6

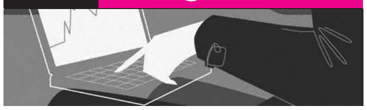
Confidence Intervals

LEARNING OBJECTIVES

After studying this chapter, you should be able to:

- Explain confidence intervals.
- Compute confidence intervals for population means.
- Compute confidence intervals for population proportions.
- Compute confidence intervals for population variances.
- Compute minimum sample sizes needed for an estimation.
- Compute confidence intervals for special types of sampling methods.
- Use templates for all confidence interval and sample-size computations.

6–1 Using Statistics



The alcoholic beverage industry, like many others, has to reinvent itself every few years: from beer to wine, to wine coolers, to cocktails. In 2007 it was clear that the gin-based martini was back as a reigning libation. But which gin was best for this cocktail? *The New York Times* arranged for experts to sample 80 martinis made with different kinds of gin, to determine the best. It also wanted to estimate the average number of stars that any given martini would get—its rating by an average drinker. This is an example of statistical inference, which we study in this chapter and the following ones. In actuality here, four people sampled a total of 80 martinis and determined that the best value was Plymouth English Gin, which received 3½ stars.[1]

In the following chapters we will learn how to compare several populations. In this chapter you will learn how to estimate a parameter of a single population and also provide a *confidence interval* for such a parameter. Thus, for example, you will be able to assess the average number of stars awarded a given gin by the average martini drinker.

In the last chapter, we saw how sample statistics are used as estimators of population parameters. We defined a point estimate of a parameter as a single value obtained from the estimator. We saw that an estimator, a sample statistic, is a random variable with a certain probability distribution—its sampling distribution. A given point estimate is a single realization of the random variable. The actual estimate may or may not be close to the parameter of interest. Therefore, if we only provide a point estimate of the parameter of interest, we are not giving any information about the *accuracy* of the estimation procedure. For example, saying that the sample mean is 550 is giving a point estimate of the population mean. This estimate does not tell us how close μ may be to its estimate, 550. Suppose, on the other hand, that we also said: "We are *99% confident* that μ is in the interval [449, 551]." This conveys much more information about the possible value of μ. Now compare this interval with another one: "We are *90% confident* that μ is in the interval [400, 700]." This interval conveys less information about the possible value of μ, both because it is wider and because the level of confidence is lower. (When based on the same information, however, an interval of lower confidence level is narrower.)

A **confidence interval** is a *range of numbers* believed to include an unknown population parameter. Associated with the interval is a measure of the *confidence* we have that the interval does indeed contain the parameter of interest.

The sampling distribution of the statistic gives a *probability* associated with a range of values the statistic may take. After the sampling has taken place and a *particular estimate* has been obtained, this probability is transformed to a *level of confidence* for a range of values that may contain the unknown parameter.

In the next section, we will see how to construct confidence intervals for the population mean μ when the population standard deviation σ is known. Then we will alter this situation and see how a confidence interval for μ may be constructed without knowledge of σ. Other sections present confidence intervals in other situations.

[1] Eric Asimov, "No, Really, It Was Tough: 4 People, 80 Martinis," *The New York Times,* May 2, 2007, p. D1.

6–2 Confidence Interval for the Population Mean When the Population Standard Deviation Is Known

CHAPTER 9

The central limit theorem tells us that when we select a large random sample from any population with mean μ and standard deviation σ, the sample mean $\bar{X}$ is (at least approximately) normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. If the population itself is normal, $\bar{X}$ is normally distributed for any sample size. Recall that the standard normal random variable Z has a 0.95 probability of being within the range of values -1.96 to 1.96 (you may check this using Table 2 in Appendix C). Transforming Z to the random variable $\bar{X}$ with mean μ and standard deviation $\sigma/\sqrt{n}$, we find that—*before the sampling*—there is a 0.95 probability that $\bar{X}$ will fall within the interval:

$$\mu \pm 1.96 \frac{\sigma}{\sqrt{n}} \qquad (6\text{–}1)$$

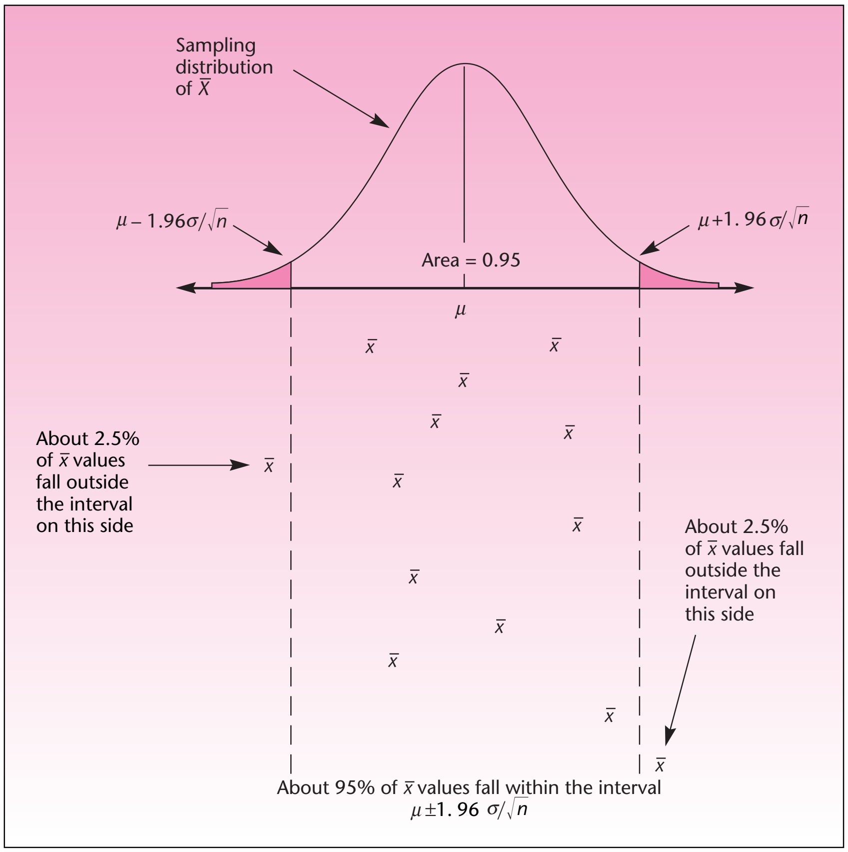
Once we have obtained our random sample, we have a particular value $\bar{x}$. This particular $\bar{x}$ either lies within the range of values specified by equation 6–1 or does not lie within this range. Since we do not know the (fixed) value of the population parameter μ, we have no way of knowing whether $\bar{x}$ is indeed within the range given in equation 6–1. Since the random sampling has already taken place and a particular $\bar{x}$ has been computed, we no longer have a random variable and may no longer talk about probabilities. We do know, however, that since the presampling probability that $\bar{X}$ will fall in the interval in equation 6–1 is 0.95, about 95% of the values of $\bar{X}$ obtained in a large number of repeated samplings will fall within the interval. Since we have a single value $\bar{x}$ that was obtained by this process, we may say that we are *95% confident that $\bar{x}$ lies within the interval.* This idea is demonstrated in Figure 6–1.

Consider a particular $\bar{x}$, and note that the distance between $\bar{x}$ and μ is the same as the distance between μ and $\bar{x}$. Thus, $\bar{x}$ falls inside the interval $\mu \pm 1.96\sigma/\sqrt{n}$ *if and only if* μ happens to be inside the interval $\bar{x} \pm 1.96\sigma/\sqrt{n}$. In a large number of repeated trials, this would happen about 95% of the time. We therefore call the interval $\bar{x} \pm 1.96\sigma/\sqrt{n}$ a *95% confidence interval for the unknown population mean* μ. This is demonstrated in Figure 6–2.

Instead of measuring a distance of $1.96\sigma/\sqrt{n}$ on either side of μ (an impossible task since μ is unknown), we measure the same distance of $1.96\sigma/\sqrt{n}$ on either side of our *known* sample mean $\bar{x}$. Since, *before the sampling,* the random interval $\bar{X} \pm 1.96\,\sigma/\sqrt{n}$ had a 0.95 probability of capturing μ, *after the sampling* we may be 95% confident that our particular interval $\bar{x} \pm 1.96\sigma/\sqrt{n}$ indeed contains the population mean μ. We cannot say that there is a 0.95 *probability* that μ is inside the interval, because the interval $\bar{x} \pm 1.96\sigma/\sqrt{n}$ is not random, and neither is μ. The population mean μ is unknown to us but is a fixed quantity—not a random variable.[2] Either μ lies inside the confidence interval (in which case the probability of this event is 1.00), or it does not (in which case the probability of the event is 0). We do know, however,

[2]We are using what is called the *classical,* or *frequentist,* interpretation of confidence intervals. An alternative view, the Bayesian approach, will be discussed in Chapter 15. The Bayesian approach allows us to treat an unknown population parameter as a random variable. As such, the unknown population mean μ may be stated to have a 0.95 *probability* of being within an interval.

FIGURE 6–1 Probability Distribution of $\bar{X}$ and Some Resulting Values of the Statistic in Repeated Samplings



that 95% of all possible intervals constructed in this manner will contain μ. Therefore, we may say that we are *95% confident* that μ lies in the particular interval we have obtained.

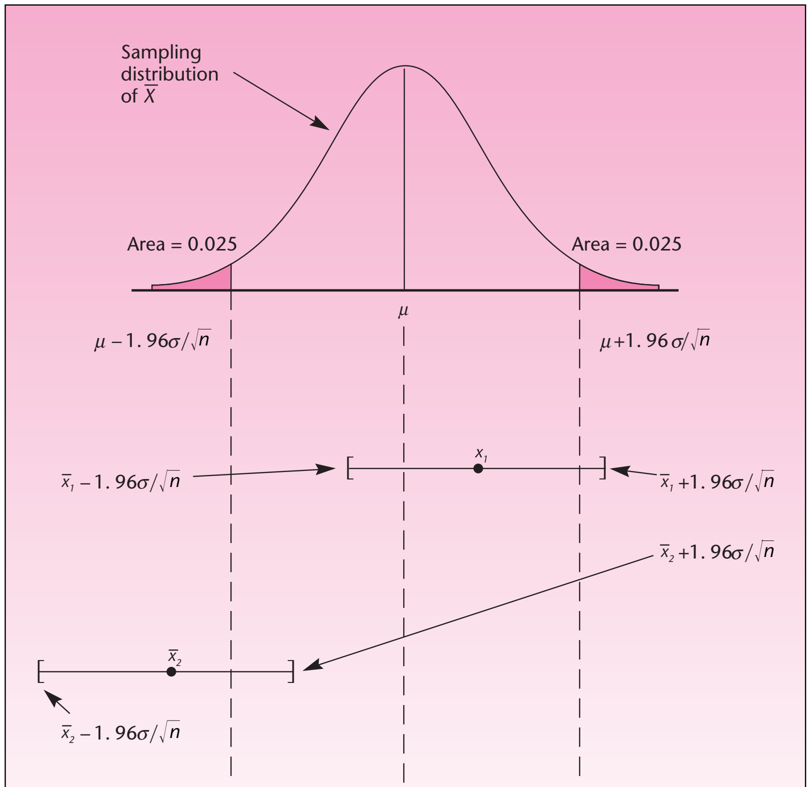
A 95% confidence interval for μ when σ is known and sampling is done from a normal population, or a large sample is used, is

$$\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}} \qquad (6\text{–}2)$$

The quantity $1.96\sigma/\sqrt{n}$ is often called the *margin of error* or the *sampling error*. Its data-derived estimate (using s instead of the unknown σ) is commonly reported.

To compute a 95% confidence interval for μ, all we need to do is substitute the values of the required entities in equation 6–2. Suppose, for example, that we are sampling from a normal population, in which case the random variable $\bar{X}$ is normally distributed for any sample size. We use a sample of size $n = 25$, and we get a sample mean $\bar{x} = 122$. Suppose we also know that the population standard deviation is $\sigma = 20$.

FIGURE 6–2 **Construction of a 95% Confidence Interval for the Population Mean μ**



The sample mean $\bar{x}_1$ falls inside the interval $\mu \pm 1.96\sigma/\sqrt{n}$. Therefore, the confidence interval based on $\bar{x}_1$, which is $\bar{x}_1 \pm 1.96\sigma/\sqrt{n}$, contains μ.

Another sample mean, $\bar{x}_2$, falls outside the interval $\mu \pm 1.96\sigma/\sqrt{n}$. Therefore, the confidence interval based on $\bar{x}_2$, which is $\bar{x}_2 \pm 1.96\sigma/\sqrt{n}$, does not contain μ.

Let us compute a 95% confidence interval for the unknown population mean μ. Using equation 6–2, we get

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}} = 122 \pm 1.96\frac{20}{\sqrt{25}} = 122 \pm 7.84 = [114.16, 129.84]$$

Thus, we may be 95% confident that the unknown population mean μ lies anywhere between the values 114.16 and 129.84.

