$$V(\bar{Y}) = V\left(\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right]\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Y^2.$$

Simplifying further, we have

$$V(\bar{Y}) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Y^2 = \frac{n\sigma_Y^2}{n^2} = \frac{\sigma_Y^2}{n}. \tag{1.5}$$

We've shown that the sampling variance of a sample average depends on the variance of the underlying observations, σ_Y^2, and the sample size, n. As you might have guessed, more data means less dispersion of sample averages in repeated samples. In fact, when the sample size is very large, there's almost no dispersion at all, because when n is large, σ_Y^2/n is small. This is the LLN at work: as n approaches infinity, the sample average approaches the population mean, and sampling variance disappears.

In practice, we often work with the standard deviation of the sample mean rather than its variance. The standard deviation of a statistic like the sample average is called its *standard error*. The standard error of the sample mean can be written as

$$SE(\bar{Y}) = \frac{\sigma_Y}{\sqrt{n}}. \tag{1.6}$$

Every estimate discussed in this book has an associated standard error. This includes sample means (for which the standard error formula appears in equation (1.6)), differences in sample means (discussed later in this appendix), regression coefficients (discussed in Chapter 2), and instrumental variables and other more sophisticated estimates. Formulas for standard errors can get complicated, but the idea remains simple. The standard error summarizes the variability in an estimate due to random sampling. Again, it's important to avoid confusing standard errors with the standard deviations of the underlying variables; the two quantities are intimately related yet measure different things.

One last step on the road to standard errors: most population quantities, including the standard deviation in the numerator of (1.6), are unknown and must be estimated. In practice,

therefore, when quantifying the sampling variance of a sample mean, we work with an *estimated standard error*. This is obtained by replacing σ_Y with $S(Y_i)$ in the formula for $SE(\bar{Y})$. Specifically, the estimated standard error of the sample mean can be written as

$$\hat{SE}(\bar{Y}) = \frac{S(Y_i)}{\sqrt{n}}.$$

We often forget the qualifier "estimated" when discussing statistics and their standard errors, but that's still what we have in mind. For example, the numbers in parentheses in Table 1.4 are estimated standard errors for the relevant differences in means.

The t-Statistic and the Central Limit Theorem

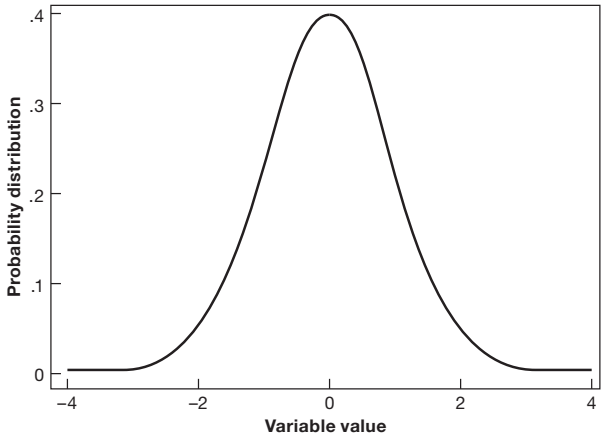
Having laid out a simple scheme to measure variability using standard errors, it remains to interpret this measure. The simplest interpretation uses a *t-statistic*. Suppose the data at hand come from a distribution for which we believe the population mean, $E[Y_i]$, takes on a particular value, μ (read this Greek letter as "mu"). This value constitutes a working *hypothesis*. A t-statistic for the sample mean under the working hypothesis that $E[Y_i] = \mu$ is constructed as

$$t(\mu) = \frac{\bar{Y} - \mu}{\hat{SE}(\bar{Y})}.$$

The working hypothesis is a reference point that is often called the *null hypothesis*. When the null hypothesis is $\mu = 0$, the t-statistic is the ratio of the sample mean to its estimated standard error.

Many people think the science of statistical inference is boring, but in fact it's nothing short of miraculous. One miraculous statistical fact is that if $E[Y_i]$ is indeed equal to μ, then—as long as the sample is large enough—the quantity $t(\mu)$ has a sampling distribution that is very close to a bell-shaped standard normal distribution, sketched in Figure 1.1. This property, which applies regardless of whether Y_i itself is normally distributed, is called the *Central Limit Theorem* (CLT). The

FIGURE 1.1
A standard normal distribution



CLT allows us to make an empirically informed decision as to whether the available data support or cast doubt on the hypothesis that $E[Y_i]$ equals μ.

