Chapter 10

Hypothesis Testing

Deciding Between Reality and Coincidence

Chapter Outline

Oh no. Not again. Your high-pressure sales contact is on the line, trying to sell you that miracle yield-enhancing additive to increase the productivity of your refinery. It looks like a good deal, but you are just not sure. You have been trying it out for a week (free, of course, for now), and—sure enough—the yield is up. But it is not up a whole lot, and, naturally, the process is variable, so it is hard to tell whether or not there is anything important going on. What you need is an objective assessment, but you know that what you will get from your contact on the phone is just another sales pitch: "The yield is up, isn't it? Well, what did I tell you? If you sign up today, we'll throw in a free engraved pen-and-pencil set! Blah blah blah." So you give the secret signal to your secretary, who says that you are in a meeting just now and will call back later.

Here is what is troubling you. Sure, the yield is up. But even if you do nothing special at all, you know that the yield fluctuates from day to day and from week to week about its long-run mean value. So the yield is up for one of two reasons: Either the additive is really working, or it is just a coincidence. After all, regardless of the additive, there is about a 50-50 chance that the week's yield would be higher than the long-term mean and about a 50-50 chance for it to be lower.

Look at this situation from the salesperson's point of view. Suppose for a moment that the additive is actually worthless and has no effect whatsoever on the yield. Next, convince managers at 100 different companies to try it out for a week. About 50 of these managers will find that their yield went down—no need to follow up those cases. But the other 50 or so will find slightly higher yields. Maybe some of these will even pay big money to continue using this worthless product.

What you need is a way of using the information gathered so far about the yield to help you determine if (on the one hand) it could reasonably be *just coincidence* that the yield was higher last week or if (on the other hand) you have convincing evidence that the additive really works. This type of separation-of-signal-from-noise is what hypothesis testing can do to help you, as a manager, filter out the unimportant random facts that reach you so that you can concentrate on important information.

Practical Business Statistics, Seventh Edition. http://dx.doi.org/10.1016/B978-0-12-804250-2.00010-9

Hypothesis testing uses data to decide between two possibilities (called *hypotheses*).[1] It can tell you whether the results you are witnessing are just coincidence (and could reasonably be due to chance) or are likely to be real. Some people think of hypothesis testing as a way of using statistics to make decisions. Taking a broader view, an executive might look at hypothesis testing as *one component* of the decision-making process. Hypothesis testing by itself probably should not be used to tell you whether to buy a product or not; nonetheless, it provides critically important information about how substantial and effective the product is.

In this chapter, you will learn how **hypothesis testing** uses data to decide between two possibilities, often to distinguish structure from mere randomness as a helpful input to executive decision making. We will define a *hypothesis* as any statement about the population; the data will help you decide which hypothesis to accept as true. There will be two hypotheses that play different roles: The *null hypothesis* represents the default, to be accepted in the absence of evidence against it; the *research hypothesis* has the burden of proof, requiring convincing evidence for its acceptance. Accepting the null hypothesis is a weak conclusion, whereas rejecting the null and accepting the research hypothesis is a strong conclusion and leads to a *statistically significant* result. Every hypothesis test can produce a *p-value* (using statistical software) that tells you how surprised you would be to learn that the null hypothesis had produced the data, with smaller p-values indicating more surprise and leading to significance. By convention, a result is *statistically significant* if $p < 0.05$, is *highly significant* if $p < 0.01$, is *very highly significant* if $p < 0.001$, and is *not significant* if $p > 0.05$.

Whenever you have an estimator together with its standard error, you may perform hypothesis testing. We use *Student's t-test* to see whether or not the population mean is equal to a reference value (a known, fixed number that does not come from the sample data); we often perform a *two-sided test* because we are interested in both sides of (larger and smaller than) the reference value. The outcome of the test is determined by checking if the sample average is farther from the reference value than random chance would reasonably allow. The test may be based either directly on the p-value (if available) or equivalently on either the two-sided confidence interval (from Chapter 9) or on the t-statistic, which measures the separation in units of standard errors (and we know that more than about two standard errors would be unlikely).

There are two types of errors that you might make when hypothesis testing. The *type I error* is committed when the null hypothesis is true, but you reject it and (wrongly) declare that your result is statistically significant; the probability of this error is controlled, conventionally at the 5% level (but you may set this *test level* or *significance level* to be other values, such as 1%, 0.1%, or perhaps even 10%). The *type II error* is committed when the research hypothesis is true, but you (wrongly) accept the null hypothesis instead and declare the result *not* to be significant; the probability of this error is not easily controlled. Note that there is no notion of the probability of a hypothesis being true because these are probabilities about the data given a hypothesis. The *assumptions for hypothesis testing* are the same as for confidence intervals: (1) the data set is a random sample from the population of interest, and (2) either the quantity being measured is approximately normal, or else the sample size is large enough that the central limit theorem ensures that the sample average is approximately normally distributed.

You will also see additional variations on hypothesis testing in this chapter. One such variation is the *one-sided test* that is better able to detect significance on the chosen side. Another variation is to test whether a new observation comes from the same population as a sample (instead of testing the mean of the population). Finally, you will see methods for testing two samples (eg, *A/B testing*) using either the *paired t-test* or the *unpaired t-test* (depending on whether or not there is a natural pairing of the two samples) to decide whether the two means are identical.

10.1 HYPOTHESES ARE NOT CREATED EQUAL!

A **hypothesis** is a statement about how the world is. It is a statement about the *population*. A hypothesis is not necessarily true; it can be either right or wrong, and you use the sample data to help you decide. When you know everything, there is no need for statistical hypothesis testing. When there is uncertainty, statistical hypothesis testing will help you learn as much as possible from the information available to you.

You will ordinarily work with a *pair* of hypotheses at a time. The data will help you decide which of the two will prevail. But the two hypotheses are not interchangeable; each one plays a different, special role. In particular, we ask whether the null hypothesis could reasonably have produced the data. If the data have a small probability (the "p-value") of occurring when the null hypothesis is true (so that p is less than the conventional 5%, threshold) then we will decide to reject the null hypothesis, accept the research hypothesis, and declare significance. Otherwise, if p is larger than 5%, we will accept the null hypothesis as a weak conclusion without declaring significance.

The Null Hypothesis

The **null hypothesis**, denoted H_0, represents the *default* statement that you will accept *unless you have convincing evidence to the contrary*. This is a very favored position. If your data are sketchy or too variable, you will end up accepting the null hypothesis because it has the "benefit of the doubt." In fact, you can end up accepting the null

1. The singular is one *hypothesis*, and the plural is two *hypotheses* (pronounced *hypotheses*).

hypothesis without really proving anything at all, putting you in a fairly weak position. Thus, it can make an important difference which of your two hypotheses you refer to as the null hypothesis.

The null hypothesis is often the *more specific* hypothesis of the two. For example, the null hypothesis might claim that the population mean is exactly equal to some known reference value or that an observed difference is just due to random chance. To see that the hypothesis of random chance is indeed more specific, note that *non*random things can have very many different kinds of structure, but randomness implies a lack of structure.

The Research Hypothesis

The **research hypothesis**, denoted H_1, is to be accepted only if there is convincing statistical evidence that would rule out the null hypothesis as a reasonable possibility. The research hypothesis is also called the **alternative hypothesis**. Accepting the research hypothesis represents a much stronger position than accepting the null hypothesis because it requires convincing evidence.

People are often interested in establishing the research hypothesis as their hidden agenda, and they set up an appropriate null hypothesis solely for the purpose of refuting it. The end result would be to show that "it's not just random, and so here's my explanation…." This is an accepted way of doing research. Since people have fairly creative imaginations, the research community has found that by requiring that the null hypothesis of pure randomness be rejected before publication of a research finding, they can effectively screen many wild ideas that have no basis in fact. This approach does not *guarantee* that all research results are true, but it does screen out many incorrect ideas.

In deciding which hypothesis should be the research hypothesis, ask yourself, "Which one has the *burden of proof*?" That is, determine which hypothesis requires the more convincing evidence before you decide to believe in it. This one will be the research hypothesis. Do not neglect your own self-interest! Feel free to shift the burden of proof onto those trying to sell you things. Make them prove their claims!

Results, Decisions, and *p*-Values

There are two possible outcomes of a hypothesis test: either "accept the null hypothesis" or "reject the null hypothesis, accept the research hypothesis, and declare significance." The result is defined to be **statistically significant** whenever you accept the research hypothesis because you have eliminated the null hypothesis as a reasonable possibility. By convention, the two possible outcomes are described as follows:

Results of a Hypothesis Test

Either:	Accept the null hypothesis, H_0, as a reasonable possibility.	A weak conclusion; not a significant result.
Or:	Reject the null hypothesis, H_0, and accept the research hypothesis, H_1	A strong conclusion; a significant result.

Note that we *never* speak of rejecting the research hypothesis. The reason has to do with the favored status of the null hypothesis as default. Accepting the null hypothesis merely implies that you do not have enough evidence to decide against it. When we decide to "accept" a null hypothesis, H_0, we should not necessarily believe that it is true, and should recognize that the research hypothesis H_1 might well *actually* be true, but because the null hypothesis might be true (and has favored status) we will accept the null hypothesis. While accepting the null hypothesis as a reasonably possible scenario that could have generated the data, we nonetheless recognize that there are many other such believable scenarios *close to* the null hypothesis that also might have generated the data. For example, when we accept the null hypothesis that claims the population mean is \$2,000, we have not usually ruled out the possibility that this mean is \$2,001 or \$1,999. For this reason, some statisticians prefer to say that we "fail to reject" the null hypothesis rather than simply say that we "accept" it.

It may help you to think of the hypotheses in terms of a criminal legal case. The null hypothesis is "innocent," and the research hypothesis is "guilty." Since our legal system is based on the principle of "innocent until proven guilty," this assignment of hypotheses makes sense. Accepting the null hypothesis of innocence says that there was not enough evidence to convict; it does not prove that the person is truly innocent. On the other hand, rejecting the null hypothesis and accepting the research hypothesis of guilt says that there is enough evidence to rule out innocence as a possibility and to convincingly establish guilt. We do not have to rule out guilt in order to find someone innocent, but we do have to rule out innocence in order to find someone guilty.

While there is a vast variety of hypothesis tests covered here and in later chapters, depending on the type of data and the chosen model, and each test has its own particular detailed calculations (and its own important intuition) there is a useful, unifying fact: Every hypothesis test can produce a *p*-value that is interpreted in the same way:

Using the *p*-Value to Perform a Hypothesis Test

If $p > 0.05$:	Accept the null hypothesis, H_0, as a reasonable possibility.
If $p < 0.05$:	Reject the null hypothesis, H_0, and accept the research hypothesis, H_1

The ***p*-value** tells you how surprised you would be to learn that the null hypothesis had produced the data, with smaller p-values indicating more surprise and leading to rejection of H_0 when p is less than the conventional 5% threshold. The p-value is computed (by statistical software) while assuming the null hypothesis is true, and tells the probability of observing your data (or data even farther from the null hypothesis). By convention, if the null hypothesis produces data like yours less than 5% of the time, this low probability is taken as evidence against the null hypothesis and leads to its rejection.[2]

Examples of Hypotheses

Following are some examples of null and research hypotheses about the population. Note in each case that they cannot both be true, and that the data will be used to decide which one to accept.

1. *The situation*: A randomly selected group of 200 people view an advertisement, and the number of people who buy the product during the next week is recorded.
   The null hypothesis: The ad has no effect. That is, the percentage of buyers among those in the general population who viewed the ad is *exactly equal* to the baseline rate for those who did *not* view the ad in the general population. This baseline rate is known to be 19.3%, based on extensive past experience.
   The research hypothesis: The ad has an effect. That is, the percentage of buyers among those in the general population who viewed the ad is *different from* the baseline rate of 19.3% representing those buyers who did *not* view the ad in the general population.
   Discussion: Note that these hypotheses are statements about the general population, not about the 200 people in the sample. The sample evidence accumulated by observing the behavior of 200 randomly selected people will help decide which hypothesis to accept. Since the null hypothesis gives an exact value for the percentage, it is more specific than the research hypothesis, which specifies a large range (ie, any percentage different from 19.3%). Also note that when you decide that an ad is effective, you will be making a strong statement since this is the research hypothesis, and you will be able to claim that the effect of the ad is *statistically significant*. It is as if you are saying "OK. If this ad works as well as we all think it does, let us give it a chance to prove it to us. Or, on the other hand, if it will be a disaster to sales, let us find that out also."
2. *The situation*: You are evaluating the yield-enhancing additive described at the start of this chapter.
   The null hypothesis: The additive has no effect on the long-run yield, an amount known from past experience.
   The research hypothesis: The additive has some effect on the long-run yield.
   Discussion: The null hypothesis is more specific. Both hypotheses refer to the population (long-run yield) and not just to the particular results of last week (the sample). Your default is that the additive has no effect, and to convince you otherwise will require a conclusive demonstration. The burden of proof is on them (the manufacturers of the additive) to show effectiveness. It is not up to you to prove to them that it is not effective.
3. *The situation*: Your firm is being sued for gender discrimination, and you are evaluating the documents filed by the other side. They include a statistical hypothesis test based on salaries of men and women that finds a "highly significant difference" on average between men's and women's salaries.
   The null hypothesis: Men's and women's salaries are equal except for random variation. That is, the population from which the men's salaries were sampled has the same mean as the population from which the women's salaries were sampled. Another way to view this idea is that the actual salary differences between men and women are not unreasonably different from what you might get if you were to put all salaries into a hat, mix them up well, and deal them out to people without regard to gender.
   The research hypothesis: The population means of men's and women's salaries are different (even before random variation is added).
   Discussion: Note the use of idealized populations here. Since these employees are not a random sample in any real sense, the hypotheses refer to an idealized population for each gender (one population of similar men's salaries, the other for the women). With the null hypothesis, the two populations have equal means, while with the research hypothesis, the means are different for the two genders. Your firm is in trouble since the null hypothesis has been rejected and the research hypothesis has been accepted. This is a strong conclusion that goes against you. But all is not necessarily lost. Do not forget that statistical methods generally tell you about the numbers only and not about why the numbers are this way. The salary differential might be due directly to gender discrimination, or it might be due to other factors, such as education, experience, and ability. A statistical hypothesis test that addresses only gender and salary cannot tell which factors are

2. When you find a small p-value, it is as though you ask yourself "Do I feel lucky?" because you would have to be lucky to see such data if the null hypothesis were true. Because we are not generally lucky all of the time (except perhaps in the movies) we then find the null hypothesis less believable. If p were one in a million, then either you were incredibly lucky (which is possible, but unlikely) or else the null hypothesis is false and the research hypothesis is true.

responsible.[3] Also, the hypothesis test results could be wrong, since errors can happen whenever statistical methods are used.

10.2 TESTING THE POPULATION MEAN AGAINST A KNOWN REFERENCE VALUE: THE *t*-TEST

The simplest case of hypothesis testing involves testing the population mean against a known reference value. This **reference value** is a known, fixed number μ_0 that does not come from the sample data. The hypotheses are as follows:

The Null and the Research Hypothesis

$$H_0: \mu = \mu_0$$

The null hypothesis H_0 claims that the unknown population mean, μ, is *exactly equal* to the known reference value, μ_0.

$$H_1: \mu \neq \mu_0$$

The research hypothesis H_1 claims that the unknown population mean, μ, is *not equal* to the known reference value, μ_0.

This is a **two-sided test** because the research hypothesis includes values for the population mean μ on both sides (smaller and larger) of the reference value, μ_0.[4] Note that there are actually *three* different numbers involved here that have something to do with an average or mean value:

μ is the unknown population mean, which you are interested in learning about.
μ_0 is the known reference value you are testing against.
$\bar{X}$ is the known sample average you will use to decide which hypothesis to accept. Of these three numbers, this is the only one that is at all random because it is computed from the sample data. Note that $\bar{X}$ estimates, and hence represents, μ.

