CHAPTER 5

Statistical Inference: Estimation

This chapter shows how to use sample data to estimate population parameters. With quantitative variables, we estimate the population mean. A study dealing with health care issues, for example, might estimate population parameters such as the mean amount of money spent on prescription drugs during the past year and the mean number of visits to a physician. With categorical variables, we estimate population proportions for the categories. The health care study might estimate the proportions of people who (have, do not have) medical insurance and the proportions who (are satisfied, are not satisfied) with their health care.

We first learn about two types of estimates of parameters. Then, Sections 5.2 and 5.3 apply them to population means and proportions. Section 5.4 finds the sample size needed to achieve the desired precision of estimation. Section 5.5 discusses estimation of medians and other parameters.

5.1 POINT AND INTERVAL ESTIMATION

There are two types of estimates of parameters:

- A ***point estimate*** is a single number that is the best guess for the parameter.
- An ***interval estimate*** is an interval of numbers around the point estimate, within which the parameter value is believed to fall.

For example, a GSS asked, "Do you believe there is a life after death?" For 1958 subjects sampled, the point estimate for the proportion of all Americans who would respond *yes* equals 0.73. An interval estimate predicts that the population proportion responding *yes* falls between 0.71 and 0.75. That is, it predicts that the point estimate of 0.73 falls within a *margin of error* of 0.02 of the true value. Thus, an interval estimate helps us gauge the probable precision of a point estimate.

The term *estimate* alone is often used as short for *point estimate*. The term *estimator* then refers to a particular type of statistic for estimating a parameter and *estimate* refers to its value for a specific sample. For example, the sample proportion is an estimator of a population proportion. The value 0.73 is the estimate for the population proportion believing in life after death.

Point Estimation of Parameters

Any particular parameter has many possible estimators. For a normal population distribution, for example, the center is the mean and the median, since that distribution

is symmetric. So, with sample data, two possible estimators of that center are the sample mean and the sample median.

Estimates are the most common statistical inference reported by the mass media. For example, a Gallup poll in January 2007 reported that 36% of the American public approved of President George W. Bush's performance in office. This is an estimate rather than a parameter, because it was based on interviewing a sample of about 1000 people rather than the entire population.

Unbiased and Efficient Point Estimators

A good estimator has a sampling distribution that (1) is centered around the parameter and (2) has as small a standard error as possible.

An estimator is ***unbiased*** if its sampling distribution centers around the parameter. Specifically, the parameter is the mean of the sampling distribution. From Section 4.4, for random sampling the mean of the sampling distribution of the sample mean $\bar{y}$ equals the population mean μ. Thus, $\bar{y}$ is an unbiased estimator of the population mean μ. Figure 5.1 illustrates. For any particular sample, the sample mean may underestimate μ or may overestimate it. If the sample mean were found repeatedly with different samples, however, the overestimates would tend to counterbalance the underestimates.

By contrast, a ***biased*** estimator tends to underestimate the parameter, on the average, or it tends to overestimate the parameter. For example, the sample range is typically smaller than the population range and it cannot be larger, because the sample minimum and maximum cannot be more extreme than the population minimum and maximum. Thus, the sample range tends to underestimate the population range. It is a biased estimator of the population range.

Suppose a population distribution is skewed to the right, as shown in Figure 5.1, and you want to estimate the population mean. If you are worried about the effects of outliers, you might decide to estimate it using the sample median rather than the sample mean. However, the population median is less than the population mean in that case, and the sample median also tends to be less than the population mean μ. So the sample median is a biased estimator of μ, tending on the average to underestimate μ. It's better to use the sample mean, perhaps calculating it after deleting any extreme outlier if it has undue influence or may reflect an error in recording the data.

A second preferable property for an estimator is a relatively small standard error. An estimator having standard error smaller than those of other estimators is said to be ***efficient***. An efficient estimator falls closer, on the average, than other estimators to

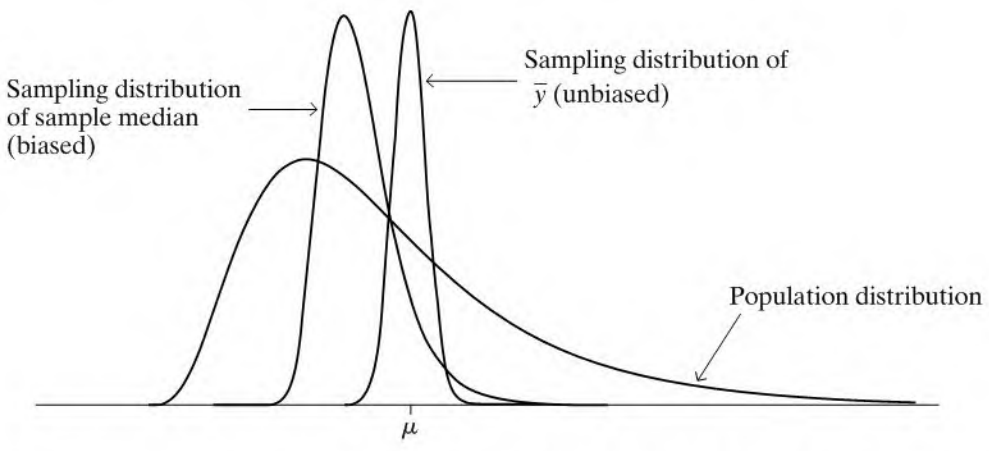

FIGURE 5.1: Sampling Distributions of Two Estimators of the Population Mean, for a Skewed Population Distribution

the parameter. For example, when a population distribution is normal, the standard error of the sample median is 25% larger than the standard error of the sample mean. The sample mean tends to be closer than the sample median to the population center. The sample mean is an efficient estimator. The sample median is inefficient.

In summary, a good estimator of a parameter is *unbiased*, or nearly so, and *efficient*. Statistical methods use estimators that possess these properties.

Estimators of Mean, Standard Deviation, and Proportion

It is common, but not necessary, to use the sample analog of a population parameter as its estimator. For instance, to estimate a population proportion, the sample proportion is an estimator that is unbiased and efficient. For estimating a population mean μ, the sample mean $\bar{y}$ is unbiased. It is efficient for the most common population distributions. Likewise, we use the sample standard deviation s as the estimator of the population standard deviation σ.

The symbol " ^ " over a parameter symbol is often used to represent an estimate of that parameter. The symbol " ^ " is called a *caret* and is usually read as *hat*. For example, $\hat{\mu}$ is read as *mu-hat*. Thus, $\hat{\mu}$ denotes an estimate of the population mean μ.

Maximum Likelihood Method of Estimation*

The most important contributions to modern statistical science were made by a British statistician and geneticist, R. A. Fisher (1890–1962). While working at an agricultural research station north of London, he developed much of the theory of point estimation as well as methodology for the design of experiments and data analysis.

For point estimation, Fisher advocated the ***maximum likelihood estimate***. This estimate is the value of the parameter that is most consistent with the observed data, in the following sense: If the parameter equaled that number (i.e., the value of the estimate), the observed data would have had greater chance of occurring than if the parameter equaled any other number. For instance, a recent survey of about 1000 adult Americans reported that the maximum likelihood estimate of the population proportion who believe in astrology is 0.37. Then the observed sample would have been more likely to occur if the population proportion equals 0.37 than if it equaled any other possible value.

For many population distributions, such as the normal, the maximum likelihood estimator of a population mean is the sample mean. The primary point estimates presented in this book are, under certain population assumptions, maximum likelihood estimates. Fisher showed that, for large samples, maximum likelihood estimators have three desirable properties:

- They are efficient, for relatively large samples: Other estimators do not have smaller standard errors and do not tend to fall closer to the parameter.
- They have little, if any, bias, with the bias diminishing as the sample size increases.
- They have approximately normal sampling distributions.

Confidence Interval Is Point Estimate ± Margin of Error

To be truly informative, an inference about a parameter should provide not only a point estimate but should also indicate how close the estimate is likely to fall to the parameter value. For example, since 1988 each year the Florida Poll conducted by Florida International University (www.fiu.edu/orgs/ipor/ffp) has asked about 1200 Floridians whether sexual relations between two adults of the same sex is wrong.

The percentage saying this is always wrong has decreased from 74% in 1988 to 54% in 2006. How accurate are these estimates? Within 2%? Within 5%? Within 10%?

The information about the precision of a point estimate determines the width of an *interval estimate* of the parameter. This consists of an interval of numbers around the point estimate. It is designed to contain the parameter with some chosen probability close to 1. Because interval estimates contain the parameter with a certain degree of confidence, they are referred to as ***confidence intervals***.

Confidence Interval

A ***confidence interval*** for a parameter is an interval of numbers within which the parameter is believed to fall. The probability that this method produces an interval that contains the parameter is called the ***confidence level***. This is a number chosen to be close to 1, such as 0.95 or 0.99.

The key to constructing a confidence interval is the sampling distribution of the point estimator. Often, the sampling distribution is approximately normal. The normal distribution then determines the probability that the estimator falls within a certain distance of the parameter. With probability about 0.95, the estimator falls within two standard errors. Almost certainly it falls within three standard errors. The smaller the standard error, the more precise the estimator tends to be.

In practice, often the sampling distribution is approximately normal. Then, to construct a confidence interval, we add and subtract from the point estimate some multiple (a z-score) of its standard error. This multiple of the standard error is the ***margin of error***. That is,

A confidence interval has the form: **Point estimate** $\pm$ **Margin of error.**

To construct a confidence interval having "95% confidence," we take the point estimate and add and subtract a margin of error that equals about two standard errors. We'll see the details in the next two sections.

5.2 CONFIDENCE INTERVAL FOR A PROPORTION

For categorical data, an observation occurs in one of a set of categories. This type of measurement occurs when the variable is nominal, such as preferred candidate (Democrat, Republican, Independent), or ordinal, such as opinion about government spending (increase, keep the same, decrease). It also occurs when inherently continuous variables are measured with categorical scales, such as when annual income has categories ($0–$24,999, $25,000–$49,999, $50,000–$74,999, at least $75,000).

To summarize categorical data, we record the *proportions* (or *percentages*) of observations in the categories. For example, a study might provide a point or interval estimate of

- The proportion of Americans who have health insurance
- The proportion of Canadians who favor independent status for Quebec
- The proportion of Australians who are unemployed

The Sample Proportion and Its Standard Error

Let π denote a population proportion.[1] Then π falls between 0 and 1. Its point estimator is the *sample proportion*. We denote the sample proportion by $\hat{\pi}$, since it estimates π.

[1] Here, π is *not* the mathematical constant, 3.1415....

Recall that the sample proportion is a mean when we let $y = 1$ for an observation in the category of interest and $y = 0$ otherwise. (See the discussion about Table 3.6 on page 44 and following Example 4.7 on page 87.) Similarly, the population proportion π is the mean μ of the probability distribution having probabilities

$$P(1) = \pi \quad \text{and} \quad P(0) = 1 - \pi.$$

The standard deviation of this probability distribution is $\sigma = \sqrt{\pi(1 - \pi)}$. (Exercise 4.55 in Chapter 4 derived this formula.) Since the formula for the standard error of a sample mean equals $\sigma_{\bar{y}} = \sigma/\sqrt{n}$, the standard error $\sigma_{\hat{\pi}}$ of the sample proportion $\hat{\pi}$ is

$$\sigma_{\hat{\pi}} = \sigma/\sqrt{n} = \sqrt{\frac{\pi(1 - \pi)}{n}}.$$

As the sample size increases, the standard error gets smaller. The sample proportion then tends to fall closer to the population proportion.

Large-Sample Confidence Interval for a Proportion

Since the sample proportion $\hat{\pi}$ is a sample mean, the Central Limit Theorem applies: For large random samples, the sampling distribution of $\hat{\pi}$ is approximately normal about the parameter π it estimates. Figure 5.2 illustrates.

Recall that 95% of a normal distribution falls within two standard deviations of the mean, or, more precisely, 1.96 standard deviations. We've just seen that the standard error of the sample proportion is $\sigma_{\hat{\pi}} = \sqrt{\pi(1 - \pi)/n}$. So, with probability 0.95, $\hat{\pi}$ falls within $1.96\sigma_{\hat{\pi}}$ units of the parameter π, that is, between $\pi - 1.96\sigma_{\hat{\pi}}$ and $\pi + 1.96\sigma_{\hat{\pi}}$, as Figure 5.2 shows.

Once the sample is selected, if $\hat{\pi}$ does fall within $1.96\sigma_{\hat{\pi}}$ units of π, then the interval from $\hat{\pi} - 1.96\sigma_{\hat{\pi}}$ to $\hat{\pi} + 1.96\sigma_{\hat{\pi}}$ contains π. See line 1 of Figure 5.2. In other words, with probability 0.95 a $\hat{\pi}$ value occurs such that the interval $\hat{\pi} \pm 1.96\sigma_{\hat{\pi}}$ contains the population proportion π.

