

## READING

# 10

## Sampling and Estimation

by Richard A. DeFusco, PhD, CFA, Dennis W. McLeavey, DBA, CFA,  
Jerald E. Pinto, PhD, CFA, and David E. Runkle, PhD, CFA

*Richard A. DeFusco, PhD, CFA, is at the University of Nebraska-Lincoln (USA). Dennis W. McLeavey, DBA, CFA, is at the University of Rhode Island (USA). Jerald E. Pinto, PhD, CFA, is at CFA Institute (USA). David E. Runkle, PhD, CFA, is at Trilogy Global Advisors (USA).*

### LEARNING OUTCOMES

Mastery	The candidate should be able to:
<input type="checkbox"/>	a. define simple random sampling and a sampling distribution;
<input type="checkbox"/>	b. explain sampling error;
<input type="checkbox"/>	c. distinguish between simple random and stratified random sampling;
<input type="checkbox"/>	d. distinguish between time-series and cross-sectional data;
<input type="checkbox"/>	e. explain the central limit theorem and its importance;
<input type="checkbox"/>	f. calculate and interpret the standard error of the sample mean;
<input type="checkbox"/>	g. identify and describe desirable properties of an estimator;
<input type="checkbox"/>	h. distinguish between a point estimate and a confidence interval estimate of a population parameter;
<input type="checkbox"/>	i. describe properties of Student's <i>t</i> -distribution and calculate and interpret its degrees of freedom;
<input type="checkbox"/>	j. calculate and interpret a confidence interval for a population mean, given a normal distribution with 1) a known population variance, 2) an unknown population variance, or 3) an unknown population variance and a large sample size;
<input type="checkbox"/>	k. describe the issues regarding selection of the appropriate sample size, data-mining bias, sample selection bias, survivorship bias, look-ahead bias, and time-period bias.

## 1

## INTRODUCTION

Each day, we observe the high, low, and close of stock market indexes from around the world. Indexes such as the S&P 500 Index and the Nikkei-Dow Jones Average are samples of stocks. Although the S&P 500 and the Nikkei do not represent the populations of US or Japanese stocks, we view them as valid indicators of the whole population's behavior. As analysts, we are accustomed to using this sample information to assess how various markets from around the world are performing. Any statistics that we compute with sample information, however, are only estimates of the underlying population parameters. A sample, then, is a subset of the population—a subset studied to infer conclusions about the population itself.

This reading explores how we sample and use sample information to estimate population parameters. In the next section, we discuss **sampling**—the process of obtaining a sample. In investments, we continually make use of the mean as a measure of central tendency of random variables, such as return and earnings per share. Even when the probability distribution of the random variable is unknown, we can make probability statements about the population mean using the central limit theorem. In Section 3, we discuss and illustrate this key result. Following that discussion, we turn to statistical estimation. Estimation seeks precise answers to the question “What is this parameter's value?”

The central limit theorem and estimation are the core of the body of methods presented in this reading. In investments, we apply these and other statistical techniques to financial data; we often interpret the results for the purpose of deciding what works and what does not work in investments. We end this reading with a discussion of the interpretation of statistical results based on financial data and the possible pitfalls in this process.

## 2

## SAMPLING

In this section, we present the various methods for obtaining information on a population (all members of a specified group) through samples (part of the population). The information on a population that we try to obtain usually concerns the value of a **parameter**, a quantity computed from or used to describe a population of data. When we use a sample to estimate a parameter, we make use of sample statistics (statistics, for short). A **statistic** is a quantity computed from or used to describe a sample of data.

We take samples for one of two reasons. In some cases, we cannot possibly examine every member of the population. In other cases, examining every member of the population would not be economically efficient. Thus, savings of time and money are two primary factors that cause an analyst to use sampling to answer a question about a population. In this section, we discuss two methods of random sampling: simple random sampling and stratified random sampling. We then define and illustrate the two types of data an analyst uses: cross-sectional data and time-series data.

### 2.1 Simple Random Sampling

Suppose a telecommunications equipment analyst wants to know how much major customers will spend on average for equipment during the coming year. One strategy is to survey the population of telecom equipment customers and inquire what their purchasing plans are. In statistical terms, the characteristics of the population

of customers' planned expenditures would then usually be expressed by descriptive measures such as the mean and variance. Surveying all companies, however, would be very costly in terms of time and money.

Alternatively, the analyst can collect a representative sample of companies and survey them about upcoming telecom equipment expenditures. In this case, the analyst will compute the sample mean expenditure,  $\bar{X}$ , a statistic. This strategy has a substantial advantage over polling the whole population because it can be accomplished more quickly and at lower cost.

Sampling, however, introduces error. The error arises because not all the companies in the population are surveyed. The analyst who decides to sample is trading time and money for sampling error.

When an analyst chooses to sample, he must formulate a sampling plan. A **sampling plan** is the set of rules used to select a sample. The basic type of sample from which we can draw statistically sound conclusions about a population is the **simple random sample** (random sample, for short).

- **Definition of Simple Random Sample.** A simple random sample is a subset of a larger population created in such a way that each element of the population has an equal probability of being selected to the subset.

The procedure of drawing a sample to satisfy the definition of a simple random sample is called **simple random sampling**. How is simple random sampling carried out? We need a method that ensures randomness—the lack of any pattern—in the selection of the sample. For a finite (limited) population, the most common method for obtaining a random sample involves the use of random numbers (numbers with assured properties of randomness). First, we number the members of the population in sequence. For example, if the population contains 500 members, we number them in sequence with three digits, starting with 001 and ending with 500. Suppose we want a simple random sample of size 50. In that case, using a computer random-number generator or a table of random numbers, we generate a series of three-digit random numbers. We then match these random numbers with the number codes of the population members until we have selected a sample of size 50.

Sometimes we cannot code (or even identify) all the members of a population. We often use **systematic sampling** in such cases. With systematic sampling, we select every  $k$ th member until we have a sample of the desired size. The sample that results from this procedure should be approximately random. Real sampling situations may require that we take an approximately random sample.

Suppose the telecommunications equipment analyst polls a random sample of telecom equipment customers to determine the average equipment expenditure. The sample mean will provide the analyst with an estimate of the population mean expenditure. Any difference between the sample mean and the population mean is called **sampling error**.

- **Definition of Sampling Error.** Sampling error is the difference between the observed value of a statistic and the quantity it is intended to estimate.

A random sample reflects the properties of the population in an unbiased way, and sample statistics, such as the sample mean, computed on the basis of a random sample are valid estimates of the underlying population parameters.

A sample statistic is a random variable. In other words, not only do the original data from the population have a distribution but so does the sample statistic.

This distribution is the statistic's **sampling distribution**.

- **Definition of Sampling Distribution of a Statistic.** The sampling distribution of a statistic is the distribution of all the distinct possible values that the statistic can assume when computed from samples of the same size randomly drawn from the same population.

In the case of the sample mean, for example, we refer to the “sampling distribution of the sample mean” or the distribution of the sample mean. We will have more to say about sampling distributions later in this reading. Next, however, we look at another sampling method that is useful in investment analysis.

## 2.2 Stratified Random Sampling

The simple random sampling method just discussed may not be the best approach in all situations. One frequently used alternative is stratified random sampling.

- **Definition of Stratified Random Sampling.** In stratified random sampling, the population is divided into subpopulations (strata) based on one or more classification criteria. Simple random samples are then drawn from each stratum in sizes proportional to the relative size of each stratum in the population. These samples are then pooled to form a stratified random sample.

In contrast to simple random sampling, stratified random sampling guarantees that population subdivisions of interest are represented in the sample. Another advantage is that estimates of parameters produced from stratified sampling have greater precision—that is, smaller variance or dispersion—than estimates obtained from simple random sampling.

Bond indexing is one area in which stratified sampling is frequently applied. **Indexing** is an investment strategy in which an investor constructs a portfolio to mirror the performance of a specified index. In pure bond indexing, also called the full-replication approach, the investor attempts to fully replicate an index by owning all the bonds in the index in proportion to their market value weights. Many bond indexes consist of thousands of issues, however, so pure bond indexing is difficult to implement. In addition, transaction costs would be high because many bonds do not have liquid markets. Although a simple random sample could be a solution to the cost problem, the sample would probably not match the index's major risk factors—interest rate sensitivity, for example. Because the major risk factors of fixed-income portfolios are well known and quantifiable, stratified sampling offers a more effective approach. In this approach, we divide the population of index bonds into groups of similar duration (interest rate sensitivity), cash flow distribution, sector, credit quality, and call exposure. We refer to each group as a stratum or cell (a term frequently used in this context). Then, we choose a sample from each stratum proportional to the relative market weighting of the stratum in the index to be replicated.

### EXAMPLE 1

#### Bond Indexes and Stratified Sampling

Suppose you are the manager of a mutual fund indexed to the Bloomberg Barclays US Government/Credit Index. You are exploring several approaches to indexing, including a stratified sampling approach. You first distinguish among agency bonds, US Treasury bonds, and investment grade corporate bonds. For each of these three groups, you define 10 maturity intervals—1 to 2 years, 2 to 3 years, 3 to 4 years, 4 to 6 years, 6 to 8 years, 8 to 10 years, 10 to 12 years,

12 to 15 years, 15 to 20 years, and 20 to 30 years—and also separate the bonds with coupons (annual interest rates) of 6 percent or less from the bonds with coupons of more than 6 percent.

- 1 How many cells or strata does this sampling plan entail?
- 2 If you use this sampling plan, what is the minimum number of issues the indexed portfolio can have?
- 3 Suppose that in selecting among the securities that qualify for selection within each cell, you apply a criterion concerning the liquidity of the security's market. Is the sample obtained random? Explain your answer.

#### **Solution to 1:**

We have 3 issuer classifications, 10 maturity classifications, and 2 coupon classifications. So, in total, this plan entails  $3(10)(2) = 60$  different strata or cells. (This answer is an application of the multiplication rule of counting discussed in the reading on probability concepts.)

#### **Solution to 2:**

You cannot have fewer than one issue for each cell, so the portfolio must include at least 60 issues.

#### **Solution to 3:**

If you apply any additional criteria to the selection of securities for the cells, not every security that might be included has an equal probability of being selected. As a result, the sampling is not random. In practice, indexing using stratified sampling usually does not strictly involve random sampling because the selection of bond issues within cells is subject to various additional criteria. Because the purpose of sampling in this application is not to make an inference about a population parameter but rather to index a portfolio, lack of randomness is not in itself a problem in this application of stratified sampling.

In the next section, we discuss the kinds of data used by financial analysts in sampling and practical issues that arise in selecting samples.

## **2.3 Time-Series and Cross-Sectional Data**

Investment analysts commonly work with both time-series and cross-sectional data. A time series is a sequence of returns collected at discrete and equally spaced intervals of time (such as a historical series of monthly stock returns). Cross-sectional data are data on some characteristic of individuals, groups, geographical regions, or companies at a single point in time. The book value per share at the end of a given year for all New York Stock Exchange-listed companies is an example of cross-sectional data.

Economic or financial theory offers no basis for determining whether a long or short time period should be selected to collect a sample. As analysts, we might have to look for subtle clues. For example, combining data from a period of fixed exchange rates with data from a period of floating exchange rates would be inappropriate. The variance of exchange rates when exchange rates were fixed would certainly be less than when rates were allowed to float. As a consequence, we would not be sampling from a population described by a single set of parameters.<sup>1</sup> Tight versus loose **monetary**

<sup>1</sup> When the mean or variance of a time series is not constant through time, the time series is not stationary.

**policy** also influences the distribution of returns to stocks; thus, combining data from tight-money and loose-money periods would be inappropriate. Example 2 illustrates the problems that can arise when sampling from more than one distribution.

### EXAMPLE 2

#### Calculating Sharpe Ratios: One or Two Years of Quarterly Data

Analysts often use the Sharpe ratio to evaluate the performance of a managed portfolio. The **Sharpe ratio** is the average return in excess of the risk-free rate divided by the standard deviation of returns. This ratio measures the excess return earned per unit of standard deviation of return.

To compute the Sharpe ratio, suppose that an analyst collects eight quarterly excess returns (i.e., total return in excess of the risk-free rate). During the first year, the investment manager of the portfolio followed a low-risk strategy, and during the second year, the manager followed a high-risk strategy. For each of these years, the analyst also tracks the quarterly excess returns of some benchmark against which the manager will be evaluated. For each of the two years, the Sharpe ratio for the benchmark is 0.21. Table 1 gives the calculation of the Sharpe ratio of the portfolio.