The hypothesis test proceeds by comparing the two known numbers $\bar{X}$ and μ_0 against each other. If they are more different than random chance could reasonably account for, then the null hypothesis $\mu = \mu_0$ will be rejected because $\bar{X}$ provides information about the unknown mean, μ. If $\bar{X}$ and μ_0 are fairly close to each other, then the null hypothesis $\mu = \mu_0$ will be accepted. But how close is close? Where will we draw the line? Closeness must be based on $S_{\bar{X}}$, since this standard error tells you about the randomness in $\bar{X}$. Thus, if $\bar{X}$ and μ_0 are a sufficient number of standard errors apart, then you have convincing evidence against μ being equal to μ_0.

There are three different ways of carrying out the hypothesis test and getting the results. The first method uses the *p*-value from computer software and is the easiest method because you may simply compare the *p*-value to the standard threshold of 0.05 or 5%. The second method uses confidence intervals, which we covered in the preceding chapter. This is the most intuitive method because (a) you already know how to construct and interpret a confidence interval, and (b) the confidence interval is directly meaningful because it is in the same units as your data (eg, dollars, people, defect rates). The third method (based on the *t*-statistic) is more traditional but less intuitive since it requires that you calculate something new that is not in the same units as your data and that must be compared to the appropriate critical *t*-value before you know the result.

It really does not matter which of the three methods (*p*-value, confidence interval or *t*-statistic) you use for hypothesis testing since they always give the same answer in the end. While the *p*-value method is easiest, you may also want to use the confidence interval method much of the time since it provides the most intuitive information about the situation. However, you will also want to know how to use the *t*-statistic method because it is still commonly used in practice. Since the three methods give the same result, any one of them may be called a *t-test*.

Using the *p*-Value: The Easy Way

If your statistical software produces a *p*-value (indicating the probability that the null hypothesis can produce data like yours, with lower probabilities indicating surprise and leading to rejection of the null hypothesis) recall that it is very easy to reach a hypothesis-testing decision, because *p* less than 0.05 leads to statistical significance while *p* greater than 0.05 does not. Please recall also that "statistically significant" is the strong conclusion declared when you reject the null hypothesis and accept the research hypothesis instead. The weak conclusion, "not statistically significant" occurs when you accept the null hypothesis, which was the default to be accepted in the absence of evidence against it.

For example, if you find that $p = 0.0371$ then you know immediately that you have a significant result because $0.0371 < 0.05$ and you may write proudly that the result is "significant ($p < 0.05$)." If, on the other hand, you find that $p = 0.862$, then you do not have a significant result because $0.0862 > 0.05$ and you may write properly that the result is "not significant ($p > 0.05$)."[5] If you find that

3. In a later chapter, you will learn how *multiple regression* can adjust for other factors (such as education and experience) and can provide an *adjusted estimate* of the effect of gender on salary while holding these other factors constant.
4. You will learn about *one-sided* hypothesis testing in Section 10.4.
5. These statements with parentheticals "significant ($p < 0.05$)" and "not significant ($p > 0.05$)" are standard accepted ways to communicate the results of a hypothesis test. You might wonder what happens when *p* seems exactly equal to 0.05. You might look for additional digits of accuracy to make the decision. If this fails and *p* is exactly equal to 0.05 (which does not happen often) we might say that the result is "borderline significant."

$p = 0.000001$ then you have very strong evidence (one in a million) against the null hypothesis because data like yours are *very* unlikely to have occurred if the null hypothesis is true. Please note that the p-value is a probability about your random *data* occurring while assuming that the null hypothesis (a statement about the world) is true. The p-value is *not* a probability about a hypothesis being true or not.

While it is easy and quick to reach a decision with the p-value method, important additional intuition about your situation can be gained by using the following methods, which reach the same answer as the p-value method.

Using the Confidence Interval: The Intuitive Way, Same Answer

Here is how to test the null hypothesis H_0: $\mu = \mu_0$ against the research hypothesis H_1:$\mu \neq \mu_0$ based on a random sample from the population. First, construct the 95% confidence interval based on $\bar{X}$ and $S_{\bar{X}}$ in the usual way (see Chapter 9). Then look to see whether or not the reference value, μ_0, is in this interval. If μ_0 is outside the confidence interval, then this reference value is not a reasonable value for the population mean, μ, and you will accept the research hypothesis; otherwise, you will accept the null hypothesis. This is illustrated in Fig. 10.2.1. There are a number of equivalent ways of describing the result of such a hypothesis test. Your decision in each case may be stated as indicated in Table 10.2.1.

Why does this method work? Remember that the confidence interval statement says that the probability that μ is in the (random) confidence interval is 0.95. Assume for a moment that the null hypothesis is true, so that $\mu = \mu_0$ exactly. Then the probability that μ_0 is in the confidence interval is also 0.95. This says that when the null hypothesis is true, you will make the correct decision in approximately 95% of all cases and be wrong only about 5% of the time. In this sense, you now have a decision-making process with exact, controlled probabilities. For a more detailed discussion of the various types of errors in hypothesis testing, please see Section 10.3.

TABLE 10.2.1 Deciding a Hypothesis Test about the Population Mean Using the Confidence Interval

If the reference value, μ_0, is in the confidence interval from $\bar{X} - tS_{\bar{X}}$ to $\bar{X} + tS_{\bar{X}}$ then:
Accept the null hypothesis, H_0, as a reasonable possibility
Do not accept the research hypothesis, H_1
The sample average, $\bar{X}$, is *not significantly different* from the reference value, μ_0
The observed difference between the sample average, $\bar{X}$, and the reference value, μ_0, could reasonably be due to random chance alone
The result is *not statistically significant* (All of the preceding statements are equivalent.)
If the reference value, μ_0, is not in the confidence interval from $\bar{X} - tS_{\bar{X}}$ to $\bar{X} + tS_{\bar{X}}$ then:
Accept the research hypothesis, H_1
Reject the null hypothesis, H_0
The sample average, $\bar{X}$, is *significantly different* from the reference value, μ_0
The observed difference between the sample average, $\bar{X}$, and the reference value, μ_0, could not reasonably be due to random chance alone
The result is *statistically significant* (All of the preceding statements are equivalent.)

Example

Does the "Yield-Increasing" Additive Really Work?

Recall the (supposedly) yield-increasing additive you were considering purchasing at the start of this chapter (with

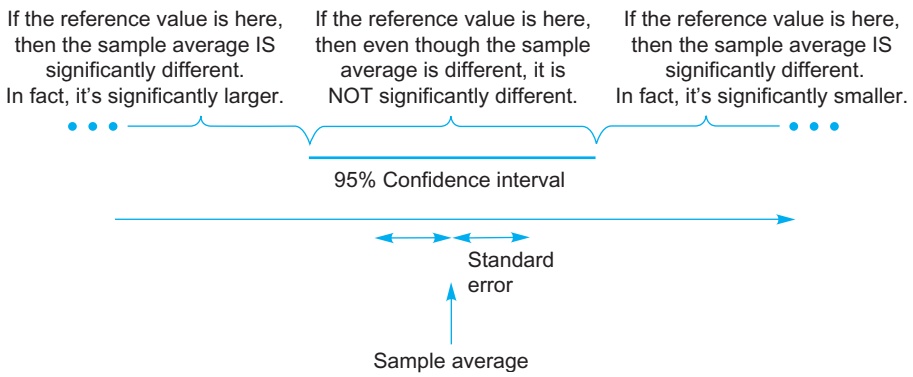

FIG. 10.2.1 A hypothesis test for the population mean can be decided based on the confidence interval. The question is whether or not the population mean could reasonably be equal to a given reference value. If the reference value is in the interval, then it is reasonably possible. If the reference value is outside the interval, then you would decide that it is not the population mean.

Example—cont'd

additional details in an example of Section 10.1). Suppose that the basic facts of the matter are as shown in Table 10.2.2. Your data set consists of $n=7$ observations of the yield taken while the additive was in use. Your population therefore should be all possible daily yields using the additive; in particular, the population mean, μ, should be the long-term mean yield achieved while using the additive (this is unknown and therefore not listed in the table). The sample average, $\bar{X}$, provides your best estimate of μ.

Indeed, it looks as if the additive is working well. The average daily yield achieved with the additive ($\bar{X}=39.6$ tons) is 7.5 tons higher than the mean daily long-term yield ($\mu_0=32.1$ tons) you expect without the additive. This is no surprise. In hypothesis testing, the reference value is almost never *exactly* equal to the observed value ($\bar{X}$ here). The question is if they are more different than random chance alone would reasonably allow. A histogram of the data, with the sample average and the reference value indicated, is shown in Fig. 10.2.2.

In preparation for hypothesis testing, you identify the hypotheses, which may be stated directly in terms of the known reference value, $\mu_0=32.1$ tons. (There is no reason to continue to use the symbolic notation μ_0 instead of its known value in the formal hypothesis statements.) The hypotheses are as follows:

TABLE 10.2.2 Basic Facts for the "Yield-Increasing" Additive

Average daily yield over the past week	$\bar{X}$	39.6 tons
Standard error	$S_{\bar{X}}$	4.2 tons
Sample size	n	7 days
Your known mean daily long-term yield (without additive)	μ_0	32.1 tons

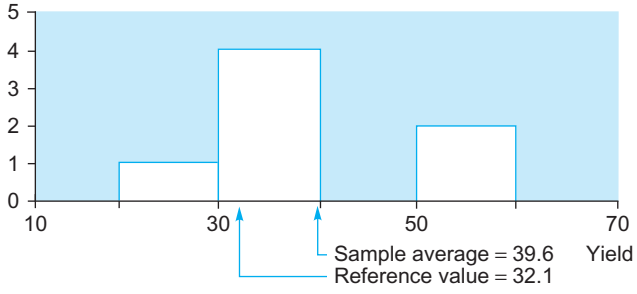

FIG. 10.2.2 A histogram of the seven yields obtained with the additive. The sample average summarizes the available data and is higher than the reference value. But is it significantly higher? The result of a hypothesis test will tell whether this sample histogram *could reasonably have come* from a population distribution whose mean is the reference value.

TABLE 10.2.3 Hypothesis Test Result for the "Yield-Increasing" Additive

Since the reference value, $\mu_0=32.1$ tons, is in the confidence interval from 29.3 to 49.9 tons
Accept the null hypothesis, H_0:$\mu=32.1$ tons, as a reasonable possibility
Do *not* accept the research hypothesis, H_1: $\mu\neq32.1$ tons
The sample average yield, $\bar{X}=39.6$, is *not significantly different* from the reference value, $\mu_0=32.1$
The observed difference between the sample average yield, $\bar{X}=39.6$, and the reference value, $\mu_0=32.1$, could reasonably be due to random chance alone
The result is *not statistically significant* (All of the preceding statements are equivalent.)

$$H_0:\mu=32.1 \text{ tons}$$

The null hypothesis claims that the unknown long-term mean daily yield with the additive, μ, is exactly *equal* to the known reference value, $\mu_0=32.1$ tons (without the additive).

$$H_1:\mu\neq32.1 \text{ tons}$$

The research hypothesis claims that the unknown long-term mean daily yield with the additive, μ, is *not equal to* the known reference value, $\mu_0=32.1$ tons (without the additive).

Next, to facilitate the hypothesis test, compute the 95% confidence interval in the usual way using the critical t-value 2.446912 for $n-1=6$ degrees of freedom:

> We are 95% sure that the long-term mean daily yield with the additive is somewhere between 29.3 and 49.9 tons.

Finally, to perform the actual hypothesis test, simply look to see whether or not the reference value, $\mu_0=32.1$ tons, is in the confidence interval.[6] It is in the interval because 32.1 is indeed between 29.3 and 49.9. That is, $29.3\leq32.1\leq49.9$ is a true statement. Your hypothesis test result is therefore not significant, as shown in Table 10.2.3.

For reference, the p-value from statistical software for this test is $p=0.124$, which is not significant because it is greater than 0.05, in agreement with the result of the confidence interval method.

The sample average daily yield with the additive, $\bar{X}=39.6$ tons; is not significantly different from the long-term mean daily yield without the additive, $\mu_0=32.1$ tons. This result is inconclusive and ambiguous. You do not have convincing evidence in favor of the additive. When you next talk to the high-pressure sales contact who is trying hard to sell you the stuff, you will have the confidence to say that *even though the yield is up, it is not up significantly*, and you are not yet convinced that the additive is worthwhile.

Does this test prove that the additive is ineffective? No. It *might* be effective; you just do not have convincing evidence one way or the other.

(*Continued*)

Example—cont'd

What else might be done to resolve the issue? Your sales contact might suggest that you use it for another month—free of charge, of course—to see if the additional information will be convincing enough. Or you might suggest this solution to your contact, if you have the nerve.

6. It would be silly to check whether or not $\bar{X}$ is in the interval, since, of course, $\bar{X}$ will always be in the confidence interval. The question here is if the known *reference value*, μ_0, is in the confidence interval.

Example

Should Your Company Sponsor the Olympics?

Why do some companies choose to pay hundreds of millions of dollars each in order to be one of the official sponsors of the Olympic Games? Research by Miyazaki and Morgan points out some of the pluses (the opportunity for marketing visibility and enhancement of the corporate image) and minuses (many consumers cannot correctly identify official sponsors and the high cost).[7] In addition, Miyazaki and Morgan performed an "event study" to see whether the market value (as measured by the stock price) of companies tends to increase or decrease significantly around the time of the official Olympic sponsorship announcement in major print media (such as the *Wall Street Journal* or the *New York Times*).

In the financial markets, over the short term, the stock price of a company generally moves up and down more or less at random, in accordance with the random walk theory of efficient markets. Therefore, the null hypothesis says that the change in a company's market value near the time of the official Olympic sponsorship announcement will be zero on average. If Olympic sponsorship adds value, then we would expect to find market value significantly increased; if sponsorship hurts value, then we would find a decrease in market value on average for these companies.

A test statistic called "CAR" (which stands for "cumulative abnormal return") is used to measure the amount of value added to the company, as a percentage over a set time period. Here are some of the results from Miyazaki and Morgan's research, for which some companies showed an increase in value and others showed a decrease:

Average change in company market value from 4 Days before an official olympic sponsorship announcement until the day of the announcement, as measured by C.A.R, along with its standard error and sample size

Average change in market value	$\bar{X}$	1.24
Standard error	$S_{\bar{X}}$	0.59
Sample size (number of firms)	n	27

The question here is: Does sponsoring the Olympics enhance a company's value? The answer will be found by performing a hypothesis test. Why not just use the fact that the average company increased its value by 1.24 (in percentage points) to say that Olympic sponsorship enhances value? Because this is a result for a *sample* of 27 companies, and it may or may not represent the larger population of Olympic sponsoring companies in general. In order to infer the effect of sponsorship on value in general, based on the average from a sample, we will use a hypothesis test.

Since we will want to be convinced before concluding that Olympic sponsorship has any effect (positive or negative), this has the burden of proof and will be the research hypothesis. The null hypothesis will claim that Olympic sponsorship has no effect. If we let μ denote the mean percentage change for the larger population of sponsoring firms (where that population consists of companies that are similar to those included in the study sample, viewing this sample as a random sample from the population), the hypotheses are as follows:

$$H_0: \mu = 0$$

The null hypothesis claims that the unknown mean effect μ of Olympic sponsorship on company value is exactly *equal* to the known reference value $\mu_0 = 0$.

$$H_1: \mu \neq 0$$

The research hypothesis claims that the unknown mean effect μ of Olympic sponsorship on company value is *not equal* to the known reference value $\mu_0 = 0$.

Next, to facilitate the hypothesis test, compute the 95% confidence interval in the usual way using a critical t-value of 2.055529 with $n = 27$ companies:

> We are 95% sure that the mean effect μ of Olympic sponsorship on company value is somewhere between 0.03 and 2.45.

Finally, to perform the hypothesis test, simply check whether or not the reference value $\mu_0 = 0$ is in the interval. It is *not* in the interval because 0 is not between 0.03 and 2.45. Your t-test result is therefore as shown in Table 10.2.4.

