that for any sample size, the *true* sampling distribution of the mean is the *t*-distribution, so when in doubt, use the *t*.

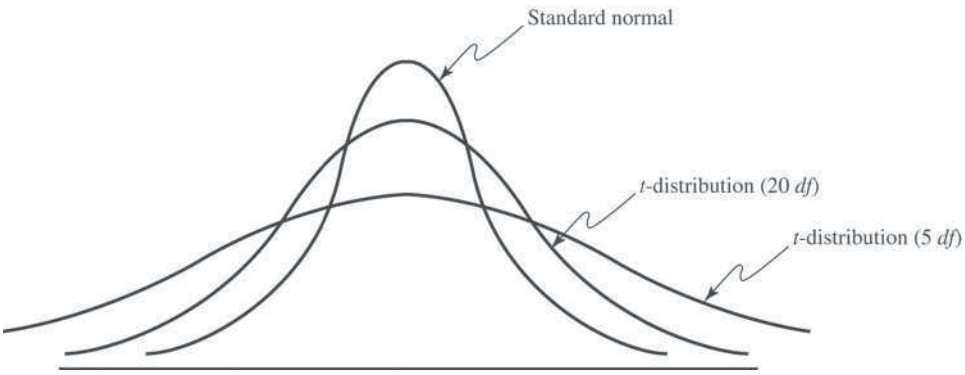
The concept of degrees of freedom can be puzzling. It can best be explained by examining the formula for the sample variance:

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

Note that to compute s^2, we first need to compute the sample mean, $\bar{x}$. If we know the value of the mean, then we need know only $n - 1$ distinct observations; the nth is completely determined. (For instance, if the mean of three values is 4 and you know that two of the values are 2 and 4, you can easily determine that the third number must be 6.) The number of sample values that are free to vary defines the number of degrees of freedom; in general, *df* equals the number of sample values minus the number of estimated parameters. Because the sample variance uses one estimated parameter, the mean, the *t*-distribution used in confidence interval calculations has $n - 1$ degrees of freedom. Because the *t*-distribution explicitly accounts for the effect of the sample size in estimating the population variance, it is the proper one to use for any sample size. However, for large samples, the difference between *t*- and *z*-values is very small, as we noted earlier.

Figure 6.6

Comparison of the *t*-Distribution to the Standard Normal Distribution



Confidence Interval for the Mean with Unknown Population Standard Deviation

The formula for a $100(1 - \alpha)\%$ confidence interval for the mean μ when the population standard deviation is unknown is

$$\bar{x} \pm t_{\alpha/2,n-1}(s/\sqrt{n}) \tag{6.3}$$

where $t_{\alpha/2,n-1}$ is the value from the *t*-distribution with $n - 1$ degrees of freedom, giving an upper-tail probability of $\alpha/2$. We may find *t*-values in Table A.2 in Appendix A at the end of the book or by using the Excel function T.INV$(1 - \alpha/2, n - 1)$ or the function T.INV.2T$(\alpha, n - 1)$. The Excel function CONFIDENCE.T(*alpha*, *standard_deviation*, *size*) can be used to compute the margin of error term, $t_{\alpha/2,n-1}(s/\sqrt{n})$; thus, the confidence interval is the sample mean $\pm$CONFIDENCE.T.

EXAMPLE 6.9 Computing a Confidence Interval with Unknown Standard Deviation

In the Excel file *Credit Approval Decisions*, a large bank has sample data used in making credit approval decisions (see Figure 6.7). Suppose that we want to find a 95% confidence interval for the mean revolving balance for the population of applicants that own a home. First, sort the data by homeowner and compute the mean and standard deviation of the revolving balance for the sample of homeowners. This results in $\bar{x}$ = \$12,630.37 and s = \$5393.38. The sample size is n = 27, so the standard error $s/\sqrt{n}$ = \$1037.96. The *t*-distribution has 26 degrees of freedom; therefore, $t_{.025,26}$ = 2.056. Using formula (6.3), the confidence interval is \$12,630.37 ± 2.056(\$1037.96) or [\$10,496, \$14,764]. The worksheet *Population Mean Sigma Unknown* in the Excel workbook *Confidence Intervals* computes this interval using the CONFIDENCE.T function to compute the margin of error in cell B10, as shown in Figure 6.8.