**Table 1 Calculation of Sharpe Ratios: Low-Risk and High-Risk Strategies**

Quarter/Measure	Year 1 Excess Returns	Year 2 Excess Returns
Quarter 1	−3%	−12%
Quarter 2	5	20
Quarter 3	−3	−12
Quarter 4	5	20
Quarterly average	1%	4%
Quarterly standard deviation	4.62%	18.48%
Sharpe ratio = 0.22 = $1/4.62 = 4/18.48$		

For the first year, during which the manager followed a low-risk strategy, the average quarterly return in excess of the risk-free rate was 1 percent with a standard deviation of 4.62 percent. The Sharpe ratio is thus  $1/4.62 = 0.22$ . The second year's results mirror the first year except for the higher average return and volatility. The Sharpe ratio for the second year is  $4/18.48 = 0.22$ . The Sharpe ratio for the benchmark is 0.21 during the first and second years. Because larger Sharpe ratios are better than smaller ones (providing more return per unit of risk), the manager appears to have outperformed the benchmark.

Now, suppose the analyst believes a larger sample to be superior to a small one. She thus decides to pool the two years together and calculate a Sharpe ratio based on eight quarterly observations. The average quarterly excess return for the two years is the average of each year's average excess return. For the two-year period, the average excess return is  $(1 + 4)/2 = 2.5$  percent per quarter. The standard deviation for all eight quarters measured from the sample mean of 2.5 percent is 12.57 percent. The portfolio's Sharpe ratio for the two-year period



is now  $2.5/12.57 = 0.199$ ; the Sharpe ratio for the benchmark remains 0.21. Thus, when returns for the two-year period are pooled, the manager appears to have provided less return per unit of risk than the benchmark and less when compared with the separate yearly results.

The problem with using eight quarters of return data is that the analyst has violated the assumption that the sampled returns come from the same population. As a result of the change in the manager's investment strategy, returns in Year 2 followed a different distribution than returns in Year 1. Clearly, during Year 1, returns were generated by an underlying population with lower mean and variance than the population of the second year. Combining the results for the first and second years yielded a sample that was representative of no population. Because the larger sample did not satisfy model assumptions, any conclusions the analyst reached based on the larger sample are incorrect. For this example, she was better off using a smaller sample than a larger sample because the smaller sample represented a more homogeneous distribution of returns.

The second basic type of data is cross-sectional data.<sup>2</sup> With cross-sectional data, the observations in the sample represent a characteristic of individuals, groups, geographical regions, or companies at a single point in time. The telecommunications analyst discussed previously is essentially collecting a cross-section of planned capital expenditures for the coming year.

Whenever we sample cross-sectionally, certain assumptions must be met if we wish to summarize the data in a meaningful way. Again, a useful approach is to think of the observation of interest as a random variable that comes from some underlying population with a given mean and variance. As we collect our sample and begin to summarize the data, we must be sure that all the data do, in fact, come from the same underlying population. For example, an analyst might be interested in how efficiently companies use their inventory assets. Some companies, however, turn over their inventory more quickly than others because of differences in their operating environments (e.g., grocery stores turn over inventory more quickly than automobile manufacturers, in general). So the distribution of inventory turnover rates may not be characterized by a single distribution with a given mean and variance. Therefore, summarizing inventory turnover across all companies might be inappropriate. If random variables are generated by different underlying distributions, the sample statistics computed from combined samples are not related to one underlying population parameter. The size of the sampling error in such cases is unknown.

In instances such as these, analysts often summarize company-level data by industry. Attempting to summarize by industry partially addresses the problem of differing underlying distributions, but large corporations are likely to be in more than one industrial sector, so analysts should be sure they understand how companies are assigned to the industry groups.

Whether we deal with time-series data or cross-sectional data, we must be sure to have a random sample that is representative of the population we wish to study. With the objective of inferring information from representative samples, we now turn to the next part of this reading, which focuses on the central limit theorem as well as point and interval estimates of the population mean.

<sup>2</sup> The reader may also encounter two types of datasets that have both time-series and cross-sectional aspects. **Panel data** consist of observations through time on a single characteristic of multiple observational units. For example, the annual inflation rate of the Eurozone countries over a five-year period would represent panel data. **Longitudinal data** consist of observations on characteristic(s) of the same observational unit through time. Observations on a set of financial ratios for a single company over a 10-year period would be an example of longitudinal data. Both panel and longitudinal data may be represented by arrays (matrixes) in which successive rows represent the observations for successive time periods.

## 3

## DISTRIBUTION OF THE SAMPLE MEAN

Earlier in this reading, we presented a telecommunications equipment analyst who decided to sample in order to estimate mean planned capital expenditures by his customers. Supposing that the sample is representative of the underlying population, how can the analyst assess the sampling error in estimating the population mean? Viewed as a formula that takes a function of the random outcomes of a random variable, the sample mean is itself a random variable with a probability distribution. That probability distribution is called the statistic's sampling distribution.<sup>3</sup> To estimate how closely the sample mean can be expected to match the underlying population mean, the analyst needs to understand the sampling distribution of the mean. Fortunately, we have a result, the central limit theorem, that helps us understand the sampling distribution of the mean for many of the estimation problems we face.

## 3.1 The Central Limit Theorem

One of the most practically useful theorems in probability theory, the central limit theorem has important implications for how we construct confidence intervals and test hypotheses. Formally, it is stated as follows:

- **The Central Limit Theorem.** Given a population described by any probability distribution having mean  $\mu$  and finite variance  $\sigma^2$ , the sampling distribution of the sample mean  $\bar{X}$  computed from samples of size  $n$  from this population will be approximately normal with mean  $\mu$  (the population mean) and variance  $\sigma^2/n$  (the population variance divided by  $n$ ) when the sample size  $n$  is large.

The central limit theorem allows us to make quite precise probability statements about the population mean by using the sample mean, *whatever the distribution of the population* (so long as it has finite variance), because the sample mean follows an approximate normal distribution for large-size samples. The obvious question is, “When is a sample’s size large enough that we can assume the sample mean is normally distributed?” In general, when sample size  $n$  is greater than or equal to 30, we can assume that the sample mean is approximately normally distributed.<sup>4</sup>

The central limit theorem states that the variance of the distribution of the sample mean is  $\sigma^2/n$ . The positive square root of variance is standard deviation. The standard deviation of a sample statistic is known as the standard error of the statistic. The standard error of the sample mean is an important quantity in applying the central limit theorem in practice.

- **Definition of the Standard Error of the Sample Mean.** For sample mean  $\bar{X}$  calculated from a sample generated by a population with standard deviation  $\sigma$ , the standard error of the sample mean is given by one of two expressions:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad (1)$$

<sup>3</sup> Sometimes confusion arises because “sample mean” is also used in another sense. When we calculate the sample mean for a particular sample, we obtain a definite number, say 8. If we state that “the sample mean is 8” we are using “sample mean” in the sense of a particular outcome of sample mean as a random variable. The number 8 is of course a constant and does not have a probability distribution. In this discussion, we are not referring to “sample mean” in the sense of a constant number related to a particular sample.

<sup>4</sup> When the underlying population is very nonnormal, a sample size well in excess of 30 may be required for the normal distribution to be a good description of the sampling distribution of the mean.



when we know  $\sigma$ , the population standard deviation, or by

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} \quad (2)$$

when we do not know the population standard deviation and need to use the sample standard deviation,  $s$ , to estimate it.<sup>5</sup>

In practice, we almost always need to use Equation 2. The estimate of  $s$  is given by the square root of the sample variance,  $s^2$ , calculated as follows:

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} \quad (3)$$

We will soon see how we can use the sample mean and its standard error to make probability statements about the population mean by using the technique of confidence intervals. First, however, we provide an illustration of the central limit theorem's force.

### EXAMPLE 3

#### The Central Limit Theorem

It is remarkable that the sample mean for large sample sizes will be distributed normally regardless of the distribution of the underlying population. To illustrate the central limit theorem in action, we specify in this example a distinctly nonnormal distribution and use it to generate a large number of random samples of size 100. We then calculate the sample mean for each sample. The frequency distribution of the calculated sample means is an approximation of the sampling distribution of the sample mean for that sample size. Does that sampling distribution look like a normal distribution?

We return to the telecommunications analyst studying the capital expenditure plans of telecom businesses. Suppose that capital expenditures for communications equipment form a continuous uniform random variable with a **lower bound** equal to \$0 and an upper bound equal to \$100—for short, call this a uniform (0, 100) random variable. The probability function of this continuous uniform random variable has a rather simple shape that is anything but normal. It is a horizontal line with a vertical intercept equal to 1/100. Unlike a normal random variable, for which outcomes close to the mean are most likely, all possible outcomes are equally likely for a uniform random variable.

To illustrate the power of the central limit theorem, we conduct a Monte Carlo simulation to study the capital expenditure plans of telecom businesses.<sup>6</sup> In this simulation, we collect 200 random samples of the capital expenditures of 100 companies (200 random draws, each consisting of the capital expenditures of 100 companies with  $n = 100$ ). In each simulation trial, 100 values for capital

<sup>5</sup> We need to note a technical point: When we take a sample of size  $n$  from a finite population of size  $N$ , we apply a shrinkage factor to the estimate of the standard error of the sample mean that is called the finite population correction factor (fpc). The fpc is equal to  $[(N - n)/(N - 1)]^{1/2}$ . Thus, if  $N = 100$  and  $n = 20$ ,  $[(100 - 20)/(100 - 1)]^{1/2} = 0.898933$ . If we have estimated a standard error of, say, 20, according to Equation 1 or Equation 2, the new estimate is  $20(0.898933) = 17.978663$ . The fpc applies only when we sample from a finite population without replacement; most practitioners also do not apply the fpc if sample size  $n$  is very small relative to  $N$  (say, less than 5 percent of  $N$ ). For more information on the finite population correction factor, see Daniel and Terrell (1995).

<sup>6</sup> Monte Carlo simulation involves the use of a computer to represent the operation of a system subject to risk. An integral part of Monte Carlo simulation is the generation of a large number of random samples from a specified probability distribution or distributions.

expenditure are generated from the uniform (0, 100) distribution. For each random sample, we then compute the sample mean. We conduct 200 simulation trials in total. Because we have specified the distribution generating the samples, we know that the population mean capital expenditure is equal to  $(\$0 + \$100 \text{ million})/2 = \$50 \text{ million}$ ; the population variance of capital expenditures is equal to  $(100 - 0)^2/12 = 833.33$ ; thus, the standard deviation is \$28.87 million and the standard error is  $28.87/\sqrt{100} = 2.887$  under the central limit theorem.<sup>7</sup>

The results of this Monte Carlo experiment are tabulated in Table 2 in the form of a frequency distribution. This distribution is the estimated sampling distribution of the sample mean.

**Table 2 Frequency Distribution: 200 Random Samples of a Uniform (0,100) Random Variable**

Range of Sample Means (\$ Million)	Absolute Frequency
$42.5 \leq \bar{X} < 44$	1
$44 \leq \bar{X} < 45.5$	6
$45.5 \leq \bar{X} < 47$	22
$47 \leq \bar{X} < 48.5$	39
$48.5 \leq \bar{X} < 50$	41
$50 \leq \bar{X} < 51.5$	39
$51.5 \leq \bar{X} < 53$	23
$53 \leq \bar{X} < 54.5$	12
$54.5 \leq \bar{X} < 56$	12
$56 \leq \bar{X} < 57.5$	5

Note:  $\bar{X}$  is the mean capital expenditure for each sample.

The frequency distribution can be described as bell-shaped and centered close to the population mean of 50. The most frequent, or modal, range, with 41 observations, is 48.5 to 50. The overall average of the sample means is \$49.92, with a standard error equal to \$2.80. The calculated standard error is close to the value of 2.887 given by the central limit theorem. The discrepancy between calculated and expected values of the mean and standard deviation under the central limit theorem is a result of random chance (sampling error).

In summary, although the distribution of the underlying population is very nonnormal, the simulation has shown that a normal distribution well describes the estimated sampling distribution of the sample mean, with mean and standard error consistent with the values predicted by the central limit theorem.

<sup>7</sup> If  $a$  is the lower limit of a uniform random variable and  $b$  is the upper limit, then the random variable's mean is given by  $(a + b)/2$  and its variance is given by  $(b - a)^2/12$ . The reading on common probability distributions fully describes continuous uniform random variables.

To summarize, according to the central limit theorem, when we sample from any distribution, the distribution of the sample mean will have the following properties as long as our sample size is large:

- The distribution of the sample mean  $\bar{X}$  will be approximately normal.
- The mean of the distribution of  $\bar{X}$  will be equal to the mean of the population from which the samples are drawn.
- The variance of the distribution of  $\bar{X}$  will be equal to the variance of the population divided by the sample size.

We next discuss the concepts and tools related to estimating the population parameters, with a special focus on the population mean. We focus on the population mean because analysts are more likely to meet interval estimates for the population mean than any other type of interval estimate.