TABLE 10.2.4 Hypothesis Test Result for the Value of Becoming an Official Olympic Sponsor

Since the reference value, $\mu_0 = 0$, is not in the confidence interval from 0.03 to 2.45,
Accept the research hypothesis, $H_1: \mu \neq 0$
Reject the null hypothesis, $H_0: \mu = 0$
The sample average score, $\bar{X} = 1.24$, is *significantly different* from the reference value, $\mu_0 = 0$
The observed difference between the sample average score, $\bar{X} = 1.24$, and the reference value, $\mu_0 = 0$, could not reasonably be due to random chance alone
The result is *statistically significant* (All of the preceding statements are equivalent.)

Example—cont'd

Based on the performance of company stock, the announcement of an Olympic sponsorship has a *statistically significant positive effect* on the value of the company.[8] The result is conclusive. You do have convincing evidence that sponsoring the Olympics, in general, enhances the value of a company. Even though the effect may be small (only 1.24 percentage points), the hypothesis test has declared that it cannot be dismissed as a mere random stock price fluctuation.

Does this test absolutely prove that, if the larger population of companies sponsoring the Olympics could be studied, the resulting mean change in company stock value would be positive? Not really. Absolute proof is generally impossible in the presence of even a small amount of randomness. You have convincing evidence but not absolute proof. This says that you might be making an error in rejecting the null hypothesis and accepting the research hypothesis here, although an error is not very likely. These ever-present errors will be discussed in Section 10.3.

More recently, positive effects of event sponsorship were also found by Zarantonello and Schmitt who used *p*-value notation, along with the terms "significant" and "hypothesis" in their advertising research, which states that: "OBE [Overall Brand Equity] scores were the dependent variable, and whether the score was pre or post-event was the factor. The analysis reported a significant difference between the two measures, with pre-event OBE = 4.18 and post-event OBE = 4.49 ($p < .05$). Hypothesis 1 [post-event brand equity higher than pre-event brand equity] was thus confirmed."[9]

7. A.D. Miyazaki and A.G. Morgan, "Assessing Market Value of Event Sponsoring: Corporate Olympic Sponsorships," *Journal of Advertising Research*, January–February 2001, pp. 9–15. The "event study" methodology they used also includes careful adjustments for overall stock-market movements and for the risk level of each company, where the event is the announcement of an Olympic sponsorship.
8. You may claim a *statistically significant positive* effect because (1) the result is statistically significant and (2) the effect as measured by $\bar{X}$ is positive (ie, $\bar{X}$ is a positive number, larger than $\mu_0 = 0$).
9. L. Zarantonello and B.H. Schmitt (2013) "The Impact of Event Marketing on Brand Equity: The Mediating Roles of Brand Experience and Brand Attitude," International Journal of Advertising, Vol. 32, No. 2, p. 255-280, accessed at https://www0.gsb.columbia.edu/mygsb/faculty/research/pubfiles/5932/event_marketing_brand_equity.pdf on November 13, 2015.

If you have a binomial situation, it is straightforward to test whether the population percentage π is equal to a given reference value π_0, provided n is not too small. However, please note a source of potential confusion: the notation p would then have two very different meanings that would need to be distinguished: the binomial percentage as observed from the sample, and the p-value from the hypothesis test (if it is calculated). The situation and the procedure for the binomial case are not very different from testing a population mean because once you have an estimator and its standard error, the confidence interval and t-statistic are formed in the same way. These similarities are shown in the following table:

	Normal	Binomial
Population mean	μ	π
Reference value	μ_0	π_0
Null hypothesis	H_0: $\mu = \mu_0$	H_0: $\pi = \pi_0$
Research hypothesis	H_1: $\mu \neq \mu_0$	H_1: $\pi \neq \pi_0$
Data	$X_1, X_2, \ldots, Xn$	X occurrences out of n trials
Estimator	$\bar{X}$	$p = X/n$
Standard error	$S_{\bar{X}} = S/\sqrt{n}$	$S_p = \sqrt{p(1-p)/n}$
Confidence interval	From $\bar{X} + tS_{\bar{X}}$ to $\bar{X} + tS_{\bar{X}}$	From $p - tSp$ to $p + tSp$
t-Statistic	$t = (\bar{X} - \mu_0)/S_{\bar{X}}$	$t = (p - \pi_0)/S_p$

Example

Pushing the Limits of Production (A Binomial Situation)

One of the mysteries of producing electronic chips (for computers, smartphones, tablets, TVs, cars, refrigerators, etc.) is that you cannot tell for sure how good the results are until you test them. At that point, you find that some are unacceptable, others are fine, and some are especially good. These especially good ones are separated and sold at a premium as "extra fast" because they process information more quickly than the others.

Your goal has been to improve the production process to the point where more than 10% of the long-run production can be sold as extra-fast chips. Based on a sample of 500 recently produced chips, you plan to perform a hypothesis test to see if the 10% goal has been exceeded, if you are far short, or if it is too close to call.

Because of recent improvements to the production process, you are hopeful. There were 58 extra-fast chips, giving an estimated rate of 11.6%, which exceeds 10%. But did you *significantly* exceed the goal, or were you just lucky? You would like to know before celebrating.

Let us model this as a binomial situation in which each chip is either extra fast or not. The binomial probability π represents the probability of being extra fast. The sample size is $n = 500$, the observed count is $X = 58$, and the sample proportion is $p = 11.60\%$. The reference value is $\pi_0 = 10\%$.

Hypothesis testing for a binomial (with sufficiently large n) is really no different from testing with quantitative data. After all, in each case you have an estimate ($\bar{X}$ or p), a standard error ($S_{\bar{X}}$ or S_p), and a reference value (μ_0 or π_0). Here are the formal hypothesis statements for this binomial situation:

$$H_0 : \pi = 10\%$$

The null hypothesis claims that extra-fast chips represent 10% of production.

$$H_1 : \pi \neq 10\%$$

The research hypothesis claims that the rate is different from 10%: either higher (Hooray! Time to celebrate!) or lower (Uh-oh, time to make some adjustments!).

The 95% confidence interval is computed in the usual way for a binomial, based on the standard error

(*Continued*)

Example—cont'd

$S_p = \sqrt{p(1-p)/n} = 0.014321$ and a critical t-value of 1.964729. The interval is found to extend from 8.8% to 14.4%:

> You are 95% sure that extra-fast chips are being produced at a rate somewhere between 8.8% and 14.4% of total production.

Finally, to perform the hypothesis test, simply see whether or not the reference value $\pi_0 = 10\%$ is in the interval. It *is* in the interval because 10% is between 8.8% and 14.4%. Your t-test result is therefore as shown in Table 10.2.5.

The observed rate of production of extra-fast chips is not statistically significantly different from 10%. You do not have enough information to tell whether the rate is conclusively either higher or lower. The result is inconclusive. Although 11.6% looked like a good rate (and actually exceeds the goal of 10%), it is not significantly different from the goal. Since 11.6% may be just randomly different from the 10% goal, you do not have strong evidence that the goal has been reached.

For reference, the p-value from statistical software for this test is $p = 0.264$, which is not significant because it is greater than 0.05, in agreement with the result of the confidence interval method.

Remember that you are doing statistical inference. You are not just interested in these particular 500 chips. You would like to know about the long-run production rate for many more chips, with the machinery running as it is now. Statistical inference has told you that the rate is so close to 10% that you cannot tell whether or not the goal has been reached yet.

You might decide to collect more data from tomorrow's production to see if the added information will allow you to show that the goal has been reached (by accepting the research hypothesis, you hope, with an observed rate *significantly higher* than 10%). On the other hand, rather than just squeak by, you might want to hedge your bets by instituting some more improvements.

TABLE 10.2.5 Hypothesis Test Result for Chip Production

Since the reference value, $\pi_0 = 10\%$, is in the confidence interval from 8.8% to 14.4%
Accept the null hypothesis, H_0: $\pi = 10\%$, as a reasonable possibility
Do not accept the research hypothesis, H_1: $\pi \neq 10\%$
The sample proportion, $p = 11.6\%$, is not significantly different from the reference value, $\pi_0 = 10\%$
The observed difference between the sample proportion, $p = 11.6\%$, and the reference value, $\pi_0 = 10\%$, could reasonably be due to random chance alone
The result is *not statistically significant* (All of the preceding statements are equivalent.)

Using the *t*-Statistic: A Traditional Way, Same Answer

Yet another way to carry out a two-sided test for a population mean is to first compute the t-statistic, which is defined as $t_{\text{statistic}} = (\bar{X} - \mu_0)/S_{\bar{X}}$, and then use the critical t-value (computed in the same way as in Chapter 9) to decide which hypothesis to accept. The answer will always be the same as from the confidence interval method (and from the p-value method) so it does not matter which method you use. The hypothesis testing procedure for comparing the population mean to a reference value based on $\bar{X}$ and $S_{\bar{X}}$ (using either method) is called the **Student's *t*-test** or simply the ***t*-test**. The name *Student* was used by W. S. Gossett, Head Brewer for Guinness, when he published the first paper to use the t-distribution (which he invented) in place of the normal distribution, correcting for the use of the sample standard deviation, S, in place of the unknown population standard deviation, σ, when the sample size, n, is small.[10]

In general, hypothesis tests proceed by first computing a number called a **test statistic** based on the data that provides the best information for discriminating between the two hypotheses. Next, this test statistic (eg, the t-statistic) is compared to the appropriate **critical value** (eg, the critical t-value) to determine which hypothesis should be accepted. In situations that are more complex than just testing a population mean, it can require some creative effort (1) to come up with a test statistic that uses the sample information most efficiently and (2) to find the appropriate critical value. Either this critical value is found by theory (as is the case with the critical t-value using the t-distribution), or, increasingly in modern times, computers can be used to create a new, special critical value for each particular situation.

There are two different values referred to as t. We worked with the critical t-value in Chapter 9; this number, t_{critical}, does not reflect the sample data in any way, and might be computed using the Excel formula = TINV (1 − ConfidenceLevel, $n - 1$). The ***t*-statistic**, on the other hand, is the test statistic and represents how many standard errors there are separating μ_0 and $\bar{X}$:

The *t*-Statistic

For univariate data:

$$t_{\text{statistic}} = \frac{\bar{X} - \mu_0}{S_{\bar{X}}}$$

For a binomial situation:

$$t_{\text{statistic}} = \frac{p - \pi_0}{S_p}$$

10. Student, "The Probable Error of a Mean," *Biometrika* 6 (1908), pp. 1–25.

TABLE 10.2.6 Using the *t*-Statistic to Decide a Hypothesis Test

If the *t*-statistic is *smaller* in absolute value than the critical *t*-value ($\lvert t_{statistic}\rvert < t_{critical}$), then
Accept the null hypothesis, H_0, as a reasonable possibility
Do *not* accept the research hypothesis, H_1
The sample average, $\bar{X}$, is *not significantly different* from the reference value, μ_0
The observed difference between the sample average, $\bar{X}$, and the reference value, μ_0, could reasonably be due to random chance alone
The result is *not statistically significant* (All of the preceding statements are equivalent.)
If the *t*-statistic is *larger* in absolute value than the critical *t*-value ($\lvert t_{statistic}\rvert > t_{critical}$), then,
Accept the research hypothesis, H_1
Reject the null hypothesis, H_0
The sample average, $\bar{X}$, is *significantly different* from the reference value, μ_0
The observed difference between the sample average, $\bar{X}$, and the reference value, μ_0, could not reasonably be due to random chance alone
The result is *statistically significant.* (All of the preceding statements are equivalent.)

The *t*-test uses both of these *t* numbers, comparing the *t*-statistic computed from the data to the critical *t*-value. The result of the test is as stated in Table 10.2.6.

The *absolute value* of a number, denoted by enclosing the number between two vertical bars, is defined by removing the minus sign, if any. For example, $|3|=3$, $|-17|=17$, and $|0|=0$. A useful rule of thumb is that if the *t*-statistic is larger in absolute value than 2, reject the null hypothesis; otherwise, accept it. (Note that critical *t*-values are approximately 2 for even moderately large *n*: with $n=20$ the critical *t*-value is 2.09, with $n=60$ it is 2.00, and with very large *n* it is 1.96.) It is thus easy to scan a column of *t*-statistics and tell which are significant. For example, 6.81, −4.97, 13.83, 2.46, and −5.81 are significant *t*-statistics, whereas 1.23, −0.51, 0.02, −1.86, and 0.75 are not significant *t*-statistics. (A negative value for the *t*-statistic tells you that the sample average, $\bar{X}$, is smaller than the reference value, μ_0.)

You might wonder what to do if the *t*-statistic is *exactly equal* to the critical *t*-value. This would happen when μ_0 falls exactly at an endpoint of the confidence interval. How would you decide? Fortunately, this almost never happens. You might compute more decimal digits to decide, or you might conclude that your result is "significant, but just borderline."

Although the *t*-statistic may be easily compared to the value 2 (or to the more exact critical *t*-value) to decide significance, remember that it is not in the same measurement units as the data. Since the measurement units in the numerator and denominator of the *t*-statistic cancel each other, the result is a pure number without measurement units. It represents the distance between $\bar{X}$ and μ_0 in *standard errors* rather than in dollars, miles per gallon, people, or whatever units your data set represents.

Other than this, there is nothing really different between the *t*-statistic and confidence interval approaches. To verify this, reconsider the preceding examples.

For the example of the "yield-increasing" additive, the sample average is $\bar{X}=39.6$ tons, the standard error is $S_{\bar{X}}=4.2$ tons, the sample size is $n=7$, and the reference value is $\mu_0=32.1$ tons. The reference value *is* in the confidence interval, which extends from 29.3 to 49.9. Based on this, you accept the null hypothesis. If you had computed the *t*-statistic instead, you would have found:

$$t_{\text{statistic}}=\frac{\bar{X}-\mu_0}{S_{\bar{X}}}=\frac{39.6-32.1}{4.2}=1.785714$$

Since the absolute value of the *t*-statistic, 1.785714, is less than the critical *t*-value of 2.446912, you accept the null hypothesis. Thus, the *t*-statistic approach gives the same end result as the confidence interval approach, as it always must.

Consider, as an example, a survey in which managers were asked to rate the effect of employee stock ownership on product quality, for which the sample average score is $\bar{X}=0.35$, the standard error is $S_{\bar{X}}=0.14$, the sample size is $n=343$, and the reference value is $\mu_0=0$ which expresses a neutral opinion (neither positive nor negative, on average).[11] The reference value is *not* in the confidence interval, which extends from 0.08 to 0.62. Based on this, you accept the research hypothesis. Had you computed the *t*-statistic instead, you would have found,

$$t_{\text{statistic}}=\frac{\bar{X}-\mu_0}{S_{\bar{X}}}=\frac{0.35-0}{0.14}=2.50$$

Since the absolute value of the *t*-statistic, 2.50, is greater than the critical *t*-value of 1.9669, you accept the research hypothesis. The *t*-statistic approach gives you the same end result as the confidence interval approach, as it always must do.

11. P. B. Voos, "Managerial Perceptions of the Economic Impact of Labor Relations Programs," *Industrial and Labor Relations Review* 40 (1987), pp. 195–208.

For the binomial example involving the limits of production, there are $X = 58$ extra-fast chips out of the sample size $n = 500$, the binomial proportion is $p = 0.116$, the standard error is $S_p = 0.0143$, and the reference value is $\pi_0 = 0.10$. The reference value *is* in the confidence interval, which extends from 0.088 to 0.144. Based on this, you accept the null hypothesis. If you had computed the t-statistic instead, you would have found,

$$t_{\text{statistic}} = \frac{p - \pi_0}{S_p} = \frac{0.116 - 0.10}{0.0143} = 1.12$$

Since the absolute value of the t-statistic, 1.12, is less than the critical t-value of 1.9647, you accept the null hypothesis, reaching the same conclusion as with the confidence interval approach.

10.3 INTERPRETING A HYPOTHESIS TEST

Now that you know the mechanics involved in performing a hypothesis test and conventional ways to describe the result, it is time to learn the probability statement behind it all. Just as in the case of confidence intervals, since it is not possible to be correct 100% of the time, you end up with a statement involving the unknown population mean that is correct 95% (or 90% or 99% or 99.9%) of the time.

