Statistical Methods for the Social Sciences

Fifth Edition

Alan Agresti

University of Florida

Pearson

Chapter

5

Statistical Inference: Estimation

CHAPTER OUTLINE

5.1 Point and Interval Estimation

5.2 Confidence Interval for a Proportion

5.3 Confidence Interval for a Mean

5.4 Choice of Sample Size

5.5 Estimation Methods: Maximum Likelihood and the Bootstrap*

5.6 Chapter Summary

This chapter shows how to use sample data to estimate population parameters. With categorical variables, we estimate population proportions for the categories. For example, a study dealing with binge drinking by college students might estimate the proportion of college students who participate in binge drinking. With quantitative variables, we estimate the population mean. For example, the study might estimate the mean number of alcoholic drinks taken in a typical binge-drinking experience for the population of college students who do this.

We first learn about two types of estimates: One is a single point and the other is an interval of points, called a ***confidence interval***. We construct confidence intervals for population proportions and means by taking a point estimate and adding and subtracting a margin of error that depends on the sample size. We also learn how to find the sample size needed to achieve the desired precision of estimation. The final section presents two general-purpose methods for estimation—***maximum likelihood*** and the ***bootstrap***—that apply to nearly all other parameters, such as the population median.

5.1 Point and Interval Estimation

We use sample data to estimate a parameter in two ways:

- A ***point estimate*** is a *single number* that is the best guess for the parameter value.
- An ***interval estimate*** is an *interval of numbers* around the point estimate that we believe contains the parameter value. This interval is also called a ***confidence interval***.

For example, a General Social Survey 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, this confidence interval tells us that the point estimate of 0.73 has a *margin of error* of 0.02. Thus, an interval estimate helps us gauge the 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 particular 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

Estimates are the most common statistical inference reported by the mass media. For example, a Gallup Poll in May 2016 reported that 53% of the American public

approved of President Barack Obama's performance in office. This is an estimate rather than a parameter, because it was based on interviewing a sample of about 1500 people rather than the entire population.

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.

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 page 85, 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 this. 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, in the long run the overestimates would counterbalance the underestimates.

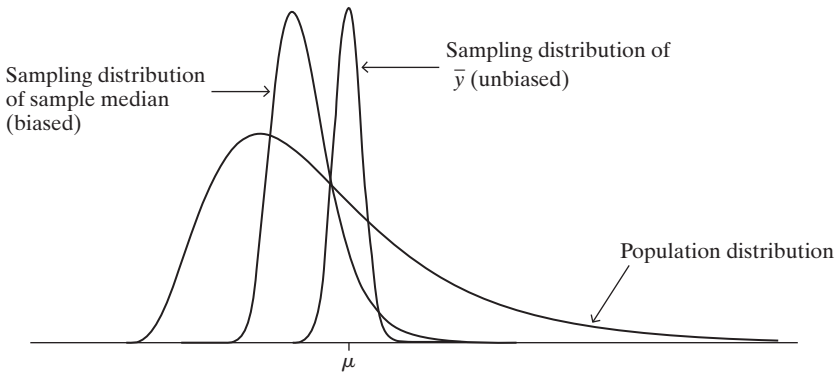

FIGURE 5.1: Sampling Distributions of Two Estimators of the Population Mean, for a Skewed Population Distribution

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 cannot be larger than the population range, 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.

A second desirable property for an estimator is a relatively small standard error. An estimator having standard error that is smaller than those of other estimators is said to be ***efficient***. An efficient estimator tends to fall closer than other estimators to 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. The final

section of this chapter introduces a general method, called *maximum likelihood*, for constructing estimators that have 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.

CONFIDENCE INTERVAL FORMED BY 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 1996 each year the Gallup Poll has asked, "Do you think marriages between same-sex couples should or should not be recognized by the law as valid, with the same rights as traditional marriages?" The percentage saying they should be valid has increased from 27% in 1996 to 60% in 2015. 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. To construct a confidence interval, we add and subtract from the point estimate a z-score multiple of its standard error. This is the ***margin of error***. That is,

Form of confidence interval: **Point estimate ± 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 how much the government should address global warming (less, the same, more). 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 lack health insurance.
- The proportion of Canadians who favor independent status for Quebec.
- The proportion of Australian young adults who have taken a "gap year," that is, a break of a year between high school and college or between college and regular employment.

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.

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 40 and following Example 4.4 on page 78.) 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[2] $\sigma = \sqrt{\pi(1-\pi)}$. Since the formula for the standard error of a sample mean is $\sigma_{\bar{y}} = \sigma/\sqrt{n}$, the standard error $\sigma_{\hat{\pi}}$ of the sample proportion $\hat{\pi}$ is

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

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

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 this. Recall that 95% of a normal distribution falls within two standard deviations of the mean, or, more precisely, 1.96 standard deviations. 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.