## POINT AND INTERVAL ESTIMATES OF THE POPULATION MEAN

## 4

Statistical inference traditionally consists of two branches, hypothesis testing and estimation. Hypothesis testing addresses the question “Is the value of this parameter (say, a population mean) equal to some specific value (0, for example)?” In this process, we have a hypothesis concerning the value of a parameter, and we seek to determine whether the evidence from a sample supports or does not support that hypothesis. We discuss hypothesis testing in detail in the reading on hypothesis testing.

The second branch of statistical inference, and the focus of this reading, is estimation. Estimation seeks an answer to the question “What is this parameter’s (for example, the population mean’s) value?” In estimating, unlike in hypothesis testing, we do not start with a hypothesis about a parameter’s value and seek to test it. Rather, we try to make the best use of the information in a sample to form one of several types of estimates of the parameter’s value. With estimation, we are interested in arriving at a rule for best calculating a single number to estimate the unknown population parameter (a point estimate). Together with calculating a point estimate, we may also be interested in calculating a range of values that brackets the unknown population parameter with some specified level of probability (a confidence interval). In Section 4.1 we discuss point estimates of parameters and then, in Section 4.2, the formulation of confidence intervals for the population mean.

### 4.1 Point Estimators

An important concept introduced in this reading is that sample statistics viewed as formulas involving random outcomes are random variables. The formulas that we use to compute the sample mean and all the other sample statistics are examples of estimation formulas or **estimators**. The particular value that we calculate from sample observations using an estimator is called an **estimate**. An estimator has a sampling distribution; an estimate is a fixed number pertaining to a given sample and thus has no sampling distribution. To take the example of the mean, the calculated value of the sample mean in a given sample, used as an estimate of the population mean, is called a **point estimate** of the population mean. As Example 3 illustrated, the formula for the sample mean can and will yield different results in repeated samples as different samples are drawn from the population.

In many applications, we have a choice among a number of possible estimators for estimating a given parameter. How do we make our choice? We often select estimators because they have one or more desirable statistical properties. Following is a brief description of three desirable properties of estimators: unbiasedness (lack of bias), efficiency, and consistency.<sup>8</sup>

- **Definition of Unbiasedness.** An unbiased estimator is one whose expected value (the mean of its sampling distribution) equals the parameter it is intended to estimate.

For example, the expected value of the sample mean,  $\bar{X}$ , equals  $\mu$ , the population mean, so we say that the sample mean is an unbiased estimator (of the population mean). The sample variance,  $s^2$ , which is calculated using a divisor of  $n - 1$  (Equation 3), is an unbiased estimator of the population variance,  $\sigma^2$ . If we were to calculate the sample variance using a divisor of  $n$ , the estimator would be biased: Its expected value would be smaller than the population variance. We would say that sample variance calculated with a divisor of  $n$  is a biased estimator of the population variance.

Whenever one unbiased estimator of a parameter can be found, we can usually find a large number of other unbiased estimators. How do we choose among alternative unbiased estimators? The criterion of efficiency provides a way to select from among unbiased estimators of a parameter.

- **Definition of Efficiency.** An unbiased estimator is efficient if no other unbiased estimator of the same parameter has a sampling distribution with smaller variance.

To explain the definition, in repeated samples we expect the estimates from an efficient estimator to be more tightly grouped around the mean than estimates from other unbiased estimators. Efficiency is an important property of an estimator.<sup>9</sup> Sample mean  $\bar{X}$  is an efficient estimator of the population mean; sample variance  $s^2$  is an efficient estimator of  $\sigma^2$ .

Recall that a statistic's sampling distribution is defined for a given sample size. Different sample sizes define different sampling distributions. For example, the variance of sampling distribution of the sample mean is smaller for larger sample sizes. Unbiasedness and efficiency are properties of an estimator's sampling distribution that hold for any size sample. An unbiased estimator is unbiased equally in a sample of size 10 and in a sample of size 1,000. In some problems, however, we cannot find estimators that have such desirable properties as unbiasedness in small samples.<sup>10</sup> In this case, statisticians may justify the choice of an estimator based on the properties of the estimator's sampling distribution in extremely large samples, the estimator's so-called asymptotic properties. Among such properties, the most important is consistency.

- **Definition of Consistency.** A consistent estimator is one for which the probability of estimates close to the value of the population parameter increases as sample size increases.

Somewhat more technically, we can define a consistent estimator as an estimator whose sampling distribution becomes concentrated on the value of the parameter it is intended to estimate as the sample size approaches infinity. The sample mean, in addition to being an efficient estimator, is also a consistent estimator of the population mean: As sample size  $n$  goes to infinity, its standard error,  $\sigma/\sqrt{n}$ , goes to 0 and its sampling distribution becomes concentrated right over the value of population mean,

<sup>8</sup> See Daniel and Terrell (1995) or Greene (2018) for a thorough treatment of the properties of estimators.

<sup>9</sup> An efficient estimator is sometimes referred to as the best unbiased estimator.

<sup>10</sup> Such problems frequently arise in regression and time-series analyses.

$\mu$ . To summarize, we can think of a consistent estimator as one that tends to produce more and more accurate estimates of the population parameter as we increase the sample's size. If an estimator is consistent, we may attempt to increase the accuracy of estimates of a population parameter by calculating estimates using a larger sample. For an inconsistent estimator, however, increasing sample size does not help to increase the probability of accurate estimates.

## 4.2 Confidence Intervals for the Population Mean

When we need a single number as an estimate of a population parameter, we make use of a point estimate. However, because of sampling error, the point estimate is not likely to equal the population parameter in any given sample. Often, a more useful approach than finding a point estimate is to find a range of values that we expect to bracket the parameter with a specified level of probability—an interval estimate of the parameter. A confidence interval fulfills this role.

- **Definition of Confidence Interval.** A confidence interval is a range for which one can assert with a given probability  $1 - \alpha$ , called the **degree of confidence**, that it will contain the parameter it is intended to estimate. This interval is often referred to as the  $100(1 - \alpha)\%$  confidence interval for the parameter.

The endpoints of a confidence interval are referred to as the lower and upper confidence limits. In this reading, we are concerned only with two-sided confidence intervals—confidence intervals for which we calculate both lower and upper limits.<sup>11</sup>

Confidence intervals are frequently given either a probabilistic interpretation or a practical interpretation. In the probabilistic interpretation, we interpret a 95 percent confidence interval for the population mean as follows. In repeated sampling, 95 percent of such confidence intervals will, in the long run, include or bracket the population mean. For example, suppose we sample from the population 1,000 times, and based on each sample, we construct a 95 percent confidence interval using the calculated sample mean. Because of random chance, these confidence intervals will vary from each other, but we expect 95 percent, or 950, of these intervals to include the unknown value of the population mean. In practice, we generally do not carry out such repeated sampling. Therefore, in the practical interpretation, we assert that we are 95 percent confident that a single 95 percent confidence interval contains the population mean. We are justified in making this statement because we know that 95 percent of all possible confidence intervals constructed in the same manner will contain the population mean. The confidence intervals that we discuss in this reading have structures similar to the following basic structure:

- **Construction of Confidence Intervals.** A  $100(1 - \alpha)\%$  confidence interval for a parameter has the following structure:

$$\text{Point estimate} \pm \text{Reliability factor} \times \text{Standard error}$$

<sup>11</sup> It is also possible to define two types of one-sided confidence intervals for a population parameter. A lower one-sided confidence interval establishes a lower limit only. Associated with such an interval is an assertion that with a specified degree of confidence the population parameter equals or exceeds the lower limit. An upper one-sided confidence interval establishes an upper limit only; the related assertion is that the population parameter is less than or equal to that upper limit, with a specified degree of confidence. Investment researchers rarely present one-sided confidence intervals, however.

where

Point estimate = a point estimate of the parameter (a value of a sample statistic)

Reliability factor = a number based on the assumed distribution of the point estimate and the degree of confidence ( $1 - \alpha$ ) for the confidence interval

Standard error = the standard error of the sample statistic providing the point estimate<sup>12</sup>

The most basic confidence interval for the population mean arises when we are sampling from a normal distribution with known variance. The reliability factor in this case is based on the standard normal distribution, which has a mean of 0 and a variance of 1. A standard normal random variable is conventionally denoted by  $Z$ . The notation  $z_\alpha$  denotes the point of the standard normal distribution such that  $\alpha$  of the probability remains in the right tail. For example, 0.05 or 5 percent of the possible values of a standard normal random variable are larger than  $z_{0.05} = 1.65$ .

Suppose we want to construct a 95 percent confidence interval for the population mean and, for this purpose, we have taken a sample of size 100 from a normally distributed population with known variance of  $\sigma^2 = 400$  (so,  $\sigma = 20$ ). We calculate a sample mean of  $\bar{X} = 25$ . Our point estimate of the population mean is, therefore, 25. If we move 1.96 standard deviations above the mean of a normal distribution, 0.025 or 2.5 percent of the probability remains in the right tail; by symmetry of the normal distribution, if we move 1.96 standard deviations below the mean, 0.025 or 2.5 percent of the probability remains in the left tail. In total, 0.05 or 5 percent of the probability is in the two tails and 0.95 or 95 percent lies in between. So,  $z_{0.025} = 1.96$  is the reliability factor for this 95 percent confidence interval. Note the relationship  $100(1 - \alpha)\%$  for the confidence interval and the  $z_{\alpha/2}$  for the reliability factor. The standard error of the sample mean, given by Equation 1, is  $\sigma_{\bar{X}} = 20/\sqrt{100} = 2$ . The confidence interval, therefore, has a lower limit of  $\bar{X} - 1.96\sigma_{\bar{X}} = 25 - 1.96(2) = 25 - 3.92 = 21.08$ . The upper limit of the confidence interval is  $\bar{X} + 1.96\sigma_{\bar{X}} = 25 + 1.96(2) = 25 + 3.92 = 28.92$ . The 95 percent confidence interval for the population mean spans 21.08 to 28.92.

- **Confidence Intervals for the Population Mean (Normally Distributed Population with Known Variance).** A  $100(1 - \alpha)\%$  confidence interval for population mean  $\mu$  when we are sampling from a normal distribution with known variance  $\sigma^2$  is given by

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (4)$$

The reliability factors for the most frequently used confidence intervals are as follows.

- **Reliability Factors for Confidence Intervals Based on the Standard Normal Distribution.** We use the following reliability factors when we construct confidence intervals based on the standard normal distribution:<sup>13</sup>
  - 90 percent confidence intervals: Use  $z_{0.05} = 1.65$

<sup>12</sup> The quantity (reliability factor)  $\times$  (standard error) is sometimes called the precision of the estimator; larger values of the product imply lower precision in estimating the population parameter.

<sup>13</sup> Most practitioners use values for  $z_{0.05}$  and  $z_{0.005}$  that are carried to two decimal places. For reference, more exact values for  $z_{0.05}$  and  $z_{0.005}$  are 1.645 and 2.575, respectively. For a quick calculation of a 95 percent confidence interval,  $z_{0.025}$  is sometimes rounded from 1.96 to 2.



- 95 percent confidence intervals: Use  $z_{0.025} = 1.96$
- 99 percent confidence intervals: Use  $z_{0.005} = 2.58$

These reliability factors highlight an important fact about all confidence intervals. As we increase the degree of confidence, the confidence interval becomes wider and gives us less precise information about the quantity we want to estimate. “The surer we want to be, the less we have to be sure of.”<sup>14</sup>

In practice, the assumption that the sampling distribution of the sample mean is at least approximately normal is frequently reasonable, either because the underlying distribution is approximately normal or because we have a large sample and the central limit theorem applies. However, rarely do we know the population variance in practice. When the population variance is unknown but the sample mean is at least approximately normally distributed, we have two acceptable ways to calculate the confidence interval for the population mean. We will soon discuss the more conservative approach, which is based on Student’s  $t$ -distribution (the  $t$ -distribution, for short).<sup>15</sup> In investment literature, it is the most frequently used approach in both estimation and hypothesis tests concerning the mean when the population variance is not known, whether sample size is small or large.

A second approach to confidence intervals for the population mean, based on the standard normal distribution, is the  $z$ -alternative. It can be used only when sample size is large. (In general, a sample size of 30 or larger may be considered large.) In contrast to the confidence interval given in Equation 4, this confidence interval uses the sample standard deviation,  $s$ , in computing the standard error of the sample mean (Equation 2).

- **Confidence Intervals for the Population Mean—The  $z$ -Alternative (Large Sample, Population Variance Unknown).** A  $100(1 - \alpha)\%$  confidence interval for population mean  $\mu$  when sampling from any distribution with unknown variance and when sample size is large is given by

$$\bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} \quad (5)$$

Because this type of confidence interval appears quite often, we illustrate its calculation in Example 4.