In business and other applications, the 95% confidence interval is commonly used. There are, however, many other possible levels of confidence. You may choose any level of confidence you wish, find the appropriate z value from the standard normal table, and use it instead of 1.96 in equation 6–2 to get an interval of the chosen level of confidence. Using the standard normal table, we find, for example, that for a

90% confidence interval we use the z value 1.645, and for a 99% confidence interval we use $z = 2.58$ (or, using an accurate interpolation, 2.576). Let us formalize the procedure and make some definitions.

We define $z_{\alpha/2}$ as the z value that cuts off a right-tail area of $\alpha/2$ under the standard normal curve.

For example, 1.96 is $z_{\alpha/2}$ for $\alpha/2 = 0.025$ because $z = 1.96$ cuts off an area of 0.025 to its right. (We find from Table 2 that for $z = 1.96$, TA $= 0.475$; therefore, the right-tail area is $\alpha/2 = 0.025$.) Now consider the two points 1.96 and -1.96. Each of them cuts off a tail area of $\alpha/2 = 0.025$ in the respective direction of its tail. The area between the two values is therefore equal to $1 - \alpha = 1 - 2(0.025) = 0.95$. The area under the curve excluding the tails, $1 - \alpha$, is called the **confidence coefficient.** (And the combined area in both tails α is called the **error probability.** This probability will be important to us in the next chapter.) The confidence coefficient multiplied by 100, expressed as a percentage, is the **confidence level.**

A $(1 - \alpha)$ 100% confidence interval for μ when σ is known and sampling is done from a normal population, or with a large sample, is

$$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{6–3}$$

Thus, for a 95% confidence interval for μ we have

$$\begin{aligned}(1 - \alpha)100\% &= 95\%\\ 1 - \alpha &= 0.95\\ \alpha &= 0.05\\ \frac{\alpha}{2} &= 0.025\end{aligned}$$

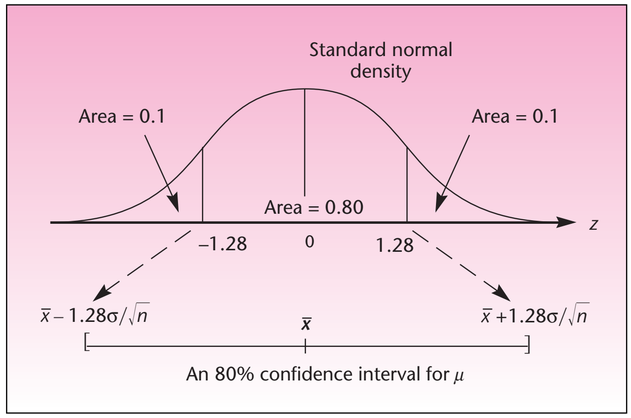
From the normal table, we find $z_{\alpha/2} = 1.96$. This is the value we substitute for $z_{\alpha/2}$ in equation 6–3.

For example, suppose we want an 80% confidence interval for μ. We have $1 - \alpha = 0.80$ and $\alpha = 0.20$; therefore, $\alpha/2 = 0.10$. We now look in the standard normal table for the value of $z_{0.10}$, that is, the z value that cuts off an area of 0.10 to its right. We have TA $= 0.5 - 0.1 = 0.4$, and from the table we find $z_{0.10} = 1.28$. The confidence interval is therefore $\bar{x} \pm 1.28\sigma/\sqrt{n}$. This is demonstrated in Figure 6–3.

Let us compute an 80% confidence interval for μ using the information presented earlier. We have $n = 25$, and $\bar{x} = 122$. We also assume $\sigma = 20$. To compute an 80% confidence interval for the unknown population mean μ, we use equation 6–3 and get

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 122 \pm 1.28\frac{20}{\sqrt{25}} = 122 \pm 5.12 = [116.88, 127.12]$$

FIGURE 6–3 Construction of an 80% Confidence Interval for μ



Comparing this interval with the 95% confidence interval for μ we computed earlier, we note that the present interval is *narrower.* This is an important property of confidence intervals.

> When sampling is from the same population, using a fixed sample size, *the higher the confidence level, the wider the interval.*

Intuitively, a wider interval has more of a presampling chance of "capturing" the unknown population parameter. If we want a 100% confidence interval for a parameter, the interval must be $[-\infty, \infty]$. The reason for this is that 100% confidence is derived from a presampling probability of 1.00 of capturing the parameter, and the only way to get such a probability using the standard normal distribution is by allowing Z to be anywhere from $-\infty$ to ∞. If we are willing to be more realistic (nothing is *certain*) and accept, say, a 99% confidence interval, our interval will be finite and based on $z = 2.58$. The width of our interval will then be $2(2.58\sigma/\sqrt{n})$. If we further reduce our confidence requirement to 95%, the width of our interval will be $2(1.96\sigma/\sqrt{n})$. Since both σ and n are fixed, the 95% interval must be narrower. The more confidence you require, the more you need to sacrifice in terms of a wider interval.

If you want both a narrow interval *and* a high degree of confidence, you need to acquire a large amount of information—take a large sample. This is so because the larger the sample size n, the narrower the interval. This makes sense in that if you buy more information, you will have less uncertainty.

> When sampling is from the same population, using a fixed confidence level, *the larger the sample size* n, *the narrower the confidence interval.*

Suppose that the 80% confidence interval developed earlier was based on a sample size $n = 2{,}500$, instead of $n = 25$. Assuming that $\bar{x}$ and σ are the same, the new confidence interval should be 10 times as narrow as the previous one (because $\sqrt{2{,}500} = 50$, which is 10 times as large as $\sqrt{25}$). Indeed, the new interval is

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 122 \pm 1.28\frac{20}{\sqrt{2{,}500}} = 122 \pm 0.512 = [121.49, 122.51]$$

This interval has width $2(0.512) = 1.024$, while the width of the interval based on a sample of size $n = 25$ is $2(5.12) = 10.24$. This demonstrates the value of information. The two confidence intervals are shown in Figure 6–4.

EXAMPLE 6–1

Comcast, the computer services company, is planning to invest heavily in online television service.[3] As part of the decision, the company wants to estimate the average number of online shows a family of four would watch per day. A random sample of $n = 100$ families is obtained, and in this sample the average number of shows viewed per day is 6.5 and the population standard deviation is known to be 3.2. Construct a 95% confidence interval for the average number of online television shows watched by the entire population of families of four.

Solution

We have

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 6.5 \pm 1.96\frac{3.2}{\sqrt{100}} = 6.5 \pm 0.6272 = [5.8728, 7.1272]$$

Thus Comcast can be 95% confident that the average family of four within its population of subscribers will watch an average daily number of online television shows between about 5.87 and 7.13.

The Template

The workbook named Estimating Mean.xls contains sheets for computing confidence intervals for population means when

1. The sample statistics are known.
2. The sample data are known.

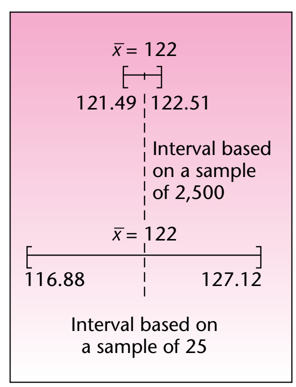
Figure 6–5 shows the first sheet. In this template, we enter the sample statistics in the top or bottom panel, depending on whether the population standard deviation σ is known or unknown.

Since the population standard deviation σ may not be known for certain, on the extreme right, not seen in the figure, there is a sensitivity analysis of the confidence interval with respect to σ. As can be seen in the plot below the panel, the half-width of the confidence interval is linearly related to σ.

Figure 6–6 shows the template to be used when the sample data are known. The data must be entered in column B. The sample size, the sample mean, and the sample standard deviation are automatically calculated and entered in cells F7, F8, F18, F19, and F20 as needed.

Note that in real-life situations we hardly ever know the population standard deviation. The following sections present more realistic applications.

FIGURE 6–4
Width of a Confidence Interval as a Function of Sample Size



[3]Ronald Grover, "Comcast Joins the Party," *BusinessWeek,* May 7, 2007, p. 26.

FIGURE 6–5 The Template for Estimating μ with Sample Statistics [Estimating Mean.xls; Sheet: Sample Stats]

Confidence Interval for μ

σ Known

Population Normal?	No	
Population Stdev.	20	σ
Sample Size	50	n
Sample Mean	122	x-bar

$(1-\alpha)$	Confidence Interval						
99%	122 ± 7.28555	=	[	114.714	,	129.286	]
95%	122 ± 5.54362	=	[	116.456	,	127.544	]
90%	122 ± 4.65235	=	[	117.348	,	126.652	]
80%	122 ± 3.62478	=	[	118.375	,	125.625	]

σ Unknown

Population Normal?	Yes	
Sample Size	15	n
Sample Mean	10.37	x-bar
Sample Stdev.	3.5	s

$(1-\alpha)$	Confidence Interval						
99%	10.37 ± 2.69016	=	[	7.67984	,	13.0602	]
95%	10.37 ± 1.93824	=	[	8.43176	,	12.3082	]
90%	10.37 ± 1.59169	=	[	8.77831	,	11.9617	]
80%	10.37 ± 1.2155	=	[	9.1545	,	11.5855	]

FIGURE 6–6 The Template for Estimating μ with Sample Data [Estimating Mean.xls; Sheet: Sample Data]

Confidence Interval for μ

Sample Data
125
124
120
121
121
128
123
119
124
120
118
119
126
123
120
124
120
117
116
121
125
123
127
120
115

σ Known

Population Normal?	No	
Population Stdev.	20	σ
Sample Size	42	n
Sample Mean	121.1667	x-bar

$(1-\alpha)$	Confidence Interval						
99%	121.1667 ± 7.94918	=	[	113.2175	,	129.1158	]
95%	121.1667 ± 6.04858	=	[	115.1181	,	127.2152	]
90%	121.1667 ± 5.07613	=	[	116.0905	,	126.2428	]
80%	121.1667 ± 3.95495	=	[	117.2117	,	125.1216	]

σ Unknown

Population Normal?	Yes	
Sample Size	42	n
Sample Mean	121.1667	x-bar
Sample Stdev.	3.54013	s

$(1-\alpha)$	Confidence Interval						
99%	121.1667 ± 1.47553	=	[	119.6911	,	122.6422	]
95%	121.1667 ± 1.10318	=	[	120.0635	,	122.2698	]
90%	121.1667 ± 0.91928	=	[	120.2474	,	122.0859	]
80%	121.1667 ± 0.71152	=	[	120.4551	,	121.8782	]

PROBLEMS

6–1. What is a confidence interval, and why is it useful? What is a confidence level?

6–2. Explain why in classical statistics describing a confidence interval in terms of probability makes no sense.

6–3. Explain how the postsampling confidence level is derived from a presampling probability.

6–4. Suppose that you computed a 95% confidence interval for a population mean. The user of the statistics claims your interval is too wide to have any meaning in the specific use for which it is intended. Discuss and compare two methods of solving this problem.

6–5. A real estate agent needs to estimate the average value of a residential property of a given size in a certain area. The real estate agent believes that the standard deviation of the property values is $\sigma = \$5{,}500.00$ and that property values are approximately normally distributed. A random sample of 16 units gives a sample mean of \$89,673.12. Give a 95% confidence interval for the average value of all properties of this kind.

6–6. In problem 6–5, suppose that a 99% confidence interval is required. Compute the new interval, and compare it with the 95% confidence interval you computed in problem 6–5.

6–7. A car manufacturer wants to estimate the average miles-per-gallon highway rating for a new model. From experience with similar models, the manufacturer believes the miles-per-gallon standard deviation is 4.6. A random sample of 100 highway runs of the new model yields a sample mean of 32 miles per gallon. Give a 95% confidence interval for the population average miles-per-gallon highway rating.

6–8. In problem 6–7, do we need to assume that the population of miles-per-gallon values is normally distributed? Explain.

6–9. A wine importer needs to report the average percentage of alcohol in bottles of French wine. From experience with previous kinds of wine, the importer believes the population standard deviation is 1.2%. The importer randomly samples 60 bottles of the new wine and obtains a sample mean $\bar{x} = 9.3\%$. Give a 90% confidence interval for the average percentage of alcohol in all bottles of the new wine.

6–10. British Petroleum has recently been investing in oil fields in the former Soviet Union.[4] Before deciding whether to buy an oilfield, the company wants to estimate the number of barrels of oil that the oilfield can supply. For a given well, the company is interested in a purchase if it can determine that the well will produce, on average, at least 1,500 barrels a day. A random sample of 30 days gives a sample mean of 1,482 and the population standard deviation is 430. Construct a 95% confidence interval. What should be the company's decision?