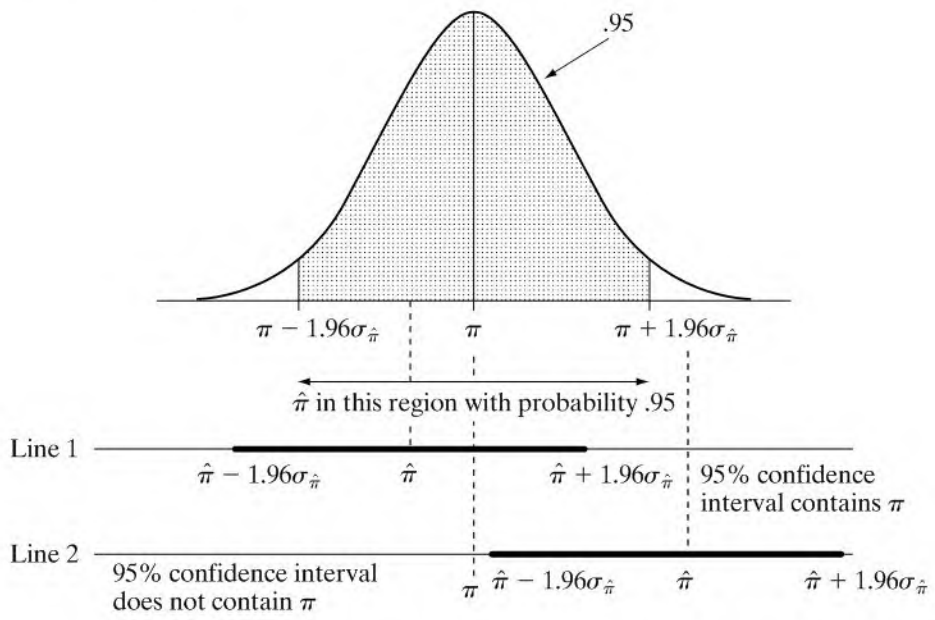

FIGURE 5.2: Sampling Distribution of $\hat{\pi}$ and Possible 95% Confidence Intervals for π

On the other hand, the probability is 0.05 that $\hat{\pi}$ does *not* fall within $1.96\sigma_{\hat{\pi}}$ of π. If that happens, then the interval from $\hat{\pi} - 1.96\sigma_{\hat{\pi}}$ to $\hat{\pi} + 1.96\sigma_{\hat{\pi}}$ does *not* contain π (see Figure 5.2, line 2). Thus, the probability is 0.05 that $\hat{\pi}$ is such that $\hat{\pi} \pm 1.96\sigma_{\hat{\pi}}$ does *not* contain π.

The interval $\hat{\pi} \pm 1.96\sigma_{\hat{\pi}}$ is an interval estimate for π with confidence level 0.95. It is called a ***95% confidence interval***. In practice, the value of the standard error $\sigma_{\hat{\pi}} = \sqrt{\pi(1 - \pi)/n}$ for this formula is unknown, because it depends on the unknown parameter π. So we estimate this standard error by substituting the sample proportion, using

$$se = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}.$$

We've used the symbol s to denote a sample standard deviation, which estimates the population standard deviation σ. In the remainder of this text, we use the symbol *se* to denote a sample estimate of a standard error.

The confidence interval formula uses this estimated standard error. In summary, the 95% confidence interval for π is

$$\hat{\pi} \pm 1.96(se), \quad \text{where} \quad se = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}.$$

EXAMPLE 5.1 Estimating Proportion Who Favor Restricting Legalized Abortion

The 2006 Florida Poll (www.fiu.edu/orgs/ipor/ffp) conducted by Florida International University asked, "In general, do you think it is appropriate for state government to make laws restricting access to abortion?" Of 1200 randomly chosen adult Floridians, 396 said *yes* and 804 said *no*. We shall estimate the population proportion who would respond *yes* to this question.

Let π represent the population proportion of adult Floridians who would respond *yes*. Of the $n = 1200$ respondents, 396 said *yes*, so $\hat{\pi} = 396/1200 = 0.33$. Then $1 - \hat{\pi} = 0.67$. That is, 33% of the sample said *yes* and 67% said *no*.

The estimated standard error of the sample proportion $\hat{\pi}$ equals

$$se = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}} = \sqrt{\frac{(0.33)(0.67)}{1200}} = \sqrt{0.000184} = 0.0136.$$

A 95% confidence interval for π is

$$\hat{\pi} \pm 1.96(se) = 0.33 \pm 1.96(0.0136) = 0.33 \pm 0.03, \quad \text{or} \quad (0.30, 0.36).$$

The population percentage supporting restricting access to abortion appears to be at least 30% but no more than 36%. All numbers in the confidence interval (0.30, 0.36) fall below 0.50. Thus, apparently fewer than half the Florida adult population support restricting access to abortion. ■

Results in such surveys vary greatly depending on the question wording and where the poll is conducted. For instance, when the 2006 GSS asked whether a pregnant woman should be able to obtain a legal abortion if the woman wants it

for any reason (variable ABANY), 1155 said *no* and 784 said *yes*. You can check that the 95% confidence interval for the population proportion saying *no* equals (0.57, 0.62).

EXAMPLE 5.2 Estimating Proportion Who "Oppose" from Proportion Who "Favor"

In the Florida Poll, for estimating the population proportion who supported restricting access to abortion, we obtained $se = 0.0136$ for the point estimate $\hat{\pi} = 0.33$. Similarly, the estimated standard error for $1 - \hat{\pi} = 0.67$, the proportion of voters who say *no* to restricting access to abortion, is

$$se = \sqrt{(1 - \hat{\pi})\hat{\pi}/n} = \sqrt{(0.67)(0.33)/1200} = 0.0136.$$

Both proportions have the same *se*.

A 95% confidence interval for the population proportion of *no* responses to restricting access to abortion is

$$0.67 \pm 1.96(0.0136) = 0.67 \pm 0.03, \quad \text{or} \quad (0.64, 0.70).$$

Now $0.64 = 1 - 0.36$ and $0.70 = 1 - 0.30$, where (0.30, 0.36) is the 95% confidence interval for π. Thus, inferences for the proportion $1 - \pi$ follow directly from those for the proportion π by subtracting each endpoint of the confidence interval from 1.0. ■

If you construct a confidence interval using a hand calculator, don't round off while doing the calculation or your answer may be affected, but do round off when you report the final answer. Likewise, in reporting results from software output, you should use only the first two or three significant digits. Report the confidence interval as (0.30, 0.36) or (0.303, 0.357) rather than (0.303395, 0.356605). Software's extra precision provides accurate calculations in finding *se* and the confidence interval. However, the extra digits are distracting in reports and not useful. They do not tell us anything extra in a practical sense about the population proportion.

Controlling the Confidence Level

With a confidence level of 0.95, that is, "95% confidence," there is a 0.05 probability that the method produces a confidence interval that does *not* contain the parameter value. In some applications, a 5% chance of an incorrect inference is unacceptable. To increase the chance of a correct inference, we use a larger confidence level, such as 0.99.

EXAMPLE 5.3 Finding a 99% Confidence Interval

For the data in Example 5.1 (page 112), let's find a 99% confidence interval for the population proportion who favor laws restricting access to abortion. Now, 99% of a normal distribution occurs within 2.58 standard deviations of the mean. So the probability is 0.99 that the sample proportion $\hat{\pi}$ falls within 2.58 standard errors of the population proportion π. A 99% confidence interval for π is $\hat{\pi} \pm 2.58(se)$.

In Example 5.1, the sample proportion was 0.33, with $se = 0.0136$. So the 99% confidence interval is

$$\hat{\pi} \pm 2.58(se) = 0.33 \pm 2.58(0.0136) = 0.33 \pm 0.04, \quad \text{or} \quad (0.29, 0.37).$$

Compared to the 95% confidence interval of (0.30, 0.36), this interval estimate is less precise, being a bit wider. To be more sure of enclosing the parameter, we must sacrifice precision of estimation by using a wider interval. ■

The general form for the confidence interval for a population proportion π is

$$\hat{\pi} \pm z(se), \quad \text{with} \quad se = \sqrt{\hat{\pi}(1 - \hat{\pi})/n},$$

where z depends on the confidence level. The higher the confidence level, the greater the chance that the confidence interval contains the parameter. High confidence levels are used in practice, so that the chance of error is small. The most common confidence level is 0.95, with 0.99 used when it is more crucial not to make an error.

The z-value multiplied by se is the *margin of error*. With greater confidence, the confidence interval is wider because the z-score in the margin of error is larger—for instance, $z = 1.96$ for 95% confidence and $z = 2.58$ for 99% confidence.

Why do we settle for anything less than 100% confidence? To be absolutely 100% certain of a correct inference, the interval must contain all possible values for π. A 100% confidence interval for the population proportion in favor of limiting access to abortion goes from 0.0 to 1.0. This is not helpful. In practice, we settle for less than perfection in order to estimate much more precisely the parameter value. In forming a confidence interval, we compromise between the desired confidence that the inference is correct and the desired precision of estimation. As one gets better, the other gets worse. This is why you would not typically see a 99.9999% confidence interval. It would usually be too wide to say much about where the population parameter falls (its z-value is 4.9).

Larger Sample Sizes Give Narrower Intervals

We'd expect to be able to estimate a population proportion π more precisely with a larger sample size. The margin of error is $z(se)$, where $se = \sqrt{\hat{\pi}(1 - \hat{\pi})/n}$. The larger the value of n, the smaller the margin of error and the narrower the interval.

To illustrate, suppose that $\hat{\pi} = 0.33$ in Example 5.1 on estimating the proportion who favor restricting legalized abortion was based on $n = 300$, only a fourth as large as the actual sample size of $n = 1200$. Then the estimated standard error of $\hat{\pi}$ is

$$se = \sqrt{\hat{\pi}(1 - \hat{\pi})/n} = \sqrt{(0.33)(0.67)/300} = 0.027,$$

twice as large as the se in Example 5.1. The resulting 95% confidence interval is

$$\hat{\pi} \pm 1.96(se) = 0.33 \pm 1.96(0.027) = 0.33 \pm 0.067.$$

This is twice as wide as the confidence interval formed from the sample of size $n = 1200$.

Since the margin of error is inversely proportional to the square root of n, and since $\sqrt{4n} = 2\sqrt{n}$, the sample size must *quadruple* in order to *double* the precision (i.e., halve the width). Section 5.4 shows how to find the sample size needed to achieve a certain precision.

In summary, this subsection and the previous one showed the following:

The **width of a confidence interval**

1. Increases as the confidence level increases
2. Decreases as the sample size increases

These properties apply to all confidence intervals, not just the one for a proportion.

Error Probability = 1 − Confidence Level

The probability that an interval estimation method yields a confidence interval that does *not* contain the parameter is called the ***error probability***. This equals 1 minus the confidence level. For confidence level 0.95, the error probability equals $1 - 0.95 = 0.05$. The Greek letter α (alpha) denotes the error probability, and $1 - \alpha$ is the confidence level. For an error probability of $\alpha = 0.05$, the confidence level equals $1 - \alpha = 0.95$.

The z-value for the confidence interval is such that the probability is α that $\hat{\pi}$ falls *more than* z standard errors from π. The z-value corresponds to a total probability of α in the two tails of a normal distribution, or $\alpha/2$ (half the error probability) in each tail. For example, for a 95% confidence interval, $\alpha = 0.05$, and the z-score is the one with probability $\alpha/2 = 0.05/2 = 0.025$ in each tail. This is $z = 1.96$.

Confidence Level Is Long-Run Proportion Correct

The confidence level for a confidence interval describes how the method performs when used over and over with many different random samples. The unknown population proportion π is a fixed number. A confidence interval constructed from any particular sample either does or does not contain π. If we repeatedly selected random samples of that size and each time constructed a 95% confidence interval, then in the long run about 95% of the intervals would contain π. This happens because about 95% of the sample proportions would fall within $1.96(se)$ of π, as does the $\hat{\pi}$ in line 1 of Figure 5.2 (page 111). Saying that a particular interval contains π with "95% confidence" signifies that *in the long run* 95% of such intervals would contain π. That is, 95% of the time the inference is correct.

Figure 5.3 shows the results of selecting ten separate samples and calculating the sample proportion for each and a 95% confidence interval for the population proportion. The confidence intervals jump around because $\hat{\pi}$ varies from sample to sample. However, nine of the ten intervals contain the population proportion π. On the average, only about 1 out of 20 times does a 95% confidence interval fail to contain the population parameter.

In practice, we select only *one* sample of some fixed size n and construct *one* confidence interval using the observations in that sample. We do not know whether that confidence interval truly contains π. Our confidence in that interval is based on long-term properties of the procedure. We can control, by our choice of the confidence level, the chance that the interval contains π. If an error probability of 0.05 makes us nervous, we can instead form a 99% confidence interval, for which the method makes an error only 1% of the time.

Large Sample Size Needed for Validity of Method

In practice, the probability that the confidence interval contains π is *approximately* equal to the chosen confidence level. The approximation is better for larger samples.

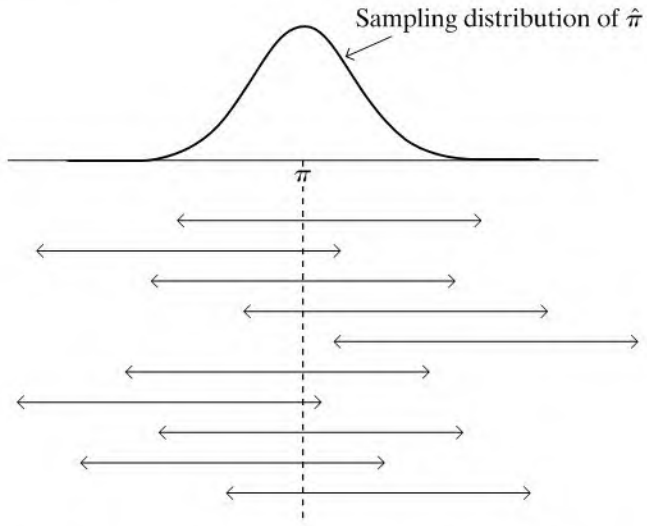

FIGURE 5.3: Ten 95% Confidence Intervals for a Population Proportion π. In the long run, only 5% of the intervals fail to contain π.