---

[1] Here, π is the Greek analog of p for proportion, *not* the mathematical constant, 3.1415....

[2] From page 72, the variance is $\sigma^2 = \sum(y-\mu)^2 P(y) = (0-\pi)^2 P(0) + (1-\pi)^2 P(1) = (0-\pi)^2(1-\pi) + (1-\pi)^2\pi$, which simplifies to $\pi(1-\pi)$. Thus, $\sigma = \sqrt{\pi(1-\pi)}$.

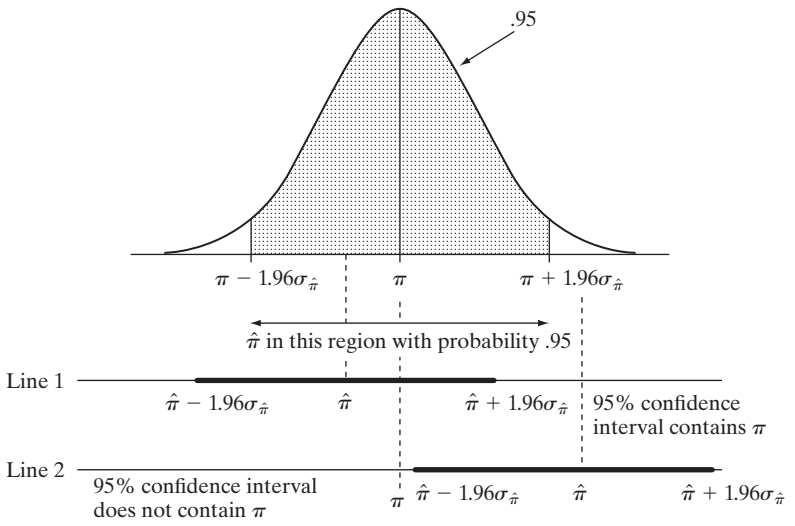

FIGURE 5.2: Sampling Distribution of $\hat{\pi}$ and Possible 95% Confidence Intervals for π

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. 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 the Proportion Who Favor Restricting Legalized Abortion For many years, the Florida Poll[3] conducted by Florida International University asked, "In general, do you think it is appropriate for state government to make laws restricting access to abortion?" In the most recent poll, 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 in the poll, 396 said *yes*, so $\hat{\pi} = 396/1200 = 0.330$. Then, $1 - \hat{\pi} = 0.670$. That is, 33% of the sample said *yes* and 67% said *no*.

[3] See `www2.fiu.edu/~ipor/ffp/abort1.htm`.

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

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

A 95% confidence interval for π is

$$\hat{\pi} \pm 1.96(se) = 0.330 \pm 1.96(0.0136) = 0.330 \pm 0.027, \quad \text{or} \quad (0.30, 0.36).$$

We conclude that 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, at the time of this poll, apparently fewer than half the Florida adult population supported restricting access to abortion. ■

You can obtain this confidence interval using software with your data file. In Stata, you can also find it directly from the summary results, by applying the `cii` command[4] to n and the count in the category of interest:

```
. cii proportions 1200 396, wald
                                                    -- Binomial Wald ---
  Variable |      Obs      Mean    Std. Err.      [95% Conf. Interval]
           |    1,200       .33    .0135739       .3033957    .3566043
```

The software R uses a confidence interval for the proportion that has a more complex formula[5] than the one we gave, so it gives slightly different results:

```
> prop.test(396, 1200)$conf.int
[1] 0.3035683 0.3575336
```

Calculators for such confidence intervals are also available with Internet applets. See Figure 5.3 for an example.

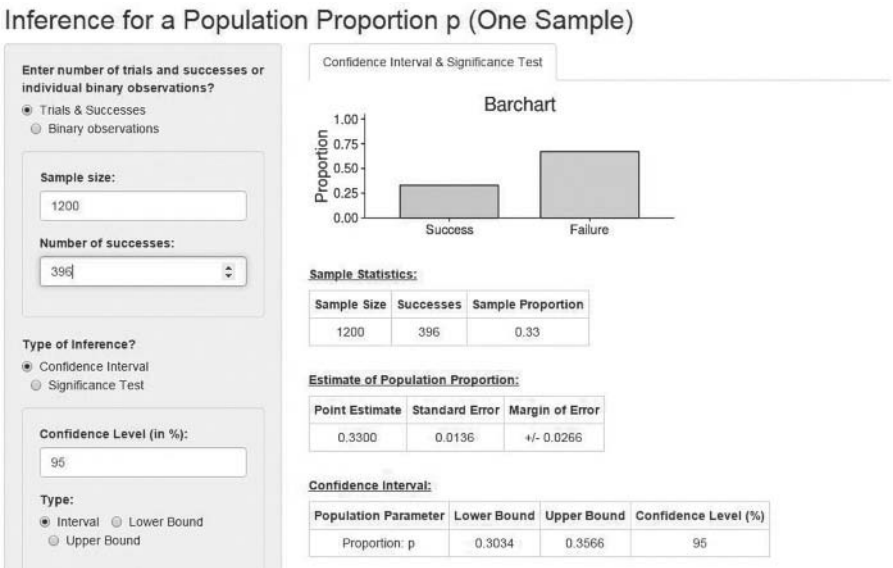

FIGURE 5.3: Applets at `www.artofstat.com/webapps.html` Perform Inference Procedures Presented in Chapters 5–9. The *Inference for a Proportion* applet can construct confidence intervals for proportions.

