56. * Refer to the formula for the variance of a probability distribution in the previous exercise. Find the standard deviation for the distribution in Problem 4.1. Can one use the Empirical Rule to interpret this standard deviation? Explain.

57. * The curve for a normal distribution with mean μ and standard deviation σ has mathematical formula

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/(2\sigma^2)}$$

(The integral of this function with respect to y between $\mu+z\sigma$ and ∞ equals the tail probability tabulated in Table A). Show that this curve is symmetric; that is, for any constant c, the curve has the same value for $y = \mu + c$ as for $y = \mu - c$.

58. * The standard error formula $\sigma_{\bar{Y}} = \sigma/\sqrt{n}$ actually treats the population size N as *infinitely* large relative to the sample size n. The formula for $\sigma_{\bar{Y}}$ for a *finite* population size N is

$$\sigma_{\bar{Y}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$$

The term $\sqrt{(N-n)/(N-1)}$ is called the ***finite population correction***.

a) When $n = 300$ students are selected from a college student body of size $N = 30,000$, show that $\sigma_{\bar{Y}} = .995\sigma/\sqrt{n}$. (In practice, n is usually small relative to N, so the correction has little influence.)

b) If $n = N$ (i.e., we sample the entire population), show that $\sigma_{\bar{Y}} = 0$. In other words, no sampling error occurs, since $\bar{Y} = \mu$.

Bibliography

Goldberg, S. (1982). *Probability in Social Science*. Boston: Birkhauser.

Moore, D., and McCabe, G. (1993). *Introduction to the Practice of Statistics*. New York: Freeman.

Olkin, I., Gleser, L., and Derman, C. (1994). *Probability Models and Applications*, 2nd ed. New York: Macmillan.

Scheaffer, R. L. (1995). *Introduction to Probability and its Applications*, 2nd ed. Belmont, CA: Wadsworth.

Scheaffer, R. L., Gnanadesikan, M., Watkins, A., and Witmer, J. (1996). *Activity–Based Statistics*. New York: Springer.

Chapter 5

Statistical Inference: Estimation

This chapter shows how to use sample data to estimate population parameters. With quantitative variables, studies usually estimate the population mean. A study dealing with health care issues in Texas, for example, might select a sample of residents to estimate such population parameters as the mean amount of money spent on health care during the past year, the mean number of visits to a physician, and the mean number of days of work missed due to illness. With qualitative variables, studies usually estimate the population proportions of measurements in the various categories. For example, the health care study might estimate the proportions of people who (have, do not have) medical insurance, the proportions who are (satisfied, not satisfied) with their access to health care, and the proportions who have (experienced, not experienced) an illness requiring hospitalization in the past year.

Statistical inference uses sample data to form two types of estimators of parameters. A ***point estimate*** consists of a single number, calculated from the data, that is the best single guess for the parameter. For example, in a recent General Social Survey, 1359 subjects were asked "Do you believe in Hell?" The point estimate for the proportion of all Americans who would respond "yes" equals .63. An ***interval estimate*** consists of a range of numbers around the point estimate, within which the parameter is believed to fall. For example, for the data just mentioned, an interval estimate predicts that the population proportion responding "yes" falls between .59 and .67; that is, it predicts that the point estimate of .63 falls within .04 of the true value. Thus, an interval estimate helps us gauge the probable accuracy of a sample point estimate.

Section 5.1 introduces point estimation. Sections 5.2 and 5.3 present interval estimates for population means and proportions. Section 5.4 shows how to determine the sample size needed to achieve the desired accuracy, and Section 5.5 discusses interval estimation of medians.

5.1 Point Estimation

The process of predicting a parameter value reduces the sample data to a single number that is the best guess about that value. The statistic that provides the prediction is called a *point estimator* of the parameter.

Point Estimator

A ***point estimator*** of a parameter is a sample statistic that predicts the value of that parameter.

For instance, to estimate a population mean, which we have denoted by μ, an obvious point estimator is the sample mean $\bar{Y}$. A good point estimator of a parameter is one with a sampling distribution that (1) is centered around the parameter and (2) has as small a standard error as possible. An estimator with property (1) is said to be *unbiased*, and an estimator with property (2) is said to be *efficient*. For simplicity, it is common to use the term "estimate" in place of point estimator.

Unbiased and Efficient Point Estimators

A point estimator is ***unbiased*** if its sampling distribution centers around the parameter, in the sense that the parameter is the mean of the distribution. For any particular sample, the point estimator may underestimate the parameter or overestimate it. If that point estimator were used repeatedly in different situations with different samples, however, it would not tend to overestimate or underestimate the parameter systematically; the overestimates would tend to counterbalance the underestimates.

From Section 4.4, for random sampling the mean of the sampling distribution of $\bar{Y}$ equals μ. Thus, $\bar{Y}$ is an unbiased estimator of the population mean μ, as Figure 5.1 illustrates. Sometimes $\bar{Y}$ falls below μ, sometimes it falls above, but it does well in this average sense.

A ***biased*** estimator, on the other hand, tends to either underestimate or overestimate the parameter, on the average. Figure 5.1 also portrays the sampling distribution of an estimator that is biased, tending on the average to underestimate μ. For instance, the sample median is a biased estimate of the population mean when the population distribution is skewed to the right. 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, on the average.

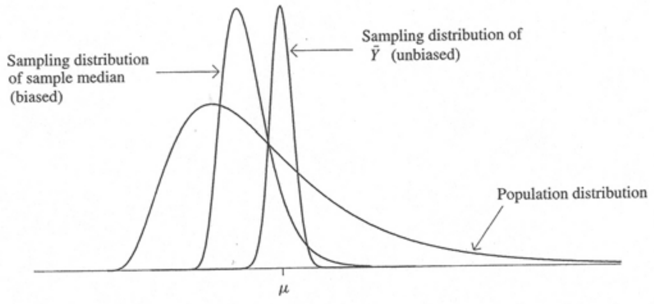

Figure 5.1 Sampling Distributions of Two Point Estimators of the Population Mean, for a Skewed Population Distribution

The concept of bias refers to the estimator's behavior in repeated sampling, not in one particular sample. Estimators are evaluated in terms of their theoretical performance in a long run of repeated samples. In practice, however, we select a single sample of fixed size to estimate a particular parameter. Statistical methods use estimators that are unbiased or for which the bias is negligible and disappears as the sample size increases.

A second preferable property for an estimator is a small sampling error compared with other estimators. An estimator whose standard error is smaller than those of other potential estimators is said to be ***efficient***. An efficient estimator is desirable because, on the average, it falls closer than other estimators to the parameter.

For instance, suppose a population distribution is normal, and we want to estimate its center, which is its mean, median, and mode. We could use the sample mean as the estimate. Or, we could use the sample median. Section 5.5 shows, though, that in sampling from a normal distribution, the sample median has a standard error that is 25% larger than the standard error of the sample mean. Thus, the sample mean tends to be closer than the sample median to the population center. In this case, the sample mean is an efficient estimator, but the sample median is inefficient.

In summary, a good estimator of a parameter is *unbiased*, or nearly so, and *efficient*. The point estimates presented in this text possess these properties.

Point Estimators of the Mean and Standard Deviation

The sample mean $\bar{Y} = \sum Y_i/n$ is the obvious point estimator of a population mean μ. In fact, $\bar{Y}$ is unbiased, and it is relatively efficient for most population distributions. It is the point estimator used in this text.

The symbol "^" over a parameter symbol represents 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 the estimate $\bar{Y}$ of the population mean μ, and $\hat{\sigma}$ denotes an estimate of the population standard deviation σ. The sample standard deviation s is the most popular point estimate of σ. That is,

$$\hat{\sigma} = s = \sqrt{\frac{\Sigma(Y_i - \bar{Y})^2}{n-1}}$$

For qualitative data, the population proportion of observations falling in some category is relevant. The usual point estimator is the sample proportion. Similarly, the point estimator of a population percentage is the sample percentage. It is common, though not necessary, to use the sample analog of a population parameter as its point estimator.

Point estimates are the most common form of inference reported by the mass media. For example, a survey in May 1996 reported that 55% of the American public approved of President Clinton's performance in office. This is a point estimate rather than parameter, since it is based on sample data rather than the entire population.

Maximum Likelihood Estimation *

As mentioned earlier, compared to other mathematical sciences, statistical science is young. Most methods described in this book were developed in the twentieth century. For instance, interval estimation methods were introduced in a series of articles beginning in 1928 by Jerzy Neyman (1894–1981) and Egon Pearson (1895-1980).

The most important contributions to modern statistical science were made by the 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 using 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 .37. Then, the observed sample would have been more likely to occur if the population proportion equals .37 than if it equaled any other possible value.

For many population distributions, such as the normal, the maximum likelihood estimate of a population mean is the sample mean. Fisher showed that, for large samples, maximum likelihood estimators have three desirable properties:

- They are efficient. One cannot find other estimators that have smaller standard errors and 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.

The point estimates presented in this book are, under certain population assumptions, maximum likelihood estimates or essentially identical to such estimates for moderate to large samples. For small samples, however, not all statisticians agree that maximum likelihood estimates are the best, particularly for problems with several parameters. Some interesting research in the past quarter century has shown conditions under which biased estimators may be better than the usual estimators such as sample means and proportions, when there are several means or proportions to estimate. See, for instance, Efron and Morris (1977).

5.2 Confidence Interval for a Mean

To be truly informative, an inference about a parameter should provide not only a point estimate but should also indicate the probable accuracy of the estimate. That is, it should describe how close that estimate is likely to fall to the true parameter value. If a study with 100 college seniors reports that the estimated mean number of sex partners that college seniors have had equals 5, we'd like to know whether that estimate of 5 is likely to be within 1 of the actual population mean, within 2, within 4, or whatever.

The accuracy of a point estimator depends on characteristics of the sampling distribution of that estimator. For example, the sampling distribution determines the probability that the estimator falls within a certain distance of the parameter. If the sampling distribution is approximately normal, then with high probability (about .95), the estimator falls within two standard errors of the parameter, and almost certainly it falls within three standard errors. The estimated standard error helps us determine the likely accuracy of the estimator. The smaller the standard error, the more accurate the estimator tends to be.

Confidence Intervals

The information about the likely accuracy of a point estimator determines the width of an ***interval estimate*** of the parameter. This consists of a range of numbers that contains the parameter with some fixed probability close to 1. Interval estimates are called ***confidence intervals***.

Confidence Interval

A ***confidence interval*** for a parameter is a range of numbers within which the parameter is believed to fall. The probability that the confidence interval contains the parameter is called the ***confidence coefficient***. This is a chosen number close to 1, such as .95 or .99.

A confidence interval is based on a point estimator and the spread of the sampling distribution of that estimator. When the sampling distribution is approximately normal, we construct a confidence interval by adding to and subtracting from the point estimate some multiple (a z-score) of its standard error. This section shows how to do this for a mean. The confidence interval has the Central Limit Theorem as its foundation, so it is appropriate whenever the sample size is large enough to apply that result, say, $n \geq 30$. (The reason for 30 as the cutoff point will be more apparent in Section 6.5, which presents an analogous method for smaller samples.)

Large-Sample Confidence Interval for a Mean

The Central Limit Theorem states that, for large random samples, the sampling distribution of $\bar{Y}$ is approximately normal. The mean of the sampling distribution equals the population mean, μ, and the standard error equals

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

Now, 95% of a normal distribution falls within two standard deviations of the mean, or, more precisely, 1.96 standard deviations. So, with probability .95, $\bar{Y}$ falls within $1.96\sigma_{\bar{Y}}$ units of the parameter μ, that is, between $\mu - 1.96\sigma_{\bar{Y}}$ and $\mu + 1.96\sigma_{\bar{Y}}$, as Figure 5.2 shows.

Now, once the sample is selected, if $\bar{Y}$ does fall within $1.96\sigma_{\bar{Y}}$ units of μ, then the interval from $\bar{Y} - 1.96\sigma_{\bar{Y}}$ to $\bar{Y} + 1.96\sigma_{\bar{Y}}$ contains μ. See line 1 of Figure 5.2. In other words, with probability .95 a $\bar{Y}$ value occurs such that the interval $\bar{Y} \pm 1.96\sigma_{\bar{Y}}$ contains the population mean μ.

On the other hand, the probability is .05 that $\bar{Y}$ does not fall within $1.96\sigma_{\bar{Y}}$ of μ. If that happens, then the interval from $\bar{Y} - 1.96\sigma_{\bar{Y}}$ to $\bar{Y} + 1.96\sigma_{\bar{Y}}$ does *not* contain μ (see Figure 5.2, line 2). Thus, the probability is .05 that $\bar{Y}$ is such that $\bar{Y} \pm 1.96\sigma_{\bar{Y}}$ does *not* contain μ.

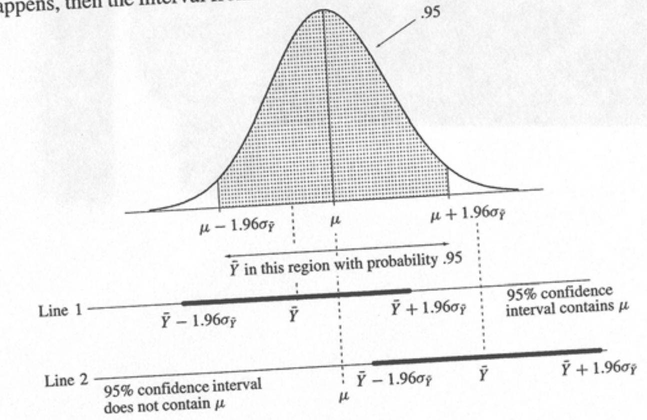
.95
$\mu - 1.96\sigma_{\bar{Y}}$
μ
$\mu + 1.96\sigma_{\bar{Y}}$
$\bar{Y}$ in this region with probability .95
Line 1
$\bar{Y} - 1.96\sigma_{\bar{Y}}$
$\bar{Y}$
$\bar{Y} + 1.96\sigma_{\bar{Y}}$
95% confidence interval contains μ
Line 2
95% confidence interval does not contain μ
$\bar{Y} - 1.96\sigma_{\bar{Y}}$
μ
$\bar{Y}$
$\bar{Y} + 1.96\sigma_{\bar{Y}}$

Figure 5.2 Sampling Distribution of $\bar{Y}$ and Possible 95% Confidence Intervals for μ

The interval $\bar{Y} \pm 1.96\sigma_{\bar{Y}}$ is an interval estimate for μ with confidence coefficient .95, called a ***95% confidence interval***. Unfortunately, the value of the standard error $\sigma_{\bar{Y}} = \sigma/\sqrt{n}$ in this formula is unknown, since the population standard deviation σ is an unknown parameter. For $n \geq 30$, a good approximation for $\sigma_{\bar{Y}}$ results from substituting the sample standard deviation s for σ in this formula. Then,

$$\hat{\sigma}_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

estimates the true standard error. One can insert this estimated standard error in the formula for a confidence interval. The error in substituting the point estimate s for σ is small when $n \geq 30$. The resulting 95% confidence interval equals

$$\bar{Y} \pm 1.96\hat{\sigma}_{\bar{Y}}, \quad \text{which is } \bar{Y} \pm 1.96\frac{s}{\sqrt{n}}$$

Example 5.1 Estimating Mean Number of Sex Partners

Recent General Social Surveys have asked respondents how many female partners they have had sex with since their 18th birthday. Over half the respondents answered 0, presumably because the question was asked of both the male and female respondents. In 1994, of those 1055 respondents who responded with a number higher than 0, the distribution was highly skewed to the right with a sample mean of 10.2 and standard deviation of 10.1. Let μ denote the mean for the population represented by this sample.