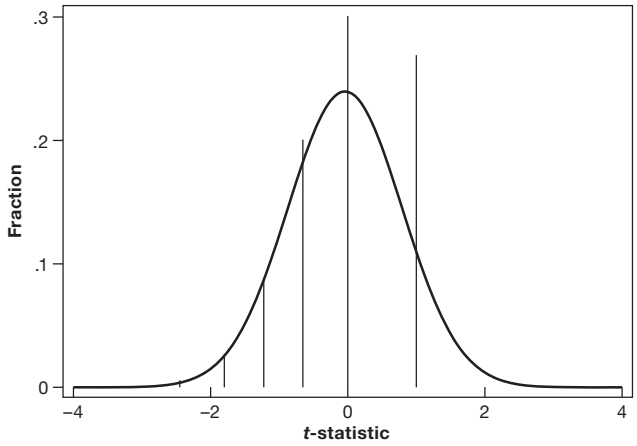
The CLT is an astonishing and powerful result. Among other things, it implies that the (large-sample) distribution of a t-statistic is independent of the distribution of the underlying data used to calculate it. For example, suppose we measure health status with a dummy variable distinguishing healthy people from sick and that 20% of the population is sick. The distribution of this dummy variable has two spikes, one of height .8 at the value 1 and one of height .2 at the value 0. The CLT tells us that with enough data, the distribution of the t-statistic is smooth and bell-shaped even though the distribution of the underlying data has only two values.

We can see the CLT in action through a sampling experiment. In sampling experiments, we use the random number generator in our computer to draw random samples of different sizes over and over again. We did this for a dummy variable that equals one 80% of the time and for samples of size 10, 40, and 100. For each sample size, we calculated the t-statistic in half a million random samples using .8 as our value of μ.

Figures 1.2–1.4 plot the distribution of 500,000 t-statistics calculated for each of the three sample sizes in our experiment, with the standard normal distribution superimposed. With only 10 observations, the sampling distribution is spiky, though the outlines of a bell-shaped curve also emerge. As the sample size increases, the fit to a normal distribution improves. With 100 observations, the standard normal is just about bang on.

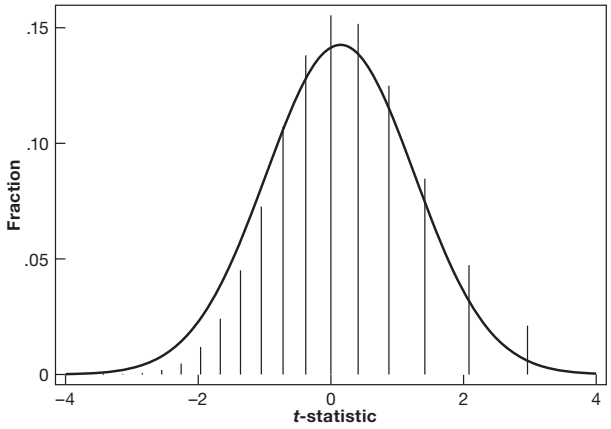
The standard normal distribution has a mean of 0 and standard deviation of 1. With any standard normal variable, values larger than ± 2 are highly unlikely. In fact, realizations larger than 2 in absolute value appear only about 5% of the time. Because the t-statistic is close to normally distributed, we similarly expect it to fall between about ± 2 most of the time. Therefore, it's customary to judge any t-statistic larger than about 2 (in absolute value) as too unlikely to be consistent with the null hypothesis used to construct it. When the null hypothesis is $\mu = 0$ and the t-statistic exceeds 2 in absolute value, we say the sample mean is *significantly different from*

FIGURE 1.2
The distribution of the t-statistic for the mean in a sample of size 10



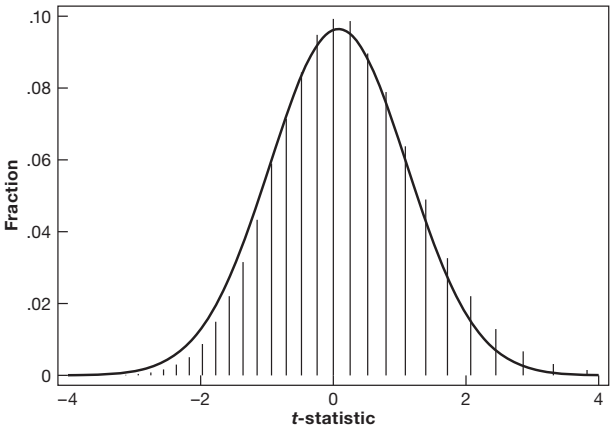
Note: This figure shows the distribution of the sample mean of a dummy variable that equals 1 with probability .8.

FIGURE 1.3
The distribution of the *t*-statistic for the mean in a sample of size 40



Note: This figure shows the distribution of the sample mean of a dummy variable that equals 1 with probability .8.

FIGURE 1.4
The distribution of the *t*-statistic for the mean in a sample of size 100



Note: This figure shows the distribution of the sample mean of a dummy variable that equals 1 with probability .8.

zero. Otherwise, it's not. Similar language is used for other values of μ as well.

We might also turn the question of statistical significance on its side: instead of checking whether the sample is consistent with a specific value of μ, we can construct the set of all values of μ that are consistent with the data. The set of such values is called a *confidence interval* for $E[Y_i]$. When calculated in repeated samples, the interval

$$\left[\bar{Y} - 2 \times \hat{SE}(\bar{Y}), \bar{Y} + 2 \times \hat{SE}(\bar{Y})\right]$$

should contain $E[Y_i]$ about 95% of the time. This interval is therefore said to be a *95% confidence interval* for the population mean. By describing the set of parameter values consistent with our data, confidence intervals provide a compact summary of the information these data contain about the population from which they were sampled.