[4] Stata calls this the *Wald* confidence interval. Here *i* following *ci* stands for *immediate*.
[5] Exercise 5.77 gives the idea behind this so-called *score* confidence interval.

Results in such surveys vary greatly depending on the question wording and where the poll is conducted. For instance, when the 2014 General Social Survey asked whether a pregnant woman should be able to obtain a legal abortion if the woman wants it *for any reason*, 907 said *no* and 746 said *yes*. The 95% confidence interval for the population proportion saying *no* equals (0.53, 0.57).

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) 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, and their validity is shaky because the sampling distribution is only *approximately* normal.

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. ■

CONTROLLING THE CONFIDENCE LEVEL BY CHOICE OF *z*-SCORE

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.

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.

Example 5.3

Finding a 99% Confidence Interval For the data in Example 5.1 (page 107), 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 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 can 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

$$\hat{\pi} \pm 1.96(se) = 0.33 \pm 1.96(0.027) = 0.33 \pm 0.053.$$

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, two factors affect the width of a confidence interval:

The width of a confidence interval

- Increases as the confidence level increases.
- Decreases as the sample size increases.

These properties apply to all confidence intervals, not only 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 $1 - 0.95 = 0.05$. In statistical inference, 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 104). 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 5.4 shows the results of selecting 10 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, 9 of the 10 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.

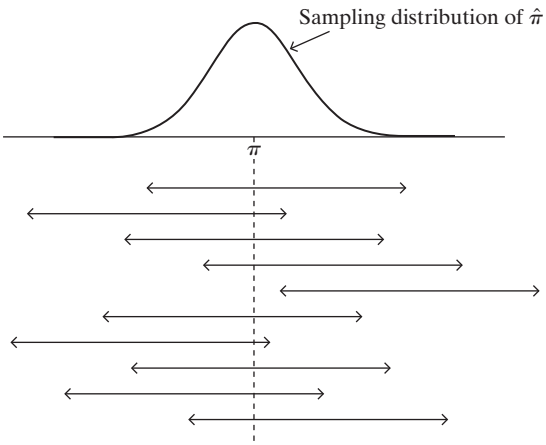

FIGURE 5.4: Ten 95% Confidence Intervals for a Population Proportion π. In the long run, only 5% of the intervals fail to contain π.

You can get a feel for this using an applet designed to illustrate the performance of confidence intervals for proportions:

- Go to `www.artofstat.com/webapps.html` and click on *Explore Coverage*. Use the *Confidence Interval for a Proportion* option.
- The default is forming a 95% confidence interval when $n = 50$ and the true parameter value is $\pi = 0.30$. To better reflect Example 5.1, set the sample size to 1200. Choose 10 samples of size 1200 each. Click on *Draw Sample*. You will see a plot of the 10 confidence intervals, with ones drawn in red that do not contain the parameter value of 0.30. The output also summarizes the number and percentage of the confidence intervals that contain $\pi = 0.30$. What is this?
- Now select 1000 for the number of samples to draw, each of size 1200. Now the proportion of the intervals that actually contain the parameter value is probably closer to 0.95.

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. 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.[6] 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 with sample proportion $\hat{\pi}$, a confidence interval for a population proportion π 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.

[6] For justification of this guideline, download the article at `www.stat.ufl.edu/~aa/ci_proportion.pdf`.

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

Point estimate ± 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 is

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

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 When the 2014 General Social Survey asked respondents how many male sex partners they have had since their 18th birthday, the 129 females in the sample between the ages of 23 and 29 reported a median of 3 and mean of 6.6. Software output summarizes the results:

Variable	n	Mean	StDev	SE Mean	95.0% CI
NUMMEN	129	6.6	13.3	1.17	(4.4, 8.8)

How did software get the standard error reported of 1.17? How do we interpret it and the confidence interval shown?

The sample standard deviation is $s = 13.3$. The sample size is $n = 129$. So, the estimated standard error of the sample mean is

$$se = s/\sqrt{n} = 13.3/\sqrt{129} = 1.17.$$

In several random samples of 129 women in this age grouping, the sample mean number of male sex partners would vary from sample to sample with a standard deviation of about 1.17.

The 95% confidence interval reported of (4.4, 8.8) is an interval estimate of μ, the mean number of male sex partners since the 18th birthday for the corresponding population. We can be 95% confident that this interval contains μ. The point estimate of μ is 6.6, and the interval estimate predicts that μ is likely to be greater than 4.4 but smaller than 8.8.