When $s = 10.1$ and $n = 1055$, the estimated standard error of the sampling distribution of $\bar{Y}$ is

$$\hat{\sigma}_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{10.1}{\sqrt{1055}} = .31$$

A 95% confidence interval for μ is

$$\bar{Y} \pm 1.96\hat{\sigma}_{\bar{Y}} = 10.2 \pm 1.96(.31) = 10.2 \pm .6, \quad \text{or } (9.6, 10.8)$$

We can be 95% confident that this interval contains μ, the population mean number of female sex partners. The point estimate of μ is 10.2, and the interval estimate predicts that μ is no smaller than 9.6 and no greater than 10.8.

The survey also asked for the number of male sex partners since the 18th birthday. Of the 1431 subjects responding with a positive number, the mean was 4.8 and the standard deviation was 6.2. You can check that the 95% confidence interval for that population mean equals (4.5, 5.1).

Keep in mind that the error allowed in these intervals refers only to sampling error. Other errors relevant for these parameters include those due to nonresponse (e.g., for the number of female partners, 270 subjects provided no response, 34 responded "don't know," and 23 refused to answer) or measurement error (lying or giving an inaccurate response). Certainly we are suspicious here, since the results are so different for males and females. The inferences may apply to a population that differs somewhat from the one of actual interest. □

Controlling the Confidence Coefficient and Error Probability

The inference just presented had a confidence coefficient of .95. In some applications, a 5% chance of an incorrect prediction is unacceptable. Increasing the chance that the confidence interval contains μ requires a larger confidence coefficient. For instance, one might construct a 99% confidence interval for μ. Now, 99% of a normal distribution occurs within 2.58 standard deviations of the mean, so the probability is .99 that $\bar{Y}$ falls within $2.58\sigma_{\bar{Y}}$ of μ. A 99% confidence interval for μ is $\bar{Y} \pm 2.58\hat{\sigma}_{\bar{Y}}$.

For the data in Example 5.1, the 99% confidence interval for μ is

$$\bar{Y} \pm 2.58\hat{\sigma}_{\bar{Y}} = 10.2 \pm 2.58(.31) = 10.2 \pm .8, \quad \text{or } (9.4, 11.0)$$

Compared to the 95% confidence interval of (9.6, 10.8), this interval estimate is less precise, being wider. This is the sacrifice for greater assurance of a correct inference.

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 μ. A 100% confidence interval for the mean number of female sex partners goes from 0 to infinity. This is not informative, and in practice we settle for a little less than perfection in order to focus more tightly on the true parameter value.

The general form for the large-sample confidence interval for the mean is

$$\bar{Y} \pm z\hat{\sigma}_{\bar{Y}}$$

where z depends on the confidence coefficient. The higher the confidence coefficient, the greater the chance that the confidence interval contains the parameter. High confidence coefficients are used in practice, so that the chance of error is small. The most common confidence level is .95, with .99 used when it is more crucial not to make an error. In summary, we have the following result:

Large-Sample Confidence Interval for μ

A large-sample confidence interval for μ is

$$\bar{Y} \pm z\hat{\sigma}_{\bar{Y}} = \bar{Y} \pm z\left(\frac{s}{\sqrt{n}}\right)$$

The z-value is such that the probability under a normal curve within z standard errors of the mean equals the confidence coefficient. For 95% and 99% confidence intervals, z equals 1.96 and 2.58.

Let's study in greater detail this formula. One multiplies the estimated standard error $\hat{\sigma}_{\bar{Y}}$ by a z-value and then adds and subtract it from $\bar{Y}$. The z-value is such that the probability within z standard errors of the mean of the normal sampling distribution equals the confidence coefficient. For example, let's find z for a 98% confidence interval. When the probability .98 falls within z standard errors of the mean, .02 falls in the two tails and .01 in the right-hand tail. Looking up .01 in the body of Table A, we find $z = 2.33$. A 98% confidence interval equals $\bar{Y} \pm 2.33\hat{\sigma}_{\bar{Y}}$, since the probability equals .98 that $\bar{Y}$ falls within 2.33 standard errors of μ.

The probability that a confidence interval does *not* contain the parameter is called the ***error probability***. This equals 1 minus the confidence coefficient. For confidence coefficient .95, the error probability equals .05. In general, the z-score for a confidence interval is the one for which the error probability falls in the two tails of a normal curve. Half the error probability falls in each tail. For instance, for a 95% confidence interval, the error probability equals .05; the z-score is the one with probability .05/2 = .025 in each tail, which is $z = 1.96$.

Let α denote the error probability. Then, $1 - \alpha$ is the confidence coefficient. For instance, for an error probability of $\alpha = .05$, the confidence coefficient equals $1 - \alpha = .95$. The z-value for the confidence interval is such that the probability is $1 - \alpha$ that $\bar{Y}$ falls within z standard errors of μ. Equivalently, the probability is α that $\bar{Y}$ falls more than z standard errors from μ. The z-value refers to a total probability of α in the two tails of a normal distribution, or $\alpha/2$ in each tail.

In reality, the probability that the confidence interval contains μ is *approximately* equal to the chosen confidence coefficient. The approximation improves for larger samples, as the sampling distribution of $\bar{Y}$ is more closely normal in form and the estimated standard error $\hat{\sigma}_{\bar{Y}}$ gets closer to the true standard error $\sigma_{\bar{Y}}$.

Properties of the Confidence Interval for a Mean

We next study the properties of confidence intervals for means. These properties also apply to confidence intervals for other parameters.

The confidence level associated with confidence intervals has a long-run relative frequency interpretation. The unknown mean μ is a fixed number. A confidence interval constructed from any particular sample either does or does not contain μ. However, 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 μ. This happens because about 95% of the sample means would fall within $1.96\sigma_{\bar{Y}}$ of μ, as does the $\bar{Y}$ in line 1 of Figure 5.2. Saying that a particular interval contains μ with "95% confidence" signifies that *in the long run* 95% of such intervals would contain μ; 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 mean for each and a 95% confidence interval for the population mean. The confidence intervals jump around because $\bar{Y}$ varies from sample to sample, but nine of

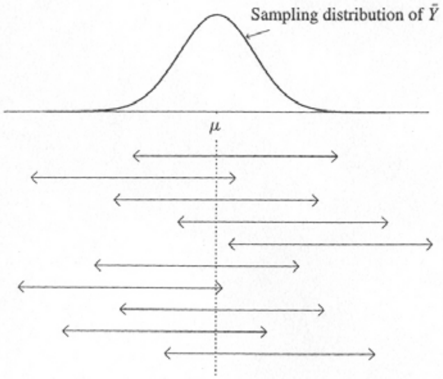

Figure 5.3 Ten 95% Confidence Intervals for μ; In the Long Run, Only 5% of the Intervals Fail to Contain μ

the ten intervals contain the population mean μ. On the average, only about 1 out of 20 times does a 95% confidence interval fail to enclose the population mean.

In practice, of course, we select just *one* sample of some fixed size n and construct one confidence interval using the observations in that sample. We do not know whether any particular 95% confidence interval truly contains μ. Our 95% confidence in that interval is based on long-term properties of the procedure. We can, though, control by our choice of the confidence coefficient the chance that the interval contains μ. If an error probability of .05 makes us nervous, we can instead form a 99% confidence interval.

Unfortunately, the greater the confidence level, the wider the confidence interval. This happens because the z-value in the formula is larger—for instance, $z = 1.96$ for 95% confidence and $z = 2.58$ for 99% confidence. To be more sure of enclosing μ, we must sacrifice precision of estimation by permitting a wider interval. In forming a confidence interval, we must often compromise between the desired precision of estimation and the desired confidence that the inference is correct; as one gets better, the other gets worse. This is why you would not typically see a 99.9999% confidence interval. Although it sounds very safe and nearly error free, it would usually be too wide to tell us much about where the population mean falls (its z-value is 4.9).

Intuitively, one should be able to estimate μ better with a larger sample size. The plus and minus part of a confidence interval is $zs/\sqrt{n}$, which is inversely proportional to the square root of the sample size. The larger the value of n, the narrower is the interval. Thus, one can improve the precision by increasing the sample size.

To illustrate, suppose that $\bar{Y} = 10.2$ and $s = 10.1$ in Example 5.1 were based on a sample of size $n = 4220$, four times the actual sample size of $n = 1055$. Then, the estimated standard error $\hat{\sigma}_{\bar{Y}}$ of the sampling distribution of $\bar{Y}$ is

$$\hat{\sigma}_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{10.1}{\sqrt{4220}} = .155$$

half as large as in that example. The resulting 95% confidence interval is

$$\bar{Y} \pm 1.96\hat{\sigma}_{\bar{Y}} = 10.2 \pm 1.96(.155) = 10.2 \pm .3, \quad \text{or } (9.9, 10.5)$$

This is half as wide as the confidence interval formed from the sample of size $n = 1055$ in Example 5.1. A confidence interval based on $n = 4220$ is half as wide as one based on $n = 1055$.

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

The width of a confidence interval
1. Increases as the confidence coefficient increases.
2. Decreases as the sample size increases.

Some statistical software can calculate confidence intervals for you. All such software reports the basic ingredients you need to construct an interval. For instance, one package reports the sample size, sample mean, sample standard deviation, and estimated standard error for Example 5.1 as:

```
  N          Mean       Std Dev       Std Err
---------------------------------------------
1055       10.233       10.069         0.310
---------------------------------------------
```

5.3 Confidence Interval for a Proportion

The last section dealt with estimating the population mean, a summary parameter for quantitative data. We now present interval estimation for qualitative data, in which each 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 \$0–20,000, \$20,000–40,000, \$40,001–75,000, over \$75,000.

One can summarize categorical data by counting the number of observations in each category. Equivalently, one can record the *proportions* of observations in the categories, which are the counts divided by the total sample size. For example, a study might provide a point or interval estimate of

- The proportion of registered voters who voted in the previous presidential election
- The proportion of Canadians who favor independent status for Quebec
- The proportion of Hispanic adults who have attended college
- The proportion of American families with income below the poverty level

Large-Sample Estimation for a Proportion

Let π denote the parameter representing a population proportion. Then, π falls between 0 and 1 (here, π is *not* the mathematical constant, 3.1415...). The point estimate of the population proportion π is the *sample proportion*. We denote the sample proportion by $\hat{\pi}$, since it estimates π. The sample proportion $\hat{\pi}$ is, in fact, an unbiased and efficient point estimator of π.

Section 4.3 noted that the proportion is a type of mean. Denote an observation by 1 if it falls in the category of interest and by 0 otherwise. Then, the sample mean is the sample proportion $\hat{\pi}$ for that category. Since the sample proportion $\hat{\pi}$ is a sample mean, its large-sample sampling distribution is approximately normal about the parameter π it estimates. Figure 5.4 illustrates.

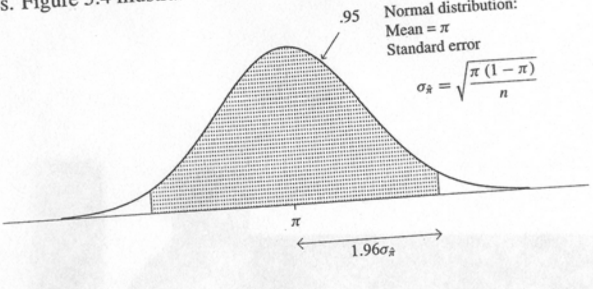

Figure 5.4 Sampling Distribution of $\hat{\pi}$

Similarly, the population proportion π is the mean μ of the probability distribution having probability π for 1 and $(1-\pi)$ for 0. The standard deviation of this probability distribution is $\sigma = \sqrt{\pi(1-\pi)}$. (Problem 4.55 derives this formula.) Since the standard error of a sample mean equals $\sigma_{\bar{Y}} = \sigma/\sqrt{n}$, the formula for the standard error $\sigma_{\hat{\pi}}$ of the sample proportion $\hat{\pi}$ is

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

That is, the standard error $\sigma_{\hat{\pi}}$ of the sampling distribution of $\hat{\pi}$ is a special case of $\sigma_{\bar{Y}}$, the standard error of the sampling distribution of the sample mean $\bar{Y}$. Again, the standard error is inversely proportional to the square root of the sample size. As the sample size increases, the standard error gets smaller, and the sample proportion tends to fall closer to the population proportion.

Like the formula $\sigma_{\bar{Y}} = \sigma/\sqrt{n}$ for the standard error of $\bar{Y}$, the formula for the standard error of $\hat{\pi}$ depends on an unknown parameter, in this case, π. In practice, we estimate this standard error using

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

This estimated standard error appears in confidence intervals. From the same reasoning shown in the previous section for the mean, a 95% confidence interval for π is

$$\hat{\pi} \pm 1.96\hat{\sigma}_{\hat{\pi}} = \hat{\pi} \pm 1.96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

Example 5.2 Estimating Proportion Favoring Legalized Abortion

The 1994 General Social Survey asked respondents, "Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if the woman wants it for any reason." Of 1934 respondents, 895 said yes and 1039 said no. We shall estimate the population proportion that would respond yes to this question.

Let π represent the population proportion that would respond yes. Of the $n = 1934$ respondents, 895 said yes, so $\hat{\pi} = 895/1934 = .46$, and $1 - \hat{\pi} = .54$. That is, 46% of those sampled said yes and 54% of those sampled said no.

The estimated standard error of the estimate $\hat{\pi}$ of π equals

$$\hat{\sigma}_{\hat{\pi}} = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = \sqrt{\frac{(.46)(.54)}{1934}} = \sqrt{.00013} = .011$$

A 95% confidence interval for π is

$$\hat{\pi} \pm 1.96\hat{\sigma}_{\hat{\pi}} = .46 \pm 1.96(.011) = .46 \pm .02, \quad \text{or } (.44, .48)$$

The population percentage that supports unrestricted access to abortion appears to be at least 44% but no more than 48%.

All numbers in the confidence interval (.44, .48) fall below .50. Thus, apparently fewer than half the population supports unrestricted access to abortion. Results in this survey varied greatly depending on the question wording. For instance, when asked whether abortion should be available if the woman becomes pregnant as a result of rape, 1616 said yes and 318 said no; you can check that the 95% confidence interval for the population proportion saying yes equals (.82, .85).

When $n = 1934$ and $\hat{\pi} = .46$, the estimated standard error of $\hat{\pi}$ is

$$\hat{\sigma}_{\hat{\pi}} = \sqrt{\hat{\pi}(1-\hat{\pi})/n} = .011$$

Similarly, the estimated standard error for $1 - \hat{\pi}$, the proportion of voters who say no to legalized abortion, is

$$\hat{\sigma}_{1-\hat{\pi}} = \sqrt{(1-\hat{\pi})\hat{\pi}/n} = \sqrt{(.54)(.46)/1934} = .011$$

A 95% confidence interval for the population proportion of negative responses is

$$.54 \pm 1.96(.011) = .54 \pm .02, \quad \text{or } (.52, .56)$$

Now $.52 = 1 - .48$ and $.56 = 1 - .44$, where (.44, .48) 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. □

Effect of Confidence Coefficient and Sample Size

The formula for the large-sample confidence interval for a proportion is $\hat{\pi} \pm z\hat{\sigma}_{\hat{\pi}}$. The z-value depends on the confidence coefficient in the same way as a confidence interval for the mean μ.