#### EXAMPLE 4

##### Confidence Interval for the Population Mean of Sharpe Ratios— $z$ -Statistic

Suppose an investment analyst takes a random sample of US equity mutual funds and calculates the average Sharpe ratio. The sample size is 100, and the average Sharpe ratio is 0.45. The sample has a standard deviation of 0.30. Calculate and interpret the 90 percent confidence interval for the population mean of all US equity mutual funds by using a reliability factor based on the standard normal distribution.

<sup>14</sup> Freund and Williams (1977), p. 266.

<sup>15</sup> The distribution of the statistic  $t$  is called Student’s  $t$ -distribution after the pen name “Student” used by W. S. Gosset, who published his work in 1908.

The reliability factor for a 90 percent confidence interval, as given earlier, is  $z_{0.05} = 1.65$ . The confidence interval will be

$$\bar{X} \pm z_{0.05} \frac{s}{\sqrt{n}} = 0.45 \pm 1.65 \frac{0.30}{\sqrt{100}} = 0.45 \pm 1.65(0.03) = 0.45 \pm 0.0495$$

The confidence interval spans 0.4005 to 0.4995, or 0.40 to 0.50, carrying two decimal places. The analyst can say with 90 percent confidence that the interval includes the population mean.

In this example, the analyst makes no specific assumption about the probability distribution describing the population. Rather, the analyst relies on the central limit theorem to produce an approximate normal distribution for the sample mean.

As Example 4 shows, even if we are unsure of the underlying population distribution, we can still construct confidence intervals for the population mean as long as the sample size is large because we can apply the central limit theorem.

We now turn to the conservative alternative, using the  $t$ -distribution, for constructing confidence intervals for the population mean when the population variance is not known. For confidence intervals based on samples from normally distributed populations with unknown variance, the theoretically correct reliability factor is based on the  $t$ -distribution. Using a reliability factor based on the  $t$ -distribution is essential for a small sample size. Using a  $t$  reliability factor is appropriate when the population variance is unknown, even when we have a large sample and could use the central limit theorem to justify using a  $z$  reliability factor. In this large sample case, the  $t$ -distribution provides more-conservative (wider) confidence intervals.

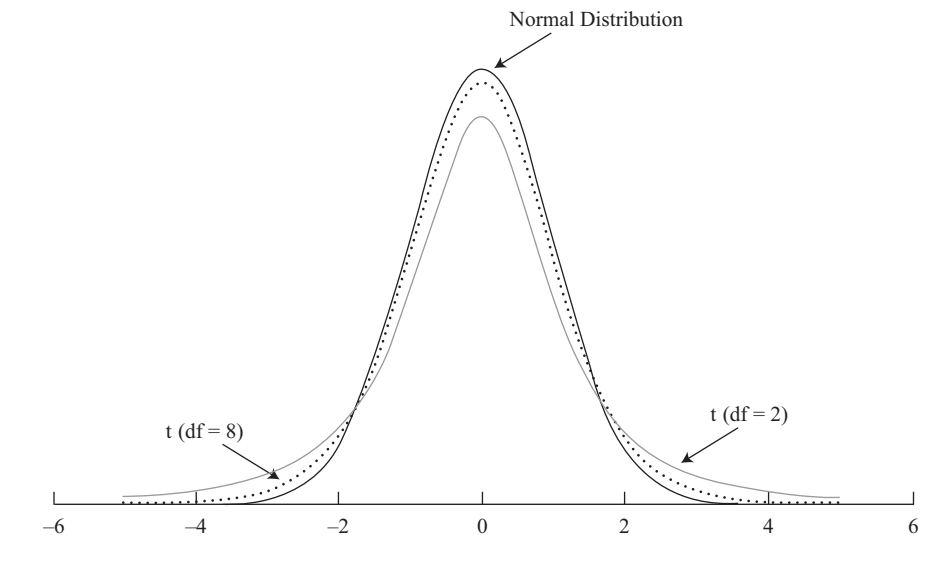
The  $t$ -distribution is a symmetrical probability distribution defined by a single parameter known as **degrees of freedom (df)**. Each value for the number of degrees of freedom defines one distribution in this family of distributions. We will shortly compare  $t$ -distributions with the standard normal distribution, but first we need to understand the concept of degrees of freedom. We can do so by examining the calculation of the sample variance.

Equation 3 gives the unbiased estimator of the sample variance that we use. The term in the denominator,  $n - 1$ , which is the sample size minus 1, is the number of degrees of freedom in estimating the population variance when using Equation 3. We also use  $n - 1$  as the number of degrees of freedom for determining reliability factors based on the  $t$ -distribution. The term “degrees of freedom” is used because in a random sample, we assume that observations are selected independently of each other. The numerator of the sample variance, however, uses the sample mean. How does the use of the sample mean affect the number of observations collected independently for the sample variance formula? With a sample of size 10 and a mean of 10 percent, for example, we can freely select only 9 observations. Regardless of the 9 observations selected, we can always find the value for the 10th observation that gives a mean equal to 10 percent. From the standpoint of the sample variance formula, then, there are 9 degrees of freedom. Given that we must first compute the sample mean from the total of  $n$  independent observations, only  $n - 1$  observations can be chosen independently for the calculation of the sample variance. The concept of degrees of freedom comes up frequently in statistics, and you will see it often in later readings.

Suppose we sample from a normal distribution. The ratio  $z = (\bar{X} - \mu) / (\sigma / \sqrt{n})$  is distributed normally with a mean of 0 and standard deviation of 1; however, the ratio  $t = (\bar{X} - \mu) / (s / \sqrt{n})$  follows the  $t$ -distribution with a mean of 0 and  $n - 1$  degrees of freedom. The ratio represented by  $t$  is not normal because  $t$  is the ratio of two random variables, the sample mean and the sample standard deviation. The definition

of the standard normal random variable involves only one random variable, the sample mean. As degrees of freedom increase, however, the  $t$ -distribution approaches the standard normal distribution. Figure 1 shows the standard normal distribution and two  $t$ -distributions, one with  $df = 2$  and one with  $df = 8$ .

**Figure 1 Student's  $t$ -Distribution versus the Standard Normal Distribution**



Of the three distributions shown in Figure 1, the standard normal distribution has tails that approach zero faster than the tails of the two  $t$ -distributions. The  $t$ -distribution is also symmetrically distributed around its mean value of zero, just like the normal distribution. As the degrees of freedom increase, the  $t$ -distribution approaches the standard normal. The  $t$ -distribution with  $df = 8$  is closer to the standard normal than the  $t$ -distribution with  $df = 2$ .

Beyond plus and minus four standard deviations from the mean, the area under the standard normal distribution appears to approach 0; both  $t$ -distributions continue to show some area under each curve beyond four standard deviations, however. The  $t$ -distributions have fatter tails, but the tails of the  $t$ -distribution with  $df = 8$  more closely resemble the normal distribution's tails. As the degrees of freedom increase, the tails of the  $t$ -distribution become less fat.

Frequently referred to values for the  $t$ -distribution are presented in tables at the end of the book. For each degree of freedom, five values are given:  $t_{0.10}$ ,  $t_{0.05}$ ,  $t_{0.025}$ ,  $t_{0.01}$ , and  $t_{0.005}$ . The values for  $t_{0.10}$ ,  $t_{0.05}$ ,  $t_{0.025}$ ,  $t_{0.01}$ , and  $t_{0.005}$  are such that, respectively, 0.10, 0.05, 0.025, 0.01, and 0.005 of the probability remains in the right tail, for the specified number of degrees of freedom.<sup>16</sup> For example, for  $df = 30$ ,  $t_{0.10} = 1.310$ ,  $t_{0.05} = 1.697$ ,  $t_{0.025} = 2.042$ ,  $t_{0.01} = 2.457$ , and  $t_{0.005} = 2.750$ .

We now give the form of confidence intervals for the population mean using the  $t$ -distribution.

- **Confidence Intervals for the Population Mean (Population Variance Unknown)— $t$ -Distribution.** If we are sampling from a population with unknown variance and either of the conditions below holds:

<sup>16</sup> The values  $t_{0.10}$ ,  $t_{0.05}$ ,  $t_{0.025}$ ,  $t_{0.01}$ , and  $t_{0.005}$  are also referred to as one-sided critical values of  $t$  at the 0.10, 0.05, 0.025, 0.01, and 0.005 significance levels, for the specified number of degrees of freedom.

- the sample is large, or
- the sample is small, but the population is normally distributed, or approximately normally distributed,

then a  $100(1 - \alpha)\%$  confidence interval for the population mean  $\mu$  is given by

$$\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \quad (6)$$

where the number of degrees of freedom for  $t_{\alpha/2}$  is  $n - 1$  and  $n$  is the sample size.

Example 5 reprises the data of Example 4 but uses the  $t$ -statistic rather than the  $z$ -statistic to calculate a confidence interval for the population mean of Sharpe ratios.

#### EXAMPLE 5

##### Confidence Interval for the Population Mean of Sharpe Ratios— $t$ -Statistic

As in Example 4, an investment analyst seeks to calculate a 90 percent confidence interval for the population mean Sharpe ratio of US equity mutual funds based on a random sample of 100 US equity mutual funds. The sample mean Sharpe ratio is 0.45, and the sample standard deviation of the Sharpe ratios is 0.30. Now recognizing that the population variance of the distribution of Sharpe ratios is unknown, the analyst decides to calculate the confidence interval using the theoretically correct  $t$ -statistic.

Because the sample size is 100,  $df = 99$ . In the tables in the back of the book, the closest value is  $df = 100$ . Using  $df = 100$  and reading down the 0.05 column, we find that  $t_{0.05} = 1.66$ . This reliability factor is slightly larger than the reliability factor  $z_{0.05} = 1.65$  that was used in Example 4. The confidence interval will be

$$\bar{X} \pm t_{0.05} \frac{s}{\sqrt{n}} = 0.45 \pm 1.66 \frac{0.30}{\sqrt{100}} = 0.45 \pm 1.66(0.03) = 0.45 \pm 0.0498$$

The confidence interval spans 0.4002 to 0.4998, or 0.40 to 0.50, carrying two decimal places. To two decimal places, the confidence interval is unchanged from the one computed in Example 4.

Table 3 summarizes the various reliability factors that we have used.

**Table 3 Basis of Computing Reliability Factors**

Sampling from:	Statistic for Small Sample Size	Statistic for Large Sample Size
Normal distribution with known variance	$z$	$z$
Normal distribution with unknown variance	$t$	$t^*$

**Table 3 (Continued)**

Sampling from:	Statistic for Small Sample Size	Statistic for Large Sample Size
Nonnormal distribution with known variance	not available	$z$
Nonnormal distribution with unknown variance	not available	$t^*$

\* Use of  $z$  also acceptable.

### 4.3 Selection of Sample Size

What choices affect the width of a confidence interval? To this point we have discussed two factors that affect width: the choice of statistic ( $t$  or  $z$ ) and the choice of degree of confidence (affecting which specific value of  $t$  or  $z$  we use). These two choices determine the reliability factor. (Recall that a confidence interval has the structure Point estimate  $\pm$  Reliability factor  $\times$  Standard error.)

The choice of sample size also affects the width of a confidence interval. All else equal, a larger sample size decreases the width of a confidence interval. Recall the expression for the standard error of the sample mean:

$$\text{Standard error of the sample mean} = \frac{\text{Sample standard deviation}}{\sqrt{\text{Sample size}}}$$

We see that the standard error varies inversely with the square root of sample size. As we increase sample size, the standard error decreases and consequently the width of the confidence interval also decreases. The larger the sample size, the greater precision with which we can estimate the population parameter.<sup>17</sup> All else equal, larger samples are good, in that sense. In practice, however, two considerations may operate against increasing sample size. First, as we saw in Example 2 concerning the Sharpe ratio, increasing the size of a sample may result in sampling from more than one population. Second, increasing sample size may involve additional expenses that outweigh the value of additional precision. Thus three issues that the analyst should weigh in selecting sample size are the need for precision, the risk of sampling from more than one population, and the expenses of different sample sizes.

#### EXAMPLE 6

##### A Money Manager Estimates Net Client Inflows

A money manager wants to obtain a 95 percent confidence interval for fund inflows and outflows over the next six months for his existing clients. He begins by calling a random sample of 10 clients and inquiring about their planned additions to and withdrawals from the fund. The manager then computes the change in cash flow for each client sampled as a percentage change in total funds placed with the manager. A positive percentage change indicates a net cash inflow to the

<sup>17</sup> A formula exists for determining the sample size needed to obtain a desired width for a confidence interval. Define  $E = \text{Reliability factor} \times \text{Standard error}$ . The smaller  $E$  is, the smaller the width of the confidence interval, because  $2E$  is the confidence interval's width. The sample size to obtain a desired value of  $E$  at a given degree of confidence  $(1 - \alpha)$  is  $n = [(t_{\alpha/2}s)/E]^2$ .

client's account, and a negative percentage change indicates a net cash outflow from the client's account. The manager weights each response by the relative size of the account within the sample and then computes a weighted average.