Pairing Off

One sample average is the loneliest number that you'll ever do. Luckily, we're usually concerned with two. We're especially keen to compare averages for subjects in experimental treatment and control groups. We reference these averages with a compact notation, writing $\bar{Y}^1$ for $Avg_n[Y_i|D_i = 1]$ and $\bar{Y}^0$ for $Avg_n[Y_i|D_i = 0]$. The treatment group mean, $\bar{Y}^1$, is the average for the n_1 observations belonging to the treatment group, with $\bar{Y}^0$ defined similarly. The total sample size is $n = n_0 + n_1$.

For our purposes, the difference between $\bar{Y}^1$ and $\bar{Y}^0$ is either an estimate of the causal effect of treatment (if Y_i is an outcome), or a check on balance (if Y_i is a covariate). To keep the discussion focused, we'll assume the former. The most important null hypothesis in this context is that treatment has no effect, in which case the two samples used to construct treatment and control averages come from the same population. On the other hand, if treatment changes outcomes, the populations from which treatment and control observations are

drawn are necessarily different. In particular, they have different means, which we denote μ^1 and μ^0.

We decide whether the evidence favors the hypothesis that $\mu^1 = \mu^0$ by looking for statistically significant differences in the corresponding sample averages. Statistically significant results provide strong evidence of a treatment effect, while results that fall short of statistical significance are consistent with the notion that the observed difference in treatment and control means is a chance finding. The expression "chance finding" in this context means that in a hypothetical experiment involving very large samples—so large that any sampling variance is effectively eliminated—we'd find treatment and control means to be the same.

Statistical significance is determined by the appropriate t-statistic. A key ingredient in any t recipe is the standard error that lives downstairs in the t ratio. The standard error for a comparison of means is the square root of the sampling variance of $\bar{Y}^1 - \bar{Y}^0$. Using the fact that the variance of a difference between two statistically independent variables is the sum of their variances, we have

$$\begin{aligned} V\left(\bar{Y}^1 - \bar{Y}^0\right) &= V\left(\bar{Y}^1\right) + V\left(\bar{Y}^0\right) \\ &= \frac{\sigma_Y^2}{n_1} + \frac{\sigma_Y^2}{n_0} = \sigma_Y^2 \left[\frac{1}{n_1} + \frac{1}{n_0}\right]. \end{aligned}$$

The second equality here uses equation (1.5), which gives the sampling variance of a single average. The standard error we need is therefore

$$SE\left(\bar{Y}^1 - \bar{Y}^0\right) = \sigma_Y \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}.$$

In deriving this expression, we've assumed that the variances of individual observations are the same in treatment and control groups. This assumption allows us to use one symbol, σ_Y^2, for the common variance. A slightly more complicated formula allows variances to differ across groups even if the means are

the same (an idea taken up again in the discussion of robust regression standard errors in the appendix to Chapter 2).[17]

Recognizing that σ_Y^2 must be estimated, in practice we work with the estimated standard error

$$\hat{SE}\left(\bar{Y}^1 - \bar{Y}^0\right) = S(Y_i)\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}, \qquad (1.7)$$

where $S(Y_i)$ is the *pooled sample standard deviation*. This is the sample standard deviation calculated using data from both treatment and control groups combined.

Under the null hypothesis that $\mu^1 - \mu^0$ is equal to the value μ, the t-statistic for a difference in means is

$$t(\mu) = \frac{\bar{Y}^1 - \bar{Y}^0 - \mu}{\hat{SE}\left(\bar{Y}^1 - \bar{Y}^0\right)}.$$

We use this t-statistic to test working hypotheses about $\mu_1 - \mu_0$ and to construct confidence intervals for this difference. When the null hypothesis is one of equal means ($\mu = 0$), the statistic $t(\mu)$ equals the difference in sample means divided by the estimated standard error of this difference. When the t-statistic is large enough to reject a difference of zero, we say the estimated difference is statistically significant. The confidence interval for a difference in means is the difference in sample means plus or minus two standard errors.

Bear in mind that t-statistics and confidence intervals have little to say about whether findings are substantively large or small. A large t-statistic arises when the estimated effect of interest is large but also when the associated standard error is small (as happens when you're blessed with a large sample). Likewise, the width of a confidence interval is determined by

[17] Using separate variances for treatment and control observations, we have

$$SE\left(\bar{Y}^1 - \bar{Y}^0\right) = \sqrt{\frac{V^1(Y_i)}{n_1} + \frac{V^0(Y_i)}{n_0}},$$

where $V^1(Y_i)$ is the variance of treated observations, and $V^0(Y_i)$ is the variance of control observations.

statistical precision as reflected in standard errors and not by the magnitude of the relationships you're trying to uncover. Conversely, *t*-statistics may be small either because the difference in the estimated averages is small or because the standard error of this difference is large. The fact that an estimated difference is not significantly different from zero need not imply that the relationship under investigation is small or unimportant. Lack of statistical significance often reflects lack of statistical precision, that is, high sampling variance. Masters are mindful of this fact when discussing econometric results.