Confidence Interval for a Proportion

For categorical variables such as gender (male or female), education (high school, college, post-graduate), and so on, we are usually interested in the *proportion* of observations in a sample that has a certain characteristic. An unbiased estimator of a population proportion π (this is not the *number* pi = 3.14159 . . .) is the statistic $\hat{p} = x/n$ (the **sample proportion**), where x is the number in the sample having the desired characteristic and n is the sample size.

	A	B	C	D	E	F
1	Credit Approval Decisions					
2						
3	Homeowner	Credit Score	Years of Credit History	Revolving Balance	Revolving Utilization	Decision
4	Y	725	20	$ 11,320	25%	Approve
5	Y	573	9	$ 7,200	70%	Reject
6	Y	677	11	$ 20,000	55%	Approve
7	N	625	15	$ 12,800	65%	Reject
8	N	527	12	$ 5,700	75%	Reject
9	Y	795	22	$ 9,000	12%	Approve
10	N	733	7	$ 35,200	20%	Approve

Figure 6.7

Portion of Excel File *Credit Approval Decisions*

Figure 6.8

Confidence Interval for Mean Revolving Balance of Homeowners

	A	B	C	D	E
1	Confidence Interval for Population Mean, Standard Deviation Unknown				
2					
3	Alpha	0.05			
4	Sample standard deviation	5393.38			
5	Sample size	27			
6	Sample average	12630.37			
7					
8	Confidence Interval	95%			
9	t-value	2.056			
10	Error	2133.55			
11	Lower	10496.82			
12	Upper	14763.92			

A $100(1 - \alpha)\%$ confidence interval for the proportion is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{6.4}$$

Notice that as with the mean, the confidence interval is the point estimate plus or minus some margin of error. In this case, $\sqrt{\hat{p}(1 - \hat{p})/n}$ is the standard error for the sampling distribution of the proportion. Excel does not have a function for computing the margin of error, but it can easily be implemented on a spreadsheet.

EXAMPLE 6.10 Computing a Confidence Interval for a Proportion

The last column in the Excel file *Insurance Survey* (see Figure 6.9) describes whether a sample of employees would be willing to pay a lower premium for a higher deductible for their health insurance. Suppose we are interested in the proportion of individuals who answered yes. We may easily confirm that 6 out of the 24 employees, or 25%, answered yes. Thus, a point estimate for the proportion answering yes is $\hat{p} = 0.25$. Using formula (6.4), we find that a 95% confidence interval for the proportion of employees answering yes is

$$0.25 \pm 1.96\sqrt{\frac{0.25(0.75)}{24}} = 0.25 \pm 0.173, \textit{ or } [0.077, 0.423]$$

The worksheet *Population Mean Sigma Unknown* in the Excel workbook *Confidence Intervals* computes this interval, as shown in Figure 6.10. Notice that this is a fairly wide confidence interval, suggesting that we have quite a bit of uncertainty as to the true value of the population proportion. This is because of the relatively small sample size.

	A	B	C	D	E	F	G
1	Insurance Survey						
2							
3	Age	Gender	Education	Marital Status	Years Employed	Satisfaction*	Premium/Deductible**
4	36	F	Some college	Divorced	4	4	N
5	55	F	Some college	Divorced	2	1	N
6	61	M	Graduate degree	Widowed	26	3	N
7	65	F	Some college	Married	9	4	N
8	53	F	Graduate degree	Married	6	4	N
9	50	F	Graduate degree	Married	10	5	N
10	28	F	College graduate	Married	4	5	N
11	62	F	College graduate	Divorced	9	3	N
12	48	M	Graduate degree	Married	6	5	N