6–11. Recently, three new airlines, MAXjet, L'Avion, and Eos, began operations selling business or first-class-only service.[5] These airlines need to estimate the highest fare a business-class traveler would pay on a New York to Paris route, roundtrip. Suppose that one of these airlines will institute its route only if it can be reasonably certain (90%) that passengers would pay \$1,800. Suppose also that a random sample of 50 passengers reveals a sample average maximum fare of \$1,700 and the population standard deviation is \$800.

a. Construct a 90% confidence interval.

b. Should the airline offer to fly this route based on your answer to part (*a*)?

[4] Jason Bush, "The Kremlin's Big Squeeze," *BusinessWeek,* April 30, 2007, p. 42.

[5] Susan Stellin, "Friendlier Skies to a Home Abroad," *The New York Times,* May 4, 2007, p. D1.

6–12. According to *Money,* the average price of a home in Albuquerque is \$165,000.[6] Suppose that the reported figure is a sample estimate based on 80 randomly chosen homes in this city, and that the population standard deviation was known to be \$55,000. Give an 80% confidence interval for the population mean home price.

6–13. A mining company needs to estimate the average amount of copper ore per ton mined. A random sample of 50 tons gives a sample mean of 146.75 pounds. The population standard deviation is assumed to be 35.2 pounds. Give a 95% confidence interval for the average amount of copper in the "population" of tons mined. Also give a 90% confidence interval and a 99% confidence interval for the average amount of copper per ton.

6–14. A new low-calorie pizza introduced by Pizza Hut has an average of 150 calories per slice. If this number is based on a random sample of 100 slices, and the population standard deviation is 30 calories, give a 90% confidence interval for the population mean.

6–15. "Small-fry" funds trade at an average of 20% discount to net asset value. If $\sigma = 8\%$ and $n = 36$, give the 95% confidence interval for average population percentage.

6–16. Suppose you have a confidence interval based on a sample of size *n*. Using the same level of confidence, how large a sample is required to produce an interval of one-half the width?

6–17. The width of a 95% confidence interval for μ is 10 units. If everything else stays the same, how wide would a 90% confidence interval be for μ?

6–3 Confidence Intervals for μ When σ Is Unknown—The *t* Distribution

In constructing confidence intervals for μ, we assume a normal population distribution or a large sample size (for normality via the central limit theorem). Until now, we have also assumed a known population standard deviation. This assumption was necessary for theoretical reasons so that we could use standard normal probabilities in constructing our intervals.

In real sampling situations, however, the population standard deviation σ is rarely known. The reason for this is that both μ and σ are population parameters. When we sample from a population with the aim of estimating its unknown mean, the other parameter of the same population, the standard deviation, is highly unlikely to be known.

The t *Distribution*

As we mentioned in Chapter 5, when the population standard deviation is not known, we may use the sample standard deviation S in its place. *If the population is normally distributed,* the standardized statistic

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}} \tag{6–4}$$

has a ***t* distribution** with $n - 1$ degrees of freedom. The degrees of freedom of the distribution are the degrees of freedom associated with the sample standard deviation S (as explained in the last chapter). The t distribution is also called *Student's distribution,* or *Student's t distribution*. What is the origin of the name *Student*?

[6]"The 100 Biggest U.S. Markets," *Money*, May 2007, p. 81.

W. S. Gossett was a scientist at the Guinness brewery in Dublin, Ireland. In 1908, Gossett discovered the distribution of the quantity in equation 6–4. He called the new distribution the *t* distribution. The Guinness brewery, however, did not allow its workers to publish findings under their own names. Therefore, Gossett published his findings under the pen name *Student*. As a result, the distribution became known also as Student's distribution.

The *t* distribution is characterized by its degrees-of-freedom parameter df. For any integer value df $= 1, 2, 3, \ldots$, there is a corresponding *t* distribution. The *t* distribution resembles the standard normal distribution *Z*: it is symmetric and bell-shaped. The *t* distribution, however, has wider tails than the *Z* distribution.

The mean of a *t* distribution is zero. For df > 2, the variance of the *t* distribution is equal to df/(df $-$ 2).

We see that the mean of *t* is the same as the mean of *Z*, but the variance of *t* is larger than the variance of *Z*. As df increases, the variance of *t* approaches 1.00, which is the variance of *Z*. Having wider tails and a larger variance than *Z* is a reflection of the fact that the *t* distribution applies to situations with a greater inherent *uncertainty*. The uncertainty comes from the fact that σ is unknown and is estimated by the *random variable S*. The *t* distribution thus reflects the uncertainty in *two* random variables, $\bar{X}$ and *S*, while *Z* reflects only an uncertainty due to $\bar{X}$. The greater uncertainty in *t* (which makes confidence intervals based on *t* wider than those based on *Z*) is the price we pay for not knowing σ and having to estimate it from our data. As df increases, the *t* distribution approaches the *Z* distribution.

Figure 6–7 shows the *t*-distribution template. We can enter any desired degrees of freedom in cell B4 and see how the distribution approaches the *Z* distribution, which is superimposed on the chart. In the range K3:O5 the template shows the critical values for all standard α values. The area to the right of the chart in this template can be used for calculating *p*-values, which we will learn in the next chapter.

FIGURE 6–7 **The *t*-Distribution Template [t.xls]**

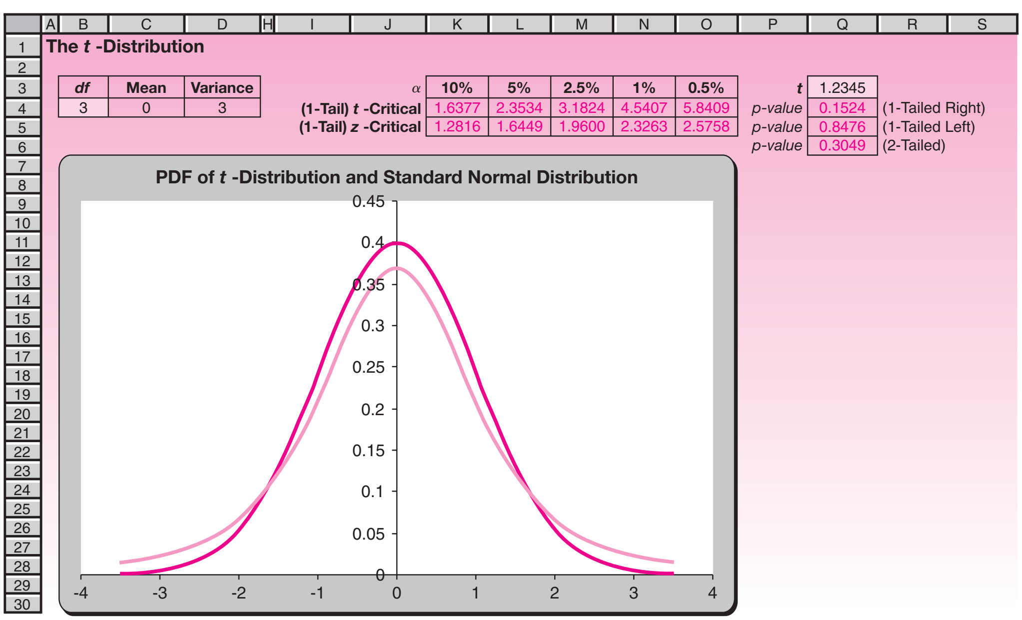


TABLE 6–1 Values and Probabilities of *t* Distributions

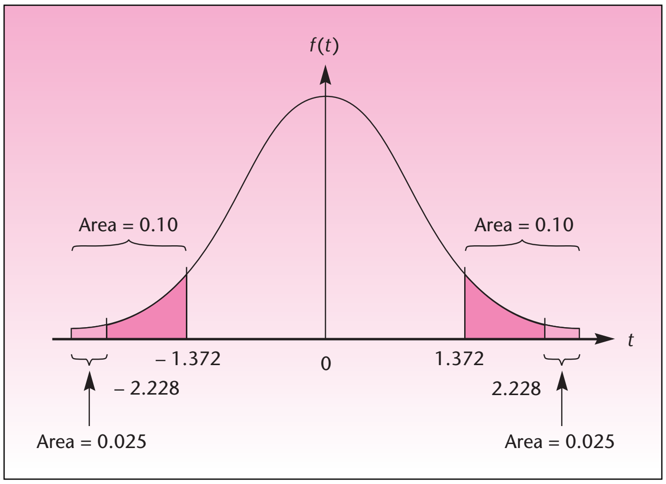
Degrees of Freedom	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.010}$	$t_{0.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

Values of t distributions for selected tail probabilities are given in Table 3 in Appendix C (reproduced here as Table 6–1). Since there are infinitely many t distributions—one for every value of the degrees-of-freedom parameter—the table contains probabilities for only some of these distributions. For each distribution, the table gives values that cut off given areas under the curve to the *right*. The t table is thus a table of values corresponding to right-tail probabilities.

Let us consider an example. A random variable with a t distribution with 10 degrees of freedom has a 0.10 probability of exceeding the value 1.372. It has a 0.025 probability of exceeding the value 2.228, and so on for the other values listed in the table. Since the t distributions are symmetric about zero, we also know, for example, that the probability that a random variable with a t distribution with 10 degrees of freedom will be less than -1.372 is 0.10. These facts are demonstrated in Figure 6–8.

As we noted earlier, the t distribution approaches the standard normal distribution as the df parameter approaches infinity. The t distribution with "infinite" degrees

FIGURE 6–8 **Table Probabilities for a Selected t Distribution (df = 10)**



of freedom is defined as the standard normal distribution. The last row in Appendix C, Table 3 (Table 6–1) corresponds to df $= \infty$, the standard normal distribution. Note that the value corresponding to a right-tail area of 0.025 in that row is 1.96, which we recognize as the appropriate z value. Similarly, the value corresponding to a right-tail area of 0.005 is 2.576, and the value corresponding to a right-tail area of 0.05 is 1.645. These, too, are values we recognize for the standard normal distribution. Look upward from the last row of the table to find cutoff values of the same right-tail probabilities for t distributions with different degrees of freedom. Suppose, for example, that we want to construct a 95% confidence interval for μ using the t distribution with 20 degrees of freedom. We may identify the value 1.96 in the last row (the appropriate z value for 95%) and then move up in the same column until we reach the row corresponding to df = 20. Here we find the required value $t_{\alpha/2} = t_{0.025} = 2.086$.

A $(1 - \alpha)$ 100% confidence interval for μ when σ is not known (assuming a normally distributed population) is

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \tag{6–5}$$

where $t_{\alpha/2}$ is the value of the t distribution with $n - 1$ degrees of freedom that cuts off a tail area of $\alpha/2$ to its right.

EXAMPLE 6–2

A stock market analyst wants to estimate the average return on a certain stock. A random sample of 15 days yields an average (annualized) return of $\bar{x} = 10.37\%$ and a standard deviation of $s = 3.5\%$. Assuming a normal population of returns, give a 95% confidence interval for the average return on this stock.

Solution

Since the sample size is $n = 15$, we need to use the t distribution with $n - 1 = 14$ degrees of freedom. In Table 3, in the row corresponding to 14 degrees of freedom and the column corresponding to a right-tail area of 0.025 (this is $\alpha/2$), we find

$t_{0.025} = 2.145$. (We could also have found this value by moving upward from 1.96 in the last row.) Using this value, we construct the 95% confidence interval as follows:

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = 10.37 \pm 2.145\frac{3.5}{\sqrt{15}} = [8.43, 12.31]$$

Thus, the analyst may be 95% sure that the average annualized return on the stock is anywhere from 8.43% to 12.31%.

Looking at the t table, we note the *convergence* of the t distributions to the Z distribution—the values in the rows preceding the last get closer and closer to the corresponding z values in the last row. Although the t distribution is the correct distribution to use whenever σ is not known (assuming the population is normal), when df is *large,* we may use the standard normal distribution as an adequate approximation to the t distribution. Thus, instead of using 1.98 in a confidence interval based on a sample of size 121 (df = 120), we will just use the z value 1.96.

We divide estimation problems into two kinds: small-sample problems and large-sample problems. Example 6–2 demonstrated the solution of a small-sample problem. In general, *large sample* will mean a sample of 30 items or more, and *small sample* will mean a sample of size less than 30. For small samples, we will use the t distribution as demonstrated above. For large samples, we will use the Z distribution as an adequate approximation. We note that the larger the sample size, the better the normal approximation. Remember, however, that this division of large and small samples is arbitrary.

Whenever σ is not known (and the population is assumed normal), the correct distribution to use is the t distribution with $n - 1$ degrees of freedom. Note, however, that for large degrees of freedom, the t distribution is approximated well by the Z distribution.

If you wish, you may always use the more accurate values obtained from the t table (when such values can be found in the table) rather than the standard normal approximation. In this chapter and elsewhere (with the exception of some examples in Chapter 14), we will assume that the population satisfies, at least approximately, a normal distribution assumption. For large samples, this assumption is less crucial.