As n increases, the sampling distribution of $\hat{\pi}$ is more closely normal in form, by the Central Limit Theorem. This is what allows us to use z-scores from the normal distribution in finding the margin of error. Also as n increases, the *estimated* standard error $se = \sqrt{\hat{\pi}(1 - \hat{\pi})/n}$ gets closer to the *true* standard error $\sigma_{\hat{\pi}} = \sqrt{\pi(1 - \pi)/n}$.

For this reason, the confidence interval formula applies with *large* random samples. How large is "large"? A general guideline states you should have at least 15 observations both in the category of interest and not in it. This is true in most social science studies. In Example 5.1, the counts in the two categories were 396 and 804, so the sample size requirement was easily satisfied. Section 5.4 and Exercise 5.77 show methods that work well when the guideline is not satisfied.

Here is a summary of the confidence interval for a proportion:

Confidence Interval for Population Proportion π

For a random sample, a confidence interval for a population proportion π based on a sample proportion $\hat{\pi}$ is

$$\hat{\pi} \pm z(se), \quad \text{which is} \quad \hat{\pi} \pm z\sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}.$$

The z-value is such that the probability under a normal curve within z standard errors of the mean equals the confidence level. For 95% and 99% confidence intervals, z equals 1.96 and 2.58. The sample size n should be sufficiently large that at least 15 observations are in the category and at least 15 are not in it.

5.3 CONFIDENCE INTERVAL FOR A MEAN

We've learned how to construct a confidence interval for a population proportion for categorical data. We now learn how to construct one for the population mean for quantitative data.

Estimated Standard Error for the Margin of Error

Like the confidence interval for a proportion, the confidence interval for a mean has the form

$$\textbf{Point estimate} \pm \textbf{Margin of error},$$

where the margin of error is a multiple of the standard error. The point estimate of the population mean μ is the sample mean, $\bar{y}$. For large random samples, by the Central Limit Theorem, the sampling distribution of $\bar{y}$ is approximately normal. So, for large samples, we can again find a margin of error by multiplying a z-score from the normal distribution times the standard error.

From Section 4.5, the standard error of the sample mean equals

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

where σ is the population standard deviation. Like the standard error of a sample proportion, this depends on an unknown parameter, in this case σ. In practice, we estimate σ by the sample standard deviation s. So, confidence intervals use the *estimated* standard error

$$se = s/\sqrt{n}.$$

EXAMPLE 5.4 Estimating Mean Number of Sex Partners

Some General Social Surveys ask respondents how many sex partners they have had since their eighteenth birthday. In 2006, when asked how many male sex partners they've had, the 231 females in the sample between the ages of 20 and 29 reported a mean of 4.96. A computer printout summarizes the results for this GSS variable (denoted by NUMMEN):

```
------------------------------------------------------------
Variable    n       Mean     StDev    SE Mean     95.0% CI
NUMMEN      231     4.96     6.81       0.45      (4.1, 5.8)
------------------------------------------------------------
```

How did software get the standard error reported? How do we interpret it and the confidence interval shown?

The sample standard deviation is $s = 6.81$. The sample size is $n = 231$. So the estimated standard error of the sample mean is

$$se = s/\sqrt{n} = 6.81/\sqrt{231} = 0.45.$$

In several random samples of 231 women in this age group, the sample mean number of male sex partners would vary from sample to sample with a standard deviation of about 0.45.

The 95% confidence interval reported of $(4.1, 5.8)$ is an interval estimate of μ, the mean number of male sex partners since the 18th birthday for the population of adult women in the U.S. of age between 20 and 29. We can be 95% confident that this interval contains μ. The point estimate of μ is 5.0, and the interval estimate predicts that μ is no smaller than 4.1 and no greater than 5.8.

This example highlights a couple of things to keep in mind in doing statistical analyses: First, the sample mean of 5.0 and standard deviation of 6.8 suggests that the

distribution of the variable NUMMEN is very highly skewed to the right. The mean may be misleading as a measure of center. In fact, a look at the entire distribution of NUMMEN in 2006 (at the GSS Web site) reveals that the median response was 3, perhaps a more useful summary. It's also worth noting that the mode was 1, with 23% of the sample.

Second, the margin of error in confidence intervals refers only to sampling error. Other potential errors include those due to nonresponse or measurement error (lying or giving an inaccurate response). If such errors are not negligible, the margin of error is actually larger than reported by software using standard statistical formulas.

Finally, as mentioned in Chapters 2 and 4, the GSS uses a multistage design that incorporates cluster sampling. [2] For this design, the estimates are not quite as precise as a simple random sample would provide. For simplicity of exposition, in this text we're acting as if the GSS were a simple random sample. ■

How did software find the margin of error for the confidence interval in the previous example? As with the proportion, for a 95% confidence interval this is roughly two times the estimated standard error. We'll next find the precise margin of error by multiplying *se* by a score that is very similar to a z-score unless n is quite small.

The *t* Distribution

We'll now learn about a confidence interval that applies for *any* random sample size. To achieve this generality, it has the disadvantage of assuming that the population distribution is normal. In that case, the sampling distribution of $\bar{y}$ is normal even for small sample sizes. (The right panel of Figure 4.14 on page 93, which showed sampling distributions for various population distributions, illustrated this.)

Suppose we knew the exact standard error of the sample mean, $\sigma_{\bar{y}} = \sigma/\sqrt{n}$. Then, with the additional assumption that the population is normal, with any n we could use the formula

$$\bar{y} \pm z\sigma_{\bar{y}}, \quad \text{which is} \quad \bar{y} \pm z\sigma/\sqrt{n},$$

for instance, with $z = 1.96$ for 95% confidence. In practice, we don't know the *population* standard deviation σ, so we don't know the *exact* standard error. Substituting the *sample* standard deviation s for σ to get the *estimated* standard error, $se = s/\sqrt{n}$, then introduces extra error. This error can be sizeable when n is small. To account for this increased error, we must replace the z-score by a slightly larger score, called a t-score. The confidence interval is then a bit wider. The t-score is like a z-score, but it comes from a bell-shaped distribution that is slightly more spread out than the standard normal distribution. This distribution is called the ***t distribution***.

Properties of the *t* Distribution

Here are the the main properties of the t distribution:

- The t distribution is bell shaped and symmetric about a mean of 0.
- The standard deviation is a bit larger than 1. The precise value depends on what is called the ***degrees of freedom***, denoted by df. The t distribution has a slightly different spread for each distinct value of df, and different t-scores apply for each df value.

[2] See sda.berkeley.edu/D3/GSS06/Doc/gs06.htm

- For inference about a population mean, the degrees of freedom equal $df = n - 1$, one less than the sample size.
- The t distribution has thicker tails and is more spread out than the standard normal distribution. The larger the df value, however, the more closely it resembles the standard normal. Figure 5.4 illustrates. When df is about 30 or more, the two distributions are nearly identical.

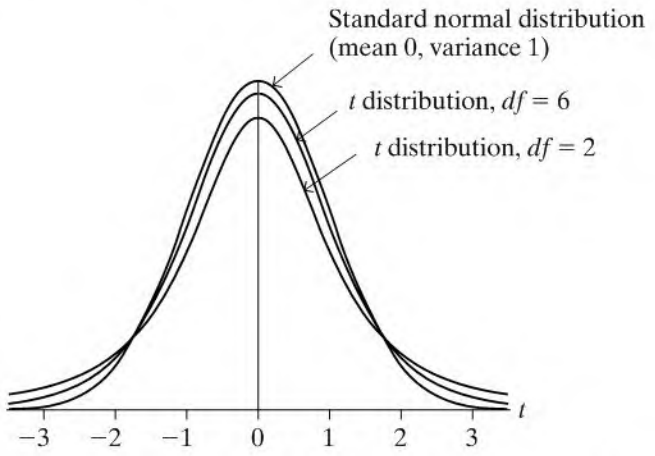

FIGURE 5.4: t Distribution Relative to Standard Normal Distribution. The t gets closer to the normal as the degrees of freedom (df) increase, and the two distributions are practically identical when $df > 30$.

- A t-score multiplied by the estimated standard error gives the margin of error for a confidence interval for the mean.

Table B at the end of the text lists t-scores from the t distribution for various right-tail probabilities. Table 5.1 is an excerpt from that table. The column labelled $t_{.025}$ has probability 0.025 in the right tail and a two-tail probability of 0.05. This is the t-score used in 95% confidence intervals.

To illustrate, when the sample size is 29, the degrees of freedom are $df = n - 1 = 28$. With $df = 28$, we see that $t_{.025} = 2.048$. This means that 2.5% of the t distribution falls in the right tail above 2.048. By symmetry, 2.5% also falls in the left tail below $-t_{.025} = -2.048$. See Figure 5.5. When $df = 28$, the probability equals 0.95 between

TABLE 5.1: Part of Table B Displaying t-Scores. The scores have right-tail probabilities of 0.100, 0.050, 0.025, 0.010, 0.005, and 0.001.

	Confidence Level					
	80%	90%	95%	98%	99%	99.8%
df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.657	318.3
10	1.372	1.812	2.228	2.764	3.169	4.144
28	1.313	1.701	2.048	2.467	2.763	3.408
30	1.310	1.697	2.042	2.457	2.750	3.385
100	1.290	1.660	1.984	2.364	2.626	3.174
infinity	1.282	1.645	1.960	2.326	2.576	3.090

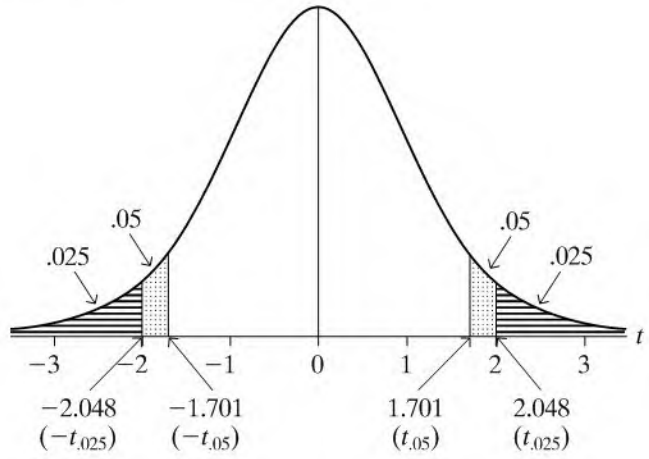

FIGURE 5.5: *t* Distribution with $df = 28$

-2.048 and 2.048. These are the t-scores for a 95% confidence interval when $n = 29$. The confidence interval is $\bar{y} \pm 2.048(se)$.

t-Scores in the Confidence Interval for a Mean

Confidence intervals for a mean resemble those for proportions, except that they use the t distribution instead of the standard normal.

Confidence Interval for Population Mean μ

For a random sample from a normal population distribution, a 95% confidence interval for μ is

$$\bar{y} \pm t_{.025}(se), \quad \text{with} \quad se = s/\sqrt{n},$$

where $df = n - 1$ for the t-score.

Like the confidence interval for a proportion, this confidence interval has margin of error that is a score multiplied by the estimated standard error. The main difference is the substitution of the t-score for the z-score. The t method also makes the assumption of a normal population distribution. This is mainly relevant for small samples. In practice, the population distribution may not be close to normal, and we discuss the importance of this assumption later in the section.

EXAMPLE 5.5 Estimating Mean Weight Change for Anorexic Girls

This example comes from an experimental study that compared various treatments for young girls suffering from anorexia, an eating disorder. For each girl, weight was measured before and after a fixed period of treatment. The variable of interest was the change in weight, that is, weight at the end of the study minus weight at the beginning of the study. The change in weight was positive if the girl gained weight and negative if she lost weight. The treatments were designed to aid weight gain. The weight changes for the 29 girls undergoing the cognitive behavioral treatment were[3]

1.7, 0.7, −0.1, −0.7, −3.5, 14.9, 3.5, 17.1, −7.6, 1.6,
11.7, 6.1, 1.1, −4.0, 20.9, −9.1, 2.1, 1.4, −0.3, −3.7,
−1.4, −0.8, 2.4, 12.6, 1.9, 3.9, 0.1, 15.4, −0.7.

[3]Courtesy of Prof. Brian Everitt, Institute of Psychiatry, London.

Software used to analyze the data reports the summary results:

Variable	Number of Cases	Mean	SD	SE of Mean
CHANGE	29	3.01	7.31	1.36

Thus, $n = 29$ girls received this treatment. Their mean weight change was $\bar{y} = 3.01$ pounds with a standard deviation (SD) of $s = 7.31$. The sample mean had an estimated standard error of $se = s/\sqrt{n} = 7.31/\sqrt{29} = 1.36$ (reported as SE of Mean).