By convention, the formal details of hypothesis testing are set up in terms of the various *errors* that can be made. The result of a hypothesis test is that we accept one of the two hypotheses based on information from the sample data. You result might be right and you might be wrong since the hypotheses are statements about the *population*, for which you have incomplete information. Generally, you will not know for sure if you are right or wrong in your choice. Of course, you hope that you are correct; however, depending on the situation, there may or may not be a useful probability statement to reassure you.

Each type of error is based on a different assumption about which hypothesis is *really* true. Of course, in reality, you will not ordinarily know which hypothesis is true, even after you have made a decision to accept one. However, to understand the results of your hypothesis test, it is helpful to put it in perspective with respect to all of the different ways the test could have come out.

Errors: Type I and Type II

If the null hypothesis is really true (even though, in reality, you will not know for sure if it is or not) but you wrongly decide to reject it and accept the research hypothesis instead, then you have committed a **type I error**, pronounced "type one error." The probability of a type I error occurring (when the null hypothesis is true) is controlled by convention at the 5% level:

$$P(\text{type I error when } H_0 \text{ is true}) = 0.05$$

It is possible to control the probability of a type I error because the null hypothesis is very specific, so there is an exact probability. For example, when you assume that the null hypothesis H_0: $\mu = \mu_0$ is true, you are assuming that you know the value of the population mean. Once you know the population mean of a normal distribution, probabilities can be easily calculated.

Testing at other levels (10%, 1%, or 0.1%, say) can be done by using a different critical t-values—for example, by working with a different confidence interval (90%, 99%, or 99.9%, respectively). You may use the following Excel formula to find the critical t-value for a given test level: = TINV (testLevel, $n - 1$), which recognizes that the test level is equal to one minus the confidence level. If you are not willing to be wrong 5% of the time when the null hypothesis is true, you might test at the 1% level instead (using the appropriate critical t-value, which would be 2.8609 for $n = 20$ and would be approximately 2.576 for a large sample size n) so that your probability of committing a type I error (when the null hypothesis is true) would only be 1%.

If the research hypothesis is really true (even though, again, you will not usually know for sure if it is or not), but you wrongly decide to accept the null hypothesis instead, you have committed a **type II error**. The probability of a type II error occurring cannot be easily controlled:

$$P(\text{type II error when } H_1 \text{ is true}) \text{ is not easily controlled}$$

It is difficult to control the probability of a type II error because, depending on the true value of μ, this probability will vary.[12] Suppose μ is very close to μ_0. Then, due to randomness in the data, it will be very difficult to tell them apart. For example, suppose the null hypothesis claims that μ is 15.00000, but μ is actually 15.00001. Then, although the research hypothesis is technically true (since $15.00000 \neq 15.00001$), in practical terms you will have much trouble telling them apart and the probability of a type II error will be approximately 95%. On the other hand, if μ is far from 15, the probability of a type II error will be nearly 0, a pleasing situation. Thus, since the probability of a type II error depends so heavily on the true value of μ, it is difficult to control. These errors are illustrated in Fig. 10.3.1.

12. In principle, the probability can be computed for each value for μ. The resulting table or graph provides the basis for what is called the *power* of the test. This is basically a *what-if* analysis, giving the type II error properties of the test under each possible value of μ.

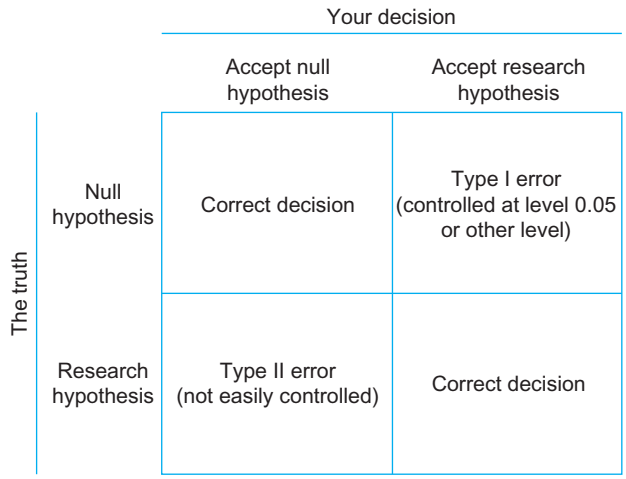

FIG. 10.3.1 Your decision to accept one of the two hypotheses may or may not be correct. Depending on which hypothesis is really true, there are two types of errors. Only the type I error is easily controlled, conventionally at the 5% level.

Assumptions Needed for Validity

You may have already suspected that some assumptions must be satisfied for the results of the hypothesis test to be valid. Since the test can be done based on the confidence interval, the assumptions for hypothesis testing are the same as the assumptions needed for confidence intervals. The **assumptions for hypothesis testing** are: (1) the data set is a random sample from the population of interest, and (2) either the quantity being measured is approximately normal, or else the sample size is large enough that the central limit theorem ensures that the sample average is approximately normally distributed.

What happens if these assumptions are not satisfied? Consider the probability of a type I error (wrongly rejecting the null hypothesis when it is actually true). This error probability will no longer be controlled at the low, manageable level of 5% (or other claimed level of your choice). Instead, the true error probability could be much higher or lower than 5%. A finding of significance then loses much of its prestige since the event "wrongly finding significance" is now a more common occurrence.

If the data set is not a random sample from the population of interest, there is essentially nothing that statistics can do for you because the required information is just not in the data, although it may be permissible to proceed under the assumption that your sample is representative of a more general situation (an idealized population).

Suppose your data set is a random sample, but the distribution is not normal. If your data distribution is so far from normal that you are concerned, you might try transforming the data (eg, using logarithms if all numbers are positive) to obtain a more normal distribution. If you decide to transform, note that you would no longer be testing the mean of the population but the mean of the *logarithm* of the population instead, and interpretation would be more complicated. Another solution would be to use a nonparametric test, to be explained in Chapter 16.

Hypotheses Have No Probabilities of Being True or False

Perhaps you have noticed that we have never said that a hypothesis is "probably" either true or false. We have always been careful either to accept or to reject a hypothesis, making a definite, exact decision each time. We talk about the errors we might make and *their* probabilities, but never about the probability of a hypothesis being true or false. The reason is simple:

There is nothing random about a hypothesis!

The null hypothesis is either true or false, depending on the value of the population mean, μ. There is no randomness involved in the population mean by itself. Similarly, the research hypothesis is also either true or false, and although you do not know which, there is no randomness involved in the hypothesis itself. The randomness comes only from the random sampling process, which gives you the data to help you decide.

Thus, your *decision* is known and random, just like a sample statistic, since it is based on the data. However, the true hypothesis is fixed but unknown, just like a population parameter.

Statistical Significance and Test Levels

By convention, a result is defined to be statistically significant if you accept the research hypothesis using a test at the 5% level (eg, based on a standard 95% confidence interval). Note that this is probably not the same use of the word *significant* that you are used to. Ordinarily, something "significant" has special importance. This is not necessarily so in statistics.

To illustrate, a lawyer once came to me deeply concerned because the other side in a lawsuit had found a *statistically significant difference* between measurements made on a door that had been involved in an accident and other, similar doors in the same building. Oh no! But after the special statistical meaning of the word *significant* was pointed out, the attorney was relieved to find that the other side had *not* shown that the door was extremely different from the others. They had only demonstrated that there was a *statistically detectable* difference. In fact, the difference was quite small. But with enough careful measurements, it was detectable! It was not just randomly different from the other doors (the null hypothesis); it was systematically different (the research hypothesis). Although the difference was statistically significant, it was not large enough to matter very much. The situation is analogous

to snowflakes; each one is truly different from the others, yet for many purposes they are essentially identical.

The moral of this story is that you should not automatically be impressed when someone boasts of a "significant result." If the word is being used in its statistical sense, it says only that random chance has been ruled out. It still remains to examine the data to see if the effect is strong enough to be important to you. Statistical methods work only with the numbers; it is up to you to use knowledge from other fields to decide the importance and relevance of the statistical results.

There is another reason you should not be overly impressed by statistically significant results. Over your lifetime, approximately 5% of the test results you will see *for situations in which the null hypothesis is really true* will be found (wrongly, due to random error) to be significant. This implies that about 1 in every 20 uninteresting situations will be declared significant by mistake (ie, by type I error). A pharmaceutical researcher once noticed that about 5% of the drugs being tested for a particularly difficult disease were found to have a significant effect. Since this is about the fraction of drugs that would be found *by mistake* to be effective, *even if none were in fact effective*, this observation suggests that the entire program might not be successful in finding a cure.

By using the appropriate critical t-value, you can perform a hypothesis test at the 10%, 5%, 1%, or 0.1% level. This **test level** or **significance level** is the probability of a type I error when the null hypothesis is in fact true.[13] When you reject the null hypothesis and accept the research hypothesis, you may claim that your result is *significant* at the 10%, 5%, 1%, or 0.1% level, depending on your choice of critical t-value. The smaller the test level for which you can find significance, the more impressive your result. For example, finding a result that is significant at the 1% level is more impressive than finding significance at the 10% or 5% level because your data are even less likely to be produced by the null hypothesis; your type I error probability is smaller and your evidence against the null hypothesis is stronger. By convention, the following phrases may be used to describe your results:

Not significant	Not significant at the conventional 5% level
Significant	Significant at the conventional 5% level
Highly significant	Significant at the 1% level
Very highly significant	Significant at the 0.1% level

What should you do if you find significance at more than one level? Celebrate! Seriously, however, the smaller the test level at which you find significance, the stronger your evidence is against the null hypothesis. You would therefore report only the *smaller* of the significance levels for which you find significance. For example, if you find significance at both the 5% and 1% levels, it would be sufficient to report only that your result is highly significant.

Whenever you find significance at one level, you will necessarily find significance at all *larger* levels.[14] Thus, a highly significant result (ie, significant at the 1% level) must *necessarily* (ie, provably, using mathematics) be significant at the 5% and 10% levels. However, it might or might not be significant at the 0.1% level.

The *p*-Value Hierarchy

As we know, every hypothesis test has a p-value, which tells you how surprised you would be to learn that the null hypothesis had produced the data, with smaller p-values indicating more surprise and leading to rejection of H_0. By convention, we reject H_0 whenever the p-value is less than 0.05. The p-value tells you the probability, assuming that the null hypothesis is true, that such data (or data showing even more differences from H_0) would be observed. Because small p-values are unlikely to arise when H_0 is true, they lead to rejection of H_0. For example, if $p < 0.001$, data with such large differences from H_0 occur less often than once in 1,000 random samples. Rather than suppose that rare 1 in 1,000 events can reasonably happen (because they do not, at least not very often), it is simpler to decide that H_0 is false and should be rejected. By convention, p-values are reported as follows:

As Reported	Interpretation
Not significant ($p > 0.05$)	Not significant at the conventional 5% level
Significant ($p < 0.05$)	Significant at the conventional 5% level but not at the 1% level
Highly significant ($p < 0.01$)	Significant at the 1% level but not at the 0.1% level
Very highly significant ($p < 0.001$)	Significant at the 0.1% level

In some fields of study, you are permitted to report a result that is significant at the 10% level. This represents an error probability of 1 in 10 of rejecting a null hypothesis that is really true; many other fields consider this an unacceptably high error rate. However, some fields recognize that the unpredictability and variability of their data make it difficult to obtain a (conventionally) significant result at the 5% level. If you are working in such a field, you

13. In more general situations, including one-sided testing, the test level or significance level is defined more carefully as the *maximum* type I error probability, maximized over all possibilities included within the null hypothesis.

14. Note that a *larger* level of significance is actually a *less* impressive result. For example, being significant at the 1% level is highly significant, whereas being significant at the (larger) 5% level is (merely) significant.

may use the following p-value statement as a possible alternative:

> Significant at the 10% level but not at the conventional 5% level ($p < 0.10$).

A p-value statement is often found inserted into text, as in "The style of music was found to have a significant effect ($p < 0.05$) on purchasing behavior." You may also see p-value statements included as footnotes either to text or to a table, as in "Productivity improved significantly[15] as a result of the new exercise program."

Most statistical software packages report an exact p-value as the result of a hypothesis test. For testing at the 5% level, if this p-value is any number less than 0.05, then the result is significant (eg, $p = 0.0358$ corresponds to a significant test result, whereas $p = 0.2083$ is not significant because 0.0358 is less than 0.05, whereas 0.2083 is not). Note that the p-value is a statistic (not a population parameter) because it can be computed based on the data (and the reference value).

Consider the example of testing whether or not the observed average yield $\bar{X} = 39.6$ is significantly different from the reference value $\mu_0 = 32.1$ tons (based on $n = 7$ observations with standard error $S_{\bar{X}} = 4.2$). The result might be reported as follows:

	n	Mean	STDEV	SE Mean	*t*	*p*-Value
Yield	7	39.629	11.120	4.203	1.79	0.12

Since the computed p-value (0.12) is more than the conventional 5% test level (ie, $0.12 > 0.05$) we have the test result "not significant ($p > 0.05$)." This result (not significant) may also be obtained by comparing the computed t-statistic (1.79) to the critical t-value of 2.446912 for this sample size of 7. The exact p-value here tells you that there is a 12% chance of seeing such a large difference (between observed mean and reference value) under the assumption that the population mean is equal to the reference value $\mu_0 = 32.1$. By convention, a 12% chance is not considered out of the ordinary, but chances of 5% or less are considered unlikely. Alternatively, you might first ask the computer for the 95% confidence interval:

	n	Mean	STDEV	SE Mean	95.0% CI
Yield	7	39.63	11.12	4.20	(29.34, 49.92)

From this output, you can see that the test is not significant because the reference value $\mu_0 = 32.1$ is within the confidence interval (29.34 to 49.92).

Next, consider testing whether or not managers, in general, view employee stock ownership as worthwhile for improving product quality, as measured on a scale from −2 (strongly not worthwhile) to +2 (strongly worthwhile). The computer output might look like this:

	n	Mean	STDEV	SE Mean	*t*	*p*-Value
Score	343	0.350	2.593	0.140	2.50	0.013

This output tells you that the p-value is $p = 0.013$. Thus, the result is significant at the 5% level (since $p < 0.05$) but is not significant at the 1% level (since $p > 0.01$). The conclusion is that managers perceive employee stock ownership as significantly worthwhile ($p < 0.05$).

If you have a binomial situation, please note that there may be two different quantities referred to as p by convention. One is the observed percentage of occurrences in the sample, $p = X/n$. The other is the p-value computed for a hypothesis test involving a particular reference value. While this may be confusing, it is standard statistical notation.

10.4 ONE-SIDED TESTING

All of the tests we have done so far are two-sided tests because they test the null hypothesis, H_0:$\mu = \mu_0$, against the research hypothesis H_1:$\mu \neq \mu_0$. This research hypothesis is two-sided because it allows for possible values for the population mean both above and below the reference value, μ_0.

However, you may not really be interested in testing whether the population mean is *different* from the reference value. You may have a special interest in it being *larger* (in some cases) or *smaller* (in other cases) than the reference value. For example, you might purchase a system only if possible long-term savings are *significantly larger* than some special number (the reference value, μ_0). Or you might be interested in claiming that your quality is high because the defect rate is *significantly smaller* than some impressively small number.

You do not need to use a one-sided test to be able to claim that the sample average is significantly larger or significantly smaller than the reference value; you may be able to use a two-sided test for this. If the two-sided test comes out significant (ie, you accept the research hypothesis), then you may base your claim of significance on whether the sample average, $\bar{X}$, is larger than or smaller than the reference value:

Using a Two-Sided Test but Reporting a One-Sided Conclusion[16]

If the two-sided test is significant and $\bar{X} > \mu_0$	The sample average, $\bar{X}$, is significantly larger than the reference value, μ_0
If the two-sided test is significant and $\bar{X} < \mu_0$	The sample average, $\bar{X}$, is significantly smaller than the reference value, μ_0

15. ($p < 0.05$).

16. Remember, the *two*-sided conclusion might be "$\bar{X}$ is significantly different from μ_0."

However, it may be advantageous to use a one-sided test. If you meet the requirements, you might be able to report a significant result using a one-sided test that would not be significant had you used a two-sided test. How is this possible? By focusing on just one side and ignoring the other, the one-sided test can better detect a difference on that side. The trade-off is that the one-sided test is incapable of detecting a difference, no matter how large, on the other side.