As a result of this process, the money manager computes a weighted average of 5.5 percent. Thus, a point estimate is that the total amount of funds under management will increase by 5.5 percent in the next six months. The standard deviation of the observations in the sample is 10 percent. A histogram of past data looks fairly close to normal, so the manager assumes the population is normal.

- 1 Calculate a 95 percent confidence interval for the population mean and interpret your findings.

The manager decides to see what the confidence interval would look like if he had used a sample size of 20 or 30 and found the same mean (5.5 percent) and standard deviation (10 percent).

- 2 Using the sample mean of 5.5 percent and standard deviation of 10 percent, compute the confidence interval for sample sizes of 20 and 30. For the sample size of 30, use Equation 6.
- 3 Interpret your results from Parts 1 and 2.

#### Solution to 1:

Because the population variance is unknown and the sample size is small, the manager must use the  $t$ -statistic in Equation 6 to calculate the confidence interval. Based on the sample size of 10,  $df = n - 1 = 10 - 1 = 9$ . For a 95 percent confidence interval, he needs to use the value of  $t_{0.025}$  for  $df = 9$ . According to the tables in Appendix B at the end of this volume, this value is 2.262. Therefore, a 95 percent confidence interval for the population mean is

$$\begin{aligned}\bar{X} \pm t_{0.025} \frac{s}{\sqrt{n}} &= 5.5\% \pm 2.262 \frac{10\%}{\sqrt{10}} \\ &= 5.5\% \pm 2.262(3.162) \\ &= 5.5\% \pm 7.15\%\end{aligned}$$

The confidence interval for the population mean spans  $-1.65$  percent to  $+12.65$  percent.<sup>18</sup> The manager can be confident at the 95 percent level that this range includes the population mean.

#### Solution to 2:

Table 4 gives the calculations for the three sample sizes.

**Table 4 The 95 Percent Confidence Interval for Three Sample Sizes**

Distribution	95% Confidence Interval	Lower Bound	Upper Bound	Relative Size
$t(n = 10)$	$5.5\% \pm 2.262(3.162)$	$-1.65\%$	$12.65\%$	100.0%
$t(n = 20)$	$5.5\% \pm 2.093(2.236)$	0.82	10.18	65.5
$t(n = 30)$	$5.5\% \pm 2.045(1.826)$	1.77	9.23	52.2

<sup>18</sup> We assumed in this example that sample size is sufficiently small compared with the size of the client base that we can disregard the finite population correction factor (mentioned in Footnote 6).



**Solution to 3:**

The width of the confidence interval decreases as we increase the sample size. This decrease is a function of the standard error becoming smaller as  $n$  increases. The reliability factor also becomes smaller as the number of degrees of freedom increases. The last column of Table 4 shows the relative size of the width of confidence intervals based on  $n = 10$  to be 100 percent. Using a sample size of 20 reduces the confidence interval's width to 65.5 percent of the interval width for a sample size of 10. Using a sample size of 30 cuts the width of the interval almost in half. Comparing these choices, the money manager would obtain the most precise results using a sample of 30.

Having covered many of the fundamental concepts of sampling and estimation, we are in a good position to focus on sampling issues of special concern to analysts. The quality of inferences depends on the quality of the data as well as on the quality of the sampling plan used. Financial data pose special problems, and sampling plans frequently reflect one or more biases. The next section of this reading discusses these issues.

## MORE ON SAMPLING

# 5

We have already seen that the selection of sample period length may raise the issue of sampling from more than one population. There are, in fact, a range of challenges to valid sampling that arise in working with financial data. In this section we discuss four such sampling-related issues: data-mining bias, sample selection bias, look-ahead bias, and time-period bias. All of these issues are important for point and interval estimation and hypothesis testing. As we will see, if the sample is biased in any way, then point and interval estimates and any other conclusions that we draw from the sample will be in error.

### 5.1 Data-Mining Bias

**Data mining** relates to overuse of the same or related data in ways that we shall describe shortly. Data-mining bias refers to the errors that arise from such misuse of data. Investment strategies that reflect data-mining biases are often not successful in the future. Nevertheless, both investment practitioners and researchers have frequently engaged in data mining. Analysts thus need to understand and guard against this problem.

Data-mining is the practice of determining a model by extensive searching through a dataset for statistically significant patterns (that is, repeatedly “drilling” in the same data until finding something that appears to work).<sup>19</sup> In exercises involving statistical significance we set a significance level, which is the probability of rejecting the hypothesis we are testing when the hypothesis is in fact correct.<sup>20</sup> Because rejecting a true hypothesis is undesirable, the investigator often sets the significance level at

<sup>19</sup> Some researchers use the term “data snooping” instead of data mining.

<sup>20</sup> To convey an understanding of data mining, it is very helpful to introduce some basic concepts related to hypothesis testing. The reading on hypothesis testing contains further discussion of significance levels and tests of significance.

a relatively small number such as 0.05 or 5 percent.<sup>21</sup> Suppose we test the hypothesis that a variable does not predict stock returns, and we test in turn 100 different variables. Let us also suppose that in truth none of the 100 variables has the ability to predict stock returns. Using a 5 percent significance level in our tests, we would still expect that 5 out of 100 variables would appear to be significant predictors of stock returns because of random chance alone. We have mined the data to find some apparently significant variables. In essence, we have explored the same data again and again until we found some after-the-fact pattern or patterns in the dataset. This is the sense in which data mining involves overuse of data. If we were to just report the significant variables, without also reporting the total number of variables that we tested that were unsuccessful as predictors, we would be presenting a very misleading picture of our findings. Our results would appear to be far more significant than they actually were, because a series of tests such as the one just described invalidates the conventional interpretation of a given significance level (such as 5 percent), according to the theory of inference.

How can we investigate the presence of data-mining bias? With most financial data, the most ready means is to conduct out-of-sample tests of the proposed variable or strategy. An **out-of-sample test** uses a sample that does not overlap the time period(s) of the sample(s) on which a variable, strategy, or model, was developed. If a variable or investment strategy is the result of data mining, it should generally not be significant in out-of-sample tests. A variable or investment strategy that is statistically and economically significant in out-of-sample tests, and that has a plausible economic basis, may be the basis for a valid investment strategy. Caution is still warranted, however. The most crucial out-of-sample test is future investment success. If the strategy becomes known to other investors, prices may adjust so that the strategy, however well tested, does not work in the future. To summarize, the analyst should be aware that many apparently profitable investment strategies may reflect data-mining bias and thus be cautious about the future applicability of published investment research results.

Untangling the extent of data mining can be complex. To assess the significance of an investment strategy, we need to know how many unsuccessful strategies were tried not only by the current investigator but also by *previous* investigators using the same or related datasets. Much research, in practice, closely builds on what other investigators have done, and so reflects intergenerational data mining, to use the terminology of McQueen and Thorley (1999). **Intergenerational data mining** involves using information developed by previous researchers using a dataset to guide current research using the same or a related dataset.<sup>22</sup> Analysts have accumulated many observations about the peculiarities of many financial datasets, and other analysts may develop models or investment strategies that will tend to be supported within a dataset based on their familiarity with the prior experience of other analysts. As a consequence, the importance of those new results may be overstated. Research has suggested that the magnitude of this type of data-mining bias may be considerable.<sup>23</sup>

With the background of the above definitions and explanations, we can understand McQueen and Thorley's (1999) cogent exploration of data mining in the context of the popular Motley Fool "Foolish Four" investment strategy. The Foolish Four strategy, first

<sup>21</sup> In terms of our previous discussion of confidence intervals, significance at the 5 percent level corresponds to a hypothesized value for a population statistic falling outside a 95 percent confidence interval based on an appropriate sample statistic (e.g., the sample mean, when the hypothesis concerns the population mean).

<sup>22</sup> The term "intergenerational" comes from viewing each round of researchers as a generation. Campbell, Lo, and MacKinlay (1997) have called intergenerational data mining "data snooping." The latter phrase, however, is commonly used as a synonym of data mining; thus McQueen and Thorley's terminology is less ambiguous. The term "intragenerational data mining" is available when we want to highlight that the reference is to an investigator's new or independent data mining.

<sup>23</sup> For example, Lo and MacKinlay (1990) concluded that the magnitude of this type of bias on tests of the capital asset pricing model was considerable.

presented in 1996, was a version of the Dow Dividend Strategy that was tuned by its developers to exhibit an even higher arithmetic mean return than the Dow Dividend Strategy over 1973 to 1993.<sup>24</sup> From 1973 to 1993, the Foolish Four portfolio had an average annual return of 25 percent, and the claim was made in print that the strategy should have similar returns in the future. As McQueen and Thorley discussed, however, the Foolish Four strategy was very much subject to data-mining bias, including bias from intergenerational data mining, as the strategy's developers exploited observations about the dataset made by earlier workers. McQueen and Thorley highlighted the data-mining issues by taking the Foolish Four portfolio one step further. They mined the data to create a "Fractured Four" portfolio that earned nearly 35 percent over 1973 to 1996, beating the Foolish Four strategy by almost 8 percentage points. Observing that all of the Foolish Four stocks did well in even years but not odd years and that the second-to-lowest-priced high-yielding stock was relatively the best-performing stock in odd years, the strategy of the Fractured Four portfolio was to hold the Foolish Four stocks with equal weights in even years and hold only the second-to-lowest-priced stock in odd years. How likely is it that a performance difference between even and odd years reflected underlying economic forces, rather than a chance pattern of the data over the particular time period? Probably, very unlikely. Unless an investment strategy reflected underlying economic forces, we would not expect it to have any value in a forward-looking sense. Because the Foolish Four strategy also partook of data mining, the same issues applied to it. McQueen and Thorley found that in an out-of-sample test over the 1949–72 period, the Foolish Four strategy had about the same mean return as buying and holding the DJIA, but with higher risk. If the higher taxes and transaction costs of the Foolish Four strategy were accounted for, the comparison would have been even more unfavorable.

McQueen and Thorley presented two signs that can warn analysts about the potential existence of data mining:

- *Too much digging/too little confidence.* The testing of many variables by the researcher is the "too much digging" warning sign of a data-mining problem. Unfortunately, many researchers do not disclose the number of variables examined in developing a model. Although the number of variables examined may not be reported, we should look closely for verbal hints that the researcher searched over many variables. The use of terms such as "we noticed (or noted) that" or "someone noticed (or noted) that," with respect to a pattern in a dataset, should raise suspicions that the researchers were trying out variables based on their own or others' observations of the data.
- *No story/no future.* The absence of an explicit economic rationale for a variable or trading strategy is the "no story" warning sign of a data-mining problem. Without a plausible economic rationale or story for why a variable should work, the variable is unlikely to have predictive power. In a demonstration exercise using an extensive search of variables in an international financial database, Leinweber (1997) found that butter production in a particular country remote from the United States explained 75 percent of the variation in US stock returns as represented by the S&P 500. Such a pattern, with no plausible economic rationale, is highly likely to be a random pattern particular to a specific time

<sup>24</sup> The Dow Dividend Strategy, also known as Dogs of the Dow Strategy, consists of holding an equally weighted portfolio of the 10 highest-yielding DJIA stocks as of the beginning of a year. At the time of McQueen and Thorley's research, the Foolish Four strategy was as follows: At the beginning of each year, the Foolish Four portfolio purchases a 4-stock portfolio from the 5 lowest-priced stocks of the 10 highest-yielding DJIA stocks. The lowest-priced stock of the five is excluded, and 40 percent is invested in the second-to-lowest-priced stock, with 20 percent weights in the remaining three.

period.<sup>25</sup> What if we do have a plausible economic explanation for a significant variable? McQueen and Thorley caution that a plausible economic rationale is a necessary but not a sufficient condition for a trading strategy to have value. As we mentioned earlier, if the strategy is publicized, market prices may adjust to reflect the new information as traders seek to exploit it; as a result, the strategy may no longer work.

## 5.2 Sample Selection Bias

When researchers look into questions of interest to analysts or portfolio managers, they may exclude certain stocks, bonds, portfolios, or time periods from the analysis for various reasons—perhaps because of data availability. When data availability leads to certain assets being excluded from the analysis, we call the resulting problem **sample selection bias**. For example, you might sample from a database that tracks only companies currently in existence. Many mutual fund databases, for instance, provide historical information about only those funds that currently exist. Databases that report historical balance sheet and income statement information suffer from the same sort of bias as the mutual fund databases: Funds or companies that are no longer in business do not appear there. So, a study that uses these types of databases suffers from a type of sample selection bias known as **survivorship bias**.