Let μ denote the population mean change in weight for the cognitive behavioral treatment, for the population represented by this sample. If this treatment has a beneficial effect, then μ is positive. Since $n = 29$, $df = n - 1 = 28$. For a 95% confidence interval, we use $t_{.025} = 2.048$. The 95% confidence interval is

$$\bar{y} \pm t_{.025}(se) = 3.01 \pm 2.048(1.36) = 3.0 \pm 2.8, \text{ or } (0.2, 5.8).$$

With 95% confidence, we infer that this interval contains the population mean weight change. It appears that the mean weight change is positive, but it may be rather small in practical terms. However, this experimental study used a volunteer sample, because it is not possible to identify and randomly sample a population of anorexic girls. Because of this, inferences are tentative and "95% confidence" in the results may be overly optimistic. The results are more convincing if researchers can argue that the sample was representative of the population. The study did employ randomization in assigning girls to three therapies (only one of which is considered here), which is reassuring for analyses conducted later in the text comparing therapies.

Another caveat about our conclusion is shown by Figure 5.6, a histogram that software shows for the data. This reveals that the data are skewed to the right. The assumption of a normal population distribution may be violated—more about that later. The median weight change is only 1.4 pounds, somewhat less than the mean of 3.0 because of the skew to the right. The sample median is another indication that the size of the effect could be small. ■

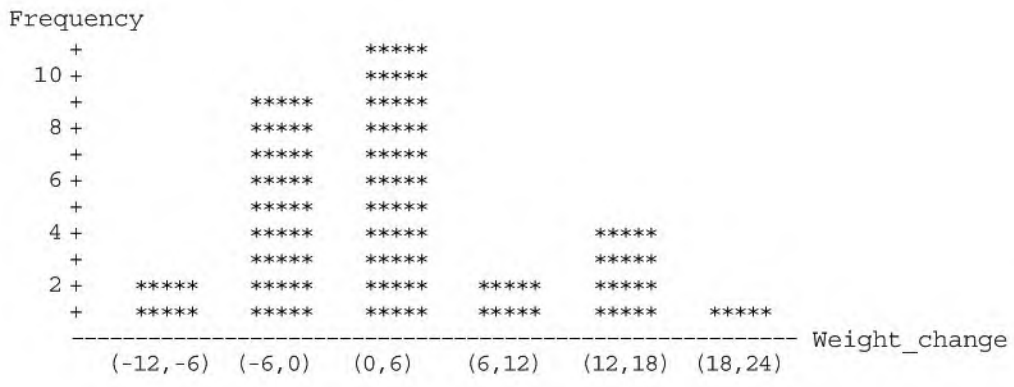

FIGURE 5.6: Histogram of Weight Change Values for Anorexia Study

Effect of Confidence Level and Sample Size

We've used the t distribution to find a 95% confidence interval. Other confidence levels use the same formula but with a different t-score.

Let α denote the error probability. This is the probability that the method yields a confidence interval that does not contain μ. The confidence interval uses the t-score

with tail probability $\alpha/2$ in each tail. For a 99% confidence interval, for instance, $\alpha = 1 - 0.99 = 0.01$, so $\alpha/2 = 0.005$, and the appropriate t-score is $t_{.005}$ for the specified df value.

To be safer in estimating the population mean weight change for the anorexia study in Example 5.5, we could instead use a 99% confidence interval. Since $df = 28$ when $n = 29$, the t-score is $t_{.005} = 2.763$. The standard error does not change. The 99% confidence interval is

$$\bar{y} \pm 2.763(se) = 3.01 \pm 2.763(1.36), \quad \text{which is} \quad (-0.7, 6.8).$$

The confidence interval is wider than the 95% interval of (0.2, 5.8). This is the cost of having greater confidence. The 99% confidence interval contains 0. This tells us it is plausible, at the 99% confidence level, that the population mean change is 0, that is, that the therapy may not result in *any* change in the mean weight.

Like the width of the confidence interval for a proportion, the width of a confidence interval for a mean also depends on the sample size n. Larger sample sizes result in narrower intervals.

Robustness for Violations of Normal Population Assumption

The assumptions for the confidence interval for a mean are (1) that randomization is used for collecting the sample, and (2) that the population distribution is normal. Under the normality assumption, the sampling distribution of $\bar{y}$ is normal even for small n. Likewise, the z-score measuring the number of standard errors that $\bar{y}$ falls from μ then has the standard normal distribution. In practice, when we use the *estimated* standard error $se = s/\sqrt{n}$ (rather than the true one, $\sigma/\sqrt{n}$), the number of se that $\bar{y}$ falls from μ has the t distribution.

The confidence interval in Example 5.5 estimated the mean weight change for anorexic girls. The histogram of the weight change data shown above is not precise when n is as small as in that example ($n = 29$), but it showed evidence of skew. Generally, the normal population assumption seems worrisome, because many variables in the social sciences have distributions that are far from normal.

A statistical method is said to be ***robust*** with respect to a particular assumption if it performs adequately even when that assumption is violated. Statisticians have shown that the confidence interval for a mean using the t distribution is robust against violations of the normal population assumption. Even if the population is not normal, confidence intervals based on the t distribution still work quite well, especially when n exceeds about 15. As the sample size gets larger, the normal population assumption becomes less important because of the Central Limit Theorem. The sampling distribution of the sample mean is then bell shaped even when the population distribution is not. The actual probability that the 95% confidence interval method provides a correct inference is close to 0.95 and gets closer as n increases.

An important case when the method does not work well is when the data are extremely skewed or contain extreme outliers. Partly this is because of the effect on the method, but also because the mean itself may not then be a representative summary of the center. For this reason, the confidence interval about number of sex partners reported in Example 5.4 (page 117) has limited usefulness.

In practice, assumptions are rarely perfectly satisfied. Thus, knowing whether a statistical method is robust when a particular assumption is violated is important. The t confidence interval method is *not* robust to violations of the randomization assumption. The t method, like all inferential statistical methods, has questionable validity if the method for producing the data did not use randomization.

Standard Normal Is *t* Distribution with *df* = Infinity

Look at the table of t-scores (Table B in the Appendix), part of which was shown in Table 5.1. As df increases, you move down the table. The t-score decreases and gets closer and closer to the z-score for a standard normal distribution. This reflects the t distribution becoming less spread out and more similar in appearance to the standard normal distribution as df increases. You can think of the standard normal distribution as a t distribution with $df = \infty$ (infinity).

For instance, when df increases from 1 to 100 in Table 5.1, the t-score $t_{.025}$ with right-tail probability equal to 0.025 decreases from 12.706 to 1.984. The z-score with right-tail probability of 0.025 for the standard normal distribution is $z = 1.96$. The t-scores are not printed for $df > 100$, but they are close to the z-scores. The last row of Table 5.1 and Table B lists the z-scores for various confidence levels, opposite $df = \infty$.

You can get t-scores for any df value using software and many calculators, so you are not restricted to Table B. For df values larger than shown in Table B (above 100), you can use a z-score to approximate the t-score. For a 95% confidence interval you will then use

$$\bar{y} \pm 1.96(se) \quad \text{instead of} \quad \bar{y} \pm t_{.025}(se).$$

You will not get *exactly* the same result that software would give, but it will be close enough for practical purposes. For instance, to get the confidence interval for the mean number of sex partners in Example 5.4, for which the GSS sample had $n = 231$, software uses the t-score for $df = 231 - 1 = 230$, which is 1.97. This is very close to the z-score of 1.96 from the standard normal distribution.

Why does the t distribution look more like the standard normal distribution as n (and hence df) increases? Because s is increasingly precise as a point estimate of σ in approximating the true standard error $\sigma/\sqrt{n}$ by $se = s/\sqrt{n}$. The additional sampling error for small samples results in the t sampling distribution being more spread out than the standard normal.

The t distribution has just celebrated its 100th anniversary. It was discovered in 1908 by the statistician and chemist W. S. Gosset. At the time, Gosset was employed by Guinness Breweries in Dublin, Ireland, designing experiments pertaining to the selection, cultivation, and treatment of barley and hops for the brewing process. Due to company policy forbidding the publishing of trade secrets, Gosset used the pseudonym *Student* in articles he wrote about his discovery. The t distribution became known as *Student's t*, a name still sometimes used today. Confidence intervals were not introduced, however, until a series of articles by Jerzy Neyman and Egon Pearson beginning in 1928.

A Caveat about Using Software

The examples in this section used output from statistical software to help us analyze data. We'll do this increasingly in future chapters as we cover methods that require substantial computation. You should use software yourself for some of the exercises to get a feel for how researchers analyze data in practice. Any particular software has a huge number of options and it's easy to *misuse* it. Just because results appear on an output window does not mean they are the correct ones or that the assumptions were sufficiently met to do that analysis.

When you start to use software for a given method, we suggest that you first use it for the example of that method in this book. Note whether you get the same results, as a way to check whether you are using the software correctly.

5.4 CHOICE OF SAMPLE SIZE

Polling organizations such as the Gallup poll take samples that typically contain about a thousand subjects. This is large enough for a sample proportion estimate to have a margin of error of about 0.03. At first glance, it seems astonishing that a sample of this size from a population of perhaps many millions is adequate for predicting outcomes of elections, summarizing opinions on controversial issues, showing relative sizes of television audiences, and so forth.

Recall that the margin of error for a confidence interval depends on the *standard error* of the point estimate. Thus, the basis for this inferential power lies in the formulas for the standard errors. As long as the sampling is properly executed, good estimates result from relatively small samples, no matter how large the population size (in fact, the methods actually treat the population size as infinite; see Exercise 57 in Chapter 4). Polling organizations use sampling methods that are more complex than simple random samples, often involving some clustering and/or stratification. However, the standard errors under their sampling plans are approximated reasonably well either by the formulas for simple random samples or by inflating those formulas by a certain factor (such as by 25%) to reflect the sample design effect.

Before data collection begins, most studies attempt to determine the sample size that will provide a certain degree of precision in estimation. A relevant measure is the value of n for which a confidence interval for the parameter has margin of error equal to some specified value. This section shows how they do this. The key results for finding the sample size are as follows:

- The *margin of error* depends directly on the *standard error* of the sampling distribution of the point estimator.
- The *standard error* itself depends on the *sample size*.

Sample Size for Estimating Proportions

To determine the sample size, we must decide on the margin of error desired. In some studies, highly precise estimation is not as important as in others. An exit poll in a close election requires a precise estimate to predict the winner. If, on the other hand, the goal is to estimate the proportion of residents of Syracuse, New York, who have health insurance, a larger margin of error might be acceptable. So we must first decide whether the margin of error should be about 0.03 (three percentage points), 0.05, or whatever.

We must also specify the *probability* with which the margin of error is achieved. For example, we might decide that the error in estimating a population proportion should not exceed 0.04, with 0.95 probability. This probability must be stated, since with any sample size the error is no more than 0.04 with *some* probability, though perhaps a small one.

The next example illustrates sample size determination for estimating a population proportion.

EXAMPLE 5.6 Sample Size for a Survey on Single-Parent Children

A social scientist wanted to estimate the proportion of school children in Boston who live with only one parent. Since her report was to be published, she wanted a reasonably precise estimate. However, since her funding was limited, she did not want to collect a larger sample than necessary. She decided to use a sample size such that, with probability 0.95, the error would not exceed 0.04. So she needed to determine n such that a 95% confidence interval for π equals $\hat{\pi} \pm 0.04$.

Since the sampling distribution of the sample proportion $\hat{\pi}$ is approximately normal, $\hat{\pi}$ falls within 1.96 standard errors of π with probability 0.95. Thus, if the sample size is such that 1.96 standard errors equals 0.04, then with probability 0.95, $\hat{\pi}$ falls within 0.04 units of π. See Figure 5.7.

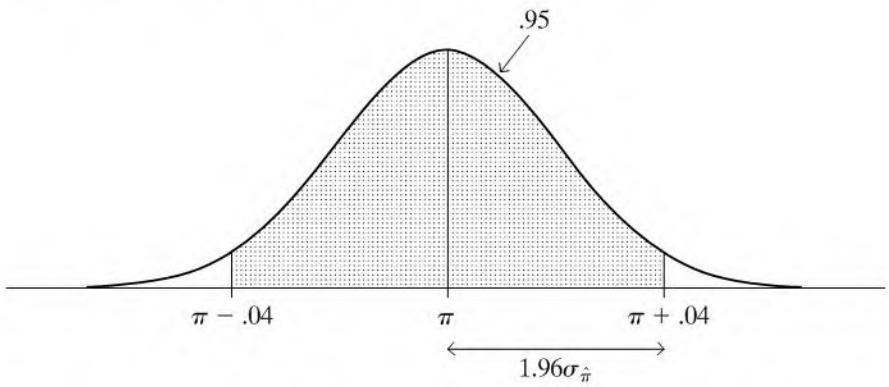

FIGURE 5.7: Sampling Distribution of $\hat{\pi}$ with the Error of Estimation No Greater than 0.04, with Probability 0.95

Recall that the true standard error is $\sigma_{\hat{\pi}} = \sqrt{\pi(1 - \pi)/n}$. How do we find the value of n that provides a value of $\sigma_{\hat{\pi}}$ for which $0.04 = 1.96\sigma_{\hat{\pi}}$? We must solve for n in the expression

$$0.04 = 1.96\sqrt{\frac{\pi(1 - \pi)}{n}}.$$

Multiplying both sides of the expression by $\sqrt{n}$ and dividing both sides by 0.04, we get

$$\sqrt{n} = \frac{1.96\sqrt{\pi(1 - \pi)}}{0.04}.$$

Squaring both sides, we obtain the result

$$n = \frac{(1.96)^2\pi(1 - \pi)}{(0.04)^2}.$$

Now we face a problem. We want to select n for the purpose of estimating the population proportion π, but this formula requires the value of π. This is because the spread of the sampling distribution depends on π. The distribution is less spread out, and it is easier to estimate π, if π is close to 0 or 1 than if it is near 0.50. Since π is unknown, we must substitute an educated guess for it in this equation to solve for n.