Figure 6.9

Portion of Excel File *Insurance Survey*

Figure 6.10

Confidence Interval for the Proportion

	A	B
1	Confidence Interval for a Proportion	
2		
3	Alpha	0.05
4	Sample proportion	0.25
5	Sample size	24
6		
7	Confidence Interval	95%
8	z-value	1.96
9	Standard error	0.088388
10	Lower	0.076762
11	Upper	0.423238

Additional Types of Confidence Intervals

Confidence intervals may be calculated for other population parameters such as a variance or standard deviation and also for differences in the means or proportions of two populations. The concepts are similar to the types of confidence intervals we have discussed, but many of the formulas are rather complex and more difficult to implement on a spreadsheet. Some advanced software packages and spreadsheet add-ins provide additional support. Therefore, we do not discuss them in this book, but we do suggest that you consult other books and statistical references should you need to use them, now that you understand the basic concepts underlying them.

Using Confidence Intervals for Decision Making

Confidence intervals can be used in many ways to support business decisions.

EXAMPLE 6.11 Drawing a Conclusion about a Population Mean Using a Confidence Interval

In packaging a commodity product such as laundry detergent, the manufacturer must ensure that the packages contain the stated amount to meet government regulations. In Example 6.8, we saw an example where the required volume is 800 milliliters, yet the sample average was only 796 milliliters. Does this indicate a serious problem? Not necessarily. The 95% confidence interval for the mean we computed in Figure 6.5 was [790.12, 801.88]. Although the sample mean is less than 800, the sample does not provide sufficient evidence to draw that conclusion that the

population mean is less than 800 because 800 is contained within the confidence interval. In fact, it is just as plausible that the population mean is 801. We cannot tell definitively because of the sampling error. However, suppose that the sample average is 792. Using the Excel worksheet *Population Mean Sigma Known* in the workbook *Confidence Intervals*, we find that the confidence interval for the mean would be [786.12, 797.88]. In this case, we would conclude that it is highly unlikely that the population mean is 800 milliliters because the confidence interval falls completely below 800; the manufacturer should check and adjust the equipment to meet the standard.

The next example shows how to interpret a confidence interval for a proportion.

EXAMPLE 6.12 Using a Confidence Interval to Predict Election Returns

Suppose that an exit poll of 1,300 voters found that 692 voted for a particular candidate in a two-person race. This represents a proportion of 53.23% of the sample. Could we conclude that the candidate will likely win the election? A 95% confidence interval for the proportion is [0.505, 0.559]. This suggests that the population proportion of voters who favor this candidate is highly likely to exceed 50%, so it is safe to predict the winner. On the other hand, suppose that only 670 of the 1,300 voters voted for the candidate, a sample proportion of 0.515. The confidence interval for the population proportion is [0.488, 0.543]. Even though the sample proportion is larger than 50%, the sampling error is large, and the confidence interval suggests that it is reasonably likely that the true population proportion could be less than 50%, so it would not be wise to predict the winner based on this information.

Prediction Intervals

Another type of interval used in estimation is a prediction interval. A **prediction interval** is one that provides a range for predicting the value of a new observation from the same population. This is different from a confidence interval, which provides an interval estimate of a population parameter, such as the mean or proportion. A confidence interval is associated with the *sampling distribution* of a statistic, but a prediction interval is associated with the distribution of the random variable itself.

When the population standard deviation is unknown, a $100(1 - \alpha)\%$ prediction interval for a new observation is

$$\bar{x} \pm t_{\alpha/2,n-1}\left(s\sqrt{1 + \frac{1}{n}}\right) \tag{6.5}$$

Note that this interval is wider than the confidence interval in formula (6.3) by virtue of the additional value of 1 under the square root. This is because, in addition to estimating the population mean, we must also account for the variability of the new observation around the mean.