A large-sample $(1 - \alpha)$ 100% confidence interval for μ is

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} \qquad (6\text{–}6)$$

We demonstrate the use of equation 6–6 in Example 6–3.

EXAMPLE 6–3

An economist wants to estimate the average amount in checking accounts at banks in a given region. A random sample of 100 accounts gives $\bar{x} = \$357.60$ and $s = \$140.00$. Give a 95% confidence interval for μ, the average amount in any checking account at a bank in the given region.

Solution

We find the 95% confidence interval for μ as follows:

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} = 357.60 \pm 1.96\frac{140}{\sqrt{100}} = [330.16, 385.04]$$

Thus, based on the data and the assumption of random sampling, the economist may be 95% confident that the average amount in checking accounts in the area is anywhere from \$330.16 to \$385.04.

PROBLEMS

6–18. A telephone company wants to estimate the average length of long-distance calls during weekends. A random sample of 50 calls gives a mean $\bar{x} = 14.5$ minutes and standard deviation $s = 5.6$ minutes. Give a 95% confidence interval and a 90% confidence interval for the average length of a long-distance phone call during weekends.

6–19. An insurance company handling malpractice cases is interested in estimating the average amount of claims against physicians of a certain specialty. The company obtains a random sample of 165 claims and finds $\bar{x} = \$16{,}530$ and $s = \$5{,}542$. Give a 95% confidence interval and a 99% confidence interval for the average amount of a claim.

6–20. The manufacturer of batteries used in small electric appliances wants to estimate the average life of a battery. A random sample of 12 batteries yields $\bar{x} = 34.2$ hours and $s = 5.9$ hours. Give a 95% confidence interval for the average life of a battery.

6–21. A tire manufacturer wants to estimate the average number of miles that may be driven on a tire of a certain type before the tire wears out. A random sample of 32 tires is chosen; the tires are driven on until they wear out, and the number of miles driven on each tire is recorded. The data, in thousands of miles, are as follows:

32, 33, 28, 37, 29, 30, 25, 27, 39, 40, 26, 26, 27, 30, 25, 30, 31, 29, 24, 36, 25, 37, 37, 20, 22, 35, 23, 28, 30, 36, 40, 41

Give a 99% confidence interval for the average number of miles that may be driven on a tire of this kind.

6–22. Digital media have recently begun to take over from print outlets.[7] A newspaper owners' association wants to estimate the average number of times a week people buy a newspaper on the street. A random sample of 100 people reveals that the sample average is 3.2 and the sample standard deviation is 2.1. Construct a 95% confidence interval for the population average.

6–23. Pier 1 Imports is a nationwide retail outlet selling imported furniture and other home items. From time to time, the company surveys its regular customers by obtaining random samples based on customer zip codes. In one mailing, customers were asked to rate a new table from Thailand on a scale of 0 to 100. The ratings of 25 randomly selected customers are as follows: 78, 85, 80, 89, 77, 50, 75, 90, 88, 100, 70, 99, 98, 55, 80, 45, 80, 76, 96, 100, 95, 90, 60, 85, 90. Give a 99% confidence interval for the rating of the table that would be given by an average member of the population of regular customers. Assume normality.

[7]"Flat Prospects: Digital Media and Globalization Shake Up an Old Industry," *The Economist,* March 17, 2007, p. 72.

6–24. An executive placement service needs to estimate the average salary of executives placed in a given industry. A random sample of 40 executives gives $\bar{x} =$ \$42,539 and $s =$ \$11,690. Give a 90% confidence interval for the average salary of an executive placed in this industry.

6–25. The following is a random sample of the wealth, in billions of U.S. dollars, of individuals listed on the *Forbes* "Billionaires" list for 2007.[8]

2.1, 5.8, 7.3, 33.0, 2.0, 8.4, 11.0, 18.4, 4.3, 4.5, 6.0, 13.3, 12.8, 3.6, 2.4, 1.0

Construct a 90% confidence interval for the average wealth in \$ billions for the people on the *Forbes* list.

6–26. For advertising purposes, the Beef Industry Council needs to estimate the average caloric content of 3-ounce top loin steak cuts. A random sample of 400 pieces gives a sample mean of 212 calories and a sample standard deviation of 38 calories. Give a 95% confidence interval for the average caloric content of a 3-ounce cut of top loin steak. Also give a 98% confidence interval for the average caloric content of a cut.

6–27. A transportation company wants to estimate the average length of time goods are in transit across the country. A random sample of 20 shipments gives $\bar{x} = 2.6$ days and $s = 0.4$ day. Give a 99% confidence interval for the average transit time.

6–28. To aid in planning the development of a tourist shopping area, a state agency wants to estimate the average dollar amount spent by a tourist in an existing shopping area. A random sample of 56 tourists gives $\bar{x} =$ \$258 and $s =$ \$85. Give a 95% confidence interval for the average amount spent by a tourist at the shopping area.

6–29. According to *Money,* the average home in Ventura County, California, sells for \$647,000.[9] Assume that this sample mean was obtained from a random sample of 200 homes in this county, and that the sample standard deviation was \$140,000. Give a 95% confidence interval for the average value of a home in Ventura County.

6–30. Citibank Visa gives its cardholders "bonus dollars," which may be spent in partial payment for gifts purchased with the Visa card. The company wants to estimate the average amount of bonus dollars that will be spent by a cardholder enrolled in the program during a year. A trial run of the program with a random sample of 225 cardholders is carried out. The results are $\bar{x} =$ \$259.60 and $s =$ \$52.00. Give a 95% confidence interval for the average amount of bonus dollars that will be spent by a cardholder during the year.

6–31. An accountant wants to estimate the average amount of an account of a service company. A random sample of 46 accounts yields $\bar{x} =$ \$16.50 and $s =$ \$2.20. Give a 95% confidence interval for the average amount of an account.

6–32. An art dealer wants to estimate the average value of works of art of a certain period and type. A random sample of 20 works of art is appraised. The sample mean is found to be \$5,139 and the sample standard deviation \$640. Give a 95% confidence interval for the average value of all works of art of this kind.

6–33. A management consulting agency needs to estimate the average number of years of experience of executives in a given branch of management. A random sample of 28 executives gives $\bar{x} = 6.7$ years and $s = 2.4$ years. Give a 99% confidence interval for the average number of years of experience for all executives in this branch.

---

[8] Luisa Krull and Allison Fass, eds., "Billionaires," *Forbes,* March 26, 2007, pp. 104–184.

[9] "The 100 Biggest U.S. Markets," *Money,* May 2007, p. 81.

6–34. The Food and Drug Administration (FDA) needs to estimate the average content of an additive in a given food product. A random sample of 75 portions of the product gives $\bar{x} = 8.9$ units and $s = 0.5$ unit. Give a 95% confidence interval for the average number of units of additive in any portion of this food product.

6–35. The management of a supermarket needs to make estimates of the average daily demand for milk. The following data are available (number of half-gallon containers sold per day): 48, 59, 45, 62, 50, 68, 57, 80, 65, 58, 79, 69. Assuming that this is a random sample of daily demand, give a 90% confidence interval for average daily demand for milk.

6–36. According to an article in *Travel & Leisure,* an average plot of land in Spain's San Martin wine-producing region yields 600 bottles of wine each year.[10] Assume this average is based on a random sample of 25 plots and that the sample standard deviation is 100 bottles. Give a 95% confidence interval for the population average number of bottles per plot.

6–37. The data on the daily consumption of fuel by a delivery truck, in gallons, recorded during 25 randomly selected working days, are as follows:

9.7, 8.9, 9.7, 10.9, 10.3, 10.1, 10.7, 10.6, 10.4, 10.6, 11.6, 11.7, 9.7, 9.7, 9.7, 9.8, 12, 10.4, 8.8, 8.9, 8.4, 9.7, 10.3, 10, 9.2

Compute a 90% confidence interval for the daily fuel consumption.

6–38. According to the Darvas Box stock trading system, a trader looks at a chart of stock prices over time and identifies box-shaped patterns. Then one buys the stock if it appears to be in the lower left corner of a box, and sells if in the upper right corner. In simulations with real data, using a sample of 376 trials, the average hold time for a stock was 41.12 days.[11] If the sample standard deviation was 12 days, give a 90% confidence interval for the average hold time in days.

6–39. Refer to the Darvas Box trading model of problem 6–38. The average profit was 11.46%.[12] If the sample standard deviation was 8.2%, give a 90% confidence interval for average profit using this trading system.

6–4 Large-Sample Confidence Intervals for the Population Proportion *p*

Sometimes interest centers on a qualitative, rather than a quantitative, variable. We may be interested in the relative frequency of occurrence of some characteristic in a population. For example, we may be interested in the proportion of people in a population who are users of some product or the proportion of defective items produced by a machine. In such cases, we want to estimate the population proportion p.

The estimator of the population proportion p is the sample proportion $\hat{P}$. In Chapter 5, we saw that when the sample size is large, $\hat{P}$ has an approximately normal sampling distribution. The mean of the sampling distribution of $\hat{P}$ is the population proportion p, and the standard deviation of the distribution of $\hat{P}$ is $\sqrt{pq/n}$, where $q = 1 - p$. Since the standard deviation of the estimator depends on the unknown population parameter, its value is also unknown to us. It turns out, however, that for large samples we may use our actual estimate $\hat{P}$ instead of the unknown parameter p in the formula for the standard deviation. We will, therefore, use $\sqrt{\hat{p}\hat{q}/n}$ as our estimate of the standard deviation of $\hat{P}$. Recall our large-sample rule of thumb: For estimating p, a sample is considered large enough when both $n \cdot p$ and $n \cdot q$ are greater

[10] Bruce Schoenfeld, "Wine: Bierzo's Bounty," *Travel & Leisure,* April 2007, p. 119.

[11] Volker Knapp, "The Darvas Box System," *Active Trader,* April 2007, p. 44.

[12] Ibid.

than 5. (We guess the value of p when determining whether the sample is large enough. As a check, we may also compute $n\hat{p}$ and $n\hat{q}$ once the sample is obtained.)

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

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \qquad (6\text{–}7)$$

where the sample proportion $\hat{p}$ is equal to the number of successes in the sample x, divided by the number of trials (the sample size) n, and $\hat{q} = 1 - \hat{p}$.

We demonstrate the use of equation 6–7 in Example 6–4.

EXAMPLE 6–4

A market research firm wants to estimate the share that foreign companies have in the U.S. market for certain products. A random sample of 100 consumers is obtained, and 34 people in the sample are found to be users of foreign-made products; the rest are users of domestic products. Give a 95% confidence interval for the share of foreign products in this market.

Solution

We have $x = 34$ and $n = 100$, so our sample estimate of the proportion is $\hat{p} = x/n = 34/100 = 0.34$. We now use equation 6–7 to obtain the confidence interval for the population proportion p. A 95% confidence interval for p is

$$\begin{aligned}\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} &= 0.34 \pm 1.96\sqrt{\frac{(0.34)(0.66)}{100}} \\ &= 0.34 \pm 1.96(0.04737) = 0.34 \pm 0.0928 \\ &= [0.2472, 0.4328]\end{aligned}$$

Thus, the firm may be 95% confident that foreign manufacturers control anywhere from 24.72% to 43.28% of the market.

Suppose the firm is not happy with such a wide confidence interval. What can be done about it? This is a problem of *value of information,* and it applies to all estimation situations. As we stated earlier, for a fixed sample size, the higher the confidence you require, the wider will be the confidence interval. The sample size is in the denominator of the standard error term, as we saw in the case of estimating μ. If we should increase n, the standard error of $\hat{P}$ will decrease, and the uncertainty about the parameter being estimated will be narrowed. If the sample size cannot be increased but you still want a narrower confidence interval, you must reduce your confidence level. Thus, for example, if the firm agrees to reduce the confidence level to 90%, z will be reduced from 1.96 to 1.645, and the confidence interval will shrink to

$$0.34 \pm 1.645(0.04737) = 0.34 \pm 0.07792 = [0.2621, 0.4179]$$

The firm may be 90% confident that the market share of foreign products is anywhere from 26.21% to 41.79%. If the firm wanted a high confidence (say 95%) *and* a narrow

FIGURE 6–9 **The Template for Estimating Population Proportions [Estimating Proportion.xls]**

Confidence Interval for Population Proportion

Sample size	100	n		
Sample proportion	0.34	p-hat		

Finite Population Correction

Population Size	2000	N
Correction Factor	0.9749	

$1-\alpha$	Confidence Interval			Finite Population Correction
99%	0.34 ± 0.1220	= [0.2180 , 0.4620]	------>	[0.2210 , 0.4590]
95%	0.34 ± 0.0928	= [0.2472 , 0.4328]	------>	[0.2495 , 0.4305]
90%	0.34 ± 0.0779	= [0.2621 , 0.4179]	------>	[0.2640 , 0.4160]
80%	0.34 ± 0.0607	= [0.2793 , 0.4007]	------>	[0.2808 , 0.3992]

confidence interval, it would have to take a larger sample. Suppose that a random sample of $n = 200$ customers gave us the same result; that is, $x = 68$, $n = 200$, and $\hat{p} = x/n = 0.34$. What would be a 95% confidence interval in this case? Using equation 6–7, we get

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.34 \pm 1.96\sqrt{\frac{(0.34)(0.66)}{200}} = [0.2743, 0.4057]$$

This interval is considerably narrower than our first 95% confidence interval, which was based on a sample of 100.