The largest possible value for $\pi(1 - \pi)$ occurs when $\pi = 0.50$. Then $\pi(1 - \pi) = 0.25$. In fact, $\pi(1 - \pi)$ is fairly close to 0.25 unless π is quite far from 0.50. For example, $\pi(1 - \pi) = 0.24$ when $\pi = 0.40$ or $\pi = 0.60$, and $\pi(1 - \pi) = 0.21$ when $\pi = 0.70$ or $\pi = 0.30$. Thus, another possible approach is to substitute $\pi = 0.50$ in the above equation. This yields

$$n = \frac{(1.96)^2\pi(1 - \pi)}{(0.04)^2} = \frac{(1.96)^2(0.50)(0.50)}{(0.04)^2} = 600.$$

This approach ensures that with confidence level 0.95, the margin of error will not exceed 0.04, no matter what the value of π. ■

Obtaining n by setting $\pi = 0.50$ is the "safe" approach. But this n value is excessively large if π is not near 0.50. Suppose that based on other studies the social scientist believed that π was no higher than 0.25. Then an adequate sample size is

$$n = \frac{(1.96)^2\pi(1 - \pi)}{(0.04)^2} = \frac{(1.96)^2(0.25)(0.75)}{(0.04)^2} = 450.$$

A sample size of 600 is larger than needed. With it, the margin of error for a 95% confidence interval would be less than 0.04.

Sample Size Formula for Estimating Proportions

We next provide a general formula for determining the sample size. Let M denote the desired margin of error. The formula also uses a general z-score (in place of 1.96) determined by the probability with which the error is no greater than M.

Sample Size for Estimating a Population Proportion π

The random sample size n having margin of error M in estimating π by the sample proportion $\hat{\pi}$ is

$$n = \pi(1 - \pi)\left(\frac{z}{M}\right)^2.$$

The z-score is the one for a confidence interval with the desired confidence level, such as $z = 1.96$ for level 0.95. You need to guess π or take the safe approach of setting $\pi = 0.50$.

To illustrate, suppose the study about single-parent children wanted to estimate the population proportion to within 0.08 with a probability of at least 0.95. Then the margin of error equals $M = 0.08$, and $z = 1.96$. The required sample size using the safe approach is

$$n = \pi(1 - \pi)\left(\frac{z}{M}\right)^2 = (0.50)(0.50)\left(\frac{1.96}{0.08}\right)^2 = 150.$$

This sample size of 150 is one-fourth the sample size of 600 necessary to guarantee a margin of error no greater than $M = 0.04$. Reducing the margin of error by a factor of one-half requires quadrupling the sample size.

Sample Size for Estimating Means

An analogous result holds for estimating a population mean μ. We want to determine how large n needs to be so that the sampling distribution of $\bar{y}$ has margin of error M. Figure 5.8 illustrates. It shows how the sampling distribution gets narrower as n increases until, at the required n, 95% falls within the chosen margin of error. A derivation using the large-sample normal sampling distribution of $\bar{y}$ yields the following result:

Sample Size for Estimating a Population Mean μ

The random sample size n having margin of error M in estimating μ by the sample mean y is

$$n = \sigma^2\left(\frac{z}{M}\right)^2.$$

The z-score is the one for a confidence interval with the desired confidence level, such as $z = 1.96$ for level 0.95. You need to guess the population standard deviation σ.

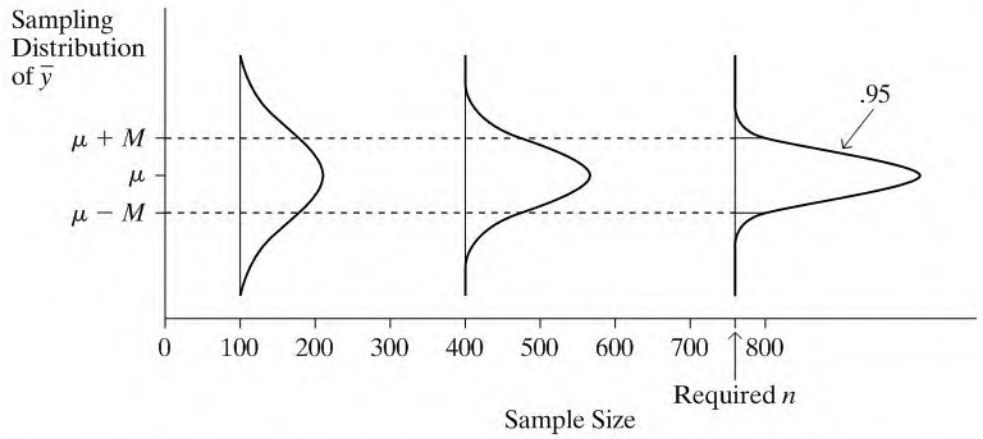

FIGURE 5.8: Determining n So That y Has Probability 0.95 of Falling within a Margin of Error of M Units of the Population Mean μ

The greater the spread of the population distribution, as measured by its standard deviation σ, the larger the sample size needed to achieve a certain margin of error. If subjects have little variation (i.e., σ is small), we need less data than if they are highly heterogenous. In practice, σ is unknown. We need to substitute an educated guess for it, perhaps based on a previous study.

A slight complication is that since we don't know σ, for inference we actually use the t distribution rather than the standard normal. But if we don't know n, we also don't know the degrees of freedom and the t-score. We saw in Table B, however, that unless df is small, the t-score is close to the z-score. So we won't worry about this complication. The approximation of replacing an unknown t-score in the sample size formula by a z-score is usually much less than that involved in having to use an educated guess for σ.

EXAMPLE 5.7 Estimating Mean Education of Native Americans

A study is planned of elderly Native Americans. Variables to be studied include educational level. How large a sample size is needed to estimate the mean number of years of attained education correct to within 1 year with probability 0.99?

Suppose the study has no prior information about the standard deviation of educational attainment for Native Americans. As a guess, perhaps nearly all values of this variable fall within a range of about 15 years, such as between 5 and 20 years. If this distribution is approximately normal, then since the range from $\mu - 3\sigma$ to $\mu + 3\sigma$ contains nearly all of a normal distribution, the range of 15 would equal about 6σ. Then, $15/6 = 2.5$ is a guess for σ.

Now, for 99% confidence, the error probability is 0.01. The z-score is the one with probability $0.01/2 = 0.005$ in each tail, which is 2.58. Since the desired margin of error is $M = 1$ year, the required sample size is

$$n = \sigma^2 \left(\frac{z}{M}\right)^2 = (2.5)^2 \left(\frac{2.58}{1}\right)^2 = 42.$$

A more cautious approach would select a larger value for σ. For example, if the range from 5 to 20 years encloses only about 95% of the education values, we could treat this as the range from $\mu - 2\sigma$ to $\mu + 2\sigma$ and set $15 = 4\sigma$. Then $\sigma = 15/4 = 3.75$ and $n = (3.75)^2(2.58/1)^2 = 94$. If σ is actually less than 3.75, the margin of error of a 99% confidence interval with $n = 94$ observations will be even less than 1. ■

These sample size formulas apply to simple and systematic random sampling. Cluster samples and complex multistage samples must usually be larger to achieve the same precision, whereas stratified samples can often be smaller. In such cases, you should seek guidance from a statistical consultant.

Other Considerations in Determining Sample Size

We have seen that the necessary sample size depends on the desired *precision* and *confidence*. Precision refers to the margin of error. Confidence refers to the probability that the confidence interval will contain the parameter. We've also seen that sample size depends on the *variability* in the population. For estimating means, the required sample size increases as σ increases. In most social surveys, large samples (1000 or more) are necessary, but for homogeneous populations (e.g., residents of nursing homes), smaller samples are often adequate due to reduced population variability.

From a practical point of view, other considerations also affect the sample size. One consideration is the *complexity of analysis* planned. The more complex the analysis, such as the more variables analyzed simultaneously, the larger the sample needed. To analyze a single variable using a mean, a relatively small sample might be adequate. Planned comparisons of several groups using complex multivariate methods, however, require a larger sample. For instance, Example 5.7 showed we may be able to estimate mean educational attainment quite well with only 42 people. But if we also wanted to compare the mean for several ethnic and racial groups and study how the mean depends on other variables such as gender, parents' income and education, and size of the community, a larger sample would be needed.

Another consideration concerns time, money, and other *resources*. Larger samples are more expensive and more time consuming. They may require greater resources than are available. For example, sample size formulas might suggest that 1000 cases provide the desired precision. Perhaps you can afford to gather only 400. Should you go ahead with the smaller sample and sacrifice precision and/or confidence, or should you give up unless you find additional resources? You may need to answer questions such as, "Is it really crucial to study all groups, or can I reduce the sample by focusing on a couple of groups?"

In summary, no simple formula can always give an appropriate sample size. While sample size is an important matter, its choice depends on resources and the analyses planned. This requires careful judgment. A final caveat: If the study is carried out poorly, or if data are never obtained for a substantial percentage of the sample, or if some subjects lie, or if some observations are incorrectly recorded by the data collector or by the statistical analyst, then the actual probability of accuracy to within the specified margin of error may be much less than intended. When someone claims to achieve a certain precision and confidence, be skeptical unless you know that the study was substantially free of such problems.

What If You Have Only a Small Sample?*

Sometimes, because of financial or ethical factors, it's just not possible to take as large a sample as we'd like. If n must be small, how does that affect the validity of confidence interval methods? The t methods for a mean can be used with any n. When n is small, though, you need to be cautious to look for extreme outliers or great departures from the normal population assumption (such as implied by highly skewed data). These can affect the results and the validity of using the mean as a summary of center.

Recall that the confidence interval formula for a proportion requires at least 15 observations of each type. Otherwise, the sampling distribution of the sample proportion need not be close to normal, and the estimate $se = \sqrt{\hat{\pi}(1 - \hat{\pi})/n}$ of the true standard error $\sqrt{\pi(1 - \pi)/n}$ may be poor. As a result, the confidence interval formula works poorly, as the next example shows.

EXAMPLE 5.8 What Proportion of Students Are Vegetarians?

For a class project, a student randomly sampled 20 fellow students at the University of Florida to estimate the proportion at that university who were vegetarians. Of the 20 students she sampled, none were vegetarians. Let π denote the population proportion of vegetarians at the university. The sample proportion was $\hat{\pi} = 0/20 = 0.0$.

When $\hat{\pi} = 0.0$, then $se = \sqrt{\hat{\pi}(1 - \hat{\pi})/n} = \sqrt{(0.0)(1.0)/20} = 0.0$. The 95% confidence interval for the population proportion of vegetarians is

$$\hat{\pi} \pm 1.96(se) = 0.0 \pm 1.96(0.0), \quad \text{which is} \quad 0.0 \pm 0.0, \quad \text{or} \quad (0.0, 0.0).$$

The student concluded she could be 95% confident that π falls between 0 and 0. But this confidence interval formula is valid only if the sample has at least 15 vegetarians and at least 15 nonvegetarians. The sample did not have at least 15 vegetarians, so the method is not appropriate. ■

For small samples, the confidence interval formula is still valid if we use it after adding 4 artificial observations, 2 of each type. The sample of size $n = 20$ in Example 5.8 had 0 vegetarians and 20 nonvegetarians. We can apply the confidence interval formula with $0 + 2 = 2$ vegetarians and $20 + 2 = 22$ nonvegetarians. The value of the sample size for the formula is then $n = 24$. Applying the formula, we get

$$\hat{\pi} = 2/24 = 0.083, \; se = \sqrt{\hat{\pi}(1 - \hat{\pi})/n} = \sqrt{(0.083)(0.917)/24} = 0.056.$$

The resulting 95% confidence interval is

$$\hat{\pi} \pm 1.96(se), \quad \text{which is} \quad 0.083 \pm 1.96(0.056), \quad \text{or} \quad (-0.03, 0.19).$$

A proportion cannot be negative, so we report the interval as (0.0, 0.19). We can be 95% confident that the proportion of vegetarians at the University of Florida is no greater than 0.19.

Why do we add 2 to the counts of the two types? The reason is that the confidence interval then approximates one based on a more complex method (described in Exercise 5.77) that does not require estimating the standard error.[4]

5.5 CONFIDENCE INTERVALS FOR MEDIAN AND OTHER PARAMETERS*

We've focused so far on estimating means and proportions. Chapter 3 showed, though, that other statistics are also useful for describing data. These other statistics also have sampling distributions. For large random samples, their sampling distributions are also approximately normal and are the basis of confidence intervals for population measures. We illustrate in this section for the median.