Dimson, Marsh, and Staunton (2002) raised the issue of survivorship bias in international indexes:

An issue that has achieved prominence is the impact of market survival on estimated long-run returns. Markets can experience not only disappointing performance but also total loss of value through confiscation, hyperinflation, nationalization, and market failure. By measuring the performance of markets that survive over long intervals, we draw inferences that are conditioned on survival. Yet, as pointed out by Brown, Goetzmann, and Ross (1995) and Goetzmann and Jorion (1999), one cannot determine in advance which markets will survive and which will perish. (p. 41)

Survivorship bias sometimes appears when we use both stock price and accounting data. For example, many studies in finance have used the ratio of a company's market price to book equity per share (i.e., the price-to-book ratio, P/B) and found that P/B is inversely related to a company's returns (see Fama and French 1992, 1993). P/B is also used to create many popular value and growth indexes. If the database that we use to collect accounting data excludes failing companies, however, a survivorship bias might result. Kothari, Shanken, and Sloan (1995) investigated just this question and argued that failing stocks would be expected to have low returns and low P/Bs. If we exclude failing stocks, then those stocks with low P/Bs that are included will have returns that are higher on average than if all stocks with low P/Bs were included. Kothari, Shanken, and Sloan suggested that this bias is responsible for the previous findings of an inverse relationship between average return and P/B.<sup>26</sup> The only advice we can offer at this point is to be aware of any biases potentially inherent in a sample. Clearly, sample selection biases can cloud the results of any study.

<sup>25</sup> In the finance literature, such a random but irrelevant-to-the-future pattern is sometimes called an artifact of the dataset.

<sup>26</sup> See Fama and French (1996, p. 80) for discussion of data snooping and survivorship bias in their tests.

A sample can also be biased because of the removal (or delisting) of a company's stock from an exchange.<sup>27</sup> For example, the Center for Research in Security Prices at the University of Chicago is a major provider of return data used in academic research. When a delisting occurs, CRSP attempts to collect returns for the delisted company, but many times, it cannot do so because of the difficulty involved; CRSP must simply list delisted company returns as missing. A study in the *Journal of Finance* by Shumway and Warther (1999) documented the bias caused by delisting for CRSP NASDAQ return data. The authors showed that delistings associated with poor company performance (e.g., bankruptcy) are missed more often than delistings associated with good or neutral company performance (e.g., merger or moving to another exchange). In addition, delistings occur more frequently for small companies.

Sample selection bias occurs even in markets where the quality and consistency of the data are quite high. Newer asset classes such as hedge funds may present even greater problems of sample selection bias. Hedge funds are a heterogeneous group of investment vehicles typically organized so as to be free from regulatory oversight. In general, hedge funds are not required to publicly disclose performance (in contrast to, say, mutual funds). Hedge funds themselves decide whether they want to be included in one of the various databases of hedge fund performance. Hedge funds with poor track records clearly may not wish to make their records public, creating a problem of self-selection bias in hedge fund databases. Further, as pointed out by Fung and Hsieh (2002), because only hedge funds with good records will volunteer to enter a database, in general, overall past hedge fund industry performance will tend to appear better than it really is. Furthermore, many hedge fund databases drop funds that go out of business, creating survivorship bias in the database. Even if the database does not drop defunct hedge funds, in the attempt to eliminate survivorship bias, the problem remains of hedge funds that stop reporting performance because of poor results.<sup>28</sup>

### 5.3 Look-Ahead Bias

A test design is subject to **look-ahead bias** if it uses information that was not available on the test date. For example, tests of trading rules that use stock market returns and accounting balance sheet data must account for look-ahead bias. In such tests, a company's book value per share is commonly used to construct the P/B variable. Although the market price of a stock is available for all market participants at the same point in time, fiscal year-end book equity per share might not become publicly available until sometime in the following quarter.

### 5.4 Time-Period Bias

A test design is subject to **time-period bias** if it is based on a time period that may make the results time-period specific. A short time series is likely to give period specific results that may not reflect a longer period. A long time series may give a more accurate picture of true investment performance; its disadvantage lies in the potential for a structural change occurring during the time frame that would result in two different return distributions. In this situation, the distribution that would reflect conditions before the change differs from the distribution that would describe conditions after the change.

<sup>27</sup> Delistings occur for a variety of reasons: merger, bankruptcy, liquidation, or migration to another exchange.

<sup>28</sup> See Fung and Hsieh (2002) and ter Horst and Verbeek (2007) for more details on the problems of interpreting hedge fund performance. Note that an offsetting type of bias may occur if successful funds stop reporting performance because they no longer want new cash inflows.

**EXAMPLE 7****Biases in Investment Research**

An analyst is reviewing the empirical evidence on historical US equity returns. She finds that value stocks (i.e., those with low P/Bs) outperformed growth stocks (i.e., those with high P/Bs) in some recent time periods. After reviewing the US market, the analyst wonders whether value stocks might be attractive in the United Kingdom. She investigates the performance of value and growth stocks in the UK market for a 14-year period. To conduct this research, the analyst does the following:

- obtains the current composition of the Financial Times Stock Exchange (FTSE) All Share Index, which is a market-capitalization-weighted index;
- eliminates the few companies that do not have December fiscal year-ends;
- uses year-end book values and market prices to rank the remaining universe of companies by P/Bs at the end of the year;
- based on these rankings, divides the universe into 10 portfolios, each of which contains an equal number of stocks;
- calculates the equal-weighted return of each portfolio and the return for the FTSE All Share Index for the 12 months following the date each ranking was made; and
- subtracts the FTSE returns from each portfolio's returns to derive excess returns for each portfolio.

Describe and discuss each of the following biases introduced by the analyst's research design:

- survivorship bias;
- look-ahead bias; and
- time-period bias.

***Survivorship Bias.***

A test design is subject to survivorship bias if it fails to account for companies that have gone bankrupt, merged, or otherwise departed the database. In this example, the analyst used the current list of FTSE stocks rather than the actual list of stocks that existed at the start of each year. To the extent that the computation of returns excluded companies removed from the index, the performance of the portfolios with the lowest P/B is subject to survivorship bias and may be overstated. At some time during the testing period, those companies not currently in existence were eliminated from testing. They would probably have had low prices (and low P/Bs) and poor returns.

***Look-Ahead Bias.***

A test design is subject to look-ahead bias if it uses information unavailable on the test date. In this example, the analyst conducted the test under the assumption that the necessary accounting information was available at the end of the fiscal year. For example, the analyst assumed that book value per share for a given fiscal year was available on 31 December of that year. Because this information is not released until several months after the close of a fiscal year, the test may have contained look-ahead bias. This bias would make a strategy based on the information appear successful, but it assumes perfect forecasting ability.



***Time-Period Bias.***

A test design is subject to time-period bias if it is based on a time period that may make the results time-period specific. Although the test covered a period extending more than 10 years, that period may be too short for testing an anomaly. Ideally, an analyst should test market anomalies over several business cycles to ensure that results are not period specific. This bias can favor a proposed strategy if the time period chosen was favorable to the strategy.

## SUMMARY

In this reading, we have presented basic concepts and results in sampling and estimation. We have also emphasized the challenges faced by analysts in appropriately using and interpreting financial data. As analysts, we should always use a critical eye when evaluating the results from any study. The quality of the sample is of the utmost importance: If the sample is biased, the conclusions drawn from the sample will be in error.

- To draw valid inferences from a sample, the sample should be random.
- In simple random sampling, each observation has an equal chance of being selected. In stratified random sampling, the population is divided into subpopulations, called strata or cells, based on one or more classification criteria; simple random samples are then drawn from each stratum.
- Stratified random sampling ensures that population subdivisions of interest are represented in the sample. Stratified random sampling also produces more-precise parameter estimates than simple random sampling.
- Time-series data are a collection of observations at equally spaced intervals of time. Cross-sectional data are observations that represent individuals, groups, geographical regions, or companies at a single point in time.
- The central limit theorem states that for large sample sizes, for any underlying distribution for a random variable, the sampling distribution of the sample mean for that variable will be approximately normal, with mean equal to the population mean for that random variable and variance equal to the population variance of the variable divided by sample size.
- Based on the central limit theorem, when the sample size is large, we can compute confidence intervals for the population mean based on the normal distribution regardless of the distribution of the underlying population. In general, a sample size of 30 or larger can be considered large.
- An estimator is a formula for estimating a parameter. An estimate is a particular value that we calculate from a sample by using an estimator.
- Because an estimator or statistic is a random variable, it is described by some probability distribution. We refer to the distribution of an estimator as its sampling distribution. The standard deviation of the sampling distribution of the sample mean is called the standard error of the sample mean.
- The desirable properties of an estimator are *unbiasedness* (the expected value of the estimator equals the population parameter), *efficiency* (the estimator has the smallest variance), and *consistency* (the probability of accurate estimates increases as sample size increases).

- The two types of estimates of a parameter are point estimates and interval estimates. A point estimate is a single number that we use to estimate a parameter. An interval estimate is a range of values that brackets the population parameter with some probability.
- A confidence interval is an interval for which we can assert with a given probability  $1 - \alpha$ , called the degree of confidence, that it will contain the parameter it is intended to estimate. This measure is often referred to as the  $100(1 - \alpha)\%$  confidence interval for the parameter.
- A  $100(1 - \alpha)\%$  confidence interval for a parameter has the following structure: Point estimate  $\pm$  Reliability factor  $\times$  Standard error, where the reliability factor is a number based on the assumed distribution of the point estimate and the degree of confidence  $(1 - \alpha)$  for the confidence interval and where standard error is the standard error of the sample statistic providing the point estimate.
- A  $100(1 - \alpha)\%$  confidence interval for population mean  $\mu$  when sampling from a normal distribution with known variance  $\sigma^2$  is given by  $\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ , where  $z_{\alpha/2}$  is the point of the standard normal distribution such that  $\alpha/2$  remains in the right tail.
- Student's  $t$ -distribution is a family of symmetrical distributions defined by a single parameter, degrees of freedom.
- A random sample of size  $n$  is said to have  $n - 1$  degrees of freedom for estimating the population variance, in the sense that there are only  $n - 1$  independent deviations from the mean on which to base the estimate.
- The degrees of freedom number for use with the  $t$ -distribution is also  $n - 1$ .
- The  $t$ -distribution has fatter tails than the standard normal distribution but converges to the standard normal distribution as degrees of freedom go to infinity.
- A  $100(1 - \alpha)\%$  confidence interval for the population mean  $\mu$  when sampling from a normal distribution with unknown variance (a  $t$ -distribution confidence interval) is given by  $\bar{X} \pm t_{\alpha/2} \left( s / \sqrt{n} \right)$ , where  $t_{\alpha/2}$  is the point of the  $t$ -distribution such that  $\alpha/2$  remains in the right tail and  $s$  is the sample standard deviation. This confidence interval can also be used, because of the central limit theorem, when dealing with a large sample from a population with unknown variance that may not be normal.
- We may use the confidence interval  $\bar{X} \pm z_{\alpha/2} \left( s / \sqrt{n} \right)$  as an alternative to the  $t$ -distribution confidence interval for the population mean when using a large sample from a population with unknown variance. The confidence interval based on the  $z$ -statistic is less conservative (narrower) than the corresponding confidence interval based on a  $t$ -distribution.
- Three issues in the selection of sample size are the need for precision, the risk of sampling from more than one population, and the expenses of different sample sizes.
- Sample data in investments can have a variety of problems. *Survivorship bias* occurs if companies are excluded from the analysis because they have gone out of business or because of reasons related to poor performance. *Data-mining bias* comes from finding models by repeatedly searching through databases for patterns. *Look-ahead bias* exists if the model uses data not available to market participants at the time the market participants act in the model. Finally, time-period bias is present if the time period used makes the results time-period specific or if the time period used includes a point of structural change.