A **one-sided *t*-test** is set up with the null hypothesis claiming that μ is on one side of μ_0 and the research hypothesis claiming that it is on the other side. (We always include the case of $\mu = \mu_0$ in the null hypothesis, which is the default. This ensures that when you accept the research hypothesis and find significance, you have a stronger conclusion: either "significantly larger than" or "significantly smaller than.")[17] The hypotheses for the two different kinds of one-sided tests are as follows:

One-Sided Testing to See If μ Is Smaller Than μ_0

$$H_0: \mu \geq \mu_0$$

The null hypothesis claims that the unknown population mean, μ, is *at least as large* as the known reference value, μ_0.

$$H_1: \mu < \mu_0$$

The research hypothesis claims that the unknown population mean, μ, is *smaller* than the known reference value, μ_0.

One-Sided Testing to See If μ Is Larger Than μ_0

$$H_0: \mu \leq \mu_0$$

The null hypothesis claims that the unknown population mean, μ, is *not larger* than the known reference value, μ_0.

$$H_1: \mu > \mu_0$$

The research hypothesis claims that the unknown population mean, μ, is *larger* than the known reference value, μ_0.

There is an important criterion you must satisfy before using a one-sided hypothesis test; it is essentially the same criterion that must be met for a one-sided confidence interval (from Chapter 9):

> In order to use a one-sided test, you must be sure that *no matter how the data had come out*, you would still have used a one-sided test on the same side ("larger than" or "smaller than") as you will use. If, had the data come out different, you might have used a one-sided test *on the other side* instead of the side you plan to use, you should use a two-sided test instead. If in doubt, use a two-sided test.

17. Technically speaking, the case of equality could be included in the research hypothesis. However, the mechanics of the test would be identical and significance would be found in exactly the same cases, but your conclusion would seem weaker.

In particular, using a one-sided test can leave you open to criticism. Since the decision of what is interesting can be a subjective one, your decision to focus only on what is interesting to you may conflict with the opinions of others you want to convince. If you need to convince people who might have a very different viewpoint (eg, regulators or opposing lawyers), you should consider using a two-sided test and giving the one-sided conclusion. On the other hand, if you need only to convince "friendly" people, with interests similar to yours (eg, within your department or firm), and you satisfy the preceding criterion, you will want to take advantage of one-sided testing.

The research hypothesis will be accepted only if there is convincing evidence against the null hypothesis. This says that you will accept the research hypothesis only when the sample average, $\bar{X}$, and the reference value, μ_0, have the same relationship as described in the research hypothesis *and* are far enough apart (namely, t_{critical} or more standard errors apart, which represents the extent of reasonable variation from a mean value). There are three different ways to implement a one-sided test, namely, using the *p*-value from statistical software, using a one-sided confidence interval or using the *t*-statistic. The one-sided critical *t*-value may be computed using Excel as either = TINV (2 * (1 − confidenceLevel), $n - 1$) for the confidence interval or, equivalently, as = TINV (2 * testLevel, $n - 1$).

Example

Launching a New Product

Suppose a break-even analysis for a new consumer product suggests that it will be successful if more than 23% of consumers are willing to try it. This 23% is the reference value, μ_0; it comes from a theoretical analysis, not from a random sample of data. To decide whether or not to launch the product, you have gathered some data from randomly selected consumers and have computed a one-sided confidence interval. Based just on the data, you expect 43.90% of consumers to try it ($\bar{X} = 43.90\%$), and your one-sided confidence interval statement is that you are 95% sure that *at least* 38.2% of consumers will be willing to try it. Since your break-even point, $\mu_0 = 23\%$, is well outside this confidence interval (and, hence, it is *not* reasonable to suppose that the mean could be 23%, because 23% is *not* at least 38.2%), you do have convincing evidence that the population mean is greater than 23%. A summary of the situation is shown in Table 10.4.1.[18]

The decision is made to accept the research hypothesis H_1 because the reference value is not in the confidence interval (ie, 23% is not "at least 38.2%").

Since the reference value, 23%, is so far below the confidence interval, perhaps you should try for a more impressive significance level. In fact, you can claim 99.9% confidence that the population mean is at least 33.1% using the critical *t*-value 3.13066 for this one-sided confidence interval. Since

TABLE 10.4.1 Testing the Percentage of Consumers Who Are Willing to Try a New Product (Confidence Interval Approach)

Null hypothesis	$H_0: \mu \leq \mu_0$	$H_0: \mu \leq 23\%$
Research hypothesis	$H_1: \mu > \mu_0$	$H_0: \mu > 23\%$
Average	$\bar{X}$	43.90%
Standard error	$S_{\bar{X}}$	3.466%
Sample size	n	205
Reference value	μ_0	23%
Confidence interval	$\bar{X} - t_{\text{critical}} S_{\bar{X}}$	"We are 95% sure that the population mean is at least 38.2%"
Decision	Accept H_1	"We expect significantly more than 23% of consumers to try our product"*

*Significant ($p < 0.05$) using a one-sided test.

Example—cont'd

the reference value, 23%, is outside even this confidence interval, the result is *very highly significant* ($p < 0.001$).

18. Because this is a binomial situation, you may substitute p in place of $\bar{X}$, S_p in place of $S_{\bar{X}}$, π in place of μ, and π_0 in place of μ_0 throughout. To see that this example is also correct as stated here; note that $\bar{X} = p$ for the data set $X_1, X_2, \ldots, X_n$ where each number is either 0 or 1 according to the response of each consumer.

How to Perform the Test

Table 10.4.2 shows how to perform a one-sided test, giving complete instructions for both types of testing situations (ie, to see if $\bar{X}$ is significantly larger than, or significantly smaller than, μ_0), performed using either the confidence interval or the t-statistic method, and giving both types of possible conclusions (significant or not) and their interpretations. A useful guiding principle is that it is significant if the reference value, μ_0, is not within the one-sided confidence interval constructed to match the direction of the research hypothesis. An alternative is to use statistical software to find the p-value for the one-sided test of your choice, then interpreting this p-value in the usual way (eg, significant if $p < 0.05$).

If you use the confidence interval method, remember that there are two different one-sided confidence interval statements. You want to choose the one that matches the side of the claim of the research hypothesis. For example,

TABLE 10.4.2 One-Sided Testing

One-Sided Testing to See If μ Is Larger than μ_0
The hypotheses being tested are $H_0: \mu \leq \mu_0$ against $H_1: \mu > \mu_0$
The confidence interval statement is "We are 95% sure that the population mean is at least as large as $\bar{X} - t_{\text{critical}} S_{\bar{X}}$."
The t-statistic is $t_{\text{statistic}} = (\bar{X} - \mu_0)/S_{\bar{X}}$ (Note: Do not use absolute values for one-sided testing)
Is $\bar{X} - t_{\text{critical}} S_{\bar{X}} \leq \mu_0$? This is the confidence interval approach, asking: Is the reference value, μ_0, inside the confidence interval? Equivalently, with the t-statistic approach: Is $t_{\text{statistic}} \leq t_{\text{critical}}$? If so, then,
Accept the null hypothesis, H_0, as a reasonable possibility
Do not accept the research hypothesis, H_1
The sample average, $\bar{X}$, is not significantly larger than the reference value, μ_0
If $\bar{X}$ is larger than μ_0, the observed difference could reasonably be due to random chance alone
The result is *not statistically significant*
Is $\bar{X} - t_{\text{critical}} S_{\bar{X}} > \mu_0$? This is the confidence interval approach, asking: Is the reference value, μ_0, outside the confidence interval? Equivalently, with the t-statistic approach: Is $t_{\text{statistic}} > t_{\text{critical}}$? If so, then,
Accept the research hypothesis, H_1
Reject the null hypothesis, H_0
The sample average, $\bar{X}$, is significantly larger than the reference value, μ_0
The observed difference between the sample average, $\bar{X}$, and the reference value, μ_0, could not reasonably be due to random chance alone
The result is *statistically significant*
One-Sided Testing to See If μ Is Smaller than μ_0
The hypotheses being tested are $H_0: \mu \geq \mu_0$ against $H_1: \mu < \mu_0$
The confidence interval statement is "We are 95% sure that the population mean is not larger than $\bar{X} + t_{\text{critical}} S_{\bar{X}}$."
The t-statistic is $t_{\text{statistic}} = (\bar{X} - \mu_0)/S_{\bar{X}}$ (Note: Do not use absolute values for one-sided testing).
Is $\bar{X} + t_{\text{critical}} S_{\bar{X}} \geq \mu_0$? This is the confidence interval approach, asking: Is the reference value, μ_0, *inside* the confidence interval? Equivalently, with the t-statistic approach: Is $t_{\text{statistic}} \geq -t_{\text{critical}}$? If so, then,
Accept the null hypothesis, H_0, as a reasonable possibility
Do not accept the research hypothesis, H_1
The sample average, $\bar{X}$, is not significantly smaller than the reference value, μ_0
If $\bar{X}$, is smaller than μ_0, the observed difference could reasonably be due to random chance alone
The result is not *statistically significant*

(Continued)

TABLE 10.4.2 One-Sided Testing—cont'd

Is $\bar{X} + t_{\text{critical}} S_{\bar{X}} < \mu_0$? This is the confidence interval approach, asking: Is the reference value, μ_0, *outside* the confidence interval? Equivalently, with the *t*-statistic approach: Is $t_{\text{statistic}} < -t_{\text{critical}}$? If so, then,
Accept the research hypothesis, H_1
Reject the null hypothesis, H_0
The sample average, $\bar{X}$, is significantly smaller than the reference value, μ_0
The observed difference between the sample average, $\bar{X}$, and the reference value, μ_0, could not reasonably be due to random chance alone
The result is *statistically significant*

if your research hypothesis is H_1: $\mu > \mu_0$, your one-sided confidence interval will consist of all values for μ that are *at least as large* as the appropriate computed number, $\bar{X} - t_{\text{critical}} S_{\bar{X}}$, using the one-sided critical t-value.

Fig. 10.4.1 shows that in order for you to decide that $\bar{X}$ is significantly larger than μ_0, the distance between them must be sufficiently large to ensure that it is not just due to random chance. Fig. 10.4.2 gives the corresponding picture for a one-sided test on the other side.

If you use the t-statistic, the test is decided by comparing $t_{\text{statistic}} = (\bar{X} - \mu_0)/S_{\bar{X}}$ either to the critical t-value, t_{critical}, or to its negative, $-t_{\text{critical}}$, depending on the side being tested (specifically, depending on whether the research hypothesis is H_1: $\mu > \mu_0$ or H_1: $\mu < \mu_0$). The idea behind the calculation is that the test is significant if the data correspond to the side of the research hypothesis and the t-statistic is large

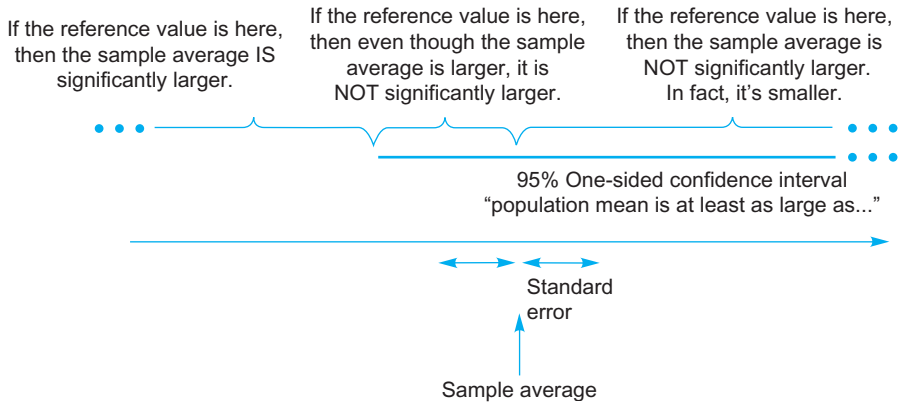

FIG. 10.4.1 Using a one-sided test to see if μ is larger than the reference value, μ_0. The one-sided confidence interval uses the same side as the research hypothesis (namely, those reasonably possible values of μ that are at least as large as the endpoint value of the interval). Only if the reference value, μ_0, is well below the sample average will you decide that the sample average is significantly larger.

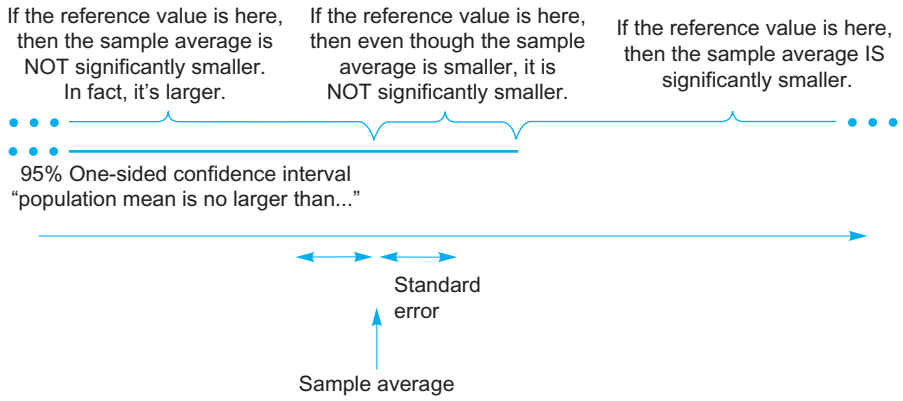

FIG. 10.4.2 Using a one-sided test to see if μ is smaller than the reference value, μ_0. The one-sided confidence interval uses the same side as the research hypothesis (namely, those reasonably possible values of μ that are smaller than or equal to the endpoint value of the interval). Only if μ_0 is well above the sample average will you decide that the sample average is significantly smaller.

in magnitude (which ensures that the difference between $\bar{X}$ and μ_0 is larger than ordinary randomness). Note that the *t*-statistic is the same for one-sided and for two-sided testing, but that you use it differently to decide significance.

Example

Launching a New Product, Revisited

For the consumer product launch example considered earlier, the launch will be profitable only if more than 23% of consumers try the product. The appropriate facts and results are as shown in Table 10.4.3. We now perform the same one-sided test we did earlier, except that we use the *t*-statistic method this time.

The decision is made to accept H_1 because $t_{\text{statistic}} > t_{\text{critical}}$; that is, we have $6.03 > 1.652357$, using the appropriate criterion from Table 10.4.2 for this research hypothesis ($H_1: \mu > \mu_0$). In the end, the result is the same (ie, significant) whether you use the one-sided confidence interval or the one-sided *t*-test. In fact, since the *t*-statistic exceeds even the one-sided critical *t*-value 3.130661 for testing at level 0.001, you may claim a *very highly significant* result ($p < 0.001$).

For reference, the *p*-value from statistical software for this one-sided test shows as $p = 3.78\text{E-}09$ which represents the number 0.00000000378, which is very highly significant because it is less than 0.001. With a *t*-statistic indicating around six standard errors of difference, it is very natural for the *p*-value to indicate such a small probability.

TABLE 10.4.3 Testing the Percentage of Consumers Who Are Willing to Try a New Product (*t*-Statistic Approach)

Null hypothesis	$H_0: \mu \le \mu_0$	$H_0: \mu \le 23\%$
Research hypothesis	$H_1: \mu > \mu_0$	$H_1: \mu > 23\%$
Average	$\bar{X}$	43.90%
Standard error	$S_{\bar{X}}$	3.466%
Sample size	n	205
Reference value	μ_0	23%
t-Statistic	$t_{\text{statistic}} = \dfrac{\bar{X} - \mu_0}{S_{\bar{X}}}$	$\dfrac{43.90 - 23}{3.466} = 6.03$
Critical value	t_{critical}	1.652357
Decision	Accept H_1	"We expect significantly more than 23% of consumers to try our product"*

**Significant ($p < 0.05$) using a one-sided test.*

Example

Will Costs Go Down?