This example highlights a couple of cautions: First, the sample mean of 6.6 and standard deviation of 13.3 suggest that the sample data distribution is very highly skewed to the right. The mean may be misleading as a measure of center. The median response of 3 is perhaps a more useful summary. It's also worth noting that the mode was 1, with 20.2% of the sample. Second, the margin of error 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 estimate and margin of error may be invalid. ■

How did software find the margin of error for this confidence interval? As with the proportion, for a 95% confidence interval this is approximately 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.

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.[7]

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, for any n the appropriate confidence interval formula is

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

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 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.
- For inference about a population mean, the degrees of freedom equal $df = n - 1$, one less than the sample size.

[7] The right panel of Figure 4.15 on page 88, which showed sampling distributions for various population distributions, illustrated this.

- 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 5.5 illustrates this. When df is about 30 or more, the two distributions are nearly identical.

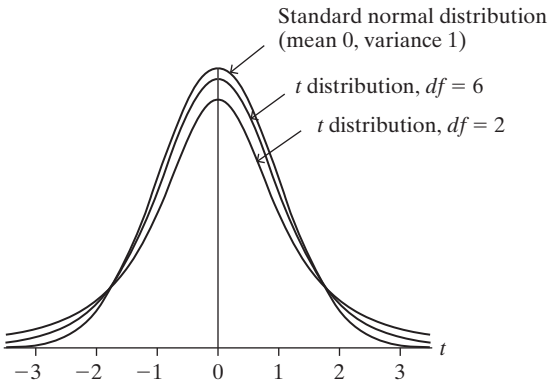

FIGURE 5.5: t Distribution Relative to Standard Normal Distribution. The t gets closer to the normal as the degrees of freedom (df) increase, and the two distributions are practically identical when $df > 30$.

- 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. The column labeled $t_{.025}$, which has probability 0.025 in the right tail and a two-tail probability of 0.05, is the t-score used in 95% confidence intervals.

TABLE 5.1: t-Scores for Various Confidence Levels and Degrees of Freedom (df). The scores, obtained with the `qt` function in R software, 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

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 5.6. When $df = 28$, the probability equals 0.95 between -2.048 and 2.048. These are the t-scores for a 95% confidence interval when $n = 29$. The confidence interval is $\bar{y} \pm 2.048(se)$.

The t-scores are also supplied by software. For example, the free software R has a function `qt` that gives the t-score for a particular cumulative probability. For

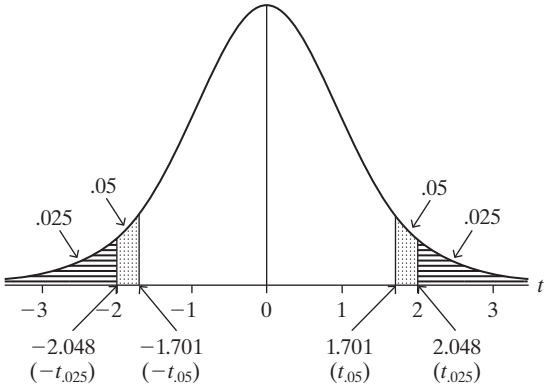

FIGURE 5.6: The t Distribution with $df = 28$

example, the right-tail probability of 0.025 corresponds to a cumulative probability of 0.975, for which the t-score when $df = 28$ is

```
> qt(0.975, 28) # q = "quantile" (percentile) for t distribution
[1] 2.048407
```

With Stata software, we can find this with the `invt` (inverse t) command:

```
. display invt(28, 0.975)
2.0484071
```

It is also possible to find t-scores with SPSS and SAS statistical software, but it is simpler to use Internet sites and statistical calculators. See Figure 5.7.

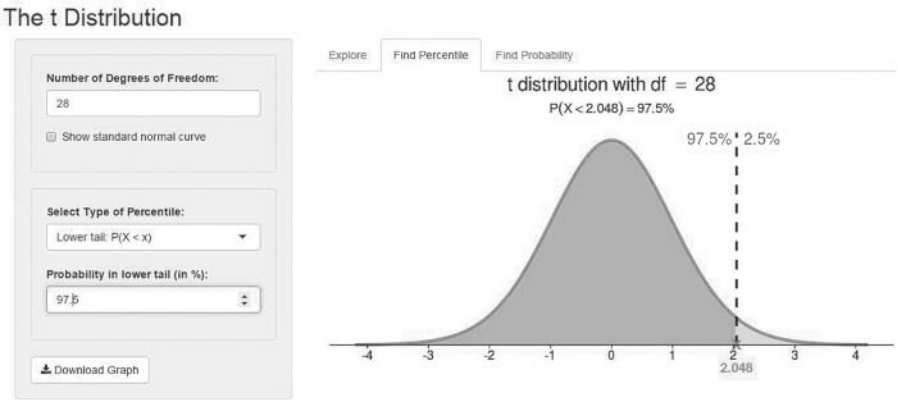

FIGURE 5.7: The *t Distribution* Applet at `www.artofstat.com/webapps.html` Can Supply t Cumulative and Tail Probabilities