## REFERENCES

- Brown, Stephen, William Goetzmann, and Stephen Ross. 1995. "Survival." *Journal of Finance*, vol. 50:853–873.
- Campbell, John, Andrew Lo, and A. Craig MacKinlay. 1997. *The Econometrics of Financial Markets*. Princeton, NJ: Princeton University Press.
- Daniel, Wayne W., and James C. Terrell. 1995. *Business Statistics for Management & Economics*, 7th edition. Boston: Houghton-Mifflin.
- Dimson, Elroy, Paul Marsh, and Mike Staunton. 2002. *Triumphs of the Optimists: 101 Years of Global Investment Returns*. Princeton, NJ: Princeton University Press.
- Fama, Eugene F., and Kenneth R. French. 1996. "Multifactor Explanations of Asset Pricing Anomalies." *Journal of Finance*, vol. 51, no. 1:55–84.
- Freund, John E., and Frank J. Williams. 1977. *Elementary Business Statistics*, 3rd edition. Englewood Cliffs, NJ: Prentice-Hall.
- Fung, William, and David Hsieh. 2002. "Hedge-Fund Benchmarks: Information Content and Biases." *Financial Analysts Journal*, vol. 58, no. 1:22–34.
- Goetzmann, William, and Philippe Jorion. 1999. "Re-Emerging Markets." *Journal of Financial and Quantitative Analysis*, vol. 34, no. 1:1–32.
- Greene, William H. 2018. *Econometric Analysis*, 8th edition. Upper Saddle River, NJ: Prentice-Hall.
- Kothari, S.P., Jay Shanken, and Richard G. Sloan. 1995. "Another Look at the Cross-Section of Expected Stock Returns." *Journal of Finance*, vol. 50, no. 1:185–224.
- Leinweber, David. 1997. *Stupid Data Mining Tricks: Over-Fitting the S&P 500*. Monograph. Pasadena, CA: First Quadrant.
- Lo, Andrew W., and A. Craig MacKinlay. 1990. "Data Snooping Biases in Tests of Financial Asset Pricing Models." *Review of Financial Studies*, vol. 3:175–208.
- McQueen, Grant, and Steven Thorley. 1999. "Mining Fools Gold." *Financial Analysts Journal*, vol. 55, no. 2:61–72.
- Shumway, Tyler, and Vincent A. Warther. 1999. "The Delisting Bias in CRSP's Nasdaq Data and Its Implications for the Size Effect." *Journal of Finance*, vol. 54, no. 6:2361–2379.
- ter Horst, Jenke, and Marno Verbeek. 2007. "Fund Liquidation, Self-selection, and Look-ahead Bias in the Hedge Fund Industry." *Review of Finance*, vol. 11:605–632.

## PRACTICE PROBLEMS

- 1 Peter Biggs wants to know how growth managers performed last year. Biggs assumes that the population cross-sectional standard deviation of growth manager returns is 6 percent and that the returns are independent across managers.
  - A How large a random sample does Biggs need if he wants the standard deviation of the sample means to be 1 percent?
  - B How large a random sample does Biggs need if he wants the standard deviation of the sample means to be 0.25 percent?
- 2 Petra Munzi wants to know how value managers performed last year. Munzi estimates that the population cross-sectional standard deviation of value manager returns is 4 percent and assumes that the returns are independent across managers.
  - A Munzi wants to build a 95 percent confidence interval for the mean return. How large a random sample does Munzi need if she wants the 95 percent confidence interval to have a total width of 1 percent?
  - B Munzi expects a cost of about \$10 to collect each observation. If she has a \$1,000 budget, will she be able to construct the confidence interval she wants?
- 3 Assume that the equity risk premium is normally distributed with a population mean of 6 percent and a population standard deviation of 18 percent. Over the last four years, equity returns (relative to the risk-free rate) have averaged –2.0 percent. You have a large client who is very upset and claims that results this poor should *never* occur. Evaluate your client's concerns.
  - A Construct a 95 percent confidence interval around the population mean for a sample of four-year returns.
  - B What is the probability of a –2.0 percent or lower average return over a four-year period?
- 4 Compare the standard normal distribution and Student's  $t$ -distribution.
- 5 Find the reliability factors based on the  $t$ -distribution for the following confidence intervals for the population mean (df = degrees of freedom,  $n$  = sample size):
  - A A 99 percent confidence interval, df = 20.
  - B A 90 percent confidence interval, df = 20.
  - C A 95 percent confidence interval,  $n$  = 25.
  - D A 95 percent confidence interval,  $n$  = 16.
- 6 Assume that monthly returns are normally distributed with a mean of 1 percent and a sample standard deviation of 4 percent. The population standard deviation is unknown. Construct a 95 percent confidence interval for the sample mean of monthly returns if the sample size is 24.
- 7 Ten analysts have given the following fiscal year earnings forecasts for a stock:

Forecast ( $X_i$ )	Number of Analysts ( $n_i$ )
1.40	1
1.43	1
1.44	3

Forecast ( $X_i$ )	Number of Analysts ( $n_i$ )
1.45	2
1.47	1
1.48	1
1.50	1


Because the sample is a small fraction of the number of analysts who follow this stock, assume that we can ignore the finite population correction factor. Assume that the analyst forecasts are normally distributed.

- A** What are the mean forecast and standard deviation of forecasts?
- B** Provide a 95 percent confidence interval for the population mean of the forecasts.
- 8** Thirteen analysts have given the following fiscal-year earnings forecasts for a stock:

Forecast ( $X_i$ )	Number of Analysts ( $n_i$ )
0.70	2
0.72	4
0.74	1
0.75	3
0.76	1
0.77	1
0.82	1

Because the sample is a small fraction of the number of analysts who follow this stock, assume that we can ignore the finite population correction factor.

- A** What are the mean forecast and standard deviation of forecasts?
- B** What aspect of the data makes us uncomfortable about using  $t$ -tables to construct confidence intervals for the population mean forecast?
- 9** Explain the differences between constructing a confidence interval when sampling from a normal population with a known population variance and sampling from a normal population with an unknown variance.
- 10** An exchange rate has a given expected future value and standard deviation.
- A** Assuming that the exchange rate is normally distributed, what are the probabilities that the exchange rate will be at least 2 or 3 standard deviations away from its mean?
- B** Assume that you do not know the distribution of exchange rates. Use Chebyshev's inequality (that at least  $1 - 1/k^2$  proportion of the observations will be within  $k$  standard deviations of the mean for any positive integer  $k$  greater than 1) to calculate the maximum probabilities that the exchange rate will be at least 2 or 3 standard deviations away from its mean.
- 11** Although he knows security returns are not independent, a colleague makes the claim that because of the central limit theorem, if we diversify across a large number of investments, the portfolio standard deviation will eventually approach zero as  $n$  becomes large. Is he correct?
- 12** Why is the central limit theorem important?
- 13** What is wrong with the following statement of the central limit theorem?



**Central Limit Theorem.** “If the random variables  $X_1, X_2, X_3, \dots, X_n$  are a random sample of size  $n$  from any distribution with finite mean  $\mu$  and variance  $\sigma^2$ , then the distribution of  $\bar{X}$  will be approximately normal, with a standard deviation of  $\sigma/\sqrt{n}$ .”

- 14 Suppose we take a random sample of 30 companies in an industry with 200 companies. We calculate the sample mean of the ratio of cash flow to total debt for the prior year. We find that this ratio is 23 percent. Subsequently, we learn that the population cash flow to total debt ratio (taking account of all 200 companies) is 26 percent. What is the explanation for the discrepancy between the sample mean of 23 percent and the population mean of 26 percent?
  - A Sampling error.
  - B Bias.
  - C A lack of consistency.
- 15 Alcorn Mutual Funds is placing large advertisements in several financial publications. The advertisements prominently display the returns of 5 of Alcorn's 30 funds for the past 1-, 3-, 5-, and 10-year periods. The results are indeed impressive, with all of the funds beating the major market indexes and a few beating them by a large margin. Is the Alcorn family of funds superior to its competitors?
- 16 Julius Spence has tested several predictive models in order to identify undervalued stocks. Spence used about 30 company-specific variables and 10 market-related variables to predict returns for about 5,000 North American and European stocks. He found that a final model using eight variables applied to telecommunications and computer stocks yields spectacular results. Spence wants you to use the model to select investments. Should you? What steps would you take to evaluate the model?
- 17 The *best* approach for creating a stratified random sample of a population involves:
  - A drawing an equal number of simple random samples from each subpopulation.
  - B selecting every  $k$ th member of the population until the desired sample size is reached.
  - C drawing simple random samples from each subpopulation in sizes proportional to the relative size of each subpopulation.
- 18 A population has a non-normal distribution with mean  $\mu$  and variance  $\sigma^2$ . The sampling distribution of the sample mean computed from samples of large size from that population will have:
  - A the same distribution as the population distribution.
  - B its mean approximately equal to the population mean.
  - C its variance approximately equal to the population variance.
- 19 A sample mean is computed from a population with a variance of 2.45. The sample size is 40. The standard error of the sample mean is *closest* to:
  - A 0.039.
  - B 0.247.
  - C 0.387.
- 20 An estimator with an expected value equal to the parameter that it is intended to estimate is described as:



- A efficient.
  - B unbiased.
  - C consistent.
- 21 If an estimator is consistent, an increase in sample size will increase the:
- A accuracy of estimates.
  - B efficiency of the estimator.
  - C unbiasedness of the estimator.
- 22 For a two-sided confidence interval, an increase in the degree of confidence will result in:
- A a wider confidence interval.
  - B a narrower confidence interval.
  - C no change in the width of the confidence interval.
- 23 As the  $t$ -distribution's degrees of freedom decrease, the  $t$ -distribution *most likely*:
- A exhibits tails that become fatter.
  - B approaches a standard normal distribution.
  - C becomes asymmetrically distributed around its mean value.
- 24 For a sample size of 17, with a mean of 116.23 and a variance of 245.55, the width of a 90% confidence interval using the appropriate  $t$ -distribution is *closest to*:
- A 13.23.
  - B 13.27.
  - C 13.68.
- 25 For a sample size of 65 with a mean of 31 taken from a normally distributed population with a variance of 529, a 99% confidence interval for the population mean will have a lower limit *closest to*:
- A 23.64.
  - B 25.41.
  - C 30.09.
- 26 An increase in sample size is *most likely* to result in a:
- A wider confidence interval.
  - B decrease in the standard error of the sample mean.
  - C lower likelihood of sampling from more than one population.
- 27 A report on long-term stock returns focused exclusively on all currently publicly traded firms in an industry is *most likely* susceptible to:
- A look-ahead bias.
  - B survivorship bias.
  - C intergenerational data mining.
- 28 Which sampling bias is *most likely* investigated with an out-of-sample test?
- A Look-ahead bias
  - B Data-mining bias
  - C Sample selection bias
- 29 Which of the following characteristics of an investment study *most likely* indicates time-period bias?
- A The study is based on a short time-series.

- B** Information not available on the test date is used.
- C** A structural change occurred prior to the start of the study's time series.

## SOLUTIONS

- 1 A** The standard deviation or standard error of the sample mean is  $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ . Substituting in the values for  $\sigma_{\bar{X}}$  and  $\sigma$ , we have  $1\% = 6\%/\sqrt{n}$ , or  $\sqrt{n} = 6$ . Squaring this value, we get a random sample of  $n = 36$ .
- B** As in Part A, the standard deviation of sample mean is  $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ . Substituting in the values for  $\sigma_{\bar{X}}$  and  $\sigma$ , we have  $0.25\% = 6\%/\sqrt{n}$ , or  $\sqrt{n} = 24$ . Squaring this value, we get a random sample of  $n = 576$ , which is substantially larger than for Part A of this question.
- 2 A** Assume the sample size will be large and thus the 95 percent confidence interval for the mean of a sample of manager returns is  $\bar{X} \pm 1.96s_{\bar{X}}$ , where  $s_{\bar{X}} = s/\sqrt{n}$ . Munzi wants the distance between the upper limit and lower limit in the confidence interval to be 1 percent, which is

$$(\bar{X} + 1.96s_{\bar{X}}) - (\bar{X} - 1.96s_{\bar{X}}) = 1\%$$

Simplifying this equation, we get  $2(1.96s_{\bar{X}}) = 1\%$ . Finally, we have  $3.92s_{\bar{X}} = 1\%$ , which gives us the standard deviation of the sample mean,  $s_{\bar{X}} = 0.255\%$ . The distribution of sample means is  $s_{\bar{X}} = s/\sqrt{n}$ . Substituting in the values for  $s_{\bar{X}}$  and  $s$ , we have  $0.255\% = 4\%/\sqrt{n}$ , or  $\sqrt{n} = 15.69$ . Squaring this value, we get a random sample of  $n = 246$ .

- B** With her budget, Munzi can pay for a sample of up to 100 observations, which is far short of the 246 observations needed. Munzi can either proceed with her current budget and settle for a wider confidence interval or she can raise her budget (to around \$2,460) to get the sample size for a 1 percent width in her confidence interval.
- 3 A** This is a small-sample problem in which the sample comes from a normal population with a known standard deviation; thus we use the  $z$ -distribution in the solution. For a 95 percent confidence interval (and 2.5 percent in each tail), the critical  $z$ -value is 1.96. For returns that are normally distributed, a 95 percent confidence interval is of the form

$$\mu \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

The lower limit is  $X_l = \mu - 1.96 \frac{\sigma}{\sqrt{n}} = 6\% - 1.96 \frac{18\%}{\sqrt{4}} = 6\% - 1.96(9\%) = -11.64\%$ .