You have tested a new system that is supposed to reduce the variable cost or unit cost of production (ie, the cost of producing each extra unit, ignoring fixed costs such as rent, which would be the same whether or not you produced the extra unit). Since the new system would involve some expenditures, you will be willing to use it only if you can be convinced that the variable cost will be less than $6.27 per unit produced.

Based on a careful examination of 30 randomly selected units that were produced using the new system, you found an average variable cost of $6.05. It looks as if your variable cost is, on average, less than the target of $6.27. But is it *significantly* less? That is, can you expect the long-run mean cost to be below $6.27, or is this just a random fluke of the particular 30 units you examined? Based on the information so far, you cannot say because you do not yet know how random the process is. Is $6.05 less than $6.27? Yes, of course. Is $6.05 *significantly* less than $6.27? You can tell only by comparing the difference to the randomness of the process, using the standard error and the critical *t*-value.

So you go back and find the standard deviation, and then you compute the standard error, $0.12. Table 10.4.4 shows a summary of the situation and the results of a one-sided test

(*Continued*)

TABLE 10.4.4 Testing the Long-Run Variable Cost

Reference value	μ_0	$6.27
Null hypothesis	$H_0: \mu \ge \mu_0$	$H_0: \mu \ge \$6.27$
Research hypothesis	$H_1: \mu < \mu_0$	$H_1: \mu < \$6.27$
Average	$\bar{X}$	$6.05
Standard error	$S_{\bar{X}}$	$0.12
Sample size	n	30
Confidence interval	$\bar{X} + t_{\text{critical}} S_{\bar{X}}$	"You are 95% sure that the long-run mean variable cost is less than $6.25"
t-Statistic	$t_{\text{statistic}} = \dfrac{\bar{X} - \mu_0}{S_{\bar{X}}}$	$\dfrac{6.05 - 6.27}{0.12} = -1.833$
−Critical *t*-value	$-t_{\text{critical}}$	−1.699127
Decision	Accept H_1	"Variable costs under the new system are significantly less than $6.27"*

**Significant ($p < 0.05$) using a one-sided test.*

Example—cont'd

(with both methods shown) of whether or not your costs are significantly lower than the required amount.

Using the confidence interval approach, you are 95% sure that the mean variable cost is less than \$6.25. You are even more sure that it is less than the required amount, \$6.27. Hence, the result is significant. Or you could simply note that the reference value (\$6.27) is not in the confidence interval, which extends only up to \$6.25.

When you use the *t*-statistic approach, the result is significant because $t_{\text{statistic}} < -t_{\text{critical}}$; that is, we have $-1.833 < -1.699127$, using the appropriate criterion from Table 10.4.2 for this research hypothesis ($H_1: \mu < \mu_0$).

You are entitled to use a one-sided test in this case because you are really interested in just that one side. If you can be convinced that the mean variable cost is less than \$6.27, then the system is worth considering. If not, then you are not interested. By using a one-sided test in this way, you are admitting that, had the system been really bad, you would not have been able to say that "variable costs are significantly *more than*…"; you would only be able to say that "they are *not significantly less*."

Had you decided to use a two-sided test, which is valid but less efficient, you would actually have found that the result is *not* significant! The two-sided confidence interval extends from \$5.80 to \$6.30, which *does* contain the reference value \$6.27. The *t*-statistic is still −1.833, but the two-sided critical *t*-value is 2.045230, which is now larger than the absolute value (1.833) of the *t*-statistic. Thus, this example shows that you can have a significant one-sided result but a nonsignificant two-sided result. This can happen only when the one-sided significance is somewhat borderline, passing the test with just a small margin, as in this example (specifically, it can happen whenever the two-sided *p*-value is less than 10%).

Should you buy the system? This is a business strategy question, not a statistical question. By all means, please use the results of the hypothesis test as one of your inputs, but consider all the other factors, such as availability of investment capital, personnel implications, and interactions with other projects. And do not forget this: Although hypothesis testing has led you to accept the research hypothesis that the variable costs are less than your threshold, this has *not* been absolutely proven; there is still room for error. You cannot give a number for the probability that your hypothesis testing decision is wrong because you do not know which hypothesis is really true. The best you can say is that *if* the new system has variable costs of exactly \$6.27, then you would wrongly decide significance (as you may have here) only 5% of the time or less.

For reference, the *p*-value from statistical software for this one-sided test is $p = 0.0385$, which is significant because it is less than 0.05. For the two-sided test, the *p*-value is exactly twice this: $p = 0.0770$, which is not significant because it is greater than 0.05. The choice of one-sided or two-sided testing makes a big difference in this situation.

Example

Can You Create Value by Changing Your Firm's Name?

When a large firm changes its name, it is a major event. The budgets for advertising the change of name and for setting up the new image can be enormous. Why do firms do it? According to the theory of financial budgeting, firms should undertake only projects that increase the value of the firm to the owners, the shareholders. If it is reasonable for a firm to spend those resources to change its name, then you should observe an increase in the firm's value as measured by the price of its stock.

A study of the change in firm value around the time of a name change announcement might use a one-sided statistical hypothesis test to see if the stock price really did go up. One of the difficulties of measuring this kind of market price reaction is that the stock market as a whole responds to many forces, and the name change should be evaluated in light of what the stock price did compared to what it should have done based on the entire stock market during that time. So if the stock market was up, you would have to find the firm's stock up an *even larger percentage* than you would ordinarily expect on such a day before deciding that the announcement has had an effect.

This kind of *event study*, using an adjustment for large-scale market forces, involves computing an *abnormal return*, which represents the rate of return an investor would have received by holding the firm's stock, minus the rate of return the investor would have expected from an investment of similar risk (but involving no name change) during the same time period.

Thus, a positive abnormal return would show that the name change caused a price rise even larger than we would have expected in the absence of the name change. This could happen because the stock market was generally up and the firm's stock was up much more. Or it could happen because the stock market was down but the firm's stock was down less than expected given the market as a whole.

One study looked at 58 corporations that changed their names and reported the methods as follows:[19]

In order to test if [the abnormal return due to the name change] is different from zero, the test statistic employed is the ratio of the average abnormal returns…to their standard deviation.… This test statistic…is distributed standard normal if n is large enough.

The study's authors are saying that they used a *t*-statistic to test the sample average against the reference value, $\mu_0 = 0$, of no abnormal returns due to a name change. The "standard deviation" they refer to is the standard error of this estimated quantity. With their sample size of $n = 58$, the one-sided critical *t*-value is 1.672029 for the 5% test level.

The results of this study are given as follows:

The mean abnormal return was found to be 0.61%, with a corresponding t-statistic…of 2.15. Thus, if the null hypothesis is that the residual returns are drawn from a population with a nonpositive mean, the one-sided null hypothesis can be rejected.

Example—cont'd

Since the study's authors rejected the null hypothesis, they accepted the research hypothesis. They showed that the stock price rises significantly as a result of a name change. Does this say that you should rush out and change your firm's name as often as possible? Well, not really. They discussed the implications as follows:

Our findings are that, for most of the firms, name changes are associated with improved performance, and that the greatest improvement tends to occur in firms that produce industrial goods and whose performance prior to the change was relatively poor.... Our findings do not support, however, the contention that the new name per se [i.e., by itself] will enhance demand for the firm's products. Rather, it seems that the act of a name change serves as a signal that other measures to improve performance such as changes in product offerings and organizational changes will be seriously and successfully undertaken.

Note that with a *t*-statistic of 2.15, they found a significant result at the 5% level (since the *t*-statistic exceeds the one-sided critical *t*-value of 1.672029). However, looking at the 1% level, since the one-sided critical *t*-value is then 2.393568, their result is significant but not highly significant. For reference, the *p*-value from statistical software for this one-sided test shows as $p = 0.0179$, which is indeed less than 0.05 but more than 0.01, in agreement with this significant but not highly significant result.

More recently, Kashmiria and Mahajanb studied the stock market reaction to corporate name changes, and found statistically significant reactions—in some cases positive, in other cases negative—based on hypothesis testing of the Cumulative Average Abnormal Return (CAAR) earned by holding the company's stock price, where the term "Abnormal" indicates that they have adjusted for movements of the market as a whole during this time in order to isolate the effect of the news of the name change:[20]

We found CAAR to be positive and significant when the type of name change was leveraging a strong brand, or when it was proactively communicating a new scope, but found CAAR to be negative and significant for name changes retroactively communicating a new scope.

19. D. Horsky and P. Swyngedouw, "Does It Pay to Change Your Company's Name? A Stock Market Perspective," *Marketing Science* 6 (1987), pp. 320–35.
20. S. Kashmiria and V. Mahajanb (2015) "The Name's the Game: Does Marketing Impact the Value of Corporate Name Changes?" *Journal of Business Research*, Vol. 68, Issue 2, p. 281–290.

10.5 TESTING WHETHER OR NOT A NEW OBSERVATION COMES FROM THE SAME POPULATION

By now, you probably have the idea that if you can construct a confidence interval, you can do a hypothesis test. This is correct. Based on the prediction interval in Chapter 9 for a new observation (instead of for the population mean), you may now quickly test whether or not this new observation came from the same population as the sample data. The null hypothesis, H_0, claims that the new observation comes from the same normally distributed population as your sample, and the research hypothesis, H_1, claims that it does not. The data set is assumed to be a random sample.

The test is fairly simple, now that you know the basics of hypothesis testing and confidence intervals. Find the prediction interval (a special kind of confidence interval) based on the sample (but not using the new observation) using the *standard error for prediction*, $S\sqrt{1+1/n}$, as explained in Chapter 9. Then get the new observation. If the new observation is *not* in the interval, you will conclude that it is significantly different from the others.

If you want to use the *t*-test method, simply compute your *t*-statistic as the new observation minus the sample average, divided by the standard error for prediction. Then proceed just as before, comparing the *t*-statistic to the critical *t*-value (with $n-1$ degrees of freedom). The *t*-statistic for testing a new observation is

$$t_{\text{statistic}} = \frac{X_{\text{new}} - \bar{X}}{S\sqrt{1+1/n}}$$

If you want a one-sided test to claim that the new observation is either significantly larger or significantly smaller than the average of the others, simply find the appropriate one-sided prediction interval or compare the *t*-statistic to the one-sided critical *t*-value.

Example

Is This System Under Control?

You are scratching your head. Usually, these art objects of molded porcelain that come out of the machine weigh about 30 pounds each. Of course, there is some variation; they do not all weigh *exactly* 30 pounds each—in fact, these "one-of-a-kind" objects are not supposed to be identical. But this is ridiculous! A piece that just came out weighs 38.31 pounds, way over the expected weight. You are wondering if the process has gone *out of control*, or if this is just a random occurrence to be expected every now and again. You would rather not adjust the machinery, since this involves shutting down the assembly line and finding the trouble; but if the line is out of control, the sooner you fix the problem, the better.

The null hypothesis claims that the system is still under control, that is, that the most recent piece is the same as the others except for the usual random variation. The research hypothesis claims that the system is out of control, and the most recent piece is significantly different from the others. Here is the information for the most recent piece as well as for a sample of ordinary pieces:

(*Continued*)

Example—cont'd

Sample size, n	19
Sample average, $\bar{X}$	31.52
Standard deviation, S	4.84
New observation, X_{new}	38.31

The standard error for prediction is

$$\text{Standard error for prediction} = S\sqrt{1+\frac{1}{n}} = 4.84\sqrt{1+\frac{1}{19}} = 4.965735$$

It would not be fair to use a one-sided test here because you would most certainly be interested in items that are greatly underweight as well as those that are overweight; either way, the system would be considered out of control. The two-sided 95% prediction interval, based on the critical t-value of 2.100922, extends from $31.52 - 2.100922 \times 4.965735 = 21.1$ to $31.52 + 2.100922 \times 4.965735 = 42.0$.

We are 95% sure that a new observation, taken from the same population as the sample, will be somewhere between 21.1 and 42.0 pounds.

Since the new observation, at 38.31 pounds, is within this prediction interval, it seems to be within the range of reasonable variation. Although it is near the high end, it is *not significantly different* from the others.

The t-statistic is less (in absolute value) than the critical value, 2.100922, confirming your decision to accept the null hypothesis and find the difference to be not significant:

$$t_{statistic} = \frac{38.31 - 31.52}{4.965735} = 1.367$$

For reference, the p-value for this test, from statistical software, is $p = 0.188$, which is not less than 0.05 and therefore not significant, in agreement with the confidence interval and the t-test methods.

In retrospect, you probably should not have been surprised at a piece weighing 38.31 pounds. Since the sample standard deviation is 4.84 pounds, you expect individuals to be about this far from the average. This piece is not even two standard deviations away from the mean and is therefore (even according to this approximate rule) within the reasonable 95% region. This quick look is just an approximation; when you use the standard error for prediction, your answer is exact because you took into account both the variation in your sample and the variation of the new observation in a mathematically correct way.

10.6 TESTING TWO SAMPLES

To test whether or not two samples are significantly different, on average, all you need to know are, (1) the appropriate standard error to use for evaluating the average difference and (2) its degrees of freedom. The problem will then be essentially identical to the methods you already know: You will be testing an observed quantity (the observed average difference) against a known reference value (zero, indicating no difference) using the appropriate standard error and critical t-value.

You will see this method repeated over and over in statistics. Whenever you have an estimated quantity and its own standard error, you can easily construct confidence intervals and do hypothesis testing. The applications get more complex (and more interesting), but the methods are just the same as the procedures you already know. Let us generalize this method now to the two-sample case, for which there are two possibilities: the two samples might be *paired* (ie, two measurements in the same units are made for each elementary unit from a single sample) or the two samples might be *unpaired* (so that the two samples represent two independent samples of elementary units).

The Paired *t*-Test

The **paired *t*-test** is used to test whether two columns of numbers are different, on average, *when there is a natural pairing between the two columns*. This is appropriate, for example, for "before/after" studies, where you have a measurement (such as a score on a test or a rating) for each person or thing both before and after some intervention (seeing an advertisement, taking a medication, adjusting the gears, etc.).

In fact, you already know how to do a paired t-test because it can be changed into a familiar one-sample problem by working with the *differences*, for example, "after" minus "before," instead of with the two lists individually. It is crucial that the data be paired; otherwise, it would not be clear how to line up the pairs when finding differences.

It is not enough to have the averages and standard deviations for each of the two groups. This would lack any indication of the important information conveyed by the pairing of the observations. Instead, you will work with the average and the standard deviation of the *differences*.

A paired t-test can be very effective even when individuals show lots of variation from one to another. Since it concentrates on *changes*, it can ignore the (potentially confusing) variation in *absolute* levels of individuals. For example, individuals could be very different from one another, and yet the changes could be very similar (eg, everyone receives a $100 pay raise). The paired t-test is not distracted by this individual variability in its methods to detect a systematic change, and we might say that the paired t-test "adjust for" or "controls for" this source of variability from one individual to another.

Again, some assumptions are required for validity of the paired t-test. The first assumption is that the elementary

units being measured are a random sample selected from the population of interest. Each elementary unit produces two measurements. Next, look at the data set consisting of the differences between these two sets of measurements. The second assumption is that the average of these differences is (at least approximately) normally distributed.

Example

Reactions to Advertising

An advertisement is being tested to determine if it is effective in creating the intended mood of relaxation. A sample of 15 people has been tested just before and just after viewing the ad. Their questionnaire included many items, but the one being considered now asked them to describe their current feelings on a scale from 1 (very tense) to 5 (completely relaxed). The results are shown in Table 10.6.1. (Note in particular that the average relaxation score increased by 0.6667, going from 2.8000 before, to 3.4667 after.)

This looks a lot like a two-sample problem, but, in a way, it is not. It can be viewed as a one-sample problem based on the *changes* in the relaxation scores. For example, person 1 went from a 3 to a 2, for a change of −1 in relaxation score. (By convention, we compute the differences as "after" minus "before" so that increases end up as positive numbers and decreases as negatives.) Computing the difference for each person, you end up with a familiar one-sample problem, as shown in Table 10.6.2.

You know exactly how to attack this kind of one-sample problem. Using the two-sided critical t-value of 2.144787, together with a sample average difference of $\bar{X} = 0.6667$ and a standard error of $S_{\bar{X}} = 0.2702$, you find

> You are 95% sure that the mean change in relaxation score for the larger population is somewhere between 0.087 and 1.25.