When proportions using small samples are estimated, the binomial distribution may be used in forming confidence intervals. Since the distribution is discrete, it may not be possible to construct an interval with an exact, prespecified confidence level such as 95% or 99%. We will not demonstrate the method here.

The Template

Figure 6–9 shows the template that can be used for computing confidence intervals for population proportions. The template also has provision for finite population correction. To make this correction, the population size N must be entered in cell N5. If the correction is not needed, it is a good idea to leave this cell blank to avoid creating a distraction.

PROBLEMS

6–40. A maker of portable exercise equipment, designed for health-conscious people who travel too frequently to use a regular athletic club, wants to estimate the proportion of traveling business people who may be interested in the product. A random sample of 120 traveling business people indicates that 28 may be interested in purchasing the portable fitness equipment. Give a 95% confidence interval for the proportion of all traveling business people who may be interested in the product.

6–41. The makers of a medicated facial skin cream are interested in determining the percentage of people in a given age group who may benefit from the ointment. A random sample of 68 people results in 42 successful treatments. Give a 99%

confidence interval for the proportion of people in the given age group who may be successfully treated with the facial cream.

6–42. According to *The Economist,* 55% of all French people of voting age are opposed to the proposed European constitution.[13] Assume that this percentage is based on a random sample of 800 French people. Give a 95% confidence interval for the population proportion in France that was against the European constitution.

6–43. According to *BusinessWeek,* many Japanese consider their cell phones toys rather than tools.[14] If a random sample of 200 Japanese cell phone owners reveals that 80 of them consider their device a toy, calculate a 90% confidence interval for the population proportion of Japanese cell phone users who feel this way.

6–44. A recent article describes the success of business schools in Europe and the demand on that continent for the MBA degree. The article reports that a survey of 280 European business positions resulted in the conclusion that only one-seventh of the positions for MBAs at European businesses are currently filled. Assuming that these numbers are exact and that the sample was randomly chosen from the entire population of interest, give a 90% confidence interval for the proportion of filled MBA positions in Europe.

6–45. According to *Fortune,* solar power now accounts for only 1% of total energy produced.[15] If this number was obtained based on a random sample of 8,000 electricity users, give a 95% confidence interval for the proportion of users of solar energy.

6–46. *Money* magazine is on a search for the indestructible suitcase.[16] If in a test of the Helium Fusion Expandable Suiter, 85 suitcases out of a sample of 100 randomly tested survived rough handling at the airport, give a 90% confidence interval for the population proportion of suitcases of this kind that would survive rough handling.

6-47. A machine produces safety devices for use in helicopters. A quality-control engineer regularly checks samples of the devices produced by the machine, and if too many of the devices are defective, the production process is stopped and the machine is readjusted. If a random sample of 52 devices yields 8 defectives, give a 98% confidence interval for the proportion of defective devices made by this machine.

6–48. Before launching its Buyers' Assurance Program, American Express wanted to estimate the proportion of cardholders who would be interested in this automatic insurance coverage plan. A random sample of 250 American Express cardholders was selected and sent questionnaires. The results were that 121 people in the sample expressed interest in the plan. Give a 99% confidence interval for the proportion of all interested American Express cardholders.

6–49. An airline wants to estimate the proportion of business passengers on a new route from New York to San Francisco. A random sample of 347 passengers on this route is selected, and 201 are found to be business travelers. Give a 90% confidence interval for the proportion of business travelers on the airline's new route.

6–50. According to the *Wall Street Journal,* the rising popularity of hedge funds and similar investment instruments has made splitting assets in cases of divorce much more difficult.[17] If a random sample of 250 divorcing couples reveals that 53 of them have great difficulties in splitting their family assets, construct a 90% confidence interval for the proportion of all couples getting divorced who encounter such problems.

---

[13] "Constitutional Conundrum," *The Economist,* March 17, 2007, p. 10.

[14] Moon Ihlwan and Kenji Hall, "New Tech, Old Habits," *BusinessWeek,* March 26, 2007 p. 49.

[15] Jigar Shah, "Question Authority," *Fortune,* March 5, 2007, p. 26.

[16] "Five Bags, Checked," *Money,* May 2007, p. 126.

[17] Rachel Emma Silverman, "Divorce: Counting Money Gets Tougher," *The Wall Street Journal,* May 5–6, 2007, p. B1.

6–51. According to *BusinessWeek,* environmental groups are making headway on American campuses. In a survey of 570 schools, 130 were found to incorporate chapters of environmental organizations.[18] Assume this is a random sample of universities and use the reported information to construct a 95% confidence interval for the proportion of all U.S. schools with environmental chapters.

6–5 Confidence Intervals for the Population Variance

In some situations, our interest centers on the population variance (or, equivalently, the population standard deviation). This happens in production processes, queuing (waiting line) processes, and other situations. As we know, the sample variance S^2 is the (unbiased) estimator of the population variance σ^2.

To compute confidence intervals for the population variance, we must learn to use a new probability distribution: the *chi-square distribution.* Chi (pronounced *ki*) is one of two X letters in the Greek alphabet and is denoted by χ. Hence, we denote the chi-square distribution by χ^2.

The chi-square distribution, like the t distribution, has associated with it a degrees-of-freedom parameter df. In the application of the chi-square distribution to estimation of the population variance, df $= n - 1$ (as with the t distribution in its application to sampling for the population mean). Unlike the t and the normal distributions, however, the chi-square distribution is not symmetric.

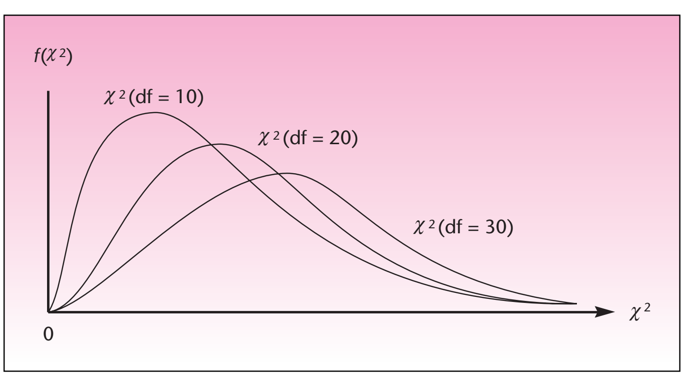
The **chi-square distribution** is the probability distribution of the sum of several independent, squared standard normal random variables.

As a sum of squares, the chi-square random variable cannot be negative and is therefore bounded on the left by zero. The resulting distribution is skewed to the right. Figure 6–10 shows several chi-square distributions with different numbers of degrees of freedom.

The mean of a chi-square distribution is equal to the degrees-of-freedom parameter df. The variance of a chi-square distribution is equal to twice the number of degrees of freedom.

Note in Figure 6–10 that as df increases, the chi-square distribution looks more and more like a normal distribution. In fact, as df increases, the chi-square distribution approaches a normal distribution with mean df and variance 2(df).

FIGURE 6–10 Several Chi-Square Distributions with Different Values of the df Parameter



[18] Heather Green, "The Greening of America's Campuses," *BusinessWeek,* April 9, 2007, p. 64.

TABLE 6–2 **Values and Probabilities of Chi-Square Distributions**

	Area in Right Tail									
df	**0.995**	**0.990**	**0.975**	**0.950**	**0.900**	**0.100**	**0.050**	**0.025**	**0.010**	**0.005**
1	$0.0^{4}393$	$0.0^{3}157$	$0.0^{3}982$	$0.0^{2}393$	0.0158	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	4.61	5.99	7.38	9.21	10.6
3	0.0717	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.3	12.8
4	0.207	0.297	0.484	0.711	1.06	7.78	9.49	11.1	13.3	14.9
5	0.412	0.554	0.831	1.15	1.61	9.24	11.1	12.8	15.1	16.7
6	0.676	0.872	1.24	1.64	2.20	10.6	12.6	14.4	16.8	18.5
7	0.989	1.24	1.69	2.17	2.83	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	14.7	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	16.0	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	5.58	17.3	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	18.5	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	7.04	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.79	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.55	22.3	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	9.31	23.5	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	10.9	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	28.4	31.4	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	13.2	29.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	14.0	30.8	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	14.8	32.0	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	15.7	33.2	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	16.5	34.4	37.7	40.6	44.3	46.9
26	11.2	12.2	13.8	15.4	17.3	35.6	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	18.1	36.7	40.1	43.2	47.0	49.6
28	12.5	13.6	15.3	16.9	18.9	37.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.0	17.7	19.8	39.1	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	20.6	40.3	43.8	47.0	50.9	53.7

Table 4 in Appendix C gives values of the chi-square distribution with different degrees of freedom, for given tail probabilities. An abbreviated version of part of the table is given as Table 6–2. We apply the chi-square distribution to problems of estimation of the population variance, using the following property.

In sampling from a normal population, the random variable

$$\chi^2 = \frac{(n - 1)s^2}{s^2} \qquad (6\text{–}8)$$

has a chi-square distribution with $n - 1$ degrees of freedom.

The distribution of the quantity in equation 6–8 leads to a confidence interval for σ^2. Since the χ^2 distribution is not symmetric, we cannot use equal values with opposite signs (such as ± 1.96 as we did with Z) and must construct the confidence interval using the two distinct tails of the distribution.

A $(1 - \alpha)$ 100% confidence interval for the population variance σ^2 (where the population is assumed normal) is

$$\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right] \quad (6\text{–}9)$$

where $\chi^2_{\alpha/2}$ is the value of the chi-square distribution with $n - 1$ degrees of freedom that cuts off an area of $\alpha/2$ to its right and $\chi^2_{1-\alpha/2}$ is the value of the distribution that cuts off an area of $\alpha/2$ to its left (equivalently, an area of $1 - \alpha/2$ to its right).

We now demonstrate the use of equation 6–9 with an example.

EXAMPLE 6–5

In an automated process, a machine fills cans of coffee. If the average amount filled is different from what it should be, the machine may be adjusted to correct the mean. If the *variance* of the filling process is too high, however, the machine is out of control and needs to be repaired. Therefore, from time to time regular checks of the variance of the filling process are made. This is done by randomly sampling filled cans, measuring their amounts, and computing the sample variance. A random sample of 30 cans gives an estimate $s^2 = 18{,}540$. Give a 95% confidence interval for the population variance σ^2.

Solution

Figure 6–11 shows the appropriate chi-square distribution with $n - 1 = 29$ degrees of freedom. From Table 6–2 we get, for df $= 29$, $\chi^2_{0.025} = 45.7$ and $\chi^2_{0.975} = 16.0$. Using these values, we compute the confidence interval as follows:

$$\left[\frac{29(18{,}540)}{45.7}, \frac{29(18{,}540)}{16.0}\right] = [11{,}765, 33{,}604]$$

We can be 95% sure that the population variance is between 11,765 and 33,604.

FIGURE 6–11 **Values and Tail Areas of a Chi-Square Distribution with 29 Degrees of Freedom**

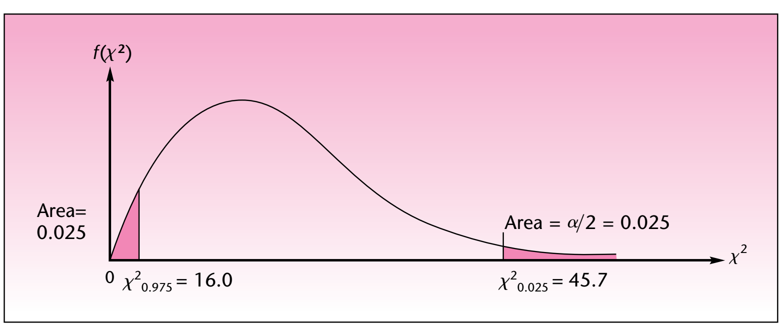


FIGURE 6–12 **The Template for Estimating Population Variances [Estimating Variance.xls; Sheet: Sample Stats]**

Confidence Interval for Population Variance

Assumption:
The population is normally distributed

Sample Size	30	n
Sample Variance	18.54	s^2

$1-\alpha$	Confidence Interval
95%	[11.7593 , 33.5052]
95%	[11.7593 , 33.5052]
90%	[12.6339 , 30.3619]
80%	[13.7553 , 27.1989]

FIGURE 6–13 **The Template for Estimating Population Variances [Estimating Variance.xls; Sheet: Sample Data]**

Confidence Interval for Population Variance

Assumption:
The population is normally distributed

Sample Data
123
125
122
120
124
123
121
121

Sample Size	25	n
Sample Variance	1.979057	s^2

$1-\alpha$	Confidence Interval
95%	[1.20662 , 3.83008]
95%	[1.20662 , 3.83008]
90%	[1.30433 , 3.4298]
80%	[1.43081 , 3.03329]

The Template

The workbook named Estimating Variance.xls provides sheets for computing confidence intervals for population variances when

1. The sample statistics are known.
2. The sample data are known.

Figure 6–12 shows the first sheet and Figure 6–13 shows the second. An assumption in both cases is that the population is normally distributed.