t-SCORES IN THE CONFIDENCE INTERVAL FOR A MEAN

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

Confidence Interval for Population Mean μ

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

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

and $df = n - 1$ 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. Besides substituting the t-score for the z-score, the t method also makes the assumption of a normal population distribution. In practice, the population distribution may not be close to normal. We discuss the importance of this assumption later in the section, where we'll find that this is mainly relevant for very small samples.

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 29 girls undergoing the cognitive behavioral treatment were[8]

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

Software used to analyze the data from a data file reports the summary results:

```
Variable |   Obs        Mean    Std. Dev.       Min        Max
  change |    29    3.006896    7.308504      -9.1       20.9
```

For the $n = 29$ girls who received this treatment, their mean weight change was $\bar{y} = 3.01$ pounds with a standard deviation of $s = 7.31$. The sample mean had an estimated standard error of $se = s/\sqrt{n} = 7.31/\sqrt{29} = 1.36$.

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

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

It is simple to do this using software. For example, with R applied to a data file having a variable called *change* for the change in weight:

```
> t.test(change, conf.level=0.95)$conf.int
[1] 0.2268902 5.7869029
```

[8] Courtesy of Prof. Brian Everitt, King's College, London; data available in `Anorexia_CB` data file at text website.

With Stata, we apply the command `ci` to the variable name. If you have only summary statistics, Stata can construct the interval using them, with the `cii` command or using a dialog box, by entering n, $\bar{y}$, and s:

```
. cii means 29 3.007 7.309

  Variable |   Obs      Mean    Std. Err.     [95% Conf. Interval]
           |    29     3.007    1.357247      .2268051    5.787195
```

SPSS output is similar (see page 148). Calculators for t confidence intervals are also available online.[9]

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 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 that compare the therapies.

Another caveat about our conclusion is shown by Figure 5.8, a histogram that software shows for the data. This reveals that the sample data distribution is skewed to the right. The assumption of a normal population distribution may be violated—more about that below. 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. ■

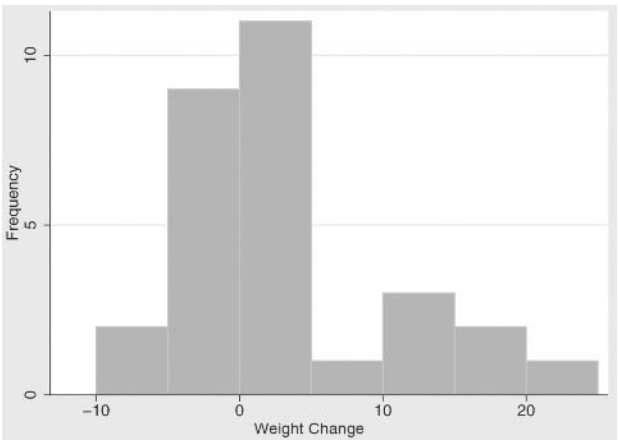

FIGURE 5.8: Histogram of Weight Change Values for Anorexia Study

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.

[9] Such as the *Inference for a Mean* applet at `www.artofstat.com/webapps.html`.

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. We then use the t-score with total probability 0.01 in the two tails, so 0.005 in each tail. Since $df = 28$ when $n = 29$, this 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 population 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) randomization for collecting the sample and (2) normal population distribution. 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.

For the anorexia study, the sample data histogram in Figure 5.8 is not a precise indication of the population distribution because n is only 29, but it showed evidence of skew. Generally, the normal population assumption seems worrisome for social science application of this statistical method, because variables often 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 contains μ 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.

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

STANDARD NORMAL IS THE t DISTRIBUTION WITH $df = \infty$

Look at a table of t-scores, such as Table 5.1 or Table B. 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 $df = \infty$. As we showed, you can get t-scores for *any* df value using software, so you are not restricted to those in Table B.

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 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. The method for constructing t confidence intervals for a mean was introduced 20 years after Gosset's discovery.

USING SOFTWARE FOR STATISTICAL METHODS

The examples in this section used output from statistical software to help us analyze the data. We'll show software output increasingly in future chapters as we cover methods that require substantial computation. You should use software yourself for some exercises and to get a feel for how researchers analyze data in practice.

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 a random sampling scheme is properly

executed, good estimates result from relatively small samples, no matter how large the population size.[10] 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. 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*.

DETERMINING SAMPLE SIZE FOR ESTIMATING PROPORTIONS

To determine the sample size, we must decide on the margin of error. Highly precise estimation is more important in some studies than 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 U.S. citizens who do not 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 is the confidence level for the confidence interval.