To illustrate, to be more cautious about possibly incorrectly predicting the population proportion favoring unrestricted abortion, we might instead use a 99% confidence interval. This equals

$$\hat{\pi} \pm 2.58\hat{\sigma}_{\hat{\pi}} = .46 \pm 2.58(.011) = .46 \pm .03, \quad \text{or } (.43, .49)$$

The confidence interval is slightly wider, (.43, .49) instead of (.44, .48), as the cost of achieving greater confidence.

Like the width of the confidence interval for a mean, the width of a confidence interval for a proportion depends on the sample size n as well as the confidence coefficient. To illustrate, suppose that 46% of a random sample of size $n = 30$ supported unrestricted abortion. Then $\hat{\sigma}_{\hat{\pi}} = \sqrt{(.46)(.54)/30} = .091$, and a 99% confidence interval for π is

$$\hat{\pi} \pm 2.58\hat{\sigma}_{\hat{\pi}} = .46 \pm 2.58(.091) = .46 \pm .23, \quad \text{or } (.23, .69)$$

In other words, if the sample proportion referred to a sample of size 30 instead of 1934, it would provide a very imprecise prediction of the population proportion. Since the interval contains values both well below and well above .50, it is plausible that a strong majority or that a weak minority of the population would support unrestricted abortion. Our conclusion from such a small sample would be ambiguous, while the conclusion from a sample as large as the General Social Survey provides is much more clear cut.

Summary of Formula and Sample Size Validity

The confidence interval for a proportion, like the one for a mean, applies for large samples. When the proportion is between about .30 and .70, the usual sample size criterion for a mean works fine. That is, one can use the method if n is at least about 30. When the proportion is less than .30 or higher than .70, the sampling distribution is skewed and requires a larger sample size to achieve normality. In this case, there should be at least ten observations both in the category of interest and not in it. When neither of these are satisfied, estimating the proportion is complex, though Problem 5.57 shows a method that usually works quite well. In Example 5.2, the sample proportion is .46 and the sample size is 895 + 1039 = 1934. The sample size requirement is easily satisfied.

We complete this section by summarizing the large-sample confidence interval for a population proportion.

Let α denote the error probability that the interval does not contain the parameter. As in the confidence interval for a mean, the z-value refers to a total probability of α in the two tails, with $\alpha/2$ in each tail.

Large-Sample Confidence Interval for Proportion π

A large-sample confidence interval for a population proportion π, based on a sample proportion $\hat{\pi}$, is

$$\hat{\pi} \pm z\hat{\sigma}_{\hat{\pi}} = \hat{\pi} \pm z\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

The z-value is such that the probability under a normal curve within z standard errors of the mean equals the confidence coefficient. For 95% and 99% confidence intervals, z equals 1.96 and 2.58. The sample size n should exceed 30, with a somewhat larger sample needed if the proportion is relatively small or large—at least ten observations in the category and at least ten not in it.

5.4 Choice of Sample Size

Samples taken by professional polling organizations, such as the Gallup poll, typically contain 1000–2000 subjects. This is large enough to obtain a sample proportion that theoretically falls within about .03 of the population value. These organizations use sampling methods that are usually more complex than simple random samples; however, the formulas for standard errors of estimates under their sampling plans are approximated reasonably well by the ones for random samples.

At first glance, it seems astonishing that a sample on the order of 1000 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. The basis for this inferential power lies in the formulas for the standard errors of the sample point estimates (which actually treat the population size as infinite; see Problem 4.58). As long as the sampling is properly executed, excellent estimates result from relatively small samples, no matter how large the population size.

Before data collection begins, most studies attempt to determine the size of the sample needed to achieve a certain degree of accuracy in estimation. A relevant measure is the value of n for which a confidence interval for the parameter is no wider than some specified width. This section studies sample size determination for estimating a mean or proportion with random samples. We use the facts that (1) the width of the confidence interval depends directly on the standard error of the sampling distribution of the estimator and (2) the standard error itself depends on the sample size.

Sample Size for Estimating Proportions

Before computing the sample size, we must first decide on the degree of *precision* desired, that is, how close the estimate should fall to the parameter. In some studies, highly precise estimation is not as important as in others. A study conducted to estimate the proportions of voters who intend to vote for each candidate in a close election requires an accurate 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 rural origins, a larger margin of error might be acceptable. So, we must first decide whether the error should be no more than .04 (four percentage points), .05, .10, or whatever.

Second, we must set the *probability* with which the specified precision is achieved. For instance, we might decide that the error in estimating a population proportion should not exceed .04, with .95 probability. This probability must be stated, since with any sample size one can have an error of no more than .04 with *some* probability, though perhaps a very small one.

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

Example 5.3 Sample Size for a Survey on Single-Parent Children

A group of social scientists wanted to estimate the proportion of school children in Boston who were living with only one parent. Since their report was to be published, they wanted a reasonably accurate estimate. However, since their funding was limited, they did not want to collect a larger sample than necessary. They decided to use a sample size such that, with probability .95, the error would not exceed .04. In other words, they wanted the sample proportion to fall within .04 of the true value, with probability .95. Thus, they wanted to determine n such that a 95% confidence interval for π equals $\hat{\pi} \pm .04$.

Since the sampling distribution of the sample proportion $\hat{\pi}$ is approximately normal, the sample proportion $\hat{\pi}$ falls within $1.96\sigma_{\hat{\pi}}$ of π with probability .95. Thus, if the sample size is such that $1.96\sigma_{\hat{\pi}} = .04$, then with probability .95, $\hat{\pi}$ falls within .04 units of π and the error of estimation does not exceed .04. See Figure 5.5.

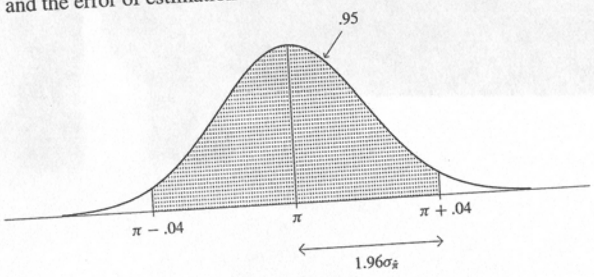

Figure 5.5 Sampling Distribution of $\hat{\pi}$ with the Error of Estimation No Greater than .04, with Probability .95

We must solve algebraically for the value of n that provides a value of $\sigma_{\hat{\pi}}$ for which $.04 = 1.96\sigma_{\hat{\pi}}$; that is, we must solve for n in the expression

$$.04 = 1.96\sqrt{\frac{\pi(1-\pi)}{n}}$$

Multiplying both sides of the expression by $\sqrt{n}$ and dividing both sides by .04, we get

$$\sqrt{n} = 1.96\frac{\sqrt{\pi(1-\pi)}}{.04}$$

Squaring both sides, we obtain the formula

$$n = \frac{(1.96)^2\pi(1-\pi)}{(.04)^2}$$

Now, we face a problem. We want to select n for the purpose of estimating the parameter π, but this formula requires the value of π. This is because the spread of the sampling distribution depends on the value of 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 .5. Since π is unknown, we must substitute an educated guess for it in this equation to obtain a numerical solution for n.

Alternatively, the largest possible value for $\pi(1-\pi)$ is .25, which occurs when $\pi = .5$. In fact, $\pi(1-\pi)$ is fairly close to .25 unless π is quite far from .5. For example, $\pi(1-\pi) = .24$ when $\pi = .4$ or $\pi = .6$, and $\pi(1-\pi) = .21$ when $\pi = .7$ or $\pi = .3$. Thus, a sample of size

$$n = \frac{(1.96)^2(.25)}{(.04)^2} = 600$$

ensures that the error will not exceed .04, with a probability of *at least* .95, no matter what the value of π. □

Obtaining n by setting $\pi(1-\pi) = .25$ is the safe and cautious approach. This n value is excessively large if π is not close to .5. Suppose, for example, that based on other studies, the social scientists believed that the proportion π of school children in Boston who were living with only one parent was no more than .25. Then an adequate sample size is

$$n = \frac{(1.96)^2\pi(1-\pi)}{(.04)^2} = \frac{(1.96)^2(.25)(.75)}{(.04)^2} = 450$$

A sample size of 600 would be larger than needed. With it, the probability would actually exceed .95 that the sample proportion falls within .04 of the true proportion.

We next provide a general formula for sample size. Let B denote the desired bound on the error (B = bound). This is the maximum distance preferred between the sample proportion and the true value, which is $B = .04$ in the example. The formula also uses a general z-value (in place of 1.96) determined by the probability with which the error is no greater than B.

Sample Size Required for Estimating a Proportion π

Let B denote the chosen bound on error. The sample size n ensuring that, with fixed probability, the error of estimation of π by the sample proportion $\hat{\pi}$ is no greater than B, is

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

The z-score is the one for a confidence interval with confidence coefficient equal to the fixed probability; for instance, $z = 1.96$ for probability .95 and $z = 2.58$ for probability .99. Using this formula requires guessing π or taking the safe but conservative approach of setting $\pi(1-\pi) = .25$.

To illustrate, suppose the study about single-parent children wanted to estimate the proportion to within .08 with a probability of at least .95. Then the bound on error equals $B = .08$, and $z = 1.96$, the z-value for a 95% confidence interval. The required sample size using the safe approach is

$$n = .25\left(\frac{z}{B}\right)^2 = .25\left(\frac{1.96}{.08}\right)^2 = 150$$

This sample size of 150 is one-fourth the sample size of 600 necessary to guarantee a 95% confidence bound of $B = .04$. Reducing the bound on error by a factor of one-half requires quadrupling the sample size.

Sample Size for Estimating Means

We next present an analogous result for quantitative data and estimating a population mean. Let μ and σ denote the population mean and standard deviation for the variable of interest. Figure 5.6 illustrates the basic problem. We want to determine how large n needs to be so that the sampling distribution of $\bar{Y}$ is sufficiently narrow that $\bar{Y}$ is very likely to fall within B units of μ. A derivation using this sampling distribution yields the following result:

Sample Size Required for Estimating a Mean μ

Let B denote the desired bound on error. The sample size n ensuring that, with fixed probability, the error of estimation of μ by $\bar{Y}$ is no greater than B, is

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

The z-score is the one for a confidence interval with confidence coefficient equal to the fixed probability.

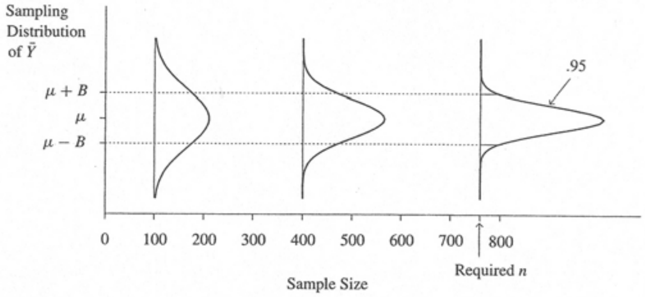

Figure 5.6 Determining n So That $\bar{Y}$ Has Probability .95 of Falling Within B Units of μ

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

Example 5.4 Estimating Mean Educational Level 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 .99?

Suppose the study has no prior information about the standard deviation of educational attainment for Native Americans. As a crude approximation, they might guess that 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 crude guess for σ. This seems plausible, since it means that about 68% of the education values would fall within 2.5 years of the mean, or within a span of 5 years.

Now, for 99% confidence, the z-score is the one with probability $.01/2 = .005$ in each tail, or $z = 2.58$. Since the desired bound on error equals $B = 1$ year, the required sample size is

$$n = \sigma^2\left(\frac{z}{B}\right)^2 = (2.5)^2\left(\frac{2.58}{1}\right)^2 = 42 \text{ subjects}$$

A more cautious approach would select for σ a number quite sure to be an upper bound for its value. For example, it is reasonable to predict that σ is no greater than 3.5, since a range of six standard deviations then extends from 0 to 21. This yields

$n = (3.5)^2(2.58/1)^2 = 81$ families. Then, if σ is actually less than 3.5, the estimate $\bar{Y}$ will fall within 1 of μ with probability even greater than .99. □

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 usually be smaller. In such cases, determination of sample size is complex, and you should seek guidance from a statistical consultant.

Other Considerations in Determining Sample Size

From a practical point of view, determining sample size is not a simple matter of plugging numbers into a formula. Several other considerations affect the number of observations needed in a study. We have just discussed two, *precision* and *confidence*. Precision refers to the width of a confidence interval, while confidence refers to the probability that the interval actually contains the estimated parameter.

A third characteristic affecting the sample size decision is the *variability* in the population for the variables measured. We have already seen this for estimating means, where the required sample size increases as σ increases. The more heterogeneous the population, the larger the sample needs to be. In the extreme case in which all population elements are alike (zero variability), a sample size of 1 can accurately represent the population. On the other hand, if there are 15 ethnic groups, age variation from 18 to 85, and wide variation in income, we would need a large sample to reflect accurately the variation in these variables. In most social surveys, large samples (1000 or more) are necessary, while for more homogeneous populations (e.g., residents of nursing homes) smaller samples are often adequate, due to reduced population variability.

A fourth consideration is the *complexity of analysis* planned. The more complex the analysis, such as the more variables one analyzes simultaneously, the larger the sample needed to make an adequate analysis. If one is to analyze a single variable using a simple measure such as a mean, a relatively small sample might be adequate. On the other hand, planned comparisons of several groups using complex multivariate methods require a larger sample. For instance, Example 5.4 showed that one can estimate mean educational attainment quite well using a sample of only 42 people. On the other hand, if one 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, IQ, and size of the community, a much larger sample would be needed, probably a thousand or more. One reason for the increase in the typical sample size of studies in recent years is the greater complexity of statistical analyses used in social science research.

Finally, a fifth consideration concerns time, money, and other *resources*. Larger samples are more expensive and more time consuming, and may require more resources than the study has available. Time, cost, and resource limitations are often the major constraints on sample size. For example, sample size formulas might suggest that 1000 cases provide the desired accuracy. Perhaps, however, we can afford to gather only 500. Should we go ahead with the smaller sample and sacrifice precision and/or confidence, or should we give up unless we find additional resources? We often must face such questions as "Is it better to have some knowledge that is not very precise, or no knowledge at all?" or "Is it really crucial to study all population groups, or can I reduce the sample by focusing on some subsets?" The costs and benefits of large samples must be weighed against the importance of the study, the need for accuracy, and the complexity of the problem and statistical analysis.

In summary, no simple formula can always determine the proper sample size. While sample size is an important matter, its choice depends on an assessment of needs and resources and requires careful judgment.

A final caveat: When we say that a sample of size 600 is adequate for estimating a proportion to within .04 with .95 confidence, this precision of .04 is a theoretical target that takes into account only sampling error. Practical problems often imply that the actual accuracy is somewhat less. If the study is carried out poorly, or if data are never obtained for a substantial percentage of the target sample, or if some subjects in the study 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 .04 may be substantially less than .95. When someone claims to achieve a certain accuracy in estimating a parameter, always be skeptical unless you know that the study was substantially free of such problems.

5.5 Confidence Intervals for a Median*

The past two chapters have emphasized sampling properties of the sample mean. Chapter 3 showed, though, that other statistics are also useful for describing data. These other statistics also have sampling distributions. Moreover, for large random samples, their sampling distributions are usually approximately normal. One can use sample data to form confidence intervals for population values of the measures. We illustrate in this section for the median.