PROBLEMS

In the following problems, assume normal populations.

6–52. The service time in queues should not have a large variance; otherwise, the queue tends to build up. A bank regularly checks service time by its tellers to determine its variance. A random sample of 22 service times (in minutes) gives $s^2 = 8$. Give a 95% confidence interval for the variance of service time at the bank.

6–53. A sensitive measuring device should not have a large variance in the errors of measurements it makes. A random sample of 41 measurement errors gives $s^2 = 102$. Give a 99% confidence interval for the variance of measurement errors.

6–54. A random sample of 60 accounts gives a sample variance of 1,228. Give a 95% confidence interval for the variance of all accounts.

6–55. In problem 6–21, give a 99% confidence interval for the variance of the number of miles that may be driven on a tire.

6–56. In problem 6–25, give a 95% confidence interval for the variance of the population of billionaires' worth in dollars.

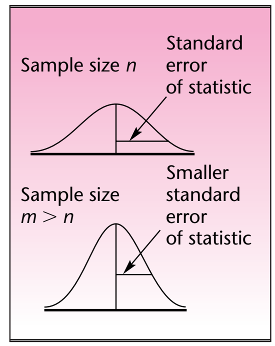
6–57. In problem 6–26, give a 95% confidence interval for the variance of the caloric content of all 3-ounce cuts of top loin steak.

6–58. In problem 6–27, give a 95% confidence interval for the variance of the transit time for all goods.

6–6 Sample-Size Determination

One of the questions a statistician is most frequently asked before any actual sampling takes place is: "How large should my sample be?" From a *statistical* point of view, the best answer to this question is: "Get as large a sample as you can afford. If possible, 'sample' the entire population." If you need to know the mean or proportion of a population, and you can sample the entire population (i.e., carry out a census), you will have all the information and will know the parameter exactly. Clearly, this is better than any estimate. This, however, is unrealistic in most situations due to economic constraints, time constraints, and other limitations. "Get as large a sample as you can afford" is the best answer if we ignore all costs, because the larger the sample, the smaller the standard error of our statistic. The smaller the standard error, the less uncertainty with which we have to contend. This is demonstrated in Figure 6–14.

FIGURE 6–14
Standard Error of a Statistic as a Function of Sample Size



When the sampling budget is limited, the question often is how to find the *minimum* sample size that will satisfy some precision requirements. In such cases, you should explain to the designer of the study that he or she must first give you answers to the following three questions:

1. How close do you want your sample estimate to be to the unknown parameter? The answer to this question is denoted by B (for "bound").
2. What do you want the confidence level to be so that the distance between the estimate and the parameter is less than or equal to B?
3. The last, and often misunderstood, question that must be answered is: What is your estimate of the variance (or standard deviation) of the population in question?

Only after you have answers to all three questions can you specify the minimum required sample size. Often the statistician is told: "How can I give you an estimate of the variance? I don't know. You are the statistician." In such cases, try to get from your client some idea about the variation in the population. If the population is approximately normal and you can get 95% bounds on the values in the *population, divide the difference between the upper and lower bounds by* 4; this will give you a rough guess of σ. Or you may take a small, inexpensive *pilot* survey and estimate σ by the sample standard deviation. Once you have obtained the three required pieces of information, all you need to do is to substitute the answers into the appropriate formula that follows:

Minimum required sample size in estimating the population mean μ is

$$n = \frac{Z^2_{\alpha/2}\sigma^2}{B^2} \qquad (6\text{–}10)$$

Minimum required sample size in estimating the population proportion p is

$$n = \frac{Z^2_{\alpha/2}pq}{B^2} \qquad (6\text{–}11)$$

Equations 6–10 and 6–11 are derived from the formulas for the corresponding confidence intervals for these population parameters based on the normal distribution. In the case of the population mean, B is the half-width of a $(1 - \alpha)$ 100% confidence interval for μ, and therefore

$$B = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{6–12}$$

Equation 6–10 is the solution of equation 6–12 for the value of n. Note that B is the margin of error. We are solving for the minimum sample size for a given margin of error.

Equation 6–11, for the minimum required sample size in estimating the population proportion, is derived in a similar way. Note that the term pq in equation 6–11 acts as the population variance in equation 6–10. To use equation 6–11, we need a guess of p, the unknown population proportion. Any prior estimate of the parameter will do. When none is available, we may take a pilot sample, or—in the absence of any information—we use the value $p = 0.5$. This value maximizes pq and thus ensures us a minimum required sample size that will work for any value of p.

EXAMPLE 6–6

A market research firm wants to conduct a survey to estimate the average amount spent on entertainment by each person visiting a popular resort. The people who plan the survey would like to be able to determine the average amount spent by all people visiting the resort to within \$120, with 95% confidence. From past operation of the resort, an estimate of the population standard deviation is $\sigma = \$400$. What is the minimum required sample size?

Solution

Using equation 6–10, the minimum required sample size is

$$n = \frac{z^2_{\alpha/2}\sigma^2}{B^2}$$

We know that $B = 120$, and σ^2 is estimated at $400^2 = 160{,}000$. Since we want 95% confidence, $z_{\alpha/2} = 1.96$. Using the equation, we get

$$n = \frac{(1.96)^2 160{,}000}{120^2} = 42.684$$

Therefore, the minimum required sample size is 43 people (we cannot sample 42.684 people, so we go to the next higher integer).

EXAMPLE 6–7

The manufacturer of a sports car wants to estimate the proportion of people in a given income bracket who are interested in the model. The company wants to know the population proportion p to within 0.10 with 99% confidence. Current company records indicate that the proportion p may be around 0.25. What is the minimum required sample size for this survey?

Using equation 6–11, we get

Solution

$$n = \frac{z_{\alpha/2}^2 pq}{B^2} = \frac{(2.576)^2(0.25)(0.75)}{0.10^2} = 124.42$$

The company should, therefore, obtain a random sample of at least 125 people. Note that a different guess of p would have resulted in a different sample size.

PROBLEMS

6–59. What is the required sample size for determining the proportion of defective items in a production process if the proportion is to be known to within 0.05 with 90% confidence? No guess as to the value of the population proportion is available.

6–60. How many test runs of the new Volvo S40 model are required for determining its average miles-per-gallon rating on the highway to within 2 miles per gallon with 95% confidence, if a guess is that the variance of the population of miles per gallon is about 100?

6–61. A company that conducts surveys of current jobs for executives wants to estimate the average salary of an executive at a given level to within $2,000 with 95% confidence. From previous surveys it is known that the variance of executive salaries is about 40,000,000. What is the minimum required sample size?

6–62. Find the minimum required sample size for estimating the average return on real estate investments to within 0.5% per year with 95% confidence. The standard deviation of returns is believed to be 2% per year.

6–63. A company believes its market share is about 14%. Find the minimum required sample size for estimating the actual market share to within 5% with 90% confidence.

6–64. Find the minimum required sample size for estimating the average number of designer shirts sold per day to within 10 units with 90% confidence if the standard deviation of the number of shirts sold per day is about 50.

6–65. Find the minimum required sample size of accounts of the Bechtel Corporation if the proportion of accounts in error is to be estimated to within 0.02 with 95% confidence. A rough guess of the proportion of accounts in error is 0.10.

6–7 The Templates

Optimizing Population Mean Estimates

Figure 6–15 shows the template that can be used for determining minimum sample size for estimating a population mean. Upon entering the three input data—confidence level desired, half-width (B) desired, and the population standard deviation—in the cells C5, C6, and C7, respectively, the minimum required sample size appears in cell C9.

Determining the Optimal Half-Width

Usually, the population standard deviation σ is not known for certain, and we would like to know how sensitive the minimum sample size is to changes in σ. In addition, there is no hard rule for deciding what the half-width B should be. Therefore, a tabulation of the minimum sample size for various values of σ and B will help us to

FIGURE 6–15 **The Template for Determining Minimum Sample Size [Sample Size.xls; Sheet: Population Mean]**

Sample Size Determination for Population Mean

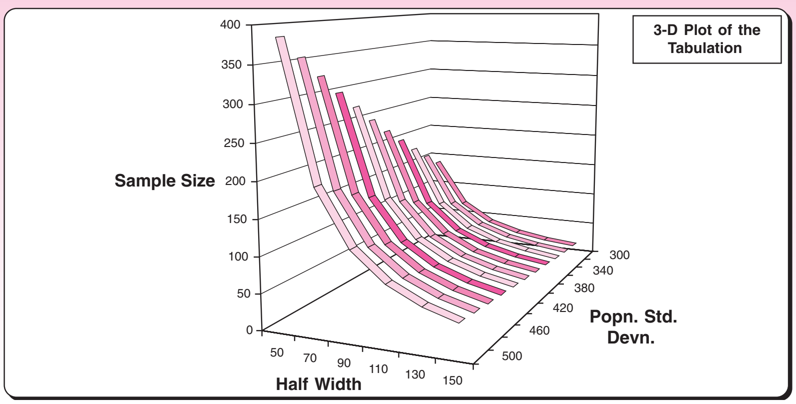
Confidence Level Desired	95%	
Half-Width Desired	120	B
Population Stdev.	400	σ
Minimum Sample Size	43	

Cost Analysis

Sampling Cost	$ 211
Error Cost	$ 720
Total Cost	$ 931

Tabulation of Minimum Sample Size

	Half Width Desired					
Popn. Std. Devn.	50	70	90	110	130	150
300	139	71	43	29	21	16
320	158	81	49	33	24	18
340	178	91	55	37	27	20
360	200	102	62	42	30	23
380	222	114	69	46	33	25
400	246	126	76	51	37	28
420	272	139	84	57	41	31
440	298	152	92	62	45	34
460	326	166	101	68	49	37
480	355	181	110	74	53	40
500	385	196	119	80	57	43



choose a suitable *B*. To create the table, enter the desired starting and ending values of σ in cells F6 and F16, respectively. Similarly, enter the desired starting and ending values for *B* in cells G5 and L5, respectively. The complete tabulation appears immediately. Note that the tabulation pertains to the confidence level entered in cell C5. If this confidence level is changed, the tabulation will be immediately updated.

To visualize the tabulated values a 3-D plot is created below the table. As can be seen from the plot, *the minimum sample size is more sensitive to the half-width* B *than the population standard deviation* σ. That is, when *B* decreases, the minimum sample size increases sharply. This sensitivity emphasizes the issue of how to choose *B*. The natural decision criterion is cost. When *B* is small, the possible error in the estimate is small and therefore the error cost is small. But a small *B* means large sample size, increasing the sampling cost. When *B* is large, the reverse is true. Thus *B* is a compromise between sampling cost and error cost. If cost data are available, an optimal *B* can be found. The template contains features that can be used to find optimal *B*.

The sampling cost formula is to be entered in cell C14. Usually, sampling cost involves a fixed cost that is independent of the sample size and a variable cost that increases linearly with the sample size. The fixed cost would include the cost of planning and organizing the sampling experiment. The variable cost would include the costs of selection, measurement, and recording of each sampled item. For example, if the fixed cost is $125 and the variable cost is $2 per sampled item, then the cost of sampling a total of 43 items will be \$125 + \$2*43 = \$211. In the template, we enter the formula "=125+2*C9" in cell C14, so that the sampling cost appears in cell C14. The instruction for entering the formula appears also in the comment attached to cell C14.

The error cost formula is entered in cell C15. In practice, error costs are more difficult to estimate than sampling costs. Usually, the error cost increases more rapidly than linearly with the amount of error. Often a quadratic formula is suitable for modeling the error cost. In the current case seen in the template, the formula "=0.05*C6^2" has been entered in cell C15. This means the error cost equals $0.05B^2$. Instructions for entering this formula appear in the comment attached to cell C15.

When proper cost formulas are entered in cells C14 and C15, cell C16 shows the total cost. It is possible to tabulate, in place of minimum sample size, the sampling cost, the error cost, or the total cost. You can select what you wish to tabulate using the drop-down box in cell G3.

Using the Solver

By manually adjusting B, we can try to minimize the total cost. Another way to find the optimal B that minimizes the total cost is to use the Solver. Unprotect the sheet and select the Solver command in the Analysis group on the Data tab and click on the Solve button. If the formulas entered in cells C14 and C15 are realistic, the Solver will find a realistic optimal B, and a message saying that an optimal B has been found will appear in a dialog box. Select Keep Solver Solution and press the OK button. In the present case, the optimal B turns out to be 70.4.