One important thing to realize also is that in formula (6.3) for a confidence interval, as n gets large, the error term tends to zero so the confidence interval converges on the mean. However, in the prediction interval formula (6.5), as n gets large, the error term converges to $t_{\alpha/2,\,n-1}(s)$, which is simply a $100(1 - \alpha)\%$ probability interval. Because we are trying to predict a new observation from the population, there will always be uncertainty.

EXAMPLE 6.13 Computing a Prediction Interval

In estimating the revolving balance in the Excel file *Credit Approval Decisions* in Example 6.9, we may use formula (6.5) to compute a 95% prediction interval for the revolving balance of a new homeowner as

$$\$12{,}630.37 \pm 2.056(\$5{,}393.38)\sqrt{1 + \frac{1}{27}}, \text{ or}$$

[$338.10, $23,922.64]

Note that compared with Example 6.9, the size of the prediction interval is considerably wider than that of the confidence interval.

Confidence Intervals and Sample Size

An important question in sampling is the size of the sample to take. Note that in all the formulas for confidence intervals, the sample size plays a critical role in determining the width of the confidence interval. As the sample size increases, the width of the confidence interval decreases, providing a more accurate estimate of the true population parameter. In many applications, we would like to control the margin of error in a confidence interval. For example, in reporting voter preferences, we might wish to ensure that the margin of error is $\pm 2\%$. Fortunately, it is relatively easy to determine the appropriate sample size needed to estimate the population parameter within a specified level of precision.

The formulas for determining sample sizes to achieve a given margin of error are based on the confidence interval half-widths. For example, consider the confidence interval for the mean with a known population standard deviation we introduced in formula (6.2):

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Suppose we want the width of the confidence interval on either side of the mean (i.e., the margin of error) to be at most E. In other words,

$$E \geq z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Solving for n, we find:

$$n \geq (z_{\alpha/2})^2\frac{\sigma^2}{E^2} \tag{6.6}$$

In a similar fashion, we can compute the sample size required to achieve a desired confidence interval half-width for a proportion by solving the following equation (based on formula (6.4) using the population proportion π in the margin of error term) for n:

$$E \geq z_{\alpha/2}\sqrt{\pi(1-\pi)/n}$$

This yields

$$n \geq (z_{\alpha/2})^2\frac{\pi(1-\pi)}{E^2} \tag{6.7}$$

In practice, the value of π will not be known. You could use the sample proportion from a preliminary sample as an estimate of π to plan the sample size, but this might require several iterations and additional samples to find the sample size that yields the required precision. When no information is available, the most conservative estimate is to set $\pi = 0.5$. This maximizes the quantity $\pi(1-\pi)$ in the formula, resulting in the sample size that will guarantee the required precision no matter what the true proportion is.

Figure 6.11

Confidence Interval for the Mean Using a Sample Size = 97

	A	B	C	D	E	F
1	Confidence Interval for Population Mean, Standard Deviation Known					
2						
3	Alpha	0.05				
4	Standard deviation	15				
5	Sample size	97				
6	Sample average	796				
7						
8	Confidence Interval	95%				
9	Error	2.985063				
10	Lower	793.0149				
11	Upper	798.9851				

EXAMPLE 6.14 Sample Size Determination for the Mean

In the liquid detergent example (Example 6.8), the confidence interval we computed in Figure 6.5 was [790.12, 801.88]. The width of the confidence interval is ±5.88 milliliters, which represents the sampling error. Suppose the manufacturer would like the sampling error to be at most 3 milliliters. Using formula (6.6), we may compute the required sample size as follows:

$$n \geq (z_{\alpha/2})^2 \frac{(\sigma^2)}{E^2}$$

$$= (1.96)^2 \frac{(15^2)}{3^2} = 96.04$$

Rounding up we find that that 97 samples would be needed. To verify this, Figure 6.11 shows that if a sample of 97 is used along with the same sample mean and standard deviation, the confidence interval does indeed have a sampling error of error less than 3 milliliters.