Inefficiency of Median for Normal Data

Let M denote the sample median. When the population distribution is normal and the sample is random, the standard error of M has formula similar to the one for the sample mean. Namely, the standard error equals $1.25\sigma/\sqrt{n}$. A large-sample confidence interval for the population median then has form

$$M \pm z\frac{(1.25s)}{\sqrt{n}}$$

where the z-score depends on the confidence coefficient in the usual way.

Since the population median for a normal distribution equals the population mean μ, the sample median and sample mean are both point estimates of the same number.

Chapter 6

Statistical Inference: Significance Tests

A common aim in many studies is to check whether the data agree with certain predictions. These predictions are *hypotheses* about variables measured in the study.

Hypothesis

A ***hypothesis*** is a statement about some characteristic of a variable or a collection of variables.

Hypotheses arise from the theory that drives the research. When a hypothesis relates to characteristics of a population, such as population parameters, one can use statistical methods with sample data to test its validity. Examples of hypotheses that might be tested statistically are the following: "A majority of Canadians are satisfied with their national health service," "The mean age at marriage for men in colonial America was the same in rural and urban areas," "For workers in service jobs, the mean income is lower for women than for men," and "There is a difference between Democrats and Republicans in the probabilities that they vote with their party leadership."

A ***significance test*** is a way of statistically testing a hypothesis by comparing the data to values predicted by the hypothesis. Data that fall far from the predicted values provide evidence against the hypothesis. The following example illustrates ideas behind significance tests.

Example 6.1 Testing for Gender Bias in Selecting Managers

A large supermarket chain in Florida occasionally selects some of its employees to receive management training. A group of women employees recently claimed that males are picked at a disproportionally high rate for such training. The company denied this claim (*Tampa Tribune*, April 6, 1996).

Let's consider how the women employees could statistically back up their assertion. Suppose the employee pool for potential selection for management training is half male and half female. Then, the company's claim of a lack of gender bias is a hypothesis. It states that, other things being equal, at each choice the probability of selecting a female equals 1/2 and the probability of selecting a male equals 1/2. If the employees truly are selected for management training randomly in terms of gender, about half the employees picked should be females and about half should be male. The women's claim is an alternative hypothesis that the probability of selecting a male exceeds 1/2.

Since this program began, suppose that nine of the ten employees chosen for management training have been male. Based on this evidence, we might be inclined to support the women's claim. However, we should check first to see if these results would be unlikely, if there were no gender bias. Would it be highly unlikely that at least nine of the ten employees chosen would have the same gender, if they were truly selected at random from the employee pool? Due to sampling variation, it need not happen that exactly 50% of the ten people in the sample are male. We need guidelines about how large the percentage of males must be before we can support the women's hypothesis. □

This chapter introduces statistical methods for obtaining evidence and making decisions about hypotheses. The process is statistical in the sense that it uses sample data to make inferences. In doing so, it can control the probability of an incorrect decision.

The first section of the chapter describes the elements of a significance test. The remainder of the chapter deals with significance tests about a population mean μ or a population proportion π. Sections 6.2 and 6.3 discuss the large-sample case, and Sections 6.5 and 6.6 present small-sample significance tests. Sections 6.4 and 6.7 show how to control the probability of an incorrect decision.

6.1 Elements of a Significance Test

Now let's take a closer look at what we mean by a significance test. All significance tests have five elements: assumptions, hypotheses, test statistic, P-value, and conclusion.

Assumptions

All significance tests require certain assumptions for the tests to be valid. These assumptions refer to

- The ***type of data***: Like other statistical methods, each test applies for either quantitative data or qualitative data.
- The form of the ***population distribution***: For some tests, the variable must have a particular form of distribution, such as the normal. This is primarily true for small-sample tests.
- The ***method of sampling***: The tests presented in this book require simple random sampling.
- The ***sample size***: The validity of many tests improves as the sample size increases. These tests require a certain minimum sample size for the analyses to work well.

Hypotheses

A significance test considers two hypotheses about the value of a parameter.

Null Hypothesis, Alternative Hypothesis

The ***null hypothesis*** is the hypothesis that is directly tested. This is usually a statement that the parameter has value corresponding to, in some sense, *no effect*. The ***alternative hypothesis*** is a hypothesis that contradicts the null hypothesis. This hypothesis states that the parameter falls in some alternative set of values to what the null hypothesis specifies.

Notation for Hypotheses

The symbol H_0 represents the null hypothesis, and the symbol H_a represents the alternative hypothesis.

A significance test analyzes the strength of sample evidence against the null hypothesis. The test is conducted to investigate whether the data contradict the null hypothesis, hence suggesting that the alternative hypothesis is true. The approach taken is the indirect one of *proof by contradiction*. The alternative hypothesis is judged acceptable if the sample data are inconsistent with the null hypothesis. In other words, the alternative hypothesis is supported if the null hypothesis appears to be incorrect.

The researcher usually conducts the test to gauge the amount of support for the alternative hypothesis. Thus, the alternative hypothesis is often called the ***research hypothesis***. The hypotheses are formulated *before* collecting or analyzing the data.

To illustrate, we refer to Example 6.1 about possible gender discrimination in the selection of employees of a supermarket chain for management training. The company claims that the probability that any given employee selected is male equals 1/2. This is an example of a null hypothesis, *no effect* referring to a lack of gender bias. The alternative hypothesis reflects the skeptical women employees' belief that this probability actually exceeds 1/2. We conduct the test by checking whether the sample data are inconsistent with the null hypothesis probability value of 1/2.

Test Statistic

The ***test statistic*** is a statistic calculated from the sample data to test the null hypothesis. This statistic typically involves a point estimate of the parameter to which the hypotheses refer.

For instance, to test a hypothesis about an unknown probability, one could use as test statistic the sample estimator of that probability. If nine out of ten selected trainees are male, the estimator is the sample proportion, 9/10 = .90.

P-Value

Using the sampling distribution of the test statistic, we calculate the probability that values of the statistic like the one observed would occur if the null hypothesis were true. This provides a measure of how unusual the observed test statistic value is compared to what H_0 predicts.

Specifically, we consider the set of possible test statistic values that provide *at least as much evidence* against the null hypothesis as the observed test statistic. This set is formed with reference to the alternative hypothesis; the values providing stronger evidence *against* the null hypothesis are those providing stronger evidence *in favor of* the alternative hypothesis. The ***P-value*** is the probability, if H_0 were true, that the test statistic would fall in this collection of values.

***P*-Value**

The ***P-value*** is the probability, when H_0 is true, of a test statistic value at least as contradictory to H_0 as the value actually observed. The smaller the P-value, the more strongly the data contradict H_0. The P-value is denoted by P.

The P-value summarizes the evidence in the data about the null hypothesis. A moderate to large P-value means that the data are consistent with H_0. For instance, a P-value such as .26 or .83 indicates that the observed data would not be unusual if H_0 were true. On the other hand, a P-value such as .001 means that such data would be very unlikely, if H_0 were true. This provides strong evidence against H_0.

For the gender bias example, the alternative hypothesis states that the probability of selecting a male for the managerial track exceeds 1/2. The test statistic is the sample proportion of males in the ten trainees selected; the observed value equals 9/10 = .90. The values of this test statistic providing this much or even stronger evidence against the null hypothesis and in favor of the alternative hypothesis are sample proportion values of .90 and higher. A formula from Section 6.6 calculates this probability as .011. Thus, the P-value equals $P = .011$, as shown in Figure 6.1. If the selections truly are random with respect to gender, the chance is only .011 of such an extreme sample result, namely, that nine or all ten selections for management training would be males. Other things being equal, this small P-value provides considerable, though not overwhelming, evidence of gender bias.

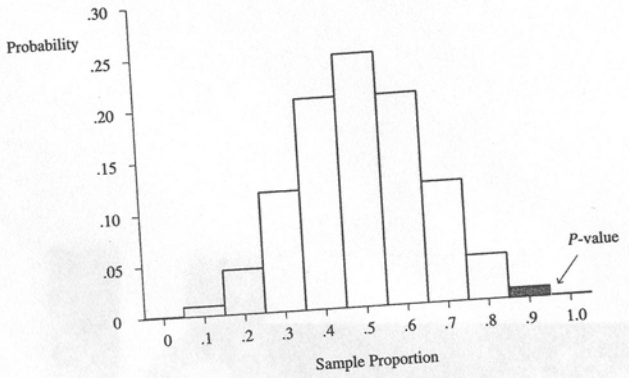

Figure 6.1 The P-value Refers to the Probability of the Observed Data or Even More Extreme Results

Generally, when a null hypothesis is true, the P-value is equally likely to fall anywhere between 0 and 1; for instance, it has a .01 chance of falling between 0.0 and .01, a .01 chance of falling between .01 and .02, a .01 chance of falling between .02 and .03, and so forth up to a .01 chance of falling between .99 and 1.00. In that case, the P-value tends to vary around an expected value of .50. By contrast, when H_0 is false, the P-value is more likely to be close to 0 than close to 1.

Conclusion

The P-value is the primary reported result of a significance test. An observer of the test results can then judge the extent of the evidence against H_0. Sometimes it is necessary to make, in addition, a formal decision about the validity of H_0. If the P-value is sufficiently small, one rejects H_0 and accepts H_a. In either case, the conclusion should include an *interpretation* of what the P-value or decision about H_0 tells us about the original question motivating the test.

Most studies require very small P-values, such as $P \leq .05$, before concluding that the data sufficiently contradict H_0 to reject it. In such cases, results are said to be *significant at the .05 level*. This means that if the null hypothesis were true, the chance of getting such extreme results as in the sample data would be no greater than 5%.

The process of making a formal decision by rejecting or not rejecting a null hypothesis is an optional part of the significance test. We defer further discussion of it until Section 6.4. Table 6.1 summarizes the elements of a significance test.

TABLE 6.1 The Five Elements of a Statistical Significance Test

1. Assumptions
   Type of data, form of population, method of sampling, sample size
2. Hypotheses
   Null hypothesis, H_0 (parameter value for "no effect")
   Alternative hypothesis, H_a (alternative parameter values)
3. Test statistic
   Compares point estimate to null hypothesized parameter value
4. P-value
   Weight of evidence about H_0; smaller P is more contradictory
5. Conclusion
   Report P-value
   Formal decision (optional; see Section 6.4)

6.2 Significance Test for a Mean

We now present a significance test about the population mean μ for quantitative variables. This test assumes that the sample size n is at least 30. It uses the fact that, for large random samples, the sampling distribution of the sample mean $\bar{Y}$ is approximately normal, no matter what distribution the variable has. The five elements of the significance test follow:

Elements of a Large-Sample Test for a Mean

1. *Assumptions*

The test requires a random sample of size $n \geq 30$. The variable measured is quantitative, and the test refers to the population mean of the variable, μ.

2. *Hypotheses*

The null hypothesis has form

$$H_0 : \mu = \mu_0$$

where μ_0 is some particular number. In other words, the hypothesized value of μ in H_0 is a single value. This usually refers to *no effect* or *no change* compared to some standard.

The alternative hypothesis refers to alternative parameter values from the one in the null hypothesis. The most common form of alternative hypothesis is

$$H_a : \mu \neq \mu_0$$

This alternative hypothesis is called ***two-sided***, since it includes values falling both below and above the value μ_0 listed in H_0.

The hypotheses $H_0 : \mu = 0$ and $H_a : \mu \neq 0$ illustrate these forms. The null hypothesis states that the population mean equals 0, and the alternative hypothesis states that the population mean equals some value other than 0.

3. Test Statistic

The sample mean $\bar{Y}$ estimates the population mean. When $n \geq 30$, the sampling distribution of $\bar{Y}$ is approximately normal about μ, with standard error $\sigma_{\bar{Y}} = \sigma/\sqrt{n}$. If H_0: $\mu = \mu_0$ is true, then the center of the sampling distribution is the number μ_0, as shown in Figure 6.2. The evidence about H_0 is the distance of the sample value $\bar{Y}$ from the null hypothesis value μ_0, relative to the standard error. A value of $\bar{Y}$ falling far out in the tail of this sampling distribution casts doubt on the validity of H_0, because it would be unlikely to observe $\bar{Y}$ very far from μ_0 if truly $\mu = \mu_0$.

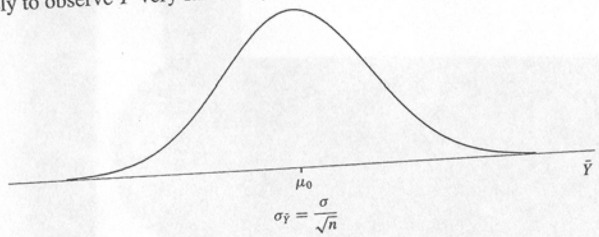

Figure 6.2 Sampling Distribution of $\bar{Y}$ if $H_0 : \mu = \mu_0$ Is True. For large random samples, it is approximately normal, centered at the null hypothesis value, μ_0.

The test statistic is the z-score

$$z = \frac{\bar{Y} - \mu_0}{\hat{\sigma}_{\bar{Y}}} = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}}$$

As in Chapter 5, we substitute the sample standard deviation s for σ to get an estimated standard error, $\hat{\sigma}_{\bar{Y}} = s/\sqrt{n}$. The test statistic counts the number of estimated standard errors that $\bar{Y}$ falls from the hypothesized value μ_0. When H_0 is true, the sampling distribution of this test statistic is approximately the *standard normal* distribution; that is, normal with mean equal to 0 and standard deviation equal to 1, as presented in Section 4.2. The farther $\bar{Y}$ falls from μ_0, the larger the absolute value of the z test statistic. Hence, the larger the value of $|z|$, the stronger the evidence against H_0.

One reason for placing a single number μ_0 in the null hypothesis H_0 should now be apparent. The calculation of the test statistic, and hence the result of the test, refers to that one value.

4. P-Value

The test statistic summarizes the sample evidence. Different tests use different test statistics, though, and it is easier to interpret the test statistic by transforming it to the probability scale of 0 to 1. The P-value does this. It describes whether the observed test statistic value is consistent with the null hypothesis, small values of P indicating inconsistency.

We calculate the P-value under the assumption that H_0 is true. That is, we give the benefit of the doubt to the null hypothesis, analyzing how likely the observed data would be if that hypothesis were true. For the alternative hypothesis H_a: $\mu \neq \mu_0$, the P-value is the probability that the z test statistic is at least as large in absolute value as the observed test statistic. This means that P is the probability of a $\bar{Y}$ value at least as far from μ_0 *in either direction* as the observed value of $\bar{Y}$. The P-value refers to the probability of the observed result or any other result that provides even stronger evidence against the null hypothesis.

A z test statistic value of 0 results when $\bar{Y} = \mu_0$. This is the z-value most consistent with H_0. The P-value is the probability of a z test statistic value at least as far from this consistent value as the one observed. In other words, P is the probability of those $\bar{Y}$ values that are at least as contradictory to $H_0 : \mu = \mu_0$ and at least as favorable to H_a: $\mu \neq \mu_0$ as the observed $\bar{Y}$, that is, at least as many standard errors distant from μ_0.

Figure 6.3 shows the sampling distribution of the z test statistic when H_0 is true. To illustrate the calculation of P, suppose $z = -1.5$. This is the z-score resulting from a sample mean $\bar{Y}$ that is 1.5 standard errors below μ_0. The P-value is the probability that $z \geq 1.5$ or $z \leq -1.5$ (i.e., $|z| \geq 1.5$). From Table A, the probability in one tail above $z = +1.5$ is .0668, so the probability in both tails, beyond $|z| = 1.5$, equals $2(.0668) = .1336$. This is the probability that the sample mean falls at least 1.5 standard errors from the true mean.