For some combinations of sampling and error cost formulas, the Solver may not yield meaningful answers. For example, it may get stuck at a value of zero for B. In such cases, the manual method must be used. At times, it may be necessary to start the Solver with different initial values for B and then take the B value that yields the least total cost.

Note that the total cost also depends on the confidence level (in cell C5) and the population standard deviation (in cell C7). If these values are changed, then the total cost will change and so will the optimal B. The new optimum must be found once again, manually or using the Solver.

Optimizing Population Proportion Estimates

Figure 6–16 shows the template that can be used to determine the minimum sample size required for estimating a population proportion. This template is almost identical to the one in Figure 6–15, which is meant for estimating the population mean. The only difference is that instead of population standard deviation, we have population proportion in cell C7.

The tabulation shows that the minimum sample size increases with population proportion p until p reaches a value of 0.5 and then starts decreasing. Thus, the worst case occurs when p is 0.5.

The formula for error cost currently in cell C15 is "=40000*C6^2." This means the cost equals $40{,}000B^2$. Notice how different this formula looks compared to the formula $0.05B^2$ we saw in the case of estimating population mean. The coefficient 40,000 is much larger than 0.05. This difference arises because B is a much smaller number in the case of proportions. The formula for sampling cost currently entered in cell C14 is "=125+2*C9" (same as the previous case).

The optimal B that minimizes the total cost in cell C16 can be found using the Solver just as in the case of estimating population mean. Select the Solver command in the Analysis group on the Data tab and press the Solve button. In the current case, the Solver finds the optimal B to be 0.07472.

FIGURE 6–16 **The Template for Determining Minimum Sample Size [Sample Size.xls; Sheet: Population Proportion]**

Sample Size Determination for Population Proportion

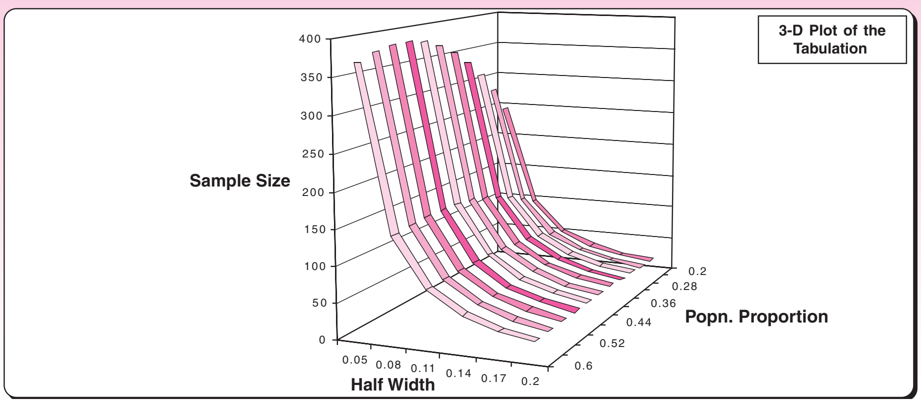
Confidence Level Desired	95%
Half-Width Desired	0.1
Population Proportion	0.25
Minimum Sample Size	73

Cost Analysis

Sampling Cost	$ 271
Error Cost	$ 400
Total Cost	$ 671

Tabulation of Minimum Sample Size

	Half Width Desired					
Popn. Proportion	0.5	0.08	0.11	0.14	0.17	0.2
0.2	246	97	51	32	22	16
0.24	281	110	58	36	25	18
0.28	310	122	65	40	27	20
0.32	335	131	70	43	29	21
0.36	355	139	74	46	31	23
0.4	369	145	77	48	32	24
0.44	379	148	79	49	33	24
0.48	384	150	80	49	34	24
0.52	384	150	80	49	34	24
0.56	379	148	79	49	33	24
0.6	369	145	77	48	32	24



6–8 Using the Computer

Using Excel Built-In Functions for Confidence Interval Estimation

In addition to the Excel templates that were described in this chapter, you can use various statistical functions of Excel to directly build the required confidence intervals. In this section we review these functions.

The function **CONFIDENCE** returns a value that you can use to construct a confidence interval for a population mean. The confidence interval is a range of values. Your sample mean *x* is at the center of this range and the range is `x ± CONFIDENCE`. In the formula `CONFIDENCE(alpha,Stdev,Size)`, *alpha* is the significance level used to compute the confidence level. Thus, the confidence level equals 100*(1 − alpha)%, or, in other words, an alpha of 0.1 indicates a 90% confidence level. *Stdev* is the population standard deviation for the data and is assumed to be known. *Size* is the sample size. The value of alpha should be between zero and one. If Size is not an integer number, it is truncated. As an example, suppose we observe that, in a sample of 40 employees, the average length of travel to work is 20 minutes with a population standard deviation of 3.5. With alpha = .05, `CONFIDENCE(.05, 3.5, 40)` returns the value 1.084641, so the corresponding confidence interval is 20 ± 1.084641, or [18.915, 21.084].

No function is available for constructing confidence intervals for population proportion or variance per se. But you can use the following useful built-in functions to build the corresponding confidence intervals.

The function **NORMSINV**, which was discussed in Chapter 4, returns the inverse of the standard normal cumulative distribution. For example, `NORMSINV(0.975)` returns the value 1.95996, or approximately the value of 1.96, which is extensively used for constructing the 95% confidence intervals on the mean. You can use this function to construct confidence intervals on the population mean or population proportion.

The function **TINV** returns the t value of the Student's t distribution as a function of the probability and the degrees of freedom. In the formula `TINV(p,df)` p is the probability associated with the two-tailed Student's t distribution, and df is the number of degrees of freedom that characterizes the distribution. Note that if df is not an integer, it is truncated. Given a value for probability p, TINV seeks that value t such that `P(|T|>t)=P(T>t or T<-t)=p`, where T is a random variable that follows the t distribution. You can use TINV for constructing a confidence interval for a population mean when the population standard deviation is not known. For example, for building a 95% confidence interval for the mean of the population assuming a sample of size 15, you need to multiply `TINV(0.05, 14)` by the sample standard deviation and divide it by the square root of sample size to construct the half width of your confidence interval.

The function **CHIINV** returns the inverse of the one-tailed probability of the chi-squared distribution. You can use this function to construct a confidence interval for the population variance. In the formula `CHIINV(p,df)`, p is a probability associated with the chi-squared distribution, and df is the number of degrees of freedom. Given a value for probability p, CHIINV seeks that value x such that $P(X > x) = p$, where X is a random variable with the chi-square distribution. As an example, `CHIINV(0.025, 10)` returns the value 20.483.

Using MINITAB for Confidence Interval Estimation

MINITAB can be used for obtaining confidence interval estimates for various population parameters. Let's start by obtaining the confidence interval for the population mean when the population standard deviation is known. In this case start by choosing Stat ► Basic Statistics ► 1-Sample Z from the menu bar. Then the 1-Sample Z dialog box will appear as shown in Figure 6–17. Enter the name of the column that contains the values of your sample. If you have summarized data of your sample, you

FIGURE 6–17 **Confidence Interval on a Population Mean Using MINITAB**

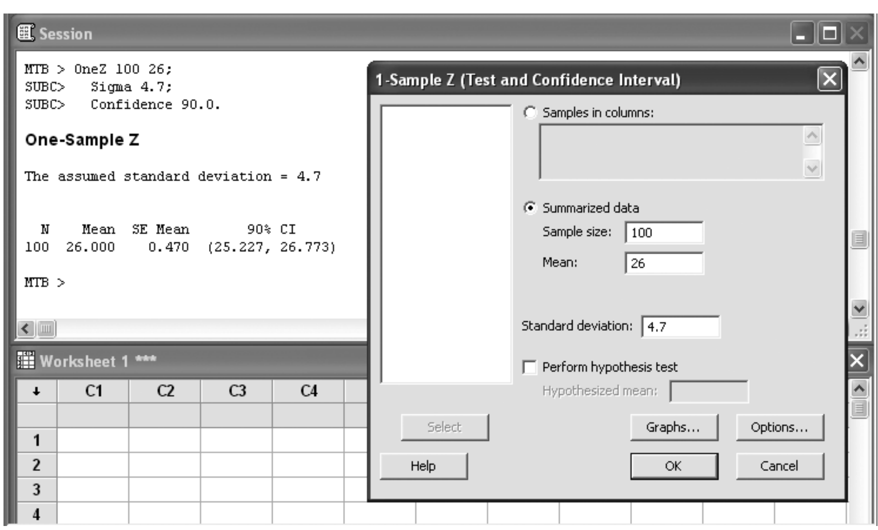
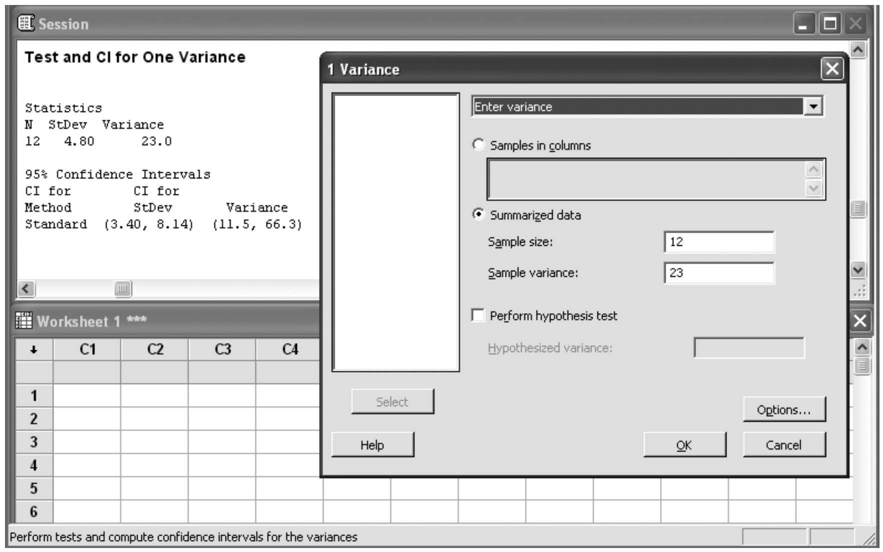


FIGURE 6–18 **Confidence Interval on Population Variance Using MINITAB**



can also enter it in the corresponding section. Enter the value of the population standard deviation in the next box. You can define the desired confidence level of your interval by clicking on the Options button. Then press the OK button. The resulting confidence interval and corresponding Session commands will appear in the Session window as seen in Figure 6–17.

For obtaining the confidence interval for the population mean when the population standard deviation is not known you need to choose Stat ▶ Basic Statistics ▶ 1-Sample t from the menu bar. The setting is the same as the previous dialog box except that you need to determine the sample standard deviation if you choose to enter summarized data of your sample.

MINITAB can also be used for obtaining a confidence interval for a population proportion by selecting Stat ▶ Basic Statistics ▶ 1Proportion from the menu bar. As before, we can define the summarized data of our sample as number of trials (sample size) and number of events (number of samples with the desired condition). After defining the required confidence level press the OK button.

MINITAB also enables you to obtain a confidence interval for the population variance. For this purpose, start by choosing Stat ▶ Basic Statistics ▶ 1 Variance from the menu bar. In the corresponding dialog box, as seen in Figure 6–18, use the drop-down menu to choose whether input values refer to standard deviation or variance. After entering the summarized data or the name of the column that contains the raw data, you can set the desired confidence level by clicking on the Options button. Then press the OK button. The final result, which will be shown in the Session window, contains confidence intervals for both population variance and standard deviation.

6–9 Summary and Review of Terms

In this chapter, we learned how to construct **confidence intervals** for population parameters. We saw how confidence intervals depend on the sampling distributions of the statistics used as estimators. We also encountered two new sampling distributions: the ***t* distribution,** used in estimating the population mean when the population standard deviation is unknown, and the **chi-square distribution,** used in estimating the population variance. The use of either distribution assumes a normal population. We saw how the new distributions, as well as the normal distribution, allow us to construct confidence intervals for population parameters. We saw how to determine the minimum required sample size for estimation.

ADDITIONAL PROBLEMS

6–66. Tradepoint, the electronic market set up to compete with the London Stock Exchange, is losing on average \$32,000 per day. A potential long-term financial partner for Tradepoint needs a confidence interval for the actual (population) average daily loss, in order to decide whether future prospects for this venture may be profitable. In particular, this potential partner wants to be confident that the true average loss at this period is not over \$35,000 per day, which it would consider hopeless. Assume the \$32,000 figure given above is based on a random sample of 10 trading days, assume a normal distribution for daily loss, and assume a standard deviation of $s = \$6,000$. Construct a 95% confidence interval for the average daily loss in this period. What decision should the potential partner make?