Of course, we generally do not know the population standard deviation prior to finding the sample size. A commonsense approach would be to take an initial sample to estimate the population standard deviation using the sample standard deviation s and determine the required sample size, collecting additional data if needed. If the half-width of the resulting confidence interval is within the required margin of error, then we clearly have achieved our goal. If not, we can use the new sample standard deviation s to determine a new sample size and collect additional data as needed. Note that if s changes significantly, we still might not have achieved the desired precision and might have to repeat the process. Usually, however, this will be unnecessary.

EXAMPLE 6.15 Sample Size Determination for a Proportion

For the voting example we discussed, suppose that we wish to determine the number of voters to poll to ensure a sampling error of at most ±2%. As we stated, when no information is available, the most conservative approach is to use 0.5 for the estimate of the true proportion. Using formula (6.7) with $\pi = 0.5$, the number of voters to poll to obtain a 95% confidence interval on the proportion of voters that choose a particular candidate with a precision of ±0.02 or less is

$$n \geq (z_{\alpha/2})^2 \frac{\pi(1 - \pi)}{E^2}$$

$$= (1.96)^2 \frac{(0.5)\,(1 - 0.5)}{0.02^2} = 2{,}401$$

Key Terms

Central limit theorem
Cluster sampling
Confidence interval
Convenience sampling
Degrees of freedom (*df*)
Estimation
Estimators
Interval estimate
Judgment sampling
Level of confidence
Nonsampling error
Point estimate
Population frame
Prediction interval
Probability interval
Sample proportion
Sampling (statistical) error
Sampling distribution of the mean
Sampling plan
Simple random sampling
Standard error of the mean
Stratified sampling
Systematic (or periodic) sampling
t-Distribution

Problems and Exercises

1. Your college or university wishes to obtain reliable information about student perceptions of administrative communication. Describe how to design a sampling plan for this situation based on your knowledge of the structure and organization of your college or university. How would you implement simple random sampling, stratified sampling, and cluster sampling for this study? What would be the pros and cons of using each of these methods?

2. Number the rows in the Excel file *Credit Risk Data* to identify each record. The bank wants to sample from this database to conduct a more-detailed audit. Use the Excel *Sampling* tool to find a simple random sample of 20 unique records.

3. Describe how to apply stratified sampling to sample from the *Credit Risk Data* file based on the different types of loans. Implement your process in Excel to choose a random sample consisting of 10% of the records for each type of loan.

4. Find the current 30 stocks that comprise the Dow Jones Industrial Average. Set up an Excel spreadsheet for their names, market capitalization, and one or two other key financial statistics (search Yahoo! Finance or a similar Web source). Using the Excel *Sampling* tool, obtain a random sample of 5 stocks, compute point estimates for the mean and standard deviation, and compare them to the population parameters.

5. Repeat the sampling experiment in Example 6.3 for sample sizes 50, 100, 250, and 500. Compare your results to the example and use the empirical rules to analyze the sampling error. For each sample, also find the standard error of the mean using formula (6.1).

6. Uncle's Pizza is doing good business in Delhi due to its prompt home delivery system. It guarantees that the pizza will be delivered within 30 minutes from the time order was placed or the order is free. The time that it takes to deliver each order on time is maintained in the Pizza Time System. Fourteen random entries from the Pizza Time System are listed.

10.1	19.6	12.2	32.6	18.2	29.5	13.2
30	10.8	14.8	22.1	15.6	45.6	15.6

   a. Find the mean for the sample.
   b. Explain if this sample can be used to estimate the average time that it takes for Uncle's Pizza to deliver the pizza.

7. A soft drink bottle filling machine is known to have a mean of 200 ml and a standard variation of 10 ml. The quality control manager took a random sample of the filled bottles and found the sample mean to be 215 ml. She assumed the sample must not be representative. Do you agree with the conclusion made by the quality control manager? Justify your answer.

8. A sample of 33 airline passengers found that the average check-in time is 2.167. Based on long-term data, the population standard deviation is known to be 0.48. Find a 95% confidence interval for the mean check-in time. Use the appropriate formula and verify your result using the *Confidence Intervals* workbook.