What is the reference value, μ_0, here? It is $\mu_0 = 0$ because a change of zero indicates *no effect* (zero effect) on relaxation in the population due to viewing the advertisement.

The hypothesis test merely involves seeing whether or not $\mu_0 = 0$ is in the confidence interval. It is not, so the result is significant; that is, 0 is not a reasonable value for the change in the population based on your data:

> Viewing of the advertisement resulted in a significant increase in relaxation ($p < 0.05$, two-sided test). For reference, statistical software produces a p-value of 0.0271 for this situation, which is indeed less than

(*Continued*)

TABLE 10.6.1 Relaxation Scores

	Before	After
Person 1	3	2
Person 2	2	2
	2	2
	4	5
	2	4
	2	1
	1	1
⋮	3	5
	3	4
	2	4
	5	5
	2	3
	4	5
	3	5
Person 15	4	4
Sample size	15	15
Average	2.8000	3.4667
Standard deviation	1.0823	1.5055

TABLE 10.6.2 Change in Score

	After-Before
Person 1	−1
Person 2	0
	0
	1
	2
	−1
⋮	0
	2
	1
	2
	0
	1
	1
	2
Person 15	0
Sample size	15
Average	0.6667
Standard deviation	1.0465
Standard error	0.2702

Example—cont'd

0.05, confirming our confidence-interval result. Of course, the *t*-statistic method produces the same result: the *t*-statistic is 2.47, which is greater (in absolute value) than the critical *t*-value of 2.144787.

A two-sided test is needed here because it is also of interest to find out if an advertisement caused a significant *reduction* in relaxation. Having found significance with the two-sided test, you may state the result as a one-sided conclusion.

For completeness, here are the underlying hypotheses: The null hypothesis, $H_0{:}\mu = 0$, claims that the population mean change in relaxation from before to after is zero; that is, there is no change in mean relaxation. The research hypothesis, $H_1{:}\mu \neq 0$, claims that there is indeed a change in mean relaxation from before to after.

Example

Data Mining to Compare Current to Previous Donations

After people skip a charitable contribution, does their next donation tend to be smaller or larger than the amount they used to give? To answer this, consider the donations database with 20,000 entries on the companion website, which we will view as a random sample from a much larger database. Recall that, at the time of the mailing, these people had given in the past but not recently. Focusing attention on the 989 donors (who did make a donation in response to the current mailing), we know the average size of their previous donations (this variable is named "AvgGift_D1" in the database for these 989 current donors) and the actual size of their current donation (named "Donation_D1").

The current donation (on average, \$15.7655 for the 989 current donors) is larger than the average of the previous donations (on average, \$11.9149, for the current donors). This suggests that it is indeed worthwhile to continue to ask for donations even when people do not respond initially, by showing that (at least among those who do respond, after a while) the donation amount increases, on average. But does it increase significantly? A hypothesis test will give the answer. This is a paired situation because each person has a current donation amount and a past average donation amount, and we are interested in their difference. The standard error of the differences is \$0.2767, and the 99.9% confidence interval statement (using a critical *t*-value of 3.300401 and the average difference of \$15.7655 − \$11.9149 = \$3.8506) is:

We are 99.9% sure that the population mean increase in donation amount, for past donors who lapsed but then resumed donating, is somewhere between \$2.94 and \$4.76.

Because the reference value 0 (representing no difference, on average) is not in this confidence interval, we conclude that the difference is significant at the 0.1% level, and the difference is therefore very highly significant ($p < 0.001$). The conclusion of this paired hypothesis test may be stated as

Past donors who lapsed but then resumed donating showed a very highly significant increase ($p < 0.001$) in their donation amount, on average, as compared to their previous average donation.

Alternatively, the *t*-statistic is $(15.7655 - 11.9149)/0.2767 = 13.9$, which is considerably larger than the critical *t*-value, 3.300401, for testing at the 0.1% level, confirming the differences as very highly significant.

Often, when large amounts of data are available, differences are found to be very highly significant. This happens because a larger sample size (all else being equal) leads to a smaller standard error after dividing the standard deviation by the square root of the large sample size *n*. In fact, in this example the *p*-value is much smaller than 0.001. When we use statistical software, the exact *p*-value is $p = 2.48\text{E}-40$, where the "E − 40" in computer output tells us to move the decimal point 40 places (to the left since −40 is a negative number) so we actually have the very small *p*-value,

$$p = 0.0000000000000000000000000000000000000000248$$

which would be reported by some statistical software as "$p = 0.0000$," obtained by rounding. This is consistent with our intuition that being 13.9 standard deviations away from the mean happens with a very small probability.

The Unpaired *t*-Test

The **unpaired *t*-test** is used to test whether two *independent* columns of numbers are different, on the average. Such columns have no natural pairings. For example, you might have data on firms in each of two industry groups, or you might want to compare samples from two different production lines. These cases cannot be reduced to just a single column of numbers; you will have to deal with both samples.

Once you find the appropriate standard error for this situation, the rest is easy. You will have an estimate (the difference between the two sample averages), its "personal" standard error, and the appropriate number of degrees of freedom. The rest, constructing the confidence interval and performing the hypothesis test, should be routine for you by now.

We have two samples, sample 1 and sample 2. The summary statistics for each sample will be denoted in a natural way, as shown in Table 10.6.3.

Here is what is new. The **standard error of the difference** indicates the sampling variability of the *difference* between the two sample averages. There are two different formulas: a large-sample formula, to be used whenever both sample sizes are 30 or larger, and a small-sample formula that is based on the assumption that the two populations

TABLE 10.6.3 Notation for Two Samples

	Sample 1	Sample 2
Sample size	n_1	n_2
Average	$\bar{X}_1$	$\bar{X}_2$
Standard deviation	S_1	S_2
Standard error	$S_{\bar{X}_1}$	$S_{\bar{X}_2}$
Average difference	$\bar{X}_2 - \bar{X}_1$	

have the same variability.[21] The large-sample formula works even when the variabilities are unequal by directly combining the two standard errors, $S_{\bar{X}_1}$ and $S_{\bar{X}_2}$, using the mathematical fact that the variance of a sum (or difference) is the sum of the variances for independent estimates. The small-sample formula includes a weighted average of the sample standard deviations to estimate the population variability (assumed equal in the two populations). The small-sample standard error has n_1+n_2-2 degrees of freedom: We start with the combined sample size, n_1+n_2, and then subtract 1 for each sample average that was estimated. Here are formulas for the standard error of the difference for each case:

Standard Error of the Difference

Large-sample situation ($n_1 \geq 30$ and $n_2 \geq 30$):

$$S_{\bar{X}_2-\bar{X}_1} = \sqrt{S_{\bar{X}_1}^2 + S_{\bar{X}_2}^2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$S_{p_2-p_1} = \sqrt{S_{p_1}^2 + S_{p_2}^2} \text{ (for two binomials)}$$

Degrees of freedom = infinity, as an approximation

Small-sample situation (equal variabilities assumed):

$$S_{\bar{X}_2-\bar{X}_1} = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

Degrees of freedom = n_1+n_2-2

For infinite degrees of freedom, the two-sided critical t-value may be computed in Excel as either NORMSINV (1/2+confidenceLevel/2) or as NORMSINV (1 − testLevel/2), while the one-sided critical t-value is either NORMSINV (confidenceLevel) or NORMSINV (1 − testLevel). For the small-sample situation, the two-sided critical t-value may be computed as either TINV (1 − confidenceLevel, DF) or TINV (testLevel, DF), while the one-sided critical t-value is either TINV (2 * (1 − confidenceLevel), DF) or TINV (2 * testLevel, DF), where "DF" is the degrees of freedom number n_1+n_2-2 in this case.

Be careful to use the correct variability measure for each formula, either the sample standard deviation or the standard error for the sample; the large-sample formula shows how to use either one. If, in the small-sample case, you are given the standard errors instead of the standard deviations, convert them to standard deviations by multiplying by the square root of the sample size for each sample. Note that in both formulas the standard deviations are squared before being combined.[22]

For the large-sample standard error formula, the estimated *variances* of the estimators $\bar{X}_1$ and $\bar{X}_2$ are added to derive the estimated variance of the difference. Taking the square root, you find the estimated standard deviation of the difference, which gives you the standard error of the difference.

For the small-sample standard error formula, the first fraction inside the square root sign combines the standard deviations using a weighted average (weighted according to the number of degrees of freedom for each one). The rest of the formula converts from the variation of *individuals* to the variation of the *average difference* by summing the reciprocal sample sizes, doing twice what you would do once to find an ordinary standard error.

The hypotheses being tested are H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 \neq \mu_2$. These may be written equivalently as H_0: $\mu_1 - \mu_2 = 0$ against H_1: $\mu_1 - \mu_2 \neq 0$. The assumptions needed in order for an unpaired two-sample t-test to be valid include the usual ones, plus one new one for the small-sample case only. First, each sample is assumed to be a random sample from its population. (There are two populations here, with each sample representing one of them independently of the other.) Second, each sample average is assumed to be approximately normally distributed (at least on average) as we have required before. Finally, for the small-sample case only, it is also assumed that the *standard deviations are equal* in the two populations: $\sigma_1 = \sigma_2$. That is, the two populations differ (if at all) only in mean value and not in terms of the variability of individuals from the mean for their population.

21. Solutions are available for the small-sample problem when variabilities are unequal, but they are more complex. One approach is presented in G. W. Snedecor and W. G. Cochran, *Statistical Methods*, 6th ed. (Ames: Iowa State University Press, 1976), p. 115.

22. Thus, the *variances* are averaged here, as happens in so many formulas like this one. This has led theoretical statisticians to concentrate their attention on the variance. However, anyone who wants to interpret such numbers in their meaningful measurement units will have to take the square root. This is why we work with the standard deviation rather than with the variance in this book. Note that their information is equivalent because either may be converted to the other.

Example

Gender Discrimination and Salaries

Your firm is being sued for gender discrimination, and you are evaluating the documents filed by the other side. They have included a statistical hypothesis test, based on salaries of men and women, that finds a "highly significant difference," on average, between men's and women's salaries. Table 10.6.4 shows a summary of their results.

TABLE 10.6.4 Salaries Arranged by Gender

	Women		Men
	$21,100		$38,700
	29,700		30,300
	26,200		32,800
	23,000		34,100
	25,800		30,700
	23,100		33,300
	21,900		34,000
	20,700		38,600
	26,900		36,900
	20,900		35,700
	24,700		26,200
	22,800		27,300
	28,100		32,100
	25,000		35,800
	27,100		26,100
			38,100
			25,500
			34,000
			37,400
			35,700
			35,700
			29,100
Sample size	15		22
Average	$24,466.7		$33,095.5
Standard deviation	$2,805.5		$4,188.8
Standard error	$724.4		$893.1
Average difference		$8,628.8	

There are 15 women and 22 men in this department; the average yearly salaries are $24,466.7 for women and $33,095.5 for men. On average, men earn $8,628.8 more than women. This is a plain, clear fact. However, the issue is whether or not this difference is within the usual random variation. Essentially, no matter how you divide this group of 37 people into two groups of sizes 15 and 22, you will find different average salaries. The question is whether such a large difference as found here could reasonably be the result of a *random* allocation of salaries to men and to women, or if there is a need for some other explanation for the apparent inequity.

Each standard deviation ($2,805.5 for women, $4,188.8 for men) indicates that individuals within each group differ from their group average by roughly this amount. There is a bit more variation among the men than the women, but not enough to keep us from going ahead with a two-sample unpaired *t*-test.

The standard errors ($724.4 for women, $893.1 for men) indicate about how far the group averages are from the means for their respective idealized populations. For example, if you view these particular 15 women as a random sample from the idealized population of women in similar circumstances, then the average women's salary of $24,466.7 (random, due to the fact that only 15 have been examined) is approximately $724.4 away from the idealized population mean.

This is clearly a two-sample *unpaired* situation. Although you might want to subtract Mary's salary from Jim's, there is no systematic way to complete the process because these are really two separate, unpaired groups.

To evaluate the average difference of $8,628.8 to see if it could be reasonably due to randomness, you need its standard error and number of degrees of freedom. Here are computations for the small-sample formula:

$$S_{\bar{X}_2-\bar{X}_1}=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$=\sqrt{\frac{(15-1)2{,}805.5^2+(22-1)4{,}188.8^2}{15+22-2}\left(\frac{1}{15}+\frac{1}{22}\right)}$$

$$=\sqrt{\frac{(14)2{,}805.5^2+(21)4{,}188.8^2}{35}(0.066667+0.045455)}$$

$$=\sqrt{13{,}675{,}959.4\times 0.112121}$$

$$=\$1{,}238.3$$

$$(\text{Degrees of freedom})=n_1+n_2-2=15+22-2=35$$

The two-sided 99.9% confidence interval is based on the critical *t*-value 3.591147. When computing this critical *t*-value, note that the *degrees of freedom* are 35 here, and that this is not the same as the sample size because more than one sample is involved; for example, you would use 35 instead of 35 − 1 in the Excel formula = TINV (0.001, 35). The confidence interval extends from $8{,}628.8 - 3.591147 \times 1{,}238.3$ to $8{,}628.8 + 3.591147 \times 1{,}238.8$:

You are 99.9% sure that the population mean salary difference is somewhere between $4,182 and $13,076.

Example—cont'd

This confidence interval does *not* include the reference value of 0, where such a reference value corresponds to no mean difference in salary between men and women in the population. Your hypothesis testing decision therefore is as follows:

> The average difference between men's and women's salaries is very highly significant ($p < 0.001$).

This result is also supported by the fact that the *t*-statistic is 8,628/1,238 = 6.97, well above the critical *t*-value of 3.591147 for testing at the 0.001 significance level. It is further supported by the statistical software of the exact *p*-value of 4.20E-08 or 0.0000000420 for this two-sample unpaired *t*-test. This *p*-value says that the probability of finding such a large average difference, if salaries were randomly assigned to men and women, is less than 1 in 23 million!

What can you conclude from this? First of all, the salary allocation between men and women is not just random. Well, it *could* be random, but only if you are willing to admit that a rare, less than 1-in-1,000 event has happened (since this is the meaning of the significance level 0.001). Second, if the salary allocation is not random, there must be some other explanation. At this point, an individual may give his or her own favorite reason as though it were completely proven by the test result. However, it is one thing to say that there is a reason and another to be able to say *what* the reason is. Statistics has ruled out random chance as a reasonable possibility. That is all. If you want to propose a reason for the observed salary difference, you are entitled to do so, but do not expect the field of statistics to back you up. Having set the stage for an explanation, the field of statistics then exits, riding off into the sunset like the Lone Ranger.

So what might cause the salary difference? One explanation is that management, in its outdated, selfish, and illegal ways, has deliberately decided to pay people less if they are women than if they are men, looking only at the person's gender. However, it might be argued that this is not the only plausible explanation. The salary difference might be due to some other factor that (1) determines salary and (2) is linked to gender. In its defense, the firm might argue that it pays solely on the basis of *education* and *experience*, and it is not to be blamed for the fact that its pool of applicants consisted of better-educated and more-experienced men as compared to the women. This argument basically shifts the blame from the firm to society in general.

This is a complicated issue. Fortunately (for the author!) the resolution of the question one way or the other will not be attempted in this book. It can be dodged by pointing out that it is not a statistical question and should be decided using expertise from another field of human endeavor. But stay tuned. This question will reappear in Chapter 12 on multiple regression (with more data) in our continuing efforts to understand the interactions among gender, salary, education, and experience.

The field of statistics can be very helpful in providing exact answers in the presence of uncertainty, but the answers are limited in their scope and much further work and thought may be required before you reach a final explanation.