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 in a single-parent family. Since her report was to be published, she wanted a reasonably precise estimate. However, her funding was limited, so 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 equal 0.04, then with probability 0.95, $\hat{\pi}$ falls within 0.04 units of π. See Figure 5.9.

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

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

[10] In fact, the methods actually treat the population size as infinite; see Exercise 4.57 in Chapter 4.

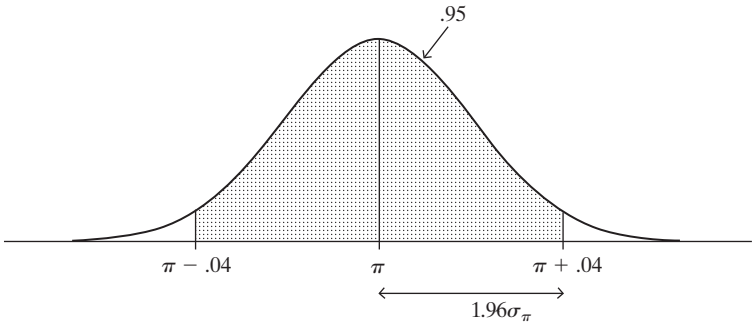

FIGURE 5.9: Sampling Distribution of $\hat{\pi}$ with the Error of Estimation No Greater than 0.04, with Probability 0.95

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, one approach merely substitutes 0.50 for π in the above equation for n. 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.

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 the chosen 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 confidence level 0.95. Then the margin of error is $M = 0.08$, 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. Calculators for this are also available online.[11]

DETERMINING SAMPLE SIZE FOR ESTIMATING MEANS

Next we find n for estimating a population mean μ. We determine how large n needs to be so that the sampling distribution of $\bar{y}$ has margin of error M. Figure 5.10 illustrates this.

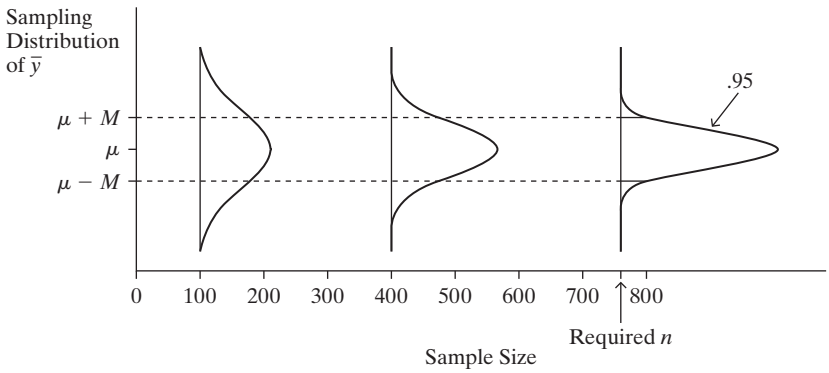

FIGURE 5.10: Determining n So that $\bar{y}$ Has Probability 0.95 of Falling within a Margin of Error of M Units of the Population Mean μ

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 $\bar{y}$ is

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

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

[11] For example, at `epitools.ausvet.com.au/content.php?page=1Proportion`.

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 heterogeneous. 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 have seen, 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 in using an educated guess for σ. Calculators for the formula for determining n are also available online.[12]

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 one year with probability 0.99?

If the study has no prior information about the standard deviation σ of educational attainment for Native Americans, we need to provide a guess. Perhaps nearly all educational attainment values fall within a range of 15 years, such as between 5 and 20 years. If the population 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 equals 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 2.58, which has probability $0.01/2 = 0.005$ in each tail. 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 $n = (3.75)^2(2.58/1)^2 = 94$. ■

OTHER CONSIDERATIONS IN DETERMINING SAMPLE SIZE

In summary, the necessary sample size depends on the desired *precision* for the margin of error, the *confidence level* for a confidence interval, and 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

[12] For example, at `http://epitools.ausvet.com.au/content.php?page=1Mean`.

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, we would need a much larger sample.

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?"

The sample size formulas of this section apply to simple 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, seek guidance from a statistical consultant.

In summary, no simple formula can always give an appropriate sample size. The needed sample size 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 observations are stated wrongly or 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 reasons, 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 is 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 of undergraduate students 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. (Recall the guidelines in the box on page 112.) 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 four artificial observations, two 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 also find this interval using some software, or with Internet applets.[13] You can find it using Stata,[14] by applying the `cii` command to n and the count in the category of interest or using a dialog box:

```
. cii proportions 20 0, agresti
                                                    -- Agresti-Coull ---
Variable |     Obs      Mean    Std. Err.    [95% Conf. Interval]
         |      20         0           0              0    .1898096
```

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 closely approximates one based on a more complex method (described in Exercise 5.77) that does not require estimating the standard error.