9. A sample of 20 international students attending an urban U.S. university found that the average amount budgeted for expenses per month was $1612.50 with a standard deviation of $1179.64. Find a 95% confidence interval for the mean monthly expense budget of the population of international students. Use the appropriate formula and verify your result using the *Confidence Intervals* workbook.

10. A sample of 25 individuals at a shopping mall found that the mean number of visits to a restaurant per week was 2.88 with a standard deviation of 1.59. Find a 99% confidence interval for the mean number of restaurant visits. Use the appropriate formula and verify your result using the *Confidence Intervals* workbook.

11. A bank sampled its customers to determine the proportion of customers who use their debit card at least once each month. A sample of 50 customers found that only 12 use their debit card monthly. Find 95% and 99% confidence intervals for the proportion of customers who use their debit card monthly. Use the appropriate formula and verify your result using the *Confidence Intervals* workbook.

12. If, based on a sample size of 850, a political candidate finds that 458 people would vote for him in a two-person race, what is the 95% confidence interval for his expected proportion of the vote? Would he be confident of winning based on this poll? Use the appropriate formula and verify your result using the *Confidence Intervals* workbook.

13. If, based on a sample size of 200, a political candidate found that 125 people would vote for her in a two-person race, what is the 99% confidence interval for her expected proportion of the vote? Would she be confident of winning based on this poll?

14. Using the data in the Excel file *Accounting Professionals*, find and interpret 95% confidence intervals for the following:

- **a.** mean years of service
- **b.** proportion of employees who have a graduate degree

15. Find the standard deviation of the total assets held by the bank in the Excel file *Credit Risk Data*.

- **a.** Treating the records in the database as a population, use your sample in Problem 2 and compute 90%, 95%, and 99% confidence intervals for the total assets held in the bank by loan applicants using formula (6.2) and any appropriate Excel functions. Explain the differences as the level of confidence increases.
- **b.** How do your confidence intervals differ if you assume that the population standard deviation is not known but estimated using your sample data?

16. The Excel file *Restaurant Sales* provides sample information on lunch, dinner, and delivery sales for a local Italian restaurant. Develop 95% confidence intervals for the mean of each of these variables, as well as total sales for weekdays and weekends. What conclusions can you reach?

17. Using the data in the worksheet *Consumer Transportation Survey*, develop 95% confidence intervals for the following:

- **a.** the proportion of individuals who are satisfied with their vehicle
- **b.** the proportion of individuals who have at least one child

18. The monthly sales of a mobile phone shop have been distributed with a standard deviation of $900. A statistical study of sales in the last nine months has found a confidence interval for the mean of monthly sales with extremes of $5663 and $6839.

- **a.** What were the average sales over the nine month period?
- **b.** What is the confidence level for this interval?

19. Using data in the Excel file *Colleges and Universities*, find 95% confidence intervals for the median SAT for each of the two groups, liberal arts colleges and research universities. Based on these confidence intervals, does there appear to be a difference in the median SAT scores between the two groups?

20. The Excel file *Baseball Attendance* shows the attendance in thousands at San Francisco Giants' baseball games for the 10 years before the Oakland A's moved to the Bay Area in 1968, as well as the combined attendance for both teams for the next 11 years. Develop 95% confidence intervals for the mean attendance of each of the two groups. Based on these confidence intervals, would you conclude that attendance has changed after the move?

21. A random sample of 100 teenagers was surveyed, and the mean number of songs that they had downloaded from the iTunes store in the past month was 9.4 with the results considered accurate is within 1.4 (18 times out of 20).

a. What percent of confidence level is the result?

b. What is the margin of error?

c. What is the confidence interval? Explain.