[4]See article by A. Agresti and B. Coull (who proposed this confidence interval), *American Statistician*, vol. 52, 1998, pp. 119–126.

Inefficiency of the Sample Median for Normal Data

When the population distribution is normal and the sample is random, the standard error of the sample median has formula similar to the one for the sample mean. The standard error equals $1.25\sigma/\sqrt{n}$.

The population median for a normal distribution equals the population mean μ. So the sample median and sample mean are both point estimates of the same number. The sample median is not as efficient as the sample mean because its standard error is 25% larger. When the population distribution is approximately normal, the sample mean is a better estimator of the center of that distribution. This is one reason the mean is more commonly used than the median in statistical inference.

When the population distribution is highly skewed, the population median is often a more useful summary than the population mean. We use the sample median to estimate the population median. However, the standard error formula $1.25\sigma/\sqrt{n}$ is valid only when the population distribution is approximately normal.

Large-Sample Confidence Interval for Median

The confidence interval for the median discussed next is valid for large samples (n at least about 20–30). It requires no assumption about the population distribution other than it is essentially continuous. Its logic utilizes ideas of this chapter.

By definition, the probability π that a randomly selected observation falls below the median is 0.50. So, for a random sample of size n, the sample proportion $\hat{\pi}$ falling below the median has mean 0.50 and standard error $\sigma_{\hat{\pi}} = \sqrt{\pi(1 - \pi)/n} = \sqrt{0.50(0.50)/n} = 0.50/\sqrt{n}$. In particular, the probability is about 0.95 that the sample proportion of observations falling below the median is within two standard errors, or $1/\sqrt{n}$, of 0.50. The sample *number* of observations falling below the median is n times the sample proportion. So the probability is about 0.95 that the number of observations falling below the median is within $n(1/\sqrt{n}) = \sqrt{n}$ of half the sample and the number of observations falling *above* the median is within $n(1/\sqrt{n}) = \sqrt{n}$ of half the sample.

Now, for an ordered sample of size n, the median is the middle measurement, which has index $(n + 1)/2$. The observation with index $(n + 1)/2 - \sqrt{n}$ is the lower endpoint of a 95% confidence interval for the median. The observation with index $(n + 1)/2 + \sqrt{n}$ is the upper endpoint.

EXAMPLE 5.9 Estimating Median Shelf Time in a Library

A librarian at the University of Florida wanted to estimate various characteristics of books in one of the university's special collections. Among the questions of interest were, "How old is a typical book in the collection?" and "How long has it been since a typical book has been checked out?" We suspected that the distributions of such variables might be heavily skewed to the right. So we used the median to describe the center.

Table 5.2 shows data on P = number of years since publication of book and C = number of years since book checked out, for a systematic random sample of 54 books from the collection. Figure 5.9 shows a SPSS box plot for the P values. The five starred values represent extreme outliers falling more than 3.0 IQR above the upper quartile. The sample median, which is 17, is more representative of the data than the sample mean of 22.6. Let's construct a 95% confidence interval for the population median of the distribution of P.

TABLE 5.2: Number of Years since Publication (*P*) and Number of Years since Checked Out (*C*) for 54 Books

C	*P*	*C*	*P*	*C*	*P*	*C*	*P*	*C*	*P*
1	3	9	9	4	4	1	18	1	5
30	30	0	17	2	7	0	12	1	13
7	19	5	5	47	47	3	15	9	17
11	140	2	19	5	8	2	10	11	18
1	5	1	22	1	11	5	19	2	3
2	97	0	10	1	21	7	7	4	19
4	4	11	11	5	20	14	14	5	43
2	19	10	10	10	10	0	18	10	17
4	13	17	71	8	19	0	17	48	48
2	19	11	11	6	6	7	20	4	4
92	92	4	44	1	5	1	54		

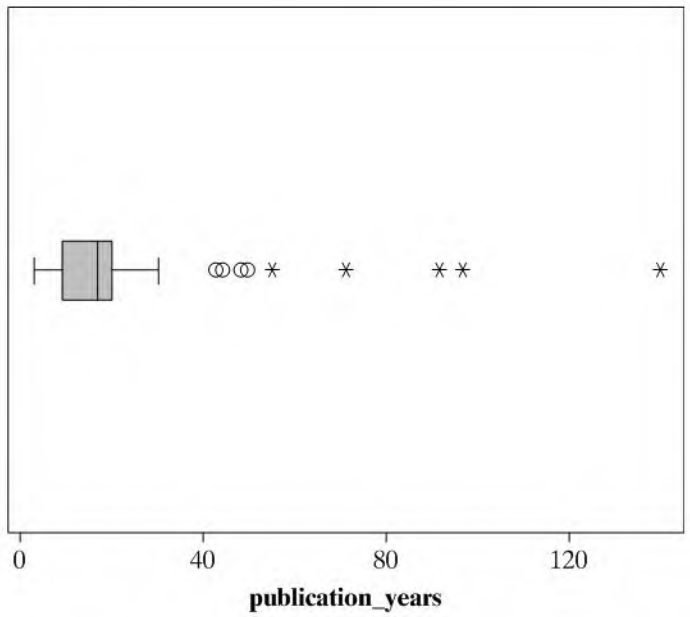

FIGURE 5.9: Box Plot for Number of Years since Publication for Sample of Library Books

For $n = 54$, the endpoints of a 95% confidence interval have indices

$$\frac{n+1}{2} \pm \sqrt{n} = \frac{54+1}{2} \pm \sqrt{54} = 27.5 \pm 7.3, \text{ or } (20.2, 34.8).$$

The confidence interval consists of the 20th smallest and 35th smallest (20th largest) values of the variable.

For a small sample such as this, it is simple to identify ordered values from a stem-and-leaf plot. Table 5.3 shows the part of this plot for the smallest 44 of the 54 observations, splitting stems into two. The 20th smallest observation equals 11 and the 35th smallest observation equals 19. The 95% confidence interval equals (11, 19). We can be 95% confident that the median time since publication is at least

TABLE 5.3: Lower Part of Stem-and-Leaf Plot for Number of Years since Publication. This does not show the long right tail of the distribution, which is not needed to find the confidence interval for the median.

Stem	Leaf
0	3 3 4 4 4
0	5 5 5 5 6 7 7 8 9
1	0 0 0 0 1 1 1 2 3 3 4
1	5 7 7 7 7 8 8 8 9 9 9 9 9 9 9
2	0 0 1 2

11 years and no greater than 19 years. To get a narrower interval, we need a larger sample size. ■

The Bootstrap

For some parameters, it is not possible to write down a confidence interval formula that works well regardless of the population distribution or sample size or sampling method. For such cases, a recent computational invention called the ***bootstrap*** is useful. This method treats the sample distribution as if it were the true population distribution. You sample n observations from this distribution, where each of the original n data points has probability $1/n$ of selection for each "new" observation. For this "new" sample of size n, you then construct the point estimate of the parameter. You repeat this sampling process a large number of times, for instance selecting 1000 separate samples of size n.

The generated sampling distribution of the point estimate values provides information about the true parameter. With the *percentile* method of bootstrapping, the 95% confidence interval for the parameter is the 95% central set of estimate values. These are the ones falling between the 2.5th percentile and 97.5th percentile of the generated sampling distribution. This is a computationally intensive process, but easily feasible with modern computing power.

5.6 CHAPTER SUMMARY

This chapter presented methods of estimation, focusing on the population mean μ for quantitative variables and the population proportion π for categorical variables.

- A ***point estimate*** is the best single guess for the parameter value. The point estimates of the population mean μ, standard deviation σ, and proportion π are the sample values, $\bar{y}$, s, and $\hat{\pi}$.
- An ***interval estimate***, called a ***confidence interval***, is an interval of numbers within which the parameter is believed to fall. Confidence intervals for a population mean μ and for a population proportion π have the form

$$\text{Point estimate} \pm \text{Margin of error},$$
$$\text{with} \quad \text{Margin of error} = \text{score} \times (se),$$

where se is the estimated standard error. The score multiplied by se is a z-score from the normal distribution for confidence intervals for proportions and a t-score from the t distribution for confidence intervals for a mean.

- The probability that the method yields an interval that contains the parameter is called the ***confidence level***. This is controlled by the choice of the z- or t-score in the margin of error. Increasing the confidence level entails the use of a larger score and, hence, the sacrifice of a wider interval.
- The ***t distribution*** looks like the standard normal distribution, having a mean of 0 but being a bit more spread out. Its spread is determined by the ***degrees of freedom***, which equal $n - 1$ for inference about a mean.
- The width of a confidence interval also depends on the standard error of the sampling distribution of the point estimate. Larger sample sizes produce smaller standard errors and narrower confidence intervals and, hence, more precise estimates.

Confidence intervals assume random sampling. For large samples, they do not need an assumption about the population distribution because the sampling distribution is roughly normal even if the population is highly nonnormal, by the Central Limit Theorem. Table 5.4 summarizes estimation methods.

TABLE 5.4: Summary of Estimation Methods for Means and Proportions

Parameter	Point Estimate	Estimated Standard Error	Confidence Interval	Sample Size to Estimate to Within M
Mean μ	$\bar{y}$	$se = \frac{s}{\sqrt{n}}$	$\bar{y} \pm t(se)$	$n = \sigma^2\left(\frac{z}{M}\right)^2$
Proportion π	$\hat{\pi}$	$se = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$	$\hat{\pi} \pm z(se)$	$n = \pi(1 - \pi)\left(\frac{z}{M}\right)^2$

Note: $z = 1.96$ for 95% confidence; for error probability α and confidence level $(1 - \alpha)$, z-score or t-score has right-tail probability $\alpha/2$ (e.g., $\alpha/2 = 0.025$ for 95% confidence).

Table 5.4 also shows formulas for the sample size needed to achieve a desired margin of error M. You must select M and the confidence level, which determines the z-score. Also, you must substitute a guess for the population standard deviation σ to determine the sample size for estimating a population mean μ. You must substitute a guess for the population proportion π to determine the sample size for estimating π. Substituting $\pi = 0.50$ guarantees that the sample size is large enough to give the desired precision and confidence.

PROBLEMS

Practicing the Basics

5.1. Of 577,006 people involved in motor vehicle accidents in Florida in a recent year, 412,878 were wearing seat belts (*Source*: Florida Department of Highway Safety and Motor Vehicles). Find a point estimate of the population proportion of Florida motorists wearing seat belts.

5.2. In response to the GSS question in 2006 about the number of hours daily spent watching TV, the responses by the seven subjects who identified themselves as Hindus were 2, 3, 2, 1, 0, 1, 4, 3.

(a) Find a point estimate of the population mean for Hindus.

(b) The margin of error for this point estimate is 1.1. Explain what this represents.

5.3. An Associated Press story (March 8, 2006) about a survey commissioned by the American Medical Association of a nationwide random sample of 644 college women or graduates ages 17 to 35 estimated that a proportion of 0.74 of women on Spring Break use drinking as an excuse for outrageous behavior, including public nudity and dancing on tables. Find the standard error of this estimate, and interpret.

5.4. A national survey conducted in July 2006 by Pew Forum on Religion & Public Life asked whether the subject favored allowing homosexual couples

to enter into civil unions—legal agreements that would give them many of the same rights as married couples. Of 2003 adults interviewed, 54% said *yes*, 42% said *no*, and 4% had no opinion. Find the estimated standard error for the sample proportion answering *yes*. Interpret it.

5.5. When polled in 2006 and asked whether Turkey should be included in the European Union if it met all conditions set by the EU, the percentage who said *yes* was 51% in Denmark ($n = 1008$) and 42% in the UK ($n = 1312$). For the Denmark result, the report stated that the margin of error is plus or minus 3.1%. Explain how they got this result.

5.6. One question (coded EQUALIZE) on a recent GSS asked, "Do you think that it should be government's responsibility to reduce income differences between the rich and the poor?" Those answering *yes* included 90 of the 142 subjects who called themselves "strong Democrat" in political party identification and 26 of the 102 who called themselves "strong Republican."

(a) Find the point estimate of the population proportion who would answer *yes* for each group.

(b) The 95% confidence interval for the population proportion of *yes* responses is (0.55, 0.71) for strong Democrats and (0.17, 0.34) for strong Republicans. Explain how to interpret the intervals.

5.7. The GSS asks whether you agree or disagree with the following statement: "It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family" (variable FEFAM). The sample proportion agreeing was 0.66 in 1977 and 0.36 in 2004 ($n = 883$).

(a) Show that the estimated standard error in 2004 was 0.016.

(b) Show that the margin of error for a 95% confidence interval using the estimate in 2004 was 0.03. Interpret.

(c) Construct the 95% confidence interval for 2004, and interpret it.

5.8. A recent GSS asked, "If the wife in a family wants children, but the husband decides that he does not want any children, is it all right for the husband to refuse to have children?" Of 598 respondents, 366 said *yes* and 232 said *no*. Show that a 99% confidence interval for the population proportion who would say *yes* is (0.56, 0.66).

5.9. In 2006, the Florida Poll conducted by Florida International University asked whether current environmental regulations are too strict or not too strict. Of 1200 respondents, 229 said they were too strict. Find and interpret a (a) 95%, (b) 99% confidence interval for a relevant parameter at the time of that survey.

5.10. When a recent GSS asked whether the government should impose strict laws to make industry do less damage to the environment, a 95% confidence interval for the population proportion responding *yes* was (0.87, 0.90). Would a 99% confidence interval be wider or shorter? Why?