5.5 Estimation Methods: Maximum Likelihood and the Bootstrap*

We've focused on estimating means and proportions, but Chapter 3 showed that other statistics are also useful for describing data. These other statistics also have sampling distributions. In this section, we introduce a standard method, called *maximum likelihood*, that statisticians use to find good estimators of parameters. We also introduce a newer method, called the *bootstrap*, that uses modern computational power to find confidence intervals in cases in which it is difficult to derive the sampling distribution.

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.

---

[13] For example, with the *Inference for a Proportion* applet at `www.artofstat.com/webapps.html`.

[14] Software calls it the *Agresti–Coull* confidence interval, because it was proposed in an article by A. Agresti and B. Coull, *American Statistician*, vol. 52 (1998), pp. 119–126.

For point estimation, Fisher proposed 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.

We illustrate this method using data from a recent survey with a random sample of 1000 adult Americans, in which a sample proportion of 0.37 said that they believed in astrology. What is the maximum likelihood estimate of the population proportion who believe in astrology? Figure 5.11 plots the probability that a random sample of size 1000 has a sample proportion of 0.37, as a function of the actual population proportion believing in astrology. The probability changes dramatically as the population proportion changes. The curve, called a *likelihood function*, suggests that such a sample would be essentially impossible if the population proportion were below about 0.32 or above about 0.42. The maximum of the curve occurs at the population proportion value of 0.37. That is, the observed sample result would have been more likely to occur if the population proportion equaled 0.37 than if it equaled any other possible value between 0 and 1. So, the maximum likelihood estimate of the population proportion who believe in astrology is 0.37. In fact, with random sampling, the maximum likelihood estimate of a population proportion is necessarily the sample proportion.

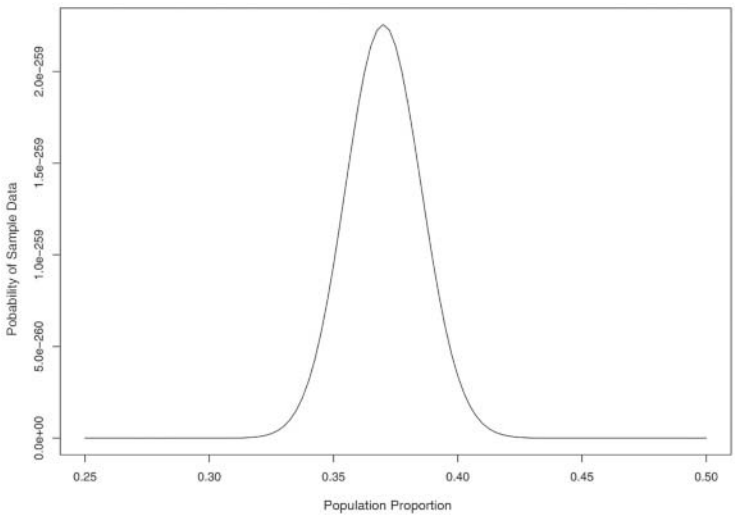

FIGURE 5.11: The Probability that Exactly 37% of a Sample of Size 1000 Believe in Astrology, Plotted as a Function of the Population Proportion Believing in Astrology. The maximum probability occurs at the population proportion value of 0.37. This is the maximum likelihood estimate.

For many population distributions, such as the normal distribution, 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.
- They are *consistent*, in the sense that as n increases they tend to get closer and closer to the unknown parameter value. In particular, they have little, if any, bias, with the bias diminishing to 0 as n increases.
- They have *approximately normal sampling distributions*.

Fisher also showed how to estimate standard errors for maximum likelihood estimators. Because their sampling distributions are approximately normal, confidence intervals for the parameters they estimate have the general form of taking the maximum likelihood estimate and then adding and subtracting a z-score multiplied by the estimated standard error. For instance, this is the method we used in Section 5.2 to find a confidence interval for a population proportion. To learn more about maximum likelihood, see Eliason (1993).

MAXIMUM LIKELIHOOD FOR MEAN, MEDIAN OF NORMAL DISTRIBUTION

When the population distribution is normal, the population mean and median are identical, because of the symmetry of the distribution. How should we estimate that common value, with the sample mean or the sample median? They are both point estimators of the same number. Fisher found that the maximum likelihood estimator is the sample mean, and that is preferred over the sample median.

In fact, for random samples, the standard error of the sample median equals $1.25\sigma/\sqrt{n}$. The sample median is not as efficient an estimator as the sample mean, because its standard error is 25% larger. When the population distribution is approximately normal, 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. We'll next learn about a general method that is useful for constructing confidence intervals even when we do not know the shape of the population distribution.

THE BOOTSTRAP

To use maximum likelihood, we need to make an assumption about the shape of the population distribution. But sometimes we do not have enough information to make a sensible assumption. In addition, some parameters do not have a confidence interval formula that works well regardless of the population distribution or sample size.