22. A study of nonfatal occupational injuries in the United States found that about 31% of all injuries in the service sector involved the back. The National Institute for Occupational Safety and Health (NIOSH) recommended conducting a comprehensive ergonomics assessment of jobs and workstations. In response to this information, Mark Glassmeyer developed a unique ergonomic handcart to help field service engineers be more productive and also to reduce back injuries from lifting parts and equipment during service calls. Using a sample of 382 field service engineers who were provided with these carts, Mark collected the following data:

	Year 1 (without Cart)	Year 2 (with Cart)
Average call time	8.27 hours	7.98 hours
Standard deviation call time	1.36 hours	1.21 hours
Proportion of back injuries	0.018	0.010

Find 95% confidence intervals for the average call times and proportion of back injuries in each year. What conclusions would you reach based on your results?

23. Using the data in the worksheet *Consumer Transportation Survey*, develop 95% and 99% prediction intervals for the following:

a. the hours per week that an individual will spend in his or her vehicle

b. the number of miles driven per week

24. The Excel file *Restaurant Sales* provides sample information on lunch, dinner, and delivery sales for a local Italian restaurant. Develop 95% prediction intervals for the daily dollar sales of each of these variables and also for the total sales dollars on a weekend day.

25. For the Excel file *Credit Approval Decisions*, find 95% confidence and prediction intervals for the credit scores and revolving balance of homeowners and nonhomeowners. How do they compare?

26. Trade associations, such as the United Dairy Farmers Association, frequently conduct surveys to identify characteristics of their membership. If this organization conducted a survey to estimate the annual per-capita consumption of milk and wanted to be 95% confident that the estimate was no more than ± 0.5 gallon away from the actual average, what sample size is needed? Past data have indicated that the standard deviation of consumption is approximately 6 gallons.

27. If a manufacturer conducted a survey among randomly selected target market households and wanted to be 95% confident that the difference between the sample estimate and the actual market share for its new product was no more than $\pm 2\%$, what sample size would be needed?

28. After regular complaints of tire blowouts on the Yamuna Expressway, in an automotive test conducted by the authorities, the average tire pressure in a sample of 62 tires was found to be 24 pounds per square inch and the standard deviation was 2.1 pound per square inch.

a. What is the estimated population standard deviation for this population?

b. Calculate the estimated standard deviation error of the mean.

29. A music company wants to know how the illegal downloading of music online affects CD sales. 600 families are randomly chosen from various parts of a particular country and the number of songs that are downloaded in an hour are noted. The sample mean is 3947 with a sample standard deviation of 104. Determine a 90% confidence interval for this data. (Assume that the population variance is not known.)

Case: Drout Advertising Research Project

The background for this case was introduced in Chapter 1. This is a continuation of the case in Chapter 4. For this part of the case, compute confidence intervals for means and proportions, and analyze the sampling errors, possibly suggesting larger sample sizes to obtain more precise estimates. Write up your findings in a formal report or add your findings to the report you completed for the case in Chapter 4, depending on your instructor's requirements.

Case: Performance Lawn Equipment

In reviewing your previous reports, several questions came to Elizabeth Burke's mind. Use point and interval estimates to help answer these questions.

1. What proportion of customers rate the company with "top box" survey responses (which is defined as scale levels 4 and 5) on quality, ease of use, price, and service in the *2012 Customer Survey* worksheet? How do these proportions differ by geographic region?
2. What estimates, with reasonable assurance, can PLE give customers for response times to customer service calls?
3. Engineering has collected data on alternative process costs for building transmissions in the worksheet *Transmission Costs*. Can you determine whether one of the proposed processes is better than the current process?
4. What would be a confidence interval for an additional sample of mower test performance as in the worksheet *Mower Test*?
5. For the data in the worksheet *Blade Weight*, what is the sampling distribution of the mean, the overall mean, and the standard error of the mean? Is a normal distribution an appropriate assumption for the sampling distribution of the mean?
6. How many blade weights must be measured to find a 95% confidence interval for the mean blade weight with a sampling error of at most 0.2? What if the sampling error is specified as 0.1?

Answer these questions and summarize your results in a formal report to Ms. Burke.