5.11. State the z-score used in finding a confidence interval for a proportion with confidence level

(a) 0.98

(b) 0.90

(c) 0.50

(d) 0.9973.

5.12. In the 2006 GSS, respondents were asked whether they favored or opposed the death penalty for people convicted of murder. Software shows results:

x	n	Sample prop	95.0% CI
1885	2815	0.6696	(0.652, 0.687)

Here, x refers to the number of the respondents who were in favor.

(a) Show how to obtain the value reported under "Sample prop."

(b) Can you conclude that more than half of all American adults are in favor? Why?

(c) Find a 95% confidence interval for the proportion of American adults who *opposed* the death penalty from the confidence interval shown for the proportion in favor.

5.13. The GSS has asked respondents, "Do you think the use of marijuana should be made legal or not?" View results for all years at sda.berkeley.edu/GSS by entering the variables GRASS and YEAR.

(a) Of the respondents in 2004, what proportion said *legal* and what proportion said *not legal*?

(b) Is there enough evidence to conclude whether a majority or a minority of the population support legalization? Explain your reasoning.

(c) Describe any trend you see since about 1986 in the proportion favoring legalization.

5.14. When the 2000 GSS asked whether human beings developed from earlier species of animals (variable SCITEST4), 53.8% of 1095 respondents answered that this was probably or definitely not true. Find a 99% confidence interval for the corresponding population proportion, and indicate whether you can conclude that a majority of Americans felt this way.

5.15. A study by the U.S. National Center for Health Statistics provided a point estimate of 25.5% for the percentage of adult Americans who were currently smokers. The sample size was 42,000. Assuming that this sample has the characteristics of a random sample, construct and interpret a 99% confidence interval for the population proportion of smokers. (*Note*: When n is very large, even confidence intervals with large confidence levels are narrow.)

5.16. In the exit poll discussed in the previous chapter (page 85), of the 2705 voters sampled, 56.5% said they voted for Schwarzenegger. Is there enough evidence to predict the winner of the election? Base your decision on a 95% confidence interval, stating needed assumptions for that decision.

5.17. For an exit poll of people who voted in a gubernatorial election, 40% voted for Jones and 60% for Smith. Assuming this is a random sample of all voters, construct a 99% confidence interval for the proportion of votes that Jones received, if the sample size was **(a)** 400, **(b)** 40. In each case, indicate whether you would be willing to predict the winner. Explain how and why the sample size affects the inference.

5.18. In 2003 the Harris Poll reported results of a survey about religious beliefs. Of 2201 American adults surveyed, 27% believed in reincarnation. Treating this as a random sample, a 95% confidence interval for the population proportion of American adults believing in reincarnation is (0.25, 0.29). Without doing any calculation, explain how the interval would change if the sample size had been only a fourth as large, $n = 550$.

5.19. Report the t-score that multiplies by the standard error to form a

(a) 95% confidence interval with 5 observations
(b) 95% confidence interval with 15 observations
(c) 95% confidence interval with 25 observations
(d) 95% confidence interval with $df = 25$
(e) 99% confidence interval with $df = 25$

5.20. Find and interpret the 95% confidence interval for μ, if $\bar{y} = 70$ and $s = 10$, based on a sample size of

(a) 5
(b) 20.

5.21. The 2004 GSS asked male respondents how many female partners they have had sex with since their eighteenth birthday. The median = 6 and mode = 1 (16.0% of the sample). A computer printout summarizes other results:

```
-------------------------------------------------------
Variable    n     Mean    StDev   SE Mean     95.0% CI
NUMWOMEN  1007  24.745   52.554    1.656    (21.5, 28.0)
-------------------------------------------------------
```

(a) Show how software got the standard error reported, and interpret.
(b) Interpret the reported confidence interval.
(c) State a factor that might make you skeptical about the usefulness of this confidence interval.

5.22. A GSS asked, "What do you think is the ideal number of children for a family to have?" The 497 females who responded had a median of 2, mean of 3.02, and standard deviation of 1.81.

(a) Report the point estimate of the population mean.
(b) Find and interpret the standard error of the sample mean.
(c) The 95% confidence interval is (2.9, 3.2). Interpret.
(d) Is it plausible that the population mean = 2.0? Explain.

5.23. Refer to the previous exercise. For the 397 males in the sample, the mean was 2.89 and the standard deviation was 1.77.

(a) Show that the standard error of the sample mean is 0.089.
(b) Show that the 95% confidence interval for the population mean is (2.7, 3.1), and explain what "95% confidence" means.

5.24. Example 5.5 (page 120) analyzed data from a study that compared therapies for anorexia. For the 17 girls who received the family therapy, the changes in weight during the study were

$$11, 11, 6, 9, 14, -3, 0, 7, 22, -5, -4, 13, 13, 9, 4, 6, 11.$$

(a) Verify that $\bar{y} = 7.29$ and $s = 7.18$ pounds.
(b) Verify that the standard error of the sample mean was 1.74.
(c) To use the t distribution, explain why $df = 16$ and a 95% confidence interval uses the t-score of 2.120.
(d) Let μ denote the population mean change in weight for this therapy. Verify that the 95% confidence interval for μ is (3.6, 11.0). Interpret.

5.25. The 2004 GSS asked, "On the average day about how many hours do you personally watch television?" Software reports:

```
-------------------------------------------------------
Variable    N     Mean   SE Mean     95.0% CI
TVHOURS   892    2.76      0.08    (2.60, 2.93)
-------------------------------------------------------
```

What's wrong with the interpretation, "In the long run, 95% of the time subjects watched between 2.60 and 2.93 hours of TV a day"? State the correct interpretation.

5.26. In response to the GSS question in 2006 about the number of hours daily spent watching TV, the responses by the 15 subjects who identified themselves as Buddhist were 0, 0, 0, 1, 1, 1, 2, 2, 2, 2, 2, 3, 4, 4, 5.

(a) Estimate the mean, standard deviation, and standard error.
(b) Construct a 95% confidence interval for the population mean.
(c) Specify the assumptions for the method. What can you say about their validity for these data?

5.27. The GSS has asked subjects, "How long have you lived in the city, town or community where you live now?" The responses of 1415 subjects in one

survey had a median of 16 years, a mean of 20.3 and a standard deviation of 18.2.

(a) Do you think that the population distribution is normal? Why or why not?

(b) Based on your answer in (a), can you construct a 99% confidence interval for the population mean? If not, explain why not. If so, do so and interpret.

5.28. A recent GSS asked, "How many days in the past 7 days have you felt sad?" The 816 women who responded had a median of 1, mean of 1.81, and standard deviation of 1.98. The 633 men who responded had a median of 1, mean of 1.42, and standard deviation of 1.83.

(a) Find a 95% confidence interval for the population mean for women. Interpret.

(b) Explain why the $\bar{y}$- and s-values suggest that this variable does not have a normal distribution. Does this cause a problem with the confidence interval method in (a)? Explain.

5.29. The 2004 GSS asked respondents how many sex partners they had in the previous 12 months. Software reports:

```
------------------------------------------------------
Variable   N     Mean   StDev  SE Mean    95.0% CI
partners  2198  1.130   1.063  0.0227   (1.09, 1.18)
------------------------------------------------------
```

(a) Interpret the confidence interval reported.

(b) Based on these results, explain why the distribution was probably skewed to the right. Explain why the skew need not cause a problem with the validity of the confidence interval, unless there are extreme outliers.

5.30. For the "Florida student survey" data file mentioned in Exercise 1.11, software reports the results for responses on the number of times a week the subject reads a newspaper:

```
------------------------------------------------------
Variable   N    Mean  Std Dev  SE Mean     95.0% CI
News       60   4.1    3.0      0.387    (3.32, 4.88)
------------------------------------------------------
```

(a) Interpret the confidence interval shown.

(b) Does it seem plausible that the population distribution of this variable is normal? Why?

(c) Explain the implications of the term "robust" regarding the normality assumption for this analysis.

5.31. The GSS asks respondents to rate their political views on a seven-point scale, where 1 = extremely liberal, 4 = moderate, and 7 = extremely conservative. A researcher analyzing data from the 2004 GSS gets software output:

```
------------------------------------------------------
Variable     N     Mean   StDev  SE Mean     99% CI
Polviews   1294    4.23   1.39    0.0387   (4.13, 4.33)
------------------------------------------------------
```

(a) Show how to construct the confidence interval from the other information provided.

(b) Would the confidence interval be wider or narrower (i) if you constructed a 95% confidence interval, (ii) if you found the 99% confidence interval only for those who called themselves *strong Democrats* on political party identification (PARTYID), for whom the mean was 3.50 with standard deviation 1.51?

(c) What assumption are you making about the scale of measurement for political ideology when you use the sample mean and standard deviation?

5.32. At sda.berkeley.edu/GSS, consider responses to the question, "On how many days in the past 7 days have you felt lonely?" (coded LONELY) for the most recent survey in which this was asked.

(a) Find a point estimate of the population mean.

(b) Construct the 95% confidence interval, and interpret.

5.33. A study estimates the mean annual family income for families living in public housing in Chicago. For a random sample of 30 families, the annual incomes (in hundreds of dollars) are

83 90 77 100 83 64 78 92 73 122
96 60 85 86 108 70 139 56 94 84
111 93 120 70 92 100 124 59 112 79.

(a) Construct a stem-and-leaf plot of the incomes. What do you predict about the shape of the population distribution?

(b) Find and interpret point estimates of μ and σ, the population mean and standard deviation.

(c) Construct and interpret a 95% confidence interval for μ.

5.34. A hospital administrator wants to estimate the mean length of stay for all inpatients in that hospital. Based on a systematic random sample of 100 records of patients for the previous year, she reports that "The sample mean was 5.3. In repeated random samples of this size, the sample mean could be expected to fall within 1.0 of the true mean about 95% of the time."

(a) Construct and interpret a 95% confidence interval for the mean.

(b) The administrator decides that this interval is too wide, and she prefers one of only half this width. How large a sample size is needed?

5.35. To estimate the proportion of traffic deaths in California last year that were alcohol related, determine the necessary sample size for the estimate to be accurate to within 0.06 with probability 0.90. Based on results of a previous study, we expect the proportion to be about 0.30.

5.36. A television network plans to predict the outcome of an election between Jacalyn Levin and Roberto

Sanchez. They will do this with an exit poll on election day. They decide to use a random sample size for which the margin of error is 0.04 for 95% confidence intervals for population proportions.

(a) What sample size should they use?

(b) If the pollsters think that the election will be close, they might use a margin of error of 0.02. How large should the sample size be? (Note that reducing the margin of error by 50% requires quadrupling n.)

5.37. A public health unit wants to sample death records for the past year in Toronto to estimate the proportion of the deaths that were due to accidents. Health officials want the estimate to be accurate to within 0.02 with probability 0.95.

(a) Find the necessary sample size if, based on previous studies, officials believe that this proportion does not exceed 0.10.

(b) Suppose that in determining the necessary sample size, officials use the safe approach that sets $\pi = 0.50$ in the appropriate formula. Then how many records need to be sampled? Compare the result to the answer in part (a), and note the reduction in sample size that occurs by making an educated guess for π.

5.38. A poll in Canada indicated that 48% of Canadians favor imposing the death penalty (Canada does not have it). A report by Amnesty International on this and related polls (www.amnesty.ca) did not report the sample size but stated, "Polls of this size are considered to be accurate within 2.5 percentage points 95% of the time." About how large was the sample size?

5.39. The June 2003 report *Views of a Changing World* conducted by the Pew Global Attitudes Project (www.people-press.org) discussed changes in views of the U.S. by other countries. In the largest Muslim nation, Indonesia, a poll conducted in May 2003 after the Iraq war began reported that 83% had an unfavorable view of America, compared to 36% a year earlier. The 2003 result was reported to have a margin of error of 3 percentage points. Find the approximate sample size for the study.

5.40. An estimate is needed of the mean acreage of farms in Manitoba, Canada. The estimate should be correct to within 100 acres with probability 0.95. A preliminary study suggests that 500 acres is a reasonable guess for the standard deviation of farm size.

(a) How large a sample of farms is required?

(b) A random sample is selected of the size found in (a). The sample has a standard deviation of 300 acres, rather than 500. What is the margin of error for a 95% confidence interval for the mean acreage of farms?

5.41. A social scientist plans a study of adult South Africans living in townships on the outskirts of Cape Town, to investigate educational attainment (the number of years of education completed) in the black community. Many of the study's potential subjects were forced to leave Cape Town in 1966 when the government passed a law forbidding blacks to live in the inner cities. Under the apartheid system, black South African children were not required to attend school, so some residents had very little education. How large a sample size is needed so that a 95% confidence interval for the mean educational attainment has margin of error equal to 1 year? There is no information about the standard deviation of educational attainment, but researchers expect that nearly all values fall between 0 and 18 years.

5.42. How large a sample size is needed to estimate the mean annual income of Native Americans correct to within \$1000 with probability 0.99? Suppose there is no prior information about the standard deviation of annual income of Native Americans, but we guess that about 95% of their incomes are between \$6000 and \$50,000 and that this distribution of incomes is approximately mound shaped.