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 and approximates by simulation the unknown sampling distribution. To do this, the method samples n observations, with replacement, from the sample distribution. That is, each of the original n data points has probability $1/n$ of selection for each "new" observation. This new sample of size n has its own point estimate of the parameter. The bootstrap method repeats this sampling process a large number of times, for instance, selecting 1000 separate samples of size n and 1000 corresponding point estimate values.

This type of empirically generated sampling distribution of the point estimate values provides information about the true parameter. For example, it generates a standard error for the point estimate we found with the actual data. This standard error is the sample standard deviation of the point estimate values from the simulations. It also generates a confidence interval for the parameter, for example, by the interval of values between the 2.5 and 97.5 percentiles of the simulated point estimate values. This is a computationally intensive process, but easily feasible with modern computing power.

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?" She suspected that the distributions of these variables were heavily skewed to the right, so she chose the median to describe the center.

Table 5.2 shows data (from the `Library` data file at the text website) 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 5.12 shows a 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.

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		

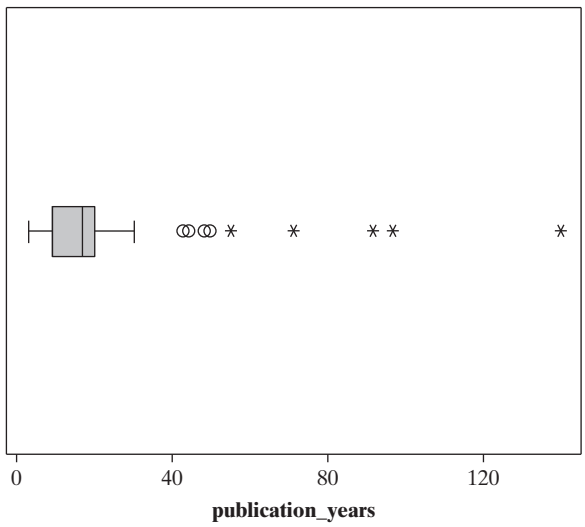

FIGURE 5.12: Box Plot for Number of Years since Publication for Sample of Library Books

What is the standard error for this sample median estimate? There is no simple formula for this when we do not assume a shape for the population distribution. However, we can use the bootstrap to find one as well as a corresponding confidence

interval. The bootstrap is available on the Internet[15] and in software. For instance, in Stata software we find

```
. bootstrap r(p50), reps(10000): summarize P,  detail

       |   Observed   Bootstrap         Normal-based
       |      Coef.   Std. Err.     [95% Conf. Interval]
 _bs_1 |         17   2.114768      12.85513    21.14487
```

to produce 10,000 replications of a bootstrap for the median (labeled by Stata as `r(p50)` for the 50th percentile) of the variable *P*. The sample median of 17 has a bootstrap standard error of 2.11 and a 95% confidence interval for the population median of (12.9, 21.1). ■

Likewise, there is not a simple formula for a confidence interval for a standard deviation unless we make rather stringent assumptions. For the library data set, in Stata we use 10,000 replications of a bootstrap for the standard deviation of the variable *P*:

```
. bootstrap r(sd), reps(10000): summarize P,  detail

       |   Observed   Bootstrap          Normal-based
       |      Coef.   Std. Err.     [95% Conf. Interval]
 _bs_1 |   25.91758   5.578261      14.98439    36.85077
```

The sample standard deviation of the time since publication of the book was 25.9 years, and a 95% bootstrap confidence interval for the population standard deviation is (15.0, 36.9).

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

Confidence intervals for a population mean μ and for a population proportion π have the form

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

where se is the estimated standard error.

[15] See the *Bootstrap* applet at `www.artofstat.com/webapps.html`.

The true standard error, which is $\sigma/\sqrt{n}$, depends on the unknown population standard deviation σ. We estimate this and use it to get an *estimated* standard error, denoted by *se*. Table 5.3 shows the formula for *se* for estimating means and proportions. 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. For the relatively large sample sizes of most social research, the t-score is essentially the same as the z-score.

- The probability that the method yields an interval that contains the parameter, called the ***confidence level***, 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*** applies for statistical inference about a mean. It 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 estimated standard error (*se*) of the sampling distribution of the point estimator. Larger sample sizes produce smaller *se* values 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 approximately normal even if the population is highly nonnormal, by the Central Limit Theorem. Confidence intervals using the t distribution apply with any n but assume a normal population distribution, although the method is ***robust*** to violations of that assumption. Table 5.3 summarizes estimation methods.

TABLE 5.3: Summary of Estimation Methods for Means and Proportions, with Margin of Error M

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: 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 and $z = 1.96$).

Table 5.3 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.

The ***maximum likelihood estimator*** is an efficient estimator that has an approximately normal sampling distribution and is commonly used in statistical inference when we are willing to make an assumption about the shape of the population distribution. The ***bootstrap*** is a resampling method that can yield standard errors and