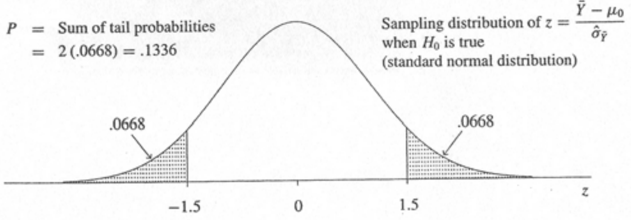

Figure 6.3 Calculation of P When $z = -1.5$, for Testing H_0: $\mu = \mu_0$ Against H_a: $\mu \neq \mu_0$. The P-Value is the two-tail probability of a more extreme result than the observed one.

One should round the calculated P-value such as .1336 to .134 or .13 before reporting it. Reporting the P-value as .1336 makes it seem as if more accuracy exists than actually does, since the sampling distribution is only *approximately* normal.

5. Conclusion

Finally, the study should report the P-value, so others can view the strength of evidence. The smaller P is, the stronger the evidence against H_0 and in favor of H_a.

Example 6.2 Political Conservatism and Liberalism

Many political commentators have remarked that since the Reagan presidential years, there has been an upsurge of political conservatism. One way to summarize political ideology in the United States is to analyze results of various items on the General Social Survey. For instance, every year that survey asks, "I'm going to show you a seven-point scale on which the political views that people might hold are arranged from extremely liberal, point 1, to extremely conservative, point 7. Where would you place yourself on this scale?" Table 6.2 shows the seven-point scale and the distribution of 627 responses among the levels for a recent survey.

TABLE 6.2 Responses of 627 Subjects to a Seven-Point Scale of Political Ideology

Response	Count
1. Extremely liberal	12
2. Liberal	66
3. Slightly liberal	109
4. Moderate, middle of road	239
5. Slightly conservative	116
6. Conservative	74
7. Extremely conservative	11

Political ideology is an ordinal scale. In some cases, the main interest in such a scale may refer to category proportions; for instance, is the population proportion who are extremely liberal different from the population proportion who are extremely conservative? More commonly, such scales are treated in a quantitative manner by assigning scores to the categories. One can then summarize responses by quantitative measures such as means, allowing us to detect the extent to which observations tend to gravitate toward the conservative or the liberal end of the scale.

If we assign the category scores shown in Table 6.2, then a mean below 4 shows a propensity toward liberalism, and a mean above 4 shows a propensity toward conservatism. We can test whether these data show much evidence of either of these by conducting a significance test about how the population mean compares to the middle value of 4.

1. *Assumptions*: The sample is randomly selected and the sample size exceeds 30, so these assumptions for a large-sample test about a mean are satisfied. We are treating political ideology as quantitative with equally-spaced scores.
2. *Hypotheses*: Let μ denote the population mean ideology, for this seven-point scale. The null hypothesis contains one specified value for μ. Since we conduct the analysis to check how, if at all, the population mean departs from the moderate response of 4, the null hypothesis is

$$H_0 : \mu = 4.0$$

The alternative hypothesis is then

$$H_a : \mu \neq 4.0$$

The null hypothesis states that, on the average, the population response is politically "moderate, middle of road." The alternative states that the mean falls in the liberal direction ($\mu < 4$) or in the conservative direction ($\mu > 4$).

3. *Test statistic*: The 627 responses in Table 6.2 are summarized by $\bar{Y} = 4.032$ and $s = 1.257$. The estimated standard error of the sampling distribution of $\bar{Y}$ is

$$\hat{\sigma}_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{1.257}{\sqrt{627}} = .050$$

The value of the test statistic is, therefore,

$$z = \frac{\bar{Y} - \mu_0}{\hat{\sigma}_{\bar{Y}}} = \frac{4.032 - 4.0}{.050} = .64$$

The sample mean falls .64 estimated standard errors above the null hypothesis value of the mean.

4. *P-value*: The P-value is the two-tail probability that z would exceed .64 in absolute value, if H_0 were true. Figure 6.4 portrays the P-value. From the normal probability table (Table A), this two-tail probability equals $P = 2(.2611) = .52$. If the population mean ideology were 4.0, then the probability equals .52 that a sample mean for $n = 627$ subjects would fall at least as far from 4.0 as the observed $\bar{Y}$ of 4.032. That is, P is the probability that $\bar{Y}$ is at least as contradictory to H_0 as the observed $\bar{Y}$.
5. *Conclusion*: To summarize the evidence about the null hypothesis, we report the P-value of $P = .52$. This value is not small, so it does not contradict the null hypothesis. If H_0 were true, the data we observed would not be unusual. It is plausible that the population mean response is 4.0, showing no tendency in the conservative or liberal direction. Generally, researchers do not regard the evidence against H_0 as strong unless P is very small, say, $P < .05$ or $P < .01$.

□

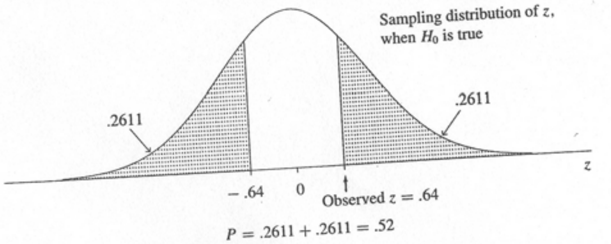

Figure 6.4 Calculation of P-value for Example 6.2. For two-sided alternatives, P is a two-tail probability.

Effect of Sample Size on *P*-values

In Example 6.2, suppose $\bar{Y} = 4.032$ and $s = 1.257$ were based on a sample of size $n = 6270$ instead of $n = 627$. The standard error then decreases, with

$$\hat{\sigma}_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{1.257}{\sqrt{6270}} = .0159$$

The test statistic increases to

$$z = \frac{\bar{Y} - \mu_0}{\hat{\sigma}_{\bar{Y}}} = \frac{4.032 - 4.0}{.0159} = 2.01$$

This has P-value equal to $P = 2(.0222) = .044$. The same difference between $\bar{Y}$ and μ_0 based on a larger sample size results in a smaller P-value. Naturally, the larger the sample size, the more certain we can be that sample deviations from H_0 are indicative of true population deviations. For a given size of effect, smaller P-values result from larger sample sizes. In particular, notice that even a small departure of the sample mean from the value in the null hypothesis can yield a small P-value if the sample size is large.

A related phenomenon holds with estimation methods. As n increases, confidence intervals for means decrease in width, leading to improved precision in estimating μ.

Correspondence Between Results of Tests and Confidence Intervals

Conclusions using significance tests are consistent with conclusions using confidence intervals. If a test says that a particular value is plausible for the parameter, then so does a confidence interval.

Example 6.3 Confidence Interval for Mean Political Ideology

An alternative inferential approach to the significance test in Example 6.2 constructs a confidence interval for the population mean political ideology. Since $\bar{Y} = 4.032$ and $\hat{\sigma}_{\bar{Y}} = .050$, a 95% confidence interval for μ is

$$\bar{Y} \pm 1.96\hat{\sigma}_{\bar{Y}} = 4.032 \pm 1.96(.050) = 4.03 \pm .10, \quad \text{or } (3.93, 4.13)$$

At the 95% confidence level, these are the plausible values for μ.

This confidence interval indicates that 4.0 is a plausible value for μ, since it falls inside the confidence interval. Thus, it is not surprising that the P-value ($P = .52$) in testing H_0: $\mu = 4.0$ against H_a: $\mu \neq 4.0$ in Example 6.2 was not small. In fact, whenever $P > .05$ in a test of H_0: $\mu = \mu_0$ against H_a: $\mu \neq \mu_0$, a 95% confidence interval for μ necessarily contains the null hypothesis value μ_0 of μ. Similarly, suppose that a confidence interval suggests that a particular number is implausible for μ, that number falling outside the confidence interval. Then, a small P-value results from testing the null hypothesis that μ equals that number. In this sense, results of confidence intervals and of two-sided significance tests are consistent. Section 6.4 discusses further the connection between the two methods. □

One-Sided Significance Tests

Two other forms of alternative hypotheses are sometimes used. They have the directional form

$$H_a : \mu > \mu_0 \quad \text{and} \quad H_a : \mu < \mu_0$$

The alternative hypothesis H_a: $\mu > \mu_0$ applies when the purpose of the test is to detect whether μ is *larger* than the particular number μ_0, whereas H_a: $\mu < \mu_0$ refers to detecting whether μ is *smaller* than that value.

The alternative hypotheses H_a: $\mu > \mu_0$ and H_a: $\mu < \mu_0$ are called ***one-sided***. They apply when the researcher predicts a deviation from H_0 in a particular direction. By contrast, the two-sided alternative H_a: $\mu \neq \mu_0$ applies when the researcher wishes to detect *any* type of deviation of μ from μ_0. This choice is made before analyzing the data.

For the one-sided alternative $H_a : \mu > \mu_0$, P is the probability of a z-score above the observed z-score (i.e., to the right of it on the real number line) when H_0 is true. Equivalently, P is the probability of a sample mean above the observed value of $\bar{Y}$. These $\bar{Y}$ values are the ones that provide at least as much evidence in favor of $H_a : \mu > \mu_0$ as the observed value. So, P equals the tail probability to the right of the observed z-score under the standard normal curve, as Figure 6.5 portrays. A z-score of .64 results in $P = .26$ for this alternative.

For H_a: $\mu < \mu_0$, P is the tail probability to the left of the observed z-score under the standard normal curve. A z-score of $-.64$ results in $P = .26$ for this alternative, and a z-score of .64 results in $P = 1 - .26 = .74$.

Example 6.4 Mean Weight Change in Anorexic Girls

This example refers to a study that compared various treatments for young girls suffering from anorexia. (The data, courtesy of Prof. Brian Everitt, Institute of Psychiatry, London, are shown in Table 12.19 in Chapter 12, where they are analyzed more fully.)

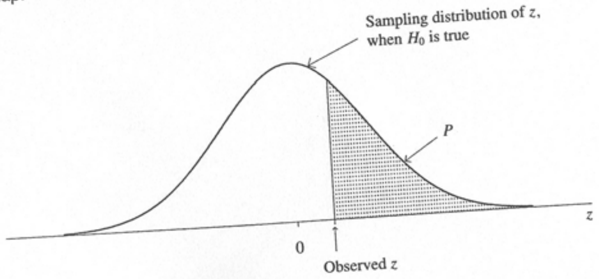

Figure 6.5 Calculation of P-value in Testing H_0: $\mu = \mu_0$ Against H_a: $\mu > \mu_0$. The P-value is the probability of values to the right of the observed test statistic.

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.

Let μ denote the population mean change in weight for the cognitive behavioral treatment, for the population represented by this sample of girls. If this treatment has beneficial effect, as expected, then μ is positive. To test for no effect of treatment versus a positive mean weight change, we test H_0: $\mu = 0$ against H_a: $\mu > 0$.

Software (SPSS) used to analyze the data reports the summary results:

```
                Number
Variable       of Cases      Mean        SD       SE of Mean
CHANGE            29         3.007      7.309       1.357
```

Thus, $n = 29$ girls received this treatment, and the mean weight change was $\bar{Y} = 3.01$ pounds and the sample standard deviation (SD) was $s = 7.31$. The sample mean had an estimated standard error (SE) of $\hat{\sigma}_{\bar{Y}} = s/\sqrt{n} = 7.31/\sqrt{29} = 1.357$. The test statistic equals

$$z = \frac{\bar{Y} - \mu_0}{\hat{\sigma}_{\bar{Y}}} = \frac{3.01 - 0}{1.36} = 2.22$$

From Table A, the one-tail P-value above 2.22 equals .0132. A P-value of .013 provides relatively strong evidence against the null hypothesis and in favor of the alternative that the mean change in weight is positive.

The sample size here ($n = 29$) is borderline for using large-sample methods. Section 6.5 shows that the same conclusion results from small-sample methods. □

Implicit One-Sided Null Hypotheses

Example 6.4 showed that if $\mu = 0$, then the probability equals .013 of observing a sample mean weight gain of 3.01 or greater for a sample of size 29. Now, suppose $\mu < 0$; that is, the true mean weight change is negative. Then the probability of observing $\bar{Y} \geq 3.01$ would be even smaller than .013. For example, a sample value of $\bar{Y} = 3.01$ is even less likely when $\mu = -5$ than when $\mu = 0$, since the sample value of 3.01 is farther out in the tail of the sampling distribution of $\bar{Y}$ when $\mu = -5$ than when $\mu = 0$. Thus, rejection of H_0: $\mu = 0$ in favor of H_a: $\mu > 0$ also inherently rejects the broader null hypothesis of H_0: $\mu \leq 0$. In other words, one concludes that $\mu = 0$ is false *and* that $\mu < 0$ is false.

The Choice of One-Sided Versus Two-Sided Tests

In practice, two-sided tests are much more common than one-sided tests. Even if a researcher predicts the direction of an effect, two-sided tests permit the detection of an effect that falls in the opposite direction. This practice coincides with the ordinary approach in estimation. Confidence intervals are two-sided, obtained by adding and subtracting some quantity from the point estimate. One can form one-sided confidence intervals, for instance concluding that a population mean is *at least* equal to 7 (i.e., between 7 and ∞). In practice, though, two-sided intervals are much more common.

In deciding whether to use a one-sided or a two-sided alternative hypothesis in a particular exercise, consider the purpose of the test. A statement such as "Test whether the mean has *changed*" suggests a two-sided alternative, to allow for increase or decrease. "Test whether the mean has *increased*" suggests the one-sided alternative, H_a: $\mu > \mu_0$.

In either the one-sided or two-sided case, both hypotheses refer to the population mean μ, not the sample mean $\bar{Y}$. Hypotheses always refer to population parameters, not sample statistics. There is no uncertainty or need to conduct statistical inference about sample statistics, since we can calculate their values exactly once we have the data.

Table 6.3 summarizes the elements of large-sample significance tests for population means.

6.3 Significance Test for a Proportion

For a qualitative variable, each measurement falls in one of a set of categories. Statistical inference refers to the proportions in the categories. For instance, one might test a hypothesis about the population proportion π planning to vote for the Democratic candidate for President. This section presents a large-sample significance test for population proportions. The test is similar to the test for a mean. It utilizes the approximate normal sampling distribution of the sample proportion $\hat{\pi}$.

Elements of the Test

1. *Assumptions*

As usual, the method assumes random sampling. The size of the sample must be sufficiently large that the sampling distribution of $\hat{\pi}$ is approximately normal. (The discussion following the examples presents sample size guidelines.)

TABLE 6.3 The Five Elements of Large-Sample Significance Tests for Population Means

1. Assumptions
   $n \geq 30$
   Random sample
   Quantitative variable
2. Hypotheses
   H_0: $\mu = \mu_0$
   H_a: $\mu \neq \mu_0$ (or H_a: $\mu > \mu_0$ or H_a: $\mu < \mu_0$)
3. Test statistic
   $z = \dfrac{\bar{Y} - \mu_0}{\hat{\sigma}_{\bar{Y}}}$ where $\hat{\sigma}_{\bar{Y}} = \dfrac{s}{\sqrt{n}}$
4. P-value
   In standard normal curve, use
   P = Two-tail probability for H_a: $\mu \neq \mu_0$
   P = Probability to right of observed z-value for H_a: $\mu > \mu_0$
   P = Probability to left of observed z-value for H_a: $\mu < \mu_0$
5. Conclusion
   Report P-value. Smaller P provides stronger evidence against H_0 and supporting H_a

2. *Hypotheses*

The null hypothesis has form

$$H_0 : \pi = \pi_0$$

where π_0 denotes a particular proportion value between 0 and 1. The most common alternative hypothesis is

$$H_a : \pi \neq \pi_0$$

This two-sided alternative states that the true proportion differs from the value in the null hypothesis. Other forms, less common, are the one-sided alternatives

$$H_a : \pi > \pi_0 \text{ and } H_a : \pi < \pi_0$$

These apply when the researcher predicts a deviation of π from π_0 in a certain direction.