5.43. An anthropologist wants to estimate the proportion of children in a tribe in the Philippines who die before reaching adulthood. For families she knew who had children born between 1980 and 1985, 3 of 30 children died before reaching adulthood. Can you use the ordinary large-sample formula to construct a 95% confidence interval for the population proportion? Why or why not? Construct an appropriate confidence interval, and interpret.

5.44. You randomly sample five students at your school to estimate the proportion of students who like tofu. None of the five students say they like it.

(a) Find the sample proportion who like it and its standard error. Does the usual interpretation of *se* make sense?

(b) Why is it not appropriate to use the ordinary confidence interval formula (from Section 5.1) for these data? Use a more appropriate approach, and interpret.

5.45. Refer to Exercise 5.33. Construct a 95% confidence interval for the median annual income of the public housing residents. Interpret.

5.46. Refer to Example 5.9 (page 130). Construct a 95% confidence interval for the median time since a book was last checked out. Interpret.

Concepts and Applications

5.47. You can use an *applet* to repeatedly generate random samples and construct confidence intervals, to

illustrate their behavior when used for many samples. To try this, go to www.prenhall.com/agresti and use the *confidence intervals for a proportion* applet. At the menu, set the population proportion value (labeled as p) to 0.50 and set the sample size to 200. Click on *Simulate*. The software will generate 100 samples of size 200 each. For each sample it displays the 95% and 99% confidence intervals for the population proportion and highlights the 95% intervals that fail to contain the parameter value. It also counts the number of intervals that contained the parameter value and the number that did not.

(a) In your simulation, what percentage of the 100 95% confidence intervals generated actually contained the parameter value? How many would be expected to contain the parameter?

(b) To get a feel for what happens "in the long run," click on *Simulate* 49 more times (50 times total). You will then have formed $50 \times 100 = 5000$ separate 95% confidence intervals. What percentage actually contained the true parameter value? You should see that close to 95% of the confidence intervals contained the true parameter.

5.48. Refer to the previous exercise. Using the *confidence interval for a proportion* applet, let's check that the confidence interval for a proportion may work poorly with small samples. Set $n = 10$ and $\pi = 0.90$. Click on *Simulate* to generate 100 random samples, each of size 10, forming confidence intervals for π for each one.

(a) How many intervals failed to contain the true value, $\pi = 0.90$? How many would you expect not to contain the true value? What does this suggest? (Notice that many of the intervals contain only the value 1.0, which happens when $\hat{\pi} = 1.0$.)

(b) To see that this is not a fluke, now click on *Simulate* 49 more times so you will have a total of 5000 confidence intervals. What percentage contain $\pi = 0.90$? (*Note*: For every interval formed, the number of *failures* is smaller than 15, so the large-sample formula is not adequate.)

(c) Using the *sampling distribution* applet at the same website, select the *Binary* population distribution. Use your mouse to click on the first bar and change the proportion of 1's in the population to 0.90. (This is the value of the parameter π.) Specify $N = 1$ random sample of size $n = 10$. Click on *Sample* and it will generate the sample of size 10 and find the sample proportion and plot it on a histogram of sample proportions. Keep clicking on *Sample* 100 times, so you will have generated sample proportions for 100 samples of size 10 each. Look at the empirical sampling distribution of the sample proportion values. Is it bell shaped and symmetric? Use this to help explain why the confidence interval performs poorly in this case.

5.49. Refer to the "Student survey" data file (Exercise 1.11 on page 8). Using software, construct and interpret a 95% confidence interval for (a) the mean weekly number of hours spent watching TV, (b) the proportion believing in life after death. Interpret.

5.50. Refer to the data file created in Exercise 1.12 (page 9). For variables chosen by your instructor, pose a research question, and conduct inferential statistical analyses using basic estimation methods. Summarize and interpret your findings, and explain how you could use them to answer the research question.

5.51. In 2006, the GSS asked about the number of hours a week spent on the World Wide Web, excluding e-mail (variable denoted WWWHR). State a research question you could address about this response variable and a relevant explanatory variable. Go to sda.berkeley.edu/GSS and analyze the data. Prepare a short report summarizing your analysis and answering the question you posed.

5.52. A recent GSS asked married respondents, "Did you live with your husband/wife before you got married?" The responses were 57 *yes*, 115 *no* for those who called themselves politically liberal; and 45 *yes*, 238 *no* for those who called themselves politicaly conservative. Analyze these data, identifying the response variable and explanatory variable. Summarize your analysis in a report of no more than 300 words.

5.53. When subjects in a recent GSS were asked whether they agreed with the following statements, the (*yes*, *no*) counts under various conditions were as follows:

- Women should take care of running their homes and leave running the country up to men: (275, 1556)
- It is better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and the family: (627, 1208)
- A preschool child is likely to suffer if her mother works: (776, 1054)

Analyze these data. Prepare a one-page report stating assumptions, showing results of description and inference, and summarizing conclusions.

5.54. The observations on TV watching for the seven Muslims in the GSS in a recent year were 0, 0, 2, 2, 2, 4, 6. A 95% confidence interval for the population

mean is (0.3, 4.3). Suppose the observation of 6 for the seventh subject was incorrectly recorded as 60. What would have been obtained for the 95% confidence interval? Compare to the interval (0.3, 4.3). How does this warn you about potential effects of outliers on confidence intervals for means?

5.55. **(a)** Explain what it means for an estimator to be unbiased.

(b) Explain why the sample range is a biased estimator of the population range. (*Hint*: How do the sample minimum and maximum compare to the population minimum and maximum? Explain why the sample range is typically less than the population range and cannot be larger.)

5.56. What is the purpose of forming a confidence interval for a parameter? What can you learn from it that you could not learn from a point estimate of the parameter?

5.57. An interval estimate for a mean is more informative than a point estimate, because with an interval estimate you can figure out the point estimate, but with the point estimate alone you have no idea how wide the interval estimate is.

(a) Explain why this statement is correct, illustrating using the reported 95% confidence interval of (4.0, 5.6) for the mean number of dates in the previous month for women at a particular college.

(b) The confidence interval in (a) used a sample size of 50. What were the sample mean and standard deviation?

5.58. Explain why confidence intervals are wider with

(a) larger confidence levels,

(b) smaller sample sizes.

5.59. Why would it be unusual to see a

(a) 99.9999%,

(b) 25% confidence interval?

5.60. Give an example of a study in which it would be important to have

(a) A high degree of confidence

(b) A high degree of precision

5.61. How does population heterogeneity affect the sample size required to estimate a population mean? Illustrate with an example.

5.62. Explain the reasoning behind the following statement: Studies about more diverse populations require larger sample sizes. Illustrate for the problem of estimating mean income for all medical doctors in the U.S. compared to estimating mean income for all entry-level employees at McDonald's restaurants in the U.S.

5.63. You would like to find the proportion of bills passed by Congress that were vetoed by the President in the last congressional session. After checking congressional records, you see that for the population of all 40 bills passed, 2 were vetoed. Does it make sense to construct a confidence interval using these data? Explain. (*Hint*: Identify the sample and population.)

5.64. The 2006 publication *Attitudes towards European Union Enlargement* from Eurobarometer states,

> The readers are reminded that survey results are *estimations*, the accuracy of which rests upon the sample size and upon the observed percentage. With samples of about 1,000 interviews, the real percentages vary within the following confidence limits:

Observed	Limits
10% or 90%	± 1.9
20%, 80%	± 2.5
30%, 70%	± 2.7
40%, 60%	± 3.0
50%	± 3.1

(a) Explain how they got 3.0 points for 40% or 60%.

(b) Explain why the margin of error differs for different observed percentages.

(c) Explain why the accuracy is the same for a particular percentage and for 100 minus that value (e.g., both 40% and 60%).

(d) Explain why it is more difficult to estimate a population proportion when it is near 0.5 than when it is near 0 or 1.

5.65. To use the large-sample confidence interval for π, you need at least 15 outcomes of each type. Show that the smallest value of n for which the method can be used is (a) 30 when $\hat{\pi} = 0.50$, (b) 50 when $\hat{\pi} = 0.30$, (c) 150 when $\hat{\pi} = 0.10$. That is, the overall n must increase as $\hat{\pi}$ moves toward 0 or 1. (When the true proportion is near 0 or 1, the sampling distribution can be highly skewed unless n is quite large.)

Select the best response in Exercises 5.66–5.69.

5.66. The reason we use a z-score from a normal distribution in constructing a large-sample confidence interval for a proportion is that

(a) For large random samples the sampling distribution of the sample proportion is approximately normal.

(b) The population distribution is normal.

(c) For large random samples the data distribution is approximately normal.

(d) If in doubt about the population distribution, it's safest to assume that it is the normal distribution.

5.67. Increasing the confidence level causes the width of a confidence interval to
(a) increase
(b) decrease
(c) stay the same.

5.68. Other things being equal, quadrupling the sample size causes the width of a confidence interval to
(a) double
(b) halve
(c) be one quarter as wide
(d) stay the same.

5.69. Based on responses of 1467 subjects in General Social Surveys, a 95% confidence interval for the mean number of close friends equals (6.8, 8.0). Which of the following interpretations is (are) correct?
(a) We can be 95% confident that $\bar{y}$ is between 6.8 and 8.0.
(b) We can be 95% confident that μ is between 6.8 and 8.0.
(c) Ninety-five percent of the values of y = number of close friends (for this sample) are between 6.8 and 8.0.
(d) If random samples of size 1467 were repeatedly selected, then 95% of the time $\bar{y}$ would fall between 6.8 and 8.0.
(e) If random samples of size 1467 were repeatedly selected, then in the long run 95% of the confidence intervals formed would contain the true value of μ.

5.70. A random sample of 50 records yields a 95% confidence interval for the mean age at first marriage of women in a certain county of 21.5 to 23.0 years. Explain what is wrong with each of the following interpretations of this interval.
(a) If random samples of 50 records were repeatedly selected, then 95% of the time the sample mean age at first marriage for women would be between 21.5 and 23.0 years.
(b) Ninety-five percent of the ages at first marriage for women in the county are between 21.5 and 23.0 years.
(c) We can be 95% confident that $\bar{y}$ is between 21.5 and 23.0 years.
(d) If we repeatedly sampled the entire population, then 95% of the time the population mean would be between 21.5 and 23.5 years.

5.71. Refer to the previous exercise. Provide the proper interpretation.

***5.72.** For a random sample of n subjects, explain why it is about 95% likely that the sample proportion has error no more than $1/\sqrt{n}$ in estimating the population proportion. (*Hint*: To show this "$1/\sqrt{n}$ Rule," find two standard errors when $\pi = 0.50$, and explain how this compares to two standard errors at other values of π.) Using this result, show that $n = 1/M^2$ is a safe sample size for estimating a proportion to within M with 95% confidence.

***5.73.** You know the sample mean of n observations. Once you know $(n - 1)$ of the observations, show that you can find the remaining one. In other words, for a given value of $\bar{y}$, the values of $(n - 1)$ observations determine the remaining one. In summarizing scores on a quantitative variable, having $(n - 1)$ degrees of freedom means that only that many observations are independent.

***5.74.** Find the standard error of the sample proportion when $\pi = 0$ or $\pi = 1$. What does this reflect?

***5.75.** Let π be the probability a randomly selected voter prefers the Republican candidate. You sample 2 people, and neither prefers, the Republican. Find the point estimate of π. Does this estimate seem sensible? Why? (The *Bayesian* estimator is an alternative one that uses a *subjective* approach, combining the sample data with your prior beliefs about π before seeing the data. For example, if you believed π was equally likely to fall anywhere from 0 to 1, the Bayesian estimate adds two observations, one of each type, thus yielding the estimate 1/4.)

5.76.** To encourage subjects to make responses on sensitive questions, the method of ***randomized response is often used. The subject is asked to flip a coin, in secret. If it is a head, the subject tosses the coin once more and reports the outcome, head or tails. If, instead, the first flip is a tail, the subject reports instead the response to the sensitive question; for instance, reporting the response *head* if the true response is *yes* and reporting the response *tail* if the true response is *no*. Let π denote the true probability of the *yes* response on the sensitive question.
(a) Explain why the numbers in Table 5.5 are the probabilities of the four possible outcomes.
(b) Let p denote the sample proportion of subjects who report *head* for the second response. Explain why $\hat{\pi} = 2p - 0.5$ estimates π.
(c) Using this approach, 200 subjects are asked whether they have ever knowingly cheated on their income tax. Report the estimate of π if the number of reported heads equals (i) 50, (ii) 70, (iii) 100, (iv) 150.

TABLE 5.5

	Second Response	
First Coin	Head	Tail
Head	0.25	0.25
Tail	$\pi/2$	$(1 - \pi)/2$

***5.77.** To construct a large-sample confidence interval for a proportion π, it is not necessary to substitute $\hat{\pi}$ for the unknown value of π in the formula for the

standard error of $\hat{\pi}$. A less approximate method finds the endpoints for a 95% interval by determining the π values that are 1.96 standard errors from the sample proportion, by solving for π in the equation

$$|\hat{\pi} - \pi| = 1.96\sqrt{\frac{\pi(1 - \pi)}{n}}.$$

For Example 5.8 (page 129) with no vegetarians in a sample of size 20, substitute $\hat{\pi}$ and n in this equation and show that the equation is satisfied at $\pi = 0$ and at $\pi = 0.161$. So the confidence interval is (0, 0.161).