The upper limit is  $X_u = \mu + 1.96 \frac{\sigma}{\sqrt{n}} = 6\% + 1.96 \frac{18\%}{\sqrt{4}} = 6\% + 1.96(9\%) = 23.64\%$ .

There is a 95 percent probability that four-year average returns will be between  $-11.64$  percent and  $+23.64$  percent.

- B** The critical  $z$ -value associated with the  $-2.0$  percent return is

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{-2\% - 6\%}{18\%/\sqrt{4}} = \frac{-8\%}{9\%} = -0.89$$

Using a normal table, the probability of a  $z$ -value less than  $-0.89$  is  $P(Z < -0.89) = 0.1867$ . Unfortunately, although your client is unhappy with the investment result, a four-year average return of  $-2.0$  percent or lower should occur 18.67 percent of the time.

- 4 (Refer to Figure 1 to help visualize the answer to this question.) Basically, only one standard normal distribution exists, but many  $t$ -distributions exist—one for every different number of degrees of freedom. The normal distribution and the  $t$ -distribution for a large number of degrees of freedom are practically the same. The lower the degrees of freedom, the flatter the  $t$ -distribution becomes. The  $t$ -distribution has less mass (lower probabilities) in the center of the distribution and more mass (higher probabilities) out in both tails. Therefore, the confidence intervals based on  $t$ -values will be wider than those based on the normal distribution. Stated differently, the probability of being within a given number of standard deviations (such as within  $\pm 1$  standard deviation or  $\pm 2$  standard deviations) is lower for the  $t$ -distribution than for the normal distribution.
- 5 **A** For a 99 percent confidence interval, the reliability factor we use is  $t_{0.005}$ ; for  $df = 20$ , this factor is 2.845.
- B** For a 90 percent confidence interval, the reliability factor we use is  $t_{0.05}$ ; for  $df = 20$ , this factor is 1.725.
- C** Degrees of freedom equals  $n - 1$ , or in this case  $25 - 1 = 24$ . For a 95 percent confidence interval, the reliability factor we use is  $t_{0.025}$ ; for  $df = 24$ , this factor is 2.064.
- D** Degrees of freedom equals  $16 - 1 = 15$ . For a 95 percent confidence interval, the reliability factor we use is  $t_{0.025}$ ; for  $df = 15$ , this factor is 2.131.
- 6 Because this is a small sample from a normal population and we have only the sample standard deviation, we use the following model to solve for the confidence interval of the population mean:

$$\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

where we find  $t_{0.025}$  (for a 95 percent confidence interval) for  $df = n - 1 = 24 - 1 = 23$ ; this value is 2.069. Our solution is  $1\% \pm 2.069(4\%)/\sqrt{24} = 1\% \pm 2.069(0.8165) = 1\% \pm 1.69$ . The 95 percent confidence interval spans the range from  $-0.69$  percent to  $+2.69$  percent.

- 7 The following table summarizes the calculations used in the answers.

Forecast ( $X_i$ )	Number of Analysts ( $n_i$ )	$X_i n_i$	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})^2 n_i$
1.40	1	1.40	-0.05	0.0025	0.0025
1.43	1	1.43	-0.02	0.0004	0.0004
1.44	3	4.32	-0.01	0.0001	0.0003
1.45	2	2.90	0.00	0.0000	0.0000
1.47	1	1.47	0.02	0.0004	0.0004
1.48	1	1.48	0.03	0.0009	0.0009
1.50	1	1.50	0.05	0.0025	0.0025
Sums	10	14.50			0.0070

- A** With  $n = 10$ ,  $\bar{X} = \sum_{i=1}^{10} X_i / n = 14.50/10 = 1.45$ . The variance is  $s^2 = \left[ \sum_{i=1}^{10} (X_i - \bar{X})^2 \right] / (n-1) = 0.0070/9 = 0.0007778$ . The sample standard deviation is  $s = \sqrt{0.0007778} = 0.02789$ .
- B** The confidence interval for the mean can be estimated by using  $\bar{X} \pm t_{\alpha/2} (s/\sqrt{n})$ . For 9 degrees of freedom, the reliability factor,  $t_{0.025}$ , equals 2.262 and the confidence interval is

$$1.45 \pm 2.262 \times 0.02789 / \sqrt{10} = 1.45 \pm 2.262(0.00882) \\ = 1.45 \pm 0.02$$

The confidence interval for the population mean ranges from 1.43 to 1.47.

- 8** The following table summarizes the calculations used in the answers.

Forecast ( $X_i$ )	Number of Analysts ( $n_i$ )	$X_i n_i$	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})^2 n_i$
0.70	2	1.40	-0.04	0.0016	0.0032
0.72	4	2.88	-0.02	0.0004	0.0016
0.74	1	0.74	0.00	0.0000	0.0000
0.75	3	2.25	0.01	0.0001	0.0003
0.76	1	0.76	0.02	0.0004	0.0004
0.77	1	0.77	0.03	0.0009	0.0009
0.82	1	0.82	0.08	0.0064	0.0064
Sums	13	9.62			0.0128

- A** With  $n = 13$ ,  $\bar{X} = \sum_{i=1}^{13} X_i / n = 9.62/13 = 0.74$ . The variance is  $s^2 = \left[ \sum_{i=1}^{13} (X_i - \bar{X})^2 \right] / (n-1) = 0.0128/12 = 0.001067$ . The sample standard deviation is  $s = \sqrt{0.001067} = 0.03266$ .
- B** The sample is small, and the distribution appears to be bimodal. We cannot compute a confidence interval for the population mean because we have probably sampled from a distribution that is not normal.
- 9** If the population variance is known, the confidence interval is

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

The confidence interval for the population mean is centered at the sample mean,  $\bar{X}$ . The population standard deviation is  $\sigma$ , and the sample size is  $n$ . The population standard deviation divided by the square root of  $n$  is the standard error of the estimate of the mean. The value of  $z$  depends on the desired degree of confidence. For a 95 percent confidence interval,  $z_{0.025} = 1.96$  and the confidence interval estimate is

$$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

If the population variance is not known, we make two changes to the technique used when the population variance is known. First, we must use the sample standard deviation instead of the population standard deviation. Second, we use the  $t$ -distribution instead of the normal distribution. The critical  $t$ -value will depend on degrees of freedom  $n - 1$ . If the sample size is large, we have the alternative of using the  $z$ -distribution with the sample standard deviation.

- 10 A** The probabilities can be taken from a normal table, in which the critical  $z$ -values are 2.00 or 3.00 and we are including the probabilities in both tails. The probabilities that the exchange rate will be at least 2 or 3 standard deviations away from the mean are

$$P(|X - \mu| \geq 2\sigma) = 0.0456$$

$$P(|X - \mu| \geq 3\sigma) = 0.0026$$

- B** With Chebyshev's inequality, the maximum probability of the exchange rate being at least  $k$  standard deviations from the mean is  $P(|X - \mu| \geq k\sigma) \leq (1/k)^2$ . The maximum probabilities of the rate being at least 2 or 3 standard deviations away from the mean are

$$P(|X - \mu| \geq 2\sigma) \leq (1/2)^2 = 0.2500$$

$$P(|X - \mu| \geq 3\sigma) \leq (1/3)^2 = 0.1111$$

The probability of the rate being outside 2 or 3 standard deviations of the mean is much smaller with a known normal distribution than when the distribution is unknown and we are relying on Chebyshev's inequality.

- 11** No. First the conclusion on the limit of zero is wrong; second, the support cited for drawing the conclusion (i.e., the central limit theorem) is not relevant in this context.
- 12** In many instances, the distribution that describes the underlying population is not normal or the distribution is not known. The central limit theorem states that if the sample size is large, regardless of the shape of the underlying population, the distribution of the sample mean is approximately normal. Therefore, even in these instances, we can still construct confidence intervals (and conduct tests of inference) as long as the sample size is large (generally  $n \geq 30$ ).
- 13** The statement makes the following mistakes:
- Given the conditions in the statement, the distribution of  $\bar{X}$  will be approximately normal only for large sample sizes.
  - The statement omits the important element of the central limit theorem that the distribution of  $\bar{X}$  will have mean  $\mu$ .
- 14** A is correct. The discrepancy arises from sampling error. Sampling error exists whenever one fails to observe every element of the population, because a sample statistic can vary from sample to sample. As stated in the reading, the sample mean is an unbiased estimator, a consistent estimator, and an efficient estimator of the population mean. Although the sample mean is an unbiased estimator of the population mean—the expected value of the sample mean equals the population mean—because of sampling error, we do not expect the sample mean to exactly equal the population mean in any one sample we may take.



- 15 No, we cannot say that Alcorn Mutual Funds as a group is superior to competitors. Alcorn Mutual Funds' advertisement may easily mislead readers because the advertisement does not show the performance of all its funds. In particular, Alcorn Mutual Funds is engaging in sample selection bias by presenting the investment results from its best-performing funds only.
- 16 Spence may be guilty of data mining. He has used so many possible combinations of variables on so many stocks, it is not surprising that he found some instances in which a model worked. In fact, it would have been more surprising if he had not found any. To decide whether to use his model, you should do two things: First, ask that the model be tested on out-of-sample data—that is, data that were not used in building the model. The model may not be successful with out-of-sample data. Second, examine his model to make sure that the relationships in the model make economic sense, have a story, and have a future.
- 17 C is correct. Stratified random sampling involves dividing a population into subpopulations based on one or more classification criteria. Then, simple random samples are drawn from each subpopulation in sizes proportional to the relative size of each subpopulation. These samples are then pooled to form a stratified random sample.
- 18 B is correct. Given a population described by any probability distribution (normal or non-normal) with finite variance, the central limit theorem states that the sampling distribution of the sample mean will be approximately normal, with the mean approximately equal to the population mean, when the sample size is large.
- 19 B is correct. Taking the square root of the known population variance to determine the population standard deviation ( $\sigma$ ) results in:

$$\sigma = \sqrt{2.45} = 1.565$$

The formula for the standard error of the sample mean ( $\sigma_X$ ), based on a known sample size ( $n$ ), is:

$$\sigma_X = \frac{\sigma}{\sqrt{n}}$$

Therefore,

$$\sigma_X = \frac{1.565}{\sqrt{40}} = 0.247$$

- 20 B is correct. An unbiased estimator is one for which the expected value equals the parameter it is intended to estimate.
- 21 A is correct. A consistent estimator is one for which the probability of estimates close to the value of the population parameter increases as sample size increases. More specifically, a consistent estimator's sampling distribution becomes concentrated on the value of the parameter it is intended to estimate as the sample size approaches infinity.
- 22 A is correct. As the degree of confidence increases (e.g., from 95% to 99%), a given confidence interval will become wider. A confidence interval is a range for which one can assert with a given probability  $1 - \alpha$ , called the degree of confidence, that it will contain the parameter it is intended to estimate.

- 23** A is correct. A standard normal distribution has tails that approach zero faster than the  $t$ -distribution. As degrees of freedom increase, the tails of the  $t$ -distribution become less fat and the  $t$ -distribution begins to look more like a standard normal distribution. But as degrees of freedom decrease, the tails of the  $t$ -distribution become fatter.
- 24** B is correct. The confidence interval is calculated using the following equation:

$$\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Sample standard deviation ( $s$ ) =  $\sqrt{245.55} = 15.670$ .

For a sample size of 17, degrees of freedom equal 16, so  $t_{0.05} = 1.746$ .

The confidence interval is calculated as

$$116.23 \pm 1.746 \frac{15.67}{\sqrt{17}} = 116.23 \pm 6.6357$$

Therefore, the interval spans 109.5943 to 122.8656, meaning its width is equal to approximately 13.271. (This interval can be alternatively calculated as  $6.6357 \times 2$ ).

- 25** A is correct. To solve, use the structure of Confidence interval = Point estimate  $\pm$  Reliability factor  $\times$  Standard error, which, for a normally distributed population with known variance, is represented by the following formula:

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

For a 99% confidence interval, use  $z_{0.005} = 2.58$ .

Also,  $\sigma = \sqrt{529} = 23$ .

Therefore, the lower limit =  $31 - 2.58 \frac{23}{\sqrt{65}} = 23.6398$ .

- 26** B is correct. All else being equal, as the sample size increases, the standard error of the sample mean decreases and the width of the confidence interval also decreases.
- 27** B is correct. A report that uses a current list of stocks does not account for firms that failed, merged, or otherwise disappeared from the public equity market in previous years. As a consequence, the report is biased. This type of bias is known as survivorship bias.
- 28** B is correct. An out-of-sample test is used to investigate the presence of data-mining bias. Such a test uses a sample that does not overlap the time period of the sample on which a variable, strategy, or model was developed.
- 29** A is correct. A short time series is likely to give period-specific results that may not reflect a longer time period.