3. *Test Statistic*

From Section 5.3, the sampling distribution of the sample proportion $\hat{\pi}$ has mean π and standard error $\sigma_{\hat{\pi}} = \sqrt{\pi(1-\pi)/n}$. When H_0 is true, $\pi = \pi_0$, so the sampling distribution has mean π_0 and standard error $\sigma_{\hat{\pi}} = \sqrt{\pi_0(1-\pi_0)/n}$.

The test statistic is

$$z = \frac{\hat{\pi} - \pi_0}{\sigma_{\hat{\pi}}} = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

This measures the distance of the sample proportion from the null hypothesis value, in standard error units.

The z test statistic has the same form as in large-sample tests for a mean, namely,

Form of z Test Statistic

$$z = \frac{\text{Estimate of parameter} - \text{null hypothesis value of parameter}}{\text{Standard error of estimator}}$$

Here, the estimate $\hat{\pi}$ of the proportion replaces the estimate $\bar{Y}$ of the mean, the hypothesized proportion π_0 replaces the hypothesized mean μ_0, and the standard error $\sigma_{\hat{\pi}}$ of the sample proportion replaces the standard error $\sigma_{\bar{Y}}$ of the sample mean. For large samples, the sampling distribution of the z test statistic is the standard normal distribution, when H_0 is true.

4. *P-Value*

The calculation of the P-value is the same as in tests for a mean. For the alternative H_a: $\pi \neq \pi_0$, P is the two-tail standard normal probability that z has absolute value larger than the absolute value of the observed z-value. See Figure 6.6. This probability is double the single-tail probability beyond the observed z-value. For a one-sided alternative, the P-value is a one-tail probability. For instance, H_a: $\pi > \pi_0$ predicts that the true proportion is larger than the null hypothesis value; its P-value is the probability to the right of the observed value of z under the standard normal curve. For $H_a : \pi < \pi_0$, the P-value is the probability to the left of the observed z-value.

5. *Conclusion*

One summarizes the test by reporting the P-value. As usual, the smaller the P-value, the more strongly the data contradict H_0 and support H_a.

As you read the examples in this section, notice the parallel between each element of the test and the corresponding element for a test about a mean.

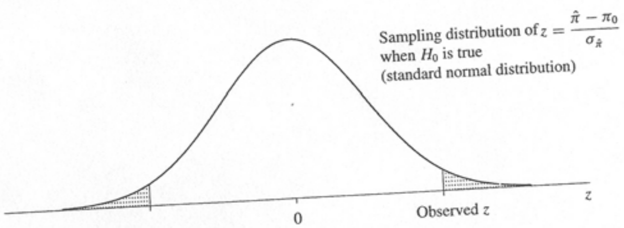

Figure 6.6 Calculation of P-value in Testing H_0: $\pi = \pi_0$ Against H_a: $\pi \neq \pi_0$. The two-sided alternative has a two-tail probability.

Example 6.5 Government Responsibility for Income Inequality

Do you think it should or should not be the government's responsibility to reduce income differences between the rich and poor? Let π denote the population proportion of American adults who believe it should be. If $\pi < .5$, this is a minority of the population, whereas if $\pi > .5$, it is a majority. One can analyze whether the sample data indicate that π is in either of these ranges by testing $H_0 : \pi = .5$ against $H_a : \pi \neq .5$.

In the 1991 General Social Survey of 1227 adults, 591 people responded that it should be the government's responsibility to reduce income differences. The estimate of π equals $591/1227 = .482$. The standard error of $\hat{\pi}$ when H_0: $\pi = .5$ is true is

$$\sigma_{\hat{\pi}} = \sqrt{\frac{\pi_0(1-\pi_0)}{n}} = \sqrt{\frac{(.5)(.5)}{1227}} = .0143$$

The value of the test statistic is, therefore,

$$z = \frac{\hat{\pi} - \pi_0}{\sigma_{\hat{\pi}}} = \frac{.482 - .50}{.0143} = -1.28$$

From Table A, the two-tail P-value for testing H_0: $\pi = .5$ against H_a: $\pi \neq .50$ is $P = 2(.1003) = .20$. If H_0 is true (i.e., if $\pi = .50$), the probability equals .20 that sample results would be as extreme in one direction or the other as in this sample. This P-value is not small, so it does not provide much evidence against H_0. It seems plausible that $\pi = .50$. With this sample, one cannot determine whether the population proportion is less than, equal to, or greater than .50. □

In calculating the standard error, we substituted the null hypothesis value π_0 for the population proportion π in the formula for the true standard error. This differs from confidence intervals, in which the sample proportion $\hat{\pi}$ substitutes for π. The parameter values of sampling distributions in tests are based on the assumption that H_0 is true, since the P-value is calculated under that assumption. This is why one uses π_0 in standard errors for tests. By contrast, the confidence interval method does not have a hypothesized value for π, so that method substitutes the point estimate $\hat{\pi}$ for π in the standard error.

Theoretically, it is not incorrect to substitute the sample proportion in the standard error for the test. One simply obtains a slightly different answer for the test statistic and P-value, but both approaches work well for large n. If one does the test that way, an advantage is that the result necessarily agrees with conclusions from confidence intervals. A disadvantage is that the normal approximation for the sampling distribution is somewhat poorer, especially for proportions close to 0 or 1.

Never "Accept H_0"

A small P-value provides evidence against H_0, since the observed sample result would be unlikely if H_0 were true. On the other hand, if the P-value is not small, the null hypothesis is plausible. In this case, the conclusion is sometimes reported as "Do not reject H_0," since the data do not contradict H_0.

When the P-value is not small, failure to reject H_0 does not mean one can "accept H_0." The population proportion has other plausible values besides the number in the null hypothesis. In addition, the failure to obtain a small P-value may be due to the sample size being too small to estimate the true proportion precisely.

For instance, Example 6.5 showed that $\hat{\pi} = .482$ for $n = 1227$ provides a P-value of .20 in testing H_0: $\pi = .5$ against H_a: $\pi \neq .50$. Thus, it is plausible that $\pi = .50$, but other values are also plausible. For instance, a 95% confidence interval for π is

$$\hat{\pi} \pm 1.96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = .482 \pm 1.96\sqrt{\frac{(.482)(.518)}{1227}} = .482 \pm .028, \text{ or } (.45, .51)$$

This interval shows a range of plausible values for π. Even though insufficient evidence exists to conclude that $\pi \neq .5$, it is improper to conclude that necessarily $\pi = .5$. The data do not contradict H_0, but we need a larger sample size to determine whether a majority or minority of the population believe that government has the responsibility to reduce income differences between the rich and poor. For instance, if $\hat{\pi} = .482$ had been based on $n = 5000$ instead of $n = 1227$, you can verify that the test statistic $z = -2.55$ and the P-value $= .01$. That P-value provides strong evidence against $H_0 : \pi = .50$ and suggests that fewer than half believe it is government's responsibility to reduce income differences. In that case, though, the 95% confidence interval for π equals (.468, .496), indicating that π is quite close to .50 in practical terms.

Of course, we could have used the confidence interval approach from the start, rather than a significance test, to gather information about the value of π. The confidence interval is more informative, since it displays the entire set of plausible values for π rather than merely indicating whether $\pi = .50$ is plausible.

Sample Size Requirement for Test

We next present a guideline about how large the sample size should be to use the large-sample test for a proportion. When π_0 is between .3 and .7, the familiar rule for means of $n \geq 30$ ensures an adequate sample size. A more general rule that applies for all π_0 is based on the normal approximation for the sampling distribution of $\hat{\pi}$, under H_0.

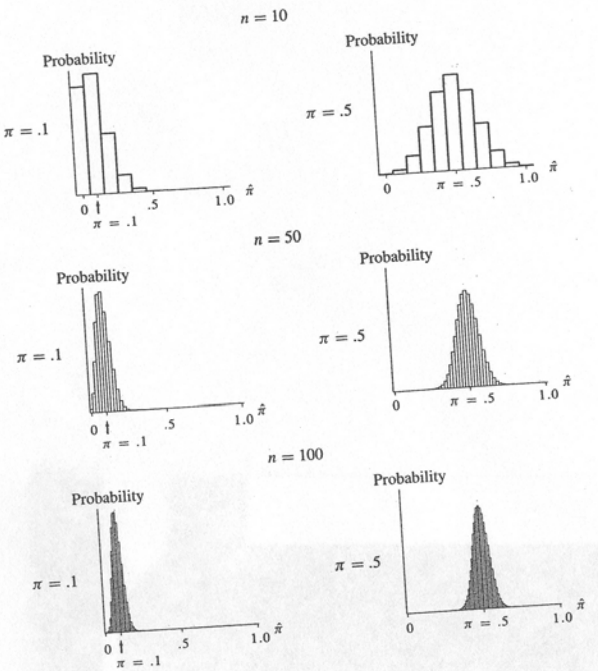

Figure 6.7 Sampling Distribution of $\hat{\pi}$ When $\pi = .1$ or .5, for $n = 10, 50, 100$

This approximation is good when

$$n > \frac{10}{\min(\pi_0, 1 - \pi_0)}$$

where the notation $\min(\pi_0, 1-\pi_0)$ denotes the minimum of the numbers π_0 and $1-\pi_0$. For example, for testing H_0: $\pi = .5$, we need $n > 10/.5 = 20$. For testing H_0: $\pi = .9$ or $H_0 : \pi = .1$, we need $n > 10/.1 = 100$.

The sampling distribution of $\hat{\pi}$ is more skewed when π is near 0 or near 1 than when π is near the middle of the range. Figure 6.7 illustrates. For instance, when $\pi = .1$, the sample proportion $\hat{\pi}$ can't fall much below .1 since it must be positive, but it could fall considerably above .1. The sample size requirement reflects the fact that a symmetric bell shape for the sampling distribution of $\hat{\pi}$ requires larger sample sizes when π is near 0 or 1 than when π is near .5. In Example 6.5, the sample size of $n = 1227$ was

more than adequate to test H_0: $\pi = .5$. One can use a small-sample test introduced in Section 6.6 when the sample size requirement is not satisfied.

Interpreting the *P*-Value

In summary, tests describe whether the data are consistent with H_0 by reporting the P-value. This is the one- or two-tail probability beyond the observed result, calculated under the assumption that H_0 is true. When the P-value is small, the data contradict H_0; the observed data would be unusual if H_0 were true.

A common error is to misinterpret the P-value as the probability that H_0 is true. Classical statistical methods apply probability statements to variables and to statistics, not to parameters. In reality, the null hypothesis H_0 is not a matter of probability; it is either true or not true, and we simply do not know which is the case. In Example 6.5, either π equals .50, or π does not equal .50. A proper interpretation for $P = .20$ is as follows: *If H_0 were true*, the probability would be .20 that the sample proportion $\hat{\pi}$ would fall at least as many standard errors from the null hypothesis value of .50 as the observed $\hat{\pi}$ does. That is, P is the probability that $\hat{\pi}$ is at least as contradictory to H_0 as the observed value, *under the assumption that H_0 is true.*

6.4 Decisions and Types of Errors in Tests of Hypotheses

In significance tests, the P-value summarizes the evidence about H_0. The smaller the P-value, the more strongly the data contradict H_0.

α-Level

It is sometimes necessary to decide whether the evidence against H_0 is strong enough to reject it. The usual approach bases the decision on whether the P-value falls below a prespecified cutoff point.To illustrate, one might reject H_0 if $P \leq .05$, but conclude that the evidence is not strong enough to reject H_0 if $P > .05$. The boundary value .05 is called the *α-level* of the test.

α-Level

The ***α-level*** is a number such that one rejects H_0 if the P-value is less than or equal to it. The α-level is also called the *significance level* of the test. The most common α-levels are .05 and .01.

Like the choice of a confidence coefficient for a confidence interval, the choice of the α-level for a test reflects how cautious the researcher wants to be. The smaller the α-level, the stronger the evidence must be to reject H_0. To avoid bias in the decision-making process, one selects the α-level *before* analyzing the data.

Table 6.4 summarizes the two possible conclusions for a test with α-level .05. The null hypothesis is either "rejected" or "not rejected." If H_0 is rejected, then H_a is accepted; in this case, H_a seems more valid than H_0. If H_0 is not rejected, then H_0 is plausible, but other parameter values are also plausible. Thus, H_0 is never "accepted." In this case, results are inconclusive, and the test does not identify either hypothesis as more valid.

TABLE 6.4 Possible Conclusions in a Test of Hypothesis with α-Level .05

	Conclusion	
P-Value	H_0	H_a
$P \leq .05$	Reject	Accept
$P > .05$	Do not reject	Do not accept

Example 6.6 Adding Decisions to Previous Examples

Example 6.2 tested the hypothesis $H_0 : \mu = 4.0$ about mean political ideology. We now use an α-level of $\alpha = .05$ to guide us in making a decision about H_0. Since the P-value equaled $P = .52$, we have $P > .05$ and insufficient evidence to reject H_0. In other words, we cannot conclude that the mean ideology in the population differs from the moderate value of 4.0.

Now consider Example 6.4 on the hypothesis $H_0 : \mu = 0$ about the mean weight gain for a sample of women suffering from anorexia. The P-value was .013. This is less than .05, so this result provides sufficient evidence to reject H_0 in favor of $H_a : \mu > 0$; we conclude that the treatment does produce an increase in mean weight. This type of conclusion is sometimes phrased as "The increase in the mean is *statistically significant* at the .05 level." Since $P = .013$ is not less than .010, the result is not significant at the .01 level. In fact, *the P-value is the smallest level for α at which the results are significant*. That is, we would reject H_0 if α were any level above .013. □

In our opinion, it is preferable to report the P-value rather than to indicate simply whether the result is significant at a particular α-level. Reporting the P-value has the advantage that the reader can tell whether the result is significant at any level. The P-values of .049 and .001 are both "significant at the .05 level," but the second case provides much stronger evidence than the first case. Likewise, P-values of .049 and .051 provide, in practical terms, the same amount of evidence about H_0. It is artificial to call one result "significant" and the other "nonsignificant."

Rejection Regions

The null hypothesis contains a single possible value for the parameter. Using the terminology "Do not reject H_0" instead of "Accept H_0" emphasizes that that value is merely one of many plausible values. Because of sampling error, there is a range of plausible values rather than just the H_0 value, so one can never accept a null hypothesis. The reason "accept H_a" terminology is permissible for the alternative hypothesis is that when the P-value is sufficiently small, the entire range of plausible values for the parameter fall within the broad range of numbers contained in H_a.

The collection of test statistic values for which the test rejects H_0 at a particular α-level is called the ***rejection region***. For example, the rejection region for a test of level $\alpha = .05$ is the set of test statistic values for which $P \leq .05$.