6–67. Landings and takeoffs at Schiphol, Holland, per month are (in 1,000s) as follows:

26, 19, 27, 30, 18, 17, 21, 28, 18, 26, 19, 20, 23, 18, 25, 29, 30, 26, 24, 22, 31, 18, 30, 19

Assume a random sample of months. Give a 95% confidence interval for the average monthly number of takeoffs and landings.

6–68. Thomas Stanley, who surveyed 200 millionaires in the United States for his book *The Millionaire Mind,* found that those in that bracket had an average net worth of \$9.2 million. The sample variance was 1.3 million $\$^2$. Assuming that the surveyed subjects are a random sample of U.S. millionaires, give a 99% confidence interval for the average net worth of U.S. millionaires.

6–69. The Java computer language, developed by Sun Microsystems, has the advantage that its programs can run on types of hardware ranging from mainframe computers all the way down to handheld computing devices or even smart phones. A test of 100 randomly selected programmers revealed that 71 preferred Java to their other most used computer languages. Construct a 95% confidence interval for the proportion of all programmers in the population from which the sample was selected who prefer Java.

6–70. According to an advertisement in *Worth,* in a survey of 68 multifamily office companies, the average client company had assets of \$55.3 million.[19] If this is a result of a random sample and the sample standard deviation was \$21.6 million, give a 95% confidence interval for the population average asset value.

6–71. According to the *Wall Street Journal,* an average of 44 tons of carbon dioxide will be saved per year if new, more efficient lamps are used.[20] Assume that this average is based on a random sample of 15 test runs of the new lamps and that the sample standard deviation was 18 tons. Give a 90% confidence interval for average annual savings.

6–72. Finjan is a company that makes a new security product designed to protect software written in the Java programming language against hostile interference. The market for this program materialized recently when Princeton University experts

[19] Ad: "The FMO Industry," *Worth,* April 2007, p. 55.

[20] John J. Fialka and Kathryn Kranhold, "Households Would Need New Bulbs to Meet Lighting Efficiency Rules," *The Wall Street Journal,* May 5–6, 2007, p. A1.

showed that a hacker could write misleading applets that fool Java's built-in security rules. If, in 430 trials, the system was fooled 47 times, give a 95% confidence interval for p = probability of successfully fooling the machine.

6–73. Sony's new optical disk system prototype tested and claimed to be able to record an average of 1.2 hours of high-definition TV. Assume $n = 10$ trials and $\sigma = 0.2$ hour. Give a 90% confidence interval.

6–74. The average customer of the Halifax bank in Britain (of whom there are 7.6 million) received $3,600 when the institution changed from a building society to a bank. If this is based on a random sample of 500 customers with standard deviation = $800, give a 95% confidence interval for the average amount paid to any of the 7.6 million bank customers.

6–75. FinAid is a new, free Web site that helps people obtain information on 180,000 college tuition aid awards. A random sample of 500 such awards revealed that 368 were granted for reasons other than financial need. They were based on the applicant's qualifications, interests, and other variables. Construct a 95% confidence interval for the proportion of all awards on this service made for reasons other than financial need.

6–76. In May 2007, a banker was arrested and charged with insider trading after government investigators had secretly looked at a sample of nine of his many trades and found that on these trades he had made a total of $7.5 million.[21] Compute the average earning per trade. Assume also that the sample standard deviation was $0.5 million and compute a 95% confidence interval for the average earning per trade for all trades made by this banker. Use the assumption that the nine trades were randomly selected.

6–77. In problem 6–76, suppose the confidence interval contained the value 0.00. How could the banker's attorney use this information to defend his client?

6–78. A small British computer-game firm, Eidos Interactive PLC, stunned the U.S.- and Japan-dominated market for computer games when it introduced Lara Croft, an Indiana Jones-like adventuress. The successful product took two years to develop. One problem was whether Lara should have a swinging ponytail, which was decided after taking a poll. If in a random sample of 200 computer-game enthusiasts, 161 thought she should have a swinging ponytail (a computer programmer's nightmare to design), construct a 95% confidence interval for the proportion in this market. If the decision to incur the high additional programming cost was to be made if $p > 0.90$, was the right decision made (when Eidos went ahead with the ponytail)?

6–79. In a survey, *Fortune* rated companies on a 0 to 10 scale. A random sample of 10 firms and their scores is as follows.[22]

> FedEx 8.94, Walt Disney 8.76, CHS 8.67, McDonald's 7.82, CVS 6.80, Safeway 6.57, Starbucks 8.09, Sysco 7.42, Staples 6.45, HNI 7.29.

Construct a 95% confidence interval for the average rating of a company on *Fortune*'s entire list.

6–80. According to a survey published in the *Financial Times,* 56% of executives at Britain's top 500 companies are less willing than they had been five years ago to sacrifice their family lifestyle for their career. If the survey consisted of a random sample of 40 executives, give a 95% confidence interval for the proportion of executives less willing to sacrifice their family lifestyle.

6–81. Fifty years after the birth of duty-free shopping at international airports and border-crossing facilities, the European commission announced plans to end this form of business. A study by Cranfield University was carried out to estimate the average

[21] Eric Dash, "Banker Jailed in Trading on 9 Deals," *The New York Times,* May 4, 2007, p. C1.

[22] Anne Fisher, "America's Most Admired Companies," *Fortune,* March 19, 2007, pp. 88–115.

percentage rise in airline landing charges that would result as airlines try to make up for the loss of on-board duty-free shopping revenues. The study found the average increase to be 60%. If this was based on a random sample of 22 international flights and the standard deviation of increase was 25%, give a 90% confidence interval for the average increase.

6–82. When NYSE, NASDAQ, and the British government bonds market were planning to change prices of shares and bonds from powers of 2, such as 1/2, 1/4, 1/8, 1/16, 1/32, to decimals (hence $1/32 = 0.03125$), they decided to run a test. If the test run of trading rooms using the new system revealed that 80% of the traders preferred the decimal system and the sample size was 200, give a 95% confidence interval for the percentage of all traders who will prefer the new system.

6–83. A survey of 5,250 business travelers worldwide conducted by OAG Business Travel Lifestyle indicated that 91% of business travelers consider legroom the most important in-flight feature. (Angle of seat recline and food service were second and third, respectively.) Give a 95% confidence interval for the proportion of all business travelers who consider legroom the most important feature.

6–84. Use the following random sample of suitcase prices to construct a 90% confidence interval for the average suitcase price.[23]

$285, 110, 495, 119, 450, 125, 250, 320

6–85. According to *Money,* 60% of men have significant balding by age 50.[24] If this finding is based on a random sample of 1,000 men of age 50, give a 95% confidence interval for the population of men of 50 who show some balding.

6–86. An estimate of the average length of pins produced by an automatic lathe is wanted to within 0.002 inch with a 95% confidence level. σ is guessed to be 0.015 inch.

a. What is the minimum sample size?

b. If the value of σ may be anywhere between 0.010 and 0.020 inch, tabulate the minimum sample size required for σ values from 0.010 to 0.020 inch.

c. If the cost of sampling and testing n pins is $(25 + 6n)$ dollars, tabulate the costs for σ values from 0.010 to 0.020 inch.

6–87. Wells Fargo Bank, based in San Francisco, offered the option of applying for a loan over the Internet. If a random sample of 200 test runs of the service reveal an average of 8 minutes to fill in the electronic application and standard deviation = 3 minutes, construct a 75% confidence interval for μ.

6–88. An estimate of the percentage defective in a lot of pins supplied by a vendor is desired to within 1% with a 90% confidence level. The actual percentage defective is guessed to be 4%.

a. What is the minimum sample size?

b. If the actual percentage defective may be anywhere between 3% and 6%, tabulate the minimum sample size required for actual percentage defective from 3% to 6%.

c. If the cost of sampling and testing n pins is $(25 + 6n)$ dollars, tabulate the costs for the same percentage defective range as in part (*b*).

6–89. The lengths of pins produced by an automatic lathe are normally distributed. A random sample of 20 pins gives a sample mean of 0.992 inch and a sample standard deviation of 0.013 inch.

a. Give a 95% confidence interval for the average lengths of all pins produced.

b. Give a 99% confidence interval for the average lengths of all pins produced.

[23] Charles Passy, "Field Test," *Money,* May 2007, p. 127.

[24] Patricia B. Gray, "Forever Young," *Money,* March 2007, p. 94.

6–90. You take a random sample of 100 pins from the lot supplied by the vendor and test them. You find 8 of them defective. What is the 95% confidence interval for percentage defective in the lot?

6–91. A statistician estimates the 90% confidence interval for the mean of a normally distributed population as 172.58 ± 3.74 at the end of a sampling experiment, assuming a known population standard deviation. What is the 95% confidence interval?

6–92. A confidence interval for a population mean is to be estimated. The population standard deviation σ is guessed to be anywhere from 14 to 24. The half-width B desired could be anywhere from 2 to 7.

a. Tabulate the minimum sample size needed for the given ranges of σ and B.

b. If the fixed cost of sampling is \$350 and the variable cost is \$4 per sample, tabulate the sampling cost for the given ranges of σ and B.

c. If the cost of estimation error is given by the formula $10B^2$, tabulate the total cost for the given ranges of σ and B. What is the value of B that minimizes the total cost when $\sigma = 14$? What is the value of B that minimizes the total cost when $\sigma = 24$?

6–93. A marketing manager wishes to estimate the proportion of customers who prefer a new packaging of a product to the old. He guesses that 60% of the customers would prefer the new packaging. The manager wishes to estimate the proportion to within 2% with 90% confidence. What is the minimum required sample size?

6–94. According to *Money,* a survey of 1,700 executives revealed that 51% of them would likely choose a different field if they could start over.[25] Construct a 95% confidence interval for the proportion of all executives who would like to start over, assuming the sample used was random and representative of all executives.

6–95. According to *Shape,* on the average, 1/2 cup of edamame beans contains 6 grams of protein.[26] If this conclusion is based on a random sample of 50 half-cups of edamames and the sample standard deviation is 3 grams, construct a 95% confidence interval for the average amount of protein in 1/2 cup of edamames.

[25] Jean Chatzky, "To Invent the New You, Don't Bankrupt Old You," *Money,* May 2007, p. 30.

[26] Susan Learner Barr, "Smart Eating," *Shape,* June 2007, p. 202.

CASE 7 Presidential Polling

A company wants to conduct a telephone survey of randomly selected voters to estimate the proportion of voters who favor a particular candidate in a presidential election, to within 2% error with 95% confidence. It is guessed that the proportion is 53%.

1. What is the required minimum sample size?
2. The project manager assigned to the survey is not sure about the actual proportion or about the 2% error limit. The proportion may be anywhere from 40% to 60%. Construct a table for the minimum sample size required with half-width ranging from 1% to 3% and actual proportion ranging from 40% to 60%.
3. Inspect the table produced in question 2 above. Comment on the relative sensitivity of the minimum sample size to the actual proportion and to the desired half-width.

4. At what value of the actual proportion is the required sample size the maximum?
5. The cost of polling includes a fixed cost of \$425 and a variable cost of \$1.20 per person sampled, thus the cost of sampling n voters is \$(425 + 1.20n). Tabulate the cost for range of values as in question 2 above.
6. A competitor of the company that had announced results to within ±3% with 95% confidence has started to announce results to within ±2% with 95% confidence. The project manager wants to go one better by improving the company's estimate to be within ±1% with 95% confidence. What would you tell the manager?

CASE 8 Privacy Problem

A business office has private information about its customers. A manager finds it necessary to check whether the workers inadvertently give away any private information over the phone. To estimate the percentage of times that a worker does give away such information, an experiment is proposed. At a randomly selected time, a call will be placed to the office and the caller will ask several routine questions. The caller will intersperse the routine questions with three questions (attempts) that ask for private information that should not be given out. The caller will note how many attempts were made and how many times private information was given away.

The true proportion of the times that private information is given away during an attempt is guessed to be 7%. The cost of making a phone call, including the caller's wages, is \$2.25. This cost is per call, or per three attempts. Thus the cost per attempt is \$0.75. In addition, the fixed cost to design the experiment is \$380.

1. What is the minimum sample size (of attempts) if the proportion is to be estimated within ±2% with 95% confidence? What is the associated total cost?
2. What is the minimum sample size if the proportion is to be estimated within ±1% with 95% confidence? What is the associated total cost?
3. Prepare a tabulation and a plot of the total cost as the desired accuracy varies from ±1% to ±3% and the population proportion varies from 5% to 10%.
4. If the caller can make as many as five attempts in one call, what is the total cost for ±2% accuracy with 95% confidence? Assume that cost per call and the fixed cost do not change.
5. If the caller can make as many as five attempts in one call, what is the total cost for ±1% accuracy with 95% confidence? Assume that the cost per call and the fixed cost do not change.
6. What are the problems with increasing the number of attempts in one call?