For large-sample two-sided tests, for instance, the two-tail probability that forms the P-value is $\leq .05$ whenever the test statistic satisfies $|z| \geq 1.96$. In other words, the rejection region for an $\alpha = .05$ level test consists of values of z for which $|z| \geq 1.96$, that is, values of z resulting from the estimate of the parameter falling at least 1.96 standard errors from the null hypothesized value.

Type I and Type II Errors

Because of sampling error, decisions in tests of hypotheses always have some uncertainty. The decision could be erroneous, just as a confidence interval can falsely predict where the parameter falls. There are two types of potential errors, conventionally called *Type I* and *Type II* errors.

Type I and Type II Errors

A ***Type I error*** occurs when H_0 is rejected, even though it is true.
A ***Type II error*** occurs when H_0 is not rejected, even though it is false.

A decision in a test has four possible results. These refer to the two possible decisions combined with the two possible conditions for H_0. Table 6.5 shows these four results.

TABLE 6.5 The Four Possible Results of Making a Decision in a Test; Two of These Refer to Incorrect Decisions

		Decision	
		Reject H_0	*Do not reject H_0*
Condition of H_0	H_0 true	Type I error	Correct decision
	H_0 false	Correct decision	Type II error

Suppose we test H_0 at the $\alpha = .05$ level, rejecting H_0 if $P \leq .05$. For example, for the large-sample test about a mean or proportion with two-sided alternative, we reject H_0 when $|z| \geq 1.96$. These z values form the rejection region. For a continuous sampling distribution such as the normal distribution, the probability of rejecting H_0 when it is true is exactly .05, since the probability of the values in the rejection region is .05. But this is precisely the α-level.

The probability of a Type I error is the α-level for the test.

With $\alpha = .05$, if the null hypothesis is true, the probability equals .05 of making a Type I error and rejecting that (true) null hypothesis. One controls the probability of a Type I error by the choice of the α-level.

The more serious the consequences of a Type I error, the smaller α should be. For exploratory research conducted for data snooping—scanning several hypotheses to see which might warrant further investigation—one would not be too stringent (say, $\alpha =$ .10). On the other hand, suppose that the decision has serious policy implications. For instance, suppose that the alternative hypothesis states that a newly developed drug is better than the one currently used to treat a particular illness; if we reject H_0, then the new drug will be prescribed instead of the current one to treat the illness. Then, we might prefer tougher standards, such as $\alpha = .01$. In that case, the data must contradict H_0 very strongly in order to reject it, to lessen the chance of Type I error.

Relation Between *P*(Type I Error) and *P*(Type II Error)

A Type II error occurs in not rejecting H_0 even though it is false. We shall see in Section 6.7 that the probability this happens depends on just how far the actual value of the parameter falls from H_0. If the true parameter value is nearly equal to the value hypothesized in H_0, the probability of a Type II error might be quite high, whereas it would be smaller for more distant values of the parameter. The farther the true value of the parameter falls from the value specified in H_0, the less likely the sample is to fail to detect the difference and result in a Type II error.

The probability of Type I error and the probability of Type II error are inversely related. The smaller the α-level and hence the probability of Type I error, the larger the probability of Type II error. In other words, the stronger the evidence required to reject H_0 (i.e., the smaller the α-value), the more likely it becomes that we will fail to detect a real difference. If we tolerate only an extremely small chance of a Type I error, then the test may be unlikely to reject the null hypothesis even if it is false.

For a fixed probability of Type I error, we can decrease the probability of Type II error by selecting a larger sample. That is, the larger the sample size, the more likely we are to reject a false null hypothesis at a particular α-level. To keep both the probabilities of Type I and Type II errors at low levels, it may be necessary to use a very large sample size.

In making a decision in a test, we do not know whether we have made a Type I or Type II error, just as we do not know whether a particular confidence interval truly contains the unknown parameter value. However, we can control the probability of an incorrect decision for either type of inference. Although we do not know whether the conclusion in a particular test is correct, we justify the procedure in terms of the long-run proportions of Type I and Type II errors.

Except in Section 6.7, we shall not study calculation of the probability of a Type II error, since these calculations are quite complex. In practice, making a decision in a test only requires setting α, the probability of Type I error. One should realize, though, that the probability of a Type II error may be quite large when the sample size is small. In other words, the reason for not rejecting H_0 may be that for that sample size, the test simply does not have a very high chance of detecting the actual deviation from H_0.

Equivalence Between Confidence Intervals and Tests of Hypotheses

We now elaborate on the equivalence between decisions from two-sided tests and conclusions from confidence intervals, first alluded to in Example 6.3. Consider the large-sample test of

$$H_0: \mu = \mu_0 \qquad \text{versus} \qquad H_a: \mu \neq \mu_0$$

When $P \leq .05$, H_0 is rejected at the $\alpha = .05$ level. This implies that the test statistic $z = (\bar{Y} - \mu_0)/\hat{\sigma}_{\bar{Y}}$ is at least 1.96 in absolute value. That is, $\bar{Y}$ falls more than $1.96\hat{\sigma}_{\bar{Y}}$ from μ_0. But if this happens, then the 95% confidence interval for μ, namely, $\bar{Y} \pm 1.96\hat{\sigma}_{\bar{Y}}$, does not contain the null hypothesis value μ_0. See Figure 6.8.

In other words, rejecting H_0: $\mu = \mu_0$ at the $\alpha = .05$ level is equivalent to the 95% confidence interval for μ not containing μ_0. These two inference procedures are consistent. If a confidence interval indicates that a particular number μ_0 is not plausible for the value of μ, then we would reject H_0: $\mu = \mu_0$ in favor of H_a: $\mu \neq \mu_0$ with the test. The null hypothesis H_0 is rejected at the α-level equal to one minus the confidence coefficient; for instance, $\alpha = .05$ for a 95% confidence interval. This level is both the probability of Type I error for the test and the probability that the confidence interval does not contain the parameter.

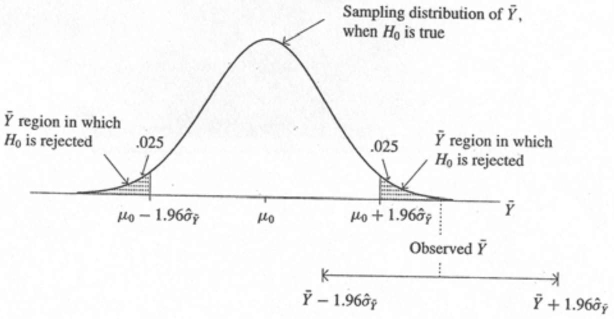
Sampling distribution of $\bar{Y}$, when H_0 is true
$\bar{Y}$ region in which H_0 is rejected
.025
.025
$\bar{Y}$ region in which H_0 is rejected
$\mu_0 - 1.96\hat{\sigma}_{\bar{Y}}$
μ_0
$\mu_0 + 1.96\hat{\sigma}_{\bar{Y}}$
$\bar{Y}$
Observed $\bar{Y}$
$\bar{Y} - 1.96\hat{\sigma}_{\bar{Y}}$
$\bar{Y} + 1.96\hat{\sigma}_{\bar{Y}}$

Figure 6.8 Relationship Between Confidence Interval and Hypothesis Test

In testing H_0: $\mu = \mu_0$ against H_a: $\mu \neq \mu_0$, suppose one rejects H_0 at the .05 α-level. Then, the 95% confidence interval for μ does not contain μ_0. The 95% confidence interval for μ consists of those μ_0 values for which one does not reject H_0: $\mu = \mu_0$ at the .05 α-level.

Example 6.2 tested a hypothesis about mean political ideology. The P-value for testing $H_0 : \mu = 4.0$ against $H_a : \mu \neq 4.0$ was $P = .52$. At the $\alpha = .05$ level, H_0: $\mu = 4.0$ is not rejected; it is plausible that $\mu = 4.0$. Example 6.3 showed that a 95% confidence interval for μ is (3.93, 4.13), which contains $\mu_0 = 4.0$. On the other hand, we saw that if the results had been based on $n = 6270$ instead of $n = 627$, then $P = .044$. In that case, since $P = .044$ is less than .05, we can reject H_0 at the $\alpha = .05$ level; that is, 4.0 is not a plausible value for μ. In fact, you can verify that a 95% confidence interval in this case equals (4.001, 4.063), not containing 4.0.

Statistical and Practical Significance

Anyone who uses significance tests should understand the distinction between *statistical* and *practical* significance. A very small P-value, such as $P = .001$, does not imply an "important" finding in any practical sense. This merely means that if H_0 were true, the observed results would be very unusual. Even if H_0 is false, though, the true value of the parameter may be close to the null hypothesized value. In this case, the difference may not be significant in practical terms. If the sample size is very large, small P-values can occur even though the difference is small.

Example 6.7 Mean Political Ideology in 1994

The mean of 4.03 for political ideology in Example 6.2 refers to a sample taken in 1978. For the General Social Survey of 1994, the counts in the seven categories were (71, 328, 378, 1049, 472, 478, 103). For a scoring of 1 through 7 for this seven-point scale, these 2879 observations have a mean of 4.17 and a standard deviation of 1.39. It appears that the mean level of conservatism increased only slightly between 1978 and 1994.

As in Example 6.2, we test whether the population mean differs from the moderate ideology score of 4.0. Then, $\hat{\sigma}_{\bar{Y}} = s/\sqrt{n} = 1.39/\sqrt{2879} = .026$, and

$$z = \frac{\bar{Y} - \mu_0}{\hat{\sigma}_{\bar{Y}}} = \frac{4.17 - 4.0}{.026} = 6.6$$

The P-value is $P = .000000000005$. There is *extremely* strong evidence that the true mean exceeds 4.0, that is, that the true mean falls on the conservative side of moderate. On the other hand, on a scale of 1 to 7, 4.17 is quite close to the moderate score of 4.0. Although the difference of .17 between the sample mean of 4.17 and the null hypothesis mean of 4.0 is highly significant statistically, the magnitude of this difference is small in practical terms. We can conclude that, on the average, the mean response on political ideology is essentially still a moderate one. □

In Example 6.2, the sample mean of 4.03 based on $n = 627$ had $P = .52$, not much evidence against H_0. On the other hand, we noted earlier that the same evidence based on $n = 6270$ yields $P = .044$. This is statistically significant at the .05 level, but not practically significant. For practical purposes, a mean political ideology of 4.03 does not differ from 4.00.

One point of this example is that larger sample sizes can provide more powerful inferences; thus, test statistics can detect deviations of smaller magnitude from H_0 than they can for smaller samples. The size of P merely measures the extent of evidence about the truth of H_0, not how far from the truth H_0 happens to be. One should always inspect the difference between the sample estimate and the hypothesized value of the parameter (e.g., between $\bar{Y}$ and μ_0, between $\hat{\pi}$ and π_0) to gauge the practical implications of a test result.

Limitations of Significance Tests Compared to Estimation

Null hypotheses such as $H_0 : \mu = \mu_0$ and $H_0 : \pi = \pi_0$ are rarely true in the social sciences. That is, rarely is the true value of the parameter *exactly* equal to the value listed in H_0. With sufficiently large samples, so that a Type II error is unlikely, these hypotheses will normally be rejected. What is more relevant is whether the true parameter value is sufficiently different from the null hypothesis value to be of importance.

Although tests of hypotheses can be useful, many social scientists and nearly all statisticians believe that significance testing is greatly overemphasized in social science research. By contrast, confidence intervals are underutilized. It is preferable to construct confidence intervals for parameters instead of performing only significance tests. A test merely indicates whether a particular parameter value is plausible; a confidence interval displays additional information, showing the entire set of plausible values. When a P-value is small, the test indicates that the parameter value in the null hypothesis is not plausible, but it tells us nothing about which potential parameter values *are* plausible. The confidence interval, on the other hand, displays those plausible values. It shows the extent to which H_0 may be false by showing whether the values in the interval are very far from the null hypothesis value. Thus, it helps us to determine whether rejection of the null hypothesis has practical importance.

To illustrate, for Example 6.7, a 95% confidence interval for μ is $\bar{Y} \pm 1.96\hat{\sigma}_{\bar{Y}} = 4.17 \pm 1.96(.026)$, or (4.12, 4.22). This indicates that the difference from the moderate score of 4.0 is of small magnitude. Although the P-value of $P = .000000000005$ provides extremely strong evidence against H_0, in practical terms the confidence interval shows that the departure from the null hypothesis is minor. On the other hand, if $\bar{Y}$ had been 5.17 (instead of 4.17), the 95% confidence interval would equal (5.12, 5.22). This indicates a more substantial difference from 4.0, the average response being close to the slightly conservative category rather than the moderate category.

A confidence interval displays the set of values that are plausible parameter values. When the P-value is not small, the confidence interval indicates whether the lack of evidence against H_0 may be due to a lack of power. A wide confidence interval containing the null hypothesis value of the parameter indicates a strong possibility of a Type II error in the test. In that case, the lack of precision of the interval estimate also indicates why it does not make sense to accept H_0, as we discussed previously. For small to moderate sample sizes, it is not unusual for a confidence interval to be wide, and this forces us to recognize the lack of precision that any inference involves.

The remainder of the text presents significance tests for a variety of situations. It is important to become familiar with the elements of these tests, if for no other reason than their frequent use in the social science literature. However, we shall also introduce confidence intervals for parameters that describe how far reality is from the hypothesized condition.

6.5 Small-Sample Inference for a Mean—The *t* Distribution

The confidence interval for means presented in Section 5.2 and the significance test for means presented in Section 6.2 both apply for large sample sizes. The large-sample assumption ensures that the sampling distribution of $\bar{Y}$ is approximately normal. It also ensures that the sample standard deviation estimate s is close to the unknown population standard deviation σ; this ensures that the estimated standard error $\hat{\sigma}_{\bar{Y}} = s/\sqrt{n}$ is sufficiently close to the true standard error $\sigma_{\bar{Y}} = \sigma/\sqrt{n}$ that one can substitute the estimate in confidence intervals and test statistics. As a rough guideline, these approximations are adequate if the sample size is at least 30.

Some studies are limited, however, to smaller sample sizes. For example, each observation may result from a long or expensive experimental procedure. A consumer group that decides to evaluate the mean repair cost resulting when a new-model automobile crashes into a brick wall at 30 miles per hour would probably not want to restrict itself to large-sample statistical methods!

The *t* Distribution

This section introduces inferential methods for small samples. The derivation of the methods assumes that the *population distribution* of the variable Y is normal. In that case, the sampling distribution of $\bar{Y}$ and of the statistic $(\bar{Y} - \mu)/\sigma_{\bar{Y}}$ is normal even for small sample sizes. (Figure 4.15 illustrated this.) However, substitution of s for σ in $\sigma_{\bar{Y}}$, as is done in practice, introduces additional variability in the sampling distribution. It is then no longer normal, but has the ***t* distribution**.

t Statistic; t Distribution

Suppose the population distribution of a variable is normal, with parameters μ and σ. Then, for a random sample of size n, the sampling distribution of the ***t* statistic**

$$t = \frac{\bar{Y} - \mu}{\hat{\sigma}_{\bar{Y}}} = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

is called the **t distribution with $(n-1)$ degrees of freedom.**

The t statistic resembles the z statistic of Section 6.2. The different symbol for the statistic here emphasizes that the sampling distribution is different and that it is valid in a different situation—when the population distribution is normal. The *degrees of freedom* for the t distribution, denoted by df, determine the precise shape of the distribution, as discussed shortly. This quantity refers to the divisor in the point estimate $s^2 = \left[\sum (Y - \bar{Y})^2\right]/(n-1)$ of σ^2, namely, $df = n - 1$.

The t distribution was discovered in 1908 by the statistician and chemist W. S. Gosset. At the time, Gosset was employed in the experimental unit of Guinness Breweries in Dublin, Ireland. He had only small samples available for several of his analyses for determining the best varieties 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 this result. The t statistic is often called *Student's t*.

Properties of the *t* Distribution

Before presenting methods of statistical inference for small samples, we list the major properties of the t distribution.

- The t distribution is bell-shaped and symmetric about 0. This property is shared by the sampling distribution of the z statistic, the standard normal distribution.
- The spread of the t distribution depends on the degrees of freedom. The standard deviation of the t distribution always exceeds 1, but decreases toward 1 as df (and hence n) increases. (The standard deviation equals $\sqrt{df/(df-2)}$. This value exceeds 1.0, but it decreases toward 1.0 as df increases.)
- Though the t distribution is bell-shaped about 0, the probability falling in the tails is higher than for the standard normal distribution. The larger the df value, however, the more closely it resembles the standard normal distribution, as Figure 6.9

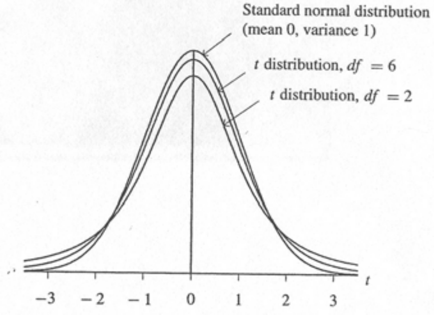

Figure 6.9 t Distribution Relative to Standard Normal Distribution. The t gets closer to the normal as the degrees of freedom (d) increase, and the two distributions are practically identical when $df > 30$.

illustrates. In the limit as df increases indefinitely, the two distributions are identical.

The change in the form of the t distribution as df increases is due to the increasing precision of s as a point estimate of σ in the standard error formula $\hat{\sigma}_{\bar{Y}} = s/\sqrt{n}$. Because s is a less accurate estimator of σ when $df < 30$, its presence in the denominator of the t statistic produces additional sampling error. This additional sampling error for small samples results in the t sampling distribution being more spread out than the standard normal sampling distribution of the large-sample z statistic, in which s is nearly identical to σ. As the sample size increases, s becomes a more accurate estimator of σ, and the t distribution becomes less disperse. When $df \geq 30$, the t distribution is so similar to the standard normal distribution that inference procedures for the mean using the t distribution are practically equivalent to those using the standard normal distribution.

- Table B at the end of the text lists values from the t distribution with various tail probabilities. Since the t distribution has a slightly different shape for each distinct value of df, different t-values apply for each df value. Table B lists the t-values only for the one-tail probabilities of .100, .050, .025, .010, and .005. The table denotes these by $t_{.100}$, $t_{.050}$, $t_{.025}$, $t_{.010}$, and $t_{.005}$. These same values refer to two-tail probabilities of .20, .10, .05, .02, and .01.

To illustrate Table B, suppose $df = 6$. Then, since $t_{.025} = 2.447$, 2.5% of the t distribution falls in the right-hand tail above 2.447. Figure 6.10 illustrates. By symmetry, 2.5% also lies in the left-hand tail below $-t_{.025} = -2.447$. When $df = 6$, the probability equals .05 that the absolute value of the t statistic exceeds 2.447.

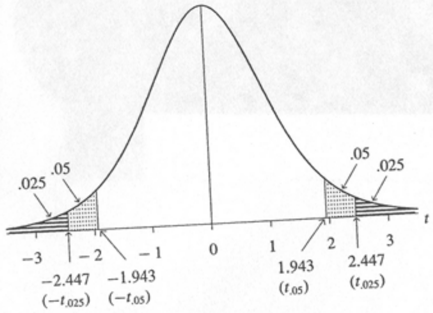

Figure 6.10 t Distribution with $df = 6$

Table B shows that as df increases, the t-score decreases to the z-score for a standard normal distribution. This reflects the t distribution becoming less disperse and more similar in appearance to the standard normal distribution as df increases. For instance, when df increases from 1 to 29 in Table B, the t-score with right-tail probability equal to .025 decreases from 12.706 to 2.045. The z-score with this right-tail probability for the standard normal distribution is $z = 1.96$. For df of about 30 or higher, the t-score is similar to the z-score of 1.96.

The last row of Table B lists the z-values for one-tail probabilities, opposite $df = \infty$ (infinity). The t-values are not printed for $df \geq 30$, since they are close to the z-values. Whenever $df \geq 30$ for a method using the t distribution, one can refer to the normal probability table (Table A) and proceed as if it uses the standard normal distribution. Computer software has the t distribution in memory for all df values, so such software does not need to use the normal approximation when $df \geq 30$.

Small-Sample Confidence Interval for a Mean

Small-sample confidence intervals and significance tests for a mean resemble those for large samples, except that they use the t distribution instead of the standard normal. We first present the confidence interval.

> **Small-Sample Confidence Interval for μ**
>
> For a random sample from a normal population distribution, a 95% confidence interval for μ is
>
> $$\bar{Y} \pm t_{.025}\hat{\sigma}_{\bar{Y}} = \bar{Y} \pm t_{.025}\left(\frac{s}{\sqrt{n}}\right)$$
>
> where $df = n - 1$ for the t-value.

The intervals use the $t_{.025}$-value, which is the t value for a right-tail probability of .025. This is because 95% of the probability for a t distribution falls between $-t_{.025}$ and $t_{.025}$. Let α denote the error probability that the confidence interval does not contain μ. For instance, for a 95% confidence interval, $\alpha = .05$. A confidence interval uses the t-score with tail probability $\alpha/2$ in each tail. For a 99% confidence interval, for instance, $\alpha = .01$, and the appropriate t-score is $t_{.005}$ for the specified df value.

Like the confidence interval in Section 5.2 for large samples, this confidence interval equals the point estimate of μ plus and minus a table value multiplied by the estimated standard error. The only difference in the formula is the substitution of the t-table value for the normal-table value, to reflect the small sample size. The t method also makes the additional assumption of a normal population distribution, which is needed for small samples. In practice, the distribution is typically not normal, and we discuss the importance of this assumption later in the section.

Example 6.8 Estimating Mean Weight Change for Anorexic Girls

Example 6.4 discussed a study that compared various treatments for young girls suffering from anorexia. The variable of interest was the change in weight from the beginning to the end of the study. For the sample of 29 girls receiving the cognitive behavioral treatment, the changes in weight were summarized by $\bar{Y} = 3.01$ and $s = 7.31$ pounds.

Example 6.4 used large-sample methods, which are of borderline acceptability with $n = 29$. Here, we use small-sample methods.

Let μ denote the population mean change in weight for this treatment. Since $n = 29$, $df = n - 1 = 28$. For a 95% confidence interval, we use $t_{.025} = 2.048$. The estimated standard error equals $\hat{\sigma} = s/\sqrt{n} = 7.31/\sqrt{29} = 1.357$. The 95% confidence interval is

$$\bar{Y} \pm t_{.025}\hat{\sigma}_{\bar{Y}} = 3.01 \pm 2.048(1.357) = 3.0 \pm 2.8, \text{ or } (0.2, 5.8)$$

We infer with 95% confidence that this interval contains the true mean weight change for this treatment. It appears that the true mean change in weight is positive, but rather small. □

Elements of a *t* Test for a Mean

We next list the five elements of a small-sample significance test for a mean.

1. *Assumptions*

A random sample is selected. The variable is quantitative and has a normal population distribution. (The method is designed for $n \leq 30$, but can be used with any size n.)

2. *Hypotheses*

The hypotheses are the same as in the large-sample test for a mean. The null hypothesis has form H_0: $\mu = \mu_0$, and the two-sided alternative hypothesis has form H_a: $\mu \neq \mu_0$. The one-sided alternative hypotheses are H_a: $\mu > \mu_0$ and H_a: $\mu < \mu_0$.

3. *Test Statistic*

The test statistic is the t statistic with $\mu = \mu_0$, namely,

$$t = \frac{\bar{Y} - \mu_0}{\hat{\sigma}_{\bar{Y}}} = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}}$$

Like the z statistic, this statistic measures the distance between the sample mean and the null hypothesis value, divided by the estimated standard error of $\bar{Y}$. If H_0 is true, the sampling distribution of the t test statistic is the t distribution with $df = n - 1$.

4. *P-Value*

The calculation of the P-value uses one or two tails in the same way as the large-sample calculation, but it uses the t distribution (Table B) instead of the standard normal distribution. For the two-sided alternative hypothesis H_a: $\mu \neq \mu_0$, P is the two-tail probability of a t-value at least as large in absolute value as the observed one, if H_0 were true. Figure 6.11 depicts the two-sided P-value. As usual, the smaller the P-value, the stronger the evidence against H_0 and in favor of H_a.

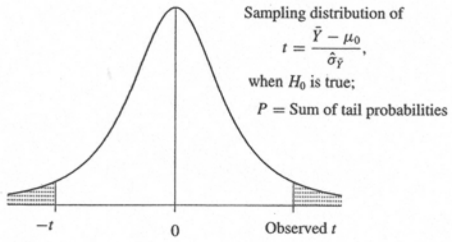

Figure 6.11 Calculation of P in Testing H_0: $\mu = \mu_0$ Against H_a: $\mu \neq \mu_0$, for Small Samples. The P-value is the total two-tail probability beyond the observed test statistic.

5. *Conclusion*

Normally, we report the P-value. For a formal decision, as in the large-sample case, reject H_0 if the P-value is no greater than some fixed α-level, such as .05 or .01.

Example 6.9 Small-Sample Test for Anorexia Data

We illustrate with a t test for the anorexia data. For the 29 observations, $\bar{Y} = 3.01$ and $\hat{\sigma}_{\bar{Y}} = 1.357$. As in Example 6.4, one might test for no effect of treatment versus a positive average weight change, by testing H_0: $\mu = 0$ against H_a: $\mu > 0$. More commonly, in practice, one would use the two-sided alternative H_a : $\mu \neq 0$. The test statistic equals

$$t = \frac{\bar{Y} - \mu_0}{\hat{\sigma}_{\bar{Y}}} = \frac{3.01 - 0}{1.357} = 2.22$$

precisely the same as the large-sample z statistic.

Now, for $n = 29$ and $df = 28$, $t = 2.048$ yields $P = .025$ for the one-sided alternative hypothesis. Since the observed $t = 2.22 > 2.048$, the one-tail P-value is $P < .025$, since a value farther out in the tail has a smaller tail probability. Moreover, $P > .01$, since Table B indicates that $t = 2.467$ has a tail probability of .01. Figure 6.12 illustrates. Table B is not detailed enough to provide the exact value of P. We could summarize the P-value for the one-sided test by reporting that $.01 < P < .025$. Table B provides enough information to determine whether the one-tailed P-value is greater than or less than .10, .05, .025, .01, and .005. If two of the tabled t-scores bracket the observed t statistic, above and below, their tail probabilities bracket the actual tail probability. For a two-sided alternative, we double the results. For instance, for these data we double the bounds of .01 and .025 to report $.02 < P < .05$.

When computer software performs the analysis, the output reports the actual P-value rather than bounds for it. Most software reports the P-value for a two-sided alternative. Table 6.6 shows how SPSS reports results for tests and confidence intervals.

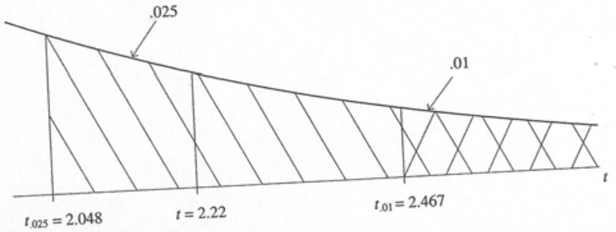

Figure 6.12 For $df = 28$, $t = 2.22$ Has a Tail Probability Between .01 and .025

For $t = 2.22$ with $df = 28$, it reports $P = .035$. The one-sided P-value is half this, or about $P = .017$. There is strong evidence against the hypothesis that the treatment has no effect. The small-sample one-sided P-value of .017 does not differ much from the value of .013 reported in Example 6.4 using large-sample methods.

TABLE 6.6

```
                Number
Variable       of Cases         Mean        SD     SE of Mean
CHANGE            29          3.0069      7.309       1.357

Mean              95% CI
          Lower         Upper       t-value     df   2-Tail Sig
3.01      .227          5.787        2.22       28      .035
```

The significance test concludes that the mean weight gain was not equal to 0. The 95% confidence interval of (0.2, 5.8) is more informative, showing just how different from 0 the true mean change is likely to be. □

Robustness for Violations of Normality Assumption

A basic assumption of the confidence interval and test using the t distribution is that the population distribution is normal. For instance, to be valid, the confidence interval in Example 6.8 and the significance test in Example 6.9 require the assumption that the true distribution of weight change for that treatment is normal. It would be impossible to verify this assumption. A histogram or stem and leaf plot can provide some information about the shape of the population distribution, but it is not precise when n is smaller than 30. In most practical problems, one has only a rough notion of the form of the population distribution when the sample size is small.

A statistical method is said to be ***robust*** if it performs adequately even when an assumption is violated. The study of the robustness of methods is important, because rarely in practice are all assumptions perfectly satisfied. Statisticians have shown that small-sample two-sided inferences for a mean using the t distribution are quite robust against violations of the assumption that the population distribution is normal. Even if the population is not normally distributed, two-sided tests and confidence intervals based on the t distribution still work quite well. The P-values and confidence coefficients are fairly accurate, the accuracy being quite good when n exceeds about 15. The test does not work so well for a one-sided test with small n when the population distribution is highly skewed. There is evidence of such skewness if you see outliers in one direction.

The results of t tests and confidence intervals are *not* robust to violations of the random sample assumption. The results may be completely invalid if the sample is not random.

Computer Software and Inference for Means

We have used the t distribution for small-sample inference about a mean and the normal distribution for large-sample inference, with $n = 30$ being a rather arbitrary dividing line. This is partly because the t table in this book (Table B) has df values only below 30, and for larger values the t-scores are practically identical to z-scores.

Computer software does not distinguish between the two cases. It uses the t distribution for all cases. It has the t distribution stored in memory for all possible df values, so it is not limited to $n < 30$. The advantage of using the t methods is that they account for the extra variability due to estimating σ by s. Though they make the extra assumption of a normal population, this is unneeded for $n > 30$; the sampling distribution of $\bar{Y}$ is then approximately normal regardless of the shape of the population, by the Central Limit Theorem (Sec. 4.4). Of course, when $n > 30$, you will get nearly identical results if you use z-scores instead of t.

Significance tests that a parameter equals a particular value μ_0 are often artificial. It is rare that we learn much by testing a hypothesis about a single population mean. The next chapter presents more realistic tests, involving comparisons of means for two populations. In most applications, we learn more by constructing a confidence interval than by performing a test. In particular, with small samples, confidence intervals are usually wide, forcing us to recognize that estimates of parameters are imprecise.

6.6 Small-Sample Inference for a Proportion—The Binomial Distribution*

The confidence interval for a population proportion π presented in Section 5.3 and the significance test presented in Section 6.3 are valid for large samples. The sampling distribution of the sample proportion $\hat{\pi}$ is then approximately normal, and one can use z-scores in tests and confidence intervals. The closer the true parameter π is to 0 or 1 for a given sample size, the more skewed the actual sampling distribution becomes, and the normal approximation may be poor (refer back to Figure 6.7). For instance,