Example

Your Productivity Versus Theirs

You have a friendly rivalry going with the manager of the other division over employee productivity. Actually, it is not entirely friendly because you both report to the same boss, who allocates resources based on performance. You would not only like to have the higher productivity, but would like it to be *significantly* higher so that there is no question about whose employees produce more.[23]

Here are summary measures of employee productivity in the two divisions:

	Your Division	Your Rival's Division
Sample size	53	61
Average	88.23	83.70
Standard deviation	11.47	9.21
Standard error	1.58	1.18
Average difference	4.53	

To evaluate the average difference of 4.53 to see if it could be reasonably due to randomness, you need its standard error. Following are computations for the large-sample formula, which is appropriate because both sample sizes are at least 30:

$$S_{\bar{X}_2-\bar{X}_1} = \sqrt{S_{\bar{X}_1}^2 + S_{\bar{X}_2}^2}$$

$$= \sqrt{1.58^2 + 1.18^2}$$

$$= \sqrt{3.8888}$$

$$= 1.9720$$

The two-sided 95% confidence interval is based on the critical *t*-value 1.959964 with infinite degrees of freedom, perhaps computed using Excel as =NORMSINV (1/2 + 0.95/2). The confidence interval extends from $4.53 - 1.959964 \times 1.9720$ to $4.53 + 1.959964 \times 1.9720$:

> You are 95% sure that the population mean productivity difference is somewhere between 0.66 and 8.40.

This confidence interval does *not* include the reference value of 0, which would indicate no mean difference in productivity between the two divisions in the (idealized) population. Thus, your hypothesis testing decision is as follows:

> The average difference between your employee productivity and that of your rival is statistically significant.

The *t*-statistic approach would, of course, have given the same answer. The *t*-statistic here is 4.53/1.9720 = 2.30, which exceeds the critical *t*-value of 1.959964. The *p*-value

(*Continued*)

Example—cont'd

of 0.0216 from statistical software also confirms this significant result.

The one-sided conclusion to a significant two-sided test may be used here. Because your division's productivity is higher, it follows that your division had *significantly higher productivity* than your rival's. Congratulations!

23. In reality, even after the hypothesis test is done, there will still be *some* questions because there is always some possibility of an error (type I or type II). What you hope to demonstrate is that the superior average productivity of your employees is not likely to be due to randomness alone.

10.7 END-OF-CHAPTER MATERIALS

Summary

Hypothesis testing uses data to decide between two possibilities (called *hypotheses*); it is often used to distinguish structure from mere randomness and should be viewed as a helpful input to executive decision making. A **hypothesis** is a statement about the population that may be either right or wrong; the data will help you decide which one (of two hypotheses) to accept as true. The **null hypothesis**, denoted H_0, represents the *default*, often a very specific case, such as pure randomness. The **research hypothesis** or **alternative hypothesis**, H_1, has the burden of proof, requiring convincing evidence against H_0 for its acceptance. Accepting the null hypothesis is a weak conclusion, whereas rejecting the null and accepting the research hypothesis is a strong conclusion and a significant result. The result is defined to be **statistically significant** whenever you accept the research hypothesis because you have eliminated the null hypothesis as a reasonable possibility. Every hypothesis test can produce a ***p*-value** (using statistical software) that tells you how surprised you would be to learn that the null hypothesis had produced the data, with smaller p-values indicating more surprise and leading to rejection of H_0 when p is less than the conventional 5% threshold.

For testing whether or not the population mean, μ, is equal to a reference value, μ_0, the hypotheses are H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$. The **reference value**, μ_0, is a known, fixed number that does not come from the sample data. This is a **two-sided test** because the research hypothesis allows possible population mean values on both sides of the reference value. This test of a population mean is known as the ***t*-test** or **Student's *t*-test**. The outcome of the test is determined by checking if the sample average, $\bar{X}$, is farther from the reference value, μ_0, than random chance would allow, if the population mean, μ, were actually equal to μ_0. Thus, the distance from $\bar{X}$, to μ_0 is compared with the standard error, $S_{\bar{X}}$, using the critical t-value. The test may be based either directly on the p-value (if available) or equivalently on either the two-sided confidence interval (from Chapter 9) or on the ***t* statistic**, defined as follows:

$$t_{\text{statistic}} = \frac{\bar{X} - \mu_0}{S_{\bar{X}}}$$

Here is how the two-sided t-test is decided, using your choice of the p-value approach, the confidence interval approach, or the t statistic approach (which always give identical results):

If $p > 0.05$, or (equivalently) the reference value, μ_0, is in the two-sided confidence interval, or (equivalently) $|t_{statistic}| < t_{critical}$, then accept the null hypothesis, H_0, as a reasonable possibility. The sample average, $\bar{X}$, is *not significantly different* from μ_0. The observed difference between $\bar{X}$ and μ_0 could reasonably be just random. The result is *not statistically significant*.

If $p < 0.05$, or (equivalently) the reference value, μ_0, is *not* in the two-sided confidence interval, or (equivalently) $|t_{statistic}| > t_{critical}$, then accept the research hypothesis, H_1, and reject the null hypothesis, H_0. The sample average, $\bar{X}$, is *significantly different* from μ_0. The observed difference between $\bar{X}$ and μ_0 could *not* reasonably be just random. The result is *statistically significant*.

By deciding the hypothesis test in this way, you are accepting the null hypothesis ($\mu = \mu_0$) whenever μ_0 appears to be a reasonably possible value for μ. When the null hypothesis is true, your probability of deciding correctly is equal to the confidence level (95% or other) used to find the critical t-value.

Table 10.7.1 shows a summary of the situation for testing either the mean of a normal distribution or the probability of occurrence for a binomial distribution.

The t statistic is an example of the general concept of a **test statistic**, which is the most helpful number that can be computed from your data for the purpose of deciding between two given hypotheses. The test statistic is compared to the appropriate **critical value**, for example, the critical t-value. A useful rule of thumb is that if the t statistic is larger in absolute value than about 2, you reject the null hypothesis; otherwise, you accept it.

Depending on which is (in reality) the true hypothesis, there are two types of errors that you might make. The **type I error** is committed when the null hypothesis is true, but you reject it and declare that your result is statistically significant. The probability of committing a type I error (when the null hypothesis is true) is controlled by your choice critical t-value, conventionally the 5% level. The **type II error** is committed when the research hypothesis is true, but you accept the null hypothesis instead and declare the result *not* to be significant. The probability of committing a type II error (when the research hypothesis is true) is not easily controlled but can (depending on the true value of μ) be anywhere between 0 and the confidence level of the test (eg, 95%). Note that each type of error is based

TABLE 10.7.1 Testing Either the Mean of a Normal Distribution or the Probability of Occurrence for a Binomial Distribution

	Normal	Binomial
Population mean	μ	π
Reference value	μ_0	π_0
Null hypothesis	$H_0: \mu = \mu_0$	$H_0: \pi = \pi_0$
Research hypothesis	$H_1: \mu \neq \mu_0$	$H_1: \pi \neq \pi_0$
Data	$X_1, X_2, \ldots, X_n$	X occurrences out of n trials
Estimator	$\bar{X}$	$p = X/n$
Standard error	$S_{\bar{X}} = S/\sqrt{n}$	$S_p = \sqrt{p(1-p)/n}$
Confidence interval	From $\bar{X} - t_{critical} S_{\bar{X}}$ to $\bar{X} + t_{critical} S_{\bar{X}}$	From $p - t_{critical} S_p$ to $p + t_{critical} S_p$
t-Statistic	$t_{statistic} = (\bar{X} - \mu_0)/S_{\bar{X}}$	$t_{statistic} = (p - \pi_0)/S_p$

on an assumption about which hypothesis is true. Since each hypothesis is either true or false depending on the population (*not* on the data), there is no notion of the probability of a hypothesis being true.

The **assumptions for hypothesis testing** are (1) the data set is a random sample from the population of interest, and (2) either the quantity being measured is approximately normal, or else the sample size is large enough that the central limit theorem ensures that the sample average is approximately normally distributed.

The **test level** or **significance level** is the probability of accepting the research hypothesis when the null hypothesis is really true (ie, committing a type I error). By convention, this level is set at 5% but may reasonably be set at 1% or 0.1% (or even 10% for some fields of study) by using the appropriate critical t-value. The p-value tells you how surprised you would be to learn that the null hypothesis had produced the data, with smaller p-values indicating more surprise and leading to rejection of H_0. By convention, we reject H_0 whenever the p-value is less than 0.05. A result is statistically significant ($p < 0.05$) if it is significant at the 5% level. Other terms used are *highly significant* ($p < 0.01$), *very highly significant* ($p < 0.001$), and *not significant* ($p > 0.05$).

A **one-sided test** is set up with the null hypothesis claiming that μ is on one side of μ_0 and the research hypothesis claiming that it is on the other side. To use a one-sided test, you must be sure that *no matter how the data had come out* you would still have used a one-sided test on the same side ("larger than" or "smaller than"). If in doubt, use a two-sided test; if it is significant, you are then entitled to state the *one*-sided conclusion. The test may be performed either by examining the p-value from statistical software, by constructing the appropriate one-sided confidence interval (matching the claim of the research hypothesis) or by using the t statistic. A significant result (accepting the research hypothesis) will be declared whenever the reference value μ_0 does *not* fall in the confidence interval. This will happen whenever $\bar{X}$ is on the side of μ_0 claimed in the research hypothesis and the absolute value of the t statistic is larger than the critical t-value. A significant result will occur whenever $t_{statistic} > t_{critical}$ (if testing H_1: $\mu > \mu_0$) or $t_{statistic} < -t_{critical}$ (if testing H_1: $\mu < \mu_0$).

For the one-sided t-test to see if μ is *larger* than μ_0, the hypotheses are H_0: $\mu \leq \mu_0$ and H_1: $\mu > \mu_0$. The confidence interval includes all values *at least* as large as $\bar{X} - t_{critical} S_{\bar{X}}$.

If μ_0 is in the confidence interval or (equivalently) $t_{statistic} \leq t_{critical}$, then accept the null hypothesis, H_0, as a reasonable possibility. The sample average, $\bar{X}$, is *not significantly larger* than μ_0. If $\bar{X}$ is larger than μ_0, the observed difference could reasonably be just random. The result is *not statistically significant*.

If μ_0 is *not* in the confidence interval or (equivalently) $t_{statistic} > t_{critical}$, then accept the research hypothesis, H_1, and reject the null hypothesis, H_0. The sample average, $\bar{X}$, is *significantly larger* than μ_0. The observed difference could *not* reasonably be just random. The result is *statistically significant*.

For the one-sided t-test to see if μ is *smaller* than μ_0, the hypotheses are H_0: $\mu \geq \mu_0$ and H_1: $\mu < \mu_0$. The confidence interval includes all values *no larger* than $\bar{X} + t_{critical} S_{\bar{X}}$.

If μ_0 is in the confidence interval or (equivalently) $t_{statistic} \geq -t_{critical}$, then accept the null hypothesis, H_0, as a reasonable possibility. The sample average, $\bar{X}$, is *not significantly smaller* than μ_0. If $\bar{X}$ is smaller than μ_0, then the observed difference could reasonably be just random. The result is *not statistically significant*.

If μ_0 is *not* in the confidence interval or (equivalently) $t_{\text{statistic}} < -t_{\text{critical}}$, then accept the research hypothesis, H_1, and reject the null hypothesis, H_0. The sample average, $\bar{X}$, is *significantly smaller* than μ_0. The observed difference could *not* reasonably be just random. The result is *statistically significant*.

Whenever you have an estimator (such as $\bar{X}$), the appropriate standard error for that estimator (such as $S_{\bar{X}}$), and an appropriate critical value (such as the critical t-value), you may construct one- or two-sided confidence intervals (at various confidence levels) and perform one- or two-sided hypothesis tests (at various significance levels).

For the test of whether a new observation came from the same population as a sample, the null hypothesis claims that it did, and the research hypothesis claims otherwise. Using the standard error for prediction, $S\sqrt{1+1/n}$, to construct the prediction interval, accept the null hypothesis if the new observation falls within the interval; otherwise, accept the research hypothesis and declare significance. Or compute the t-statistic using the following equation, and compare it to the critical t-value:

For Testing a New Observation

$$t_{\text{statistic}} = \frac{X_{\text{new}} - \bar{X}}{S\sqrt{1+1/n}}$$

Whichever method you choose (confidence interval or t statistic), you have available all of the significance levels, p-value statements, and one- or two-sided testing procedures as before.

The **paired *t*-test** is used to test whether or not two samples have the same population mean value when there is a natural pairing between the two samples—for example, "before" and "after" measurements on the same people. By working with the differences ("after" minus "before"), we reduce such a problem to the familiar one-sample t-test, using $\mu_0 = 0$ as the reference value expressing the null hypothesis of no difference in means.

The **unpaired *t*-test** is used to test whether or not two samples have the same population mean value when there is *no* natural pairing between the two samples; that is, each is an independent sample from a different population. For a two-sided test, the null hypothesis claims that the mean difference is 0. To construct confidence intervals for the mean difference and to perform the hypothesis test, you need the **standard error of the difference** (which gives the estimated standard deviation of the sample average difference) and its degrees of freedom.

For a large-sample situation ($n_1 \geq 30$ and $n_2 \geq 30$):

$$S_{\bar{X}_2 - \bar{X}_1} = \sqrt{S_{\bar{X}_1}^2 + S_{\bar{X}_2}^2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$S_{p_2 - p_1} = \sqrt{S_{p_1}^2 + S_{p_2}^2} \text{ (for two binomials)}$$

Degrees of freedom = infinity, as an approximation

For a small-sample situation (assuming equal variabilities):

$$S_{\bar{X}_2 - \bar{X}_1} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Degrees of freedom = $n_1 + n - 2$

Based on the average difference, its standard error, its number of degrees of freedom, and the reference value (0), you can construct confidence intervals and perform hypothesis tests in the usual way. Note that, in addition to the usual assumptions of random samples and normal distributions, the small-sample situation also requires that the population variabilities be equal ($\sigma_1 = \sigma_2$).

Keywords

Questions

1. **a.** What is the purpose of hypothesis testing?
   b. How is the result of a hypothesis test different from a confidence interval statement?
2. **a.** What is a hypothesis? In particular, is it a statement about the population or the sample?
   b. How is the role of the null hypothesis different from that of the research hypothesis? Which one usually includes the case of pure randomness? Which one has the burden of proof? Which one has the benefit of the doubt?
   c. Suppose you decide in favor of the null hypothesis. Is this a weak or a strong conclusion?
   d. Suppose you decide in favor of the research hypothesis. Is this a weak or a strong conclusion?
   e. "A null hypothesis can never be disproved." Comment.

3. Suppose you learn that the *p*-value for a hypothesis test is equal to 0.0217. What can you say about the result of this test?
4. a. Briefly describe the steps involved in performing a two-sided test concerning a population mean based on a confidence interval.
   b. Briefly describe the steps involved in performing a two-sided test concerning a population mean based on the *t*-statistic.
5. a. What is Student's *t*-test?
   b. Who was Student? What was his contribution?
6. a. What is the reference value? Does it come from the sample data? Is it known or unknown?
   b. What is the *t*-statistic? Does it depend on the reference value?
   c. Does the confidence interval change depending on the reference value?
7. a. What, in general, is a test statistic?
   b. Which test statistic would you use for a two-sided *t*-test?
   c. What, in general, is a critical value?
   d. Which critical value would you use for a two-sided *t*-test?
8. a. What assumptions must be satisfied for a two-sided *t*-test to be valid?
   b. Consider each assumption in turn. What happens if the assumption is not satisfied? What, if anything, can be done to fix the problem?
9. a. What is a type I error? Can it be controlled? Why or why not?
   b. What is a type II error? Can it be controlled? Why or why not?
   c. When, if ever, is it correct to say that "the null hypothesis is true with probability 0.95"?
   d. What can you say about your lifetime track record in terms of correct decisions to accept a true null hypotheses?
10. What *p*-value statement is associated with each of the following outcomes of a hypothesis test?
    a. Not significant.
    b. Significant.
    c. Highly significant.
    d. Very highly significant.
11. a. What is a one-sided test?
    b. What are the hypotheses for a one-sided test?
    c. When are you allowed to perform a one-sided test? What should you do if you are not sure if it is allowed?
    d. If you perform a one-sided test when it's really not permitted, what is the worst that can happen?
    e. Under what conditions are you permitted to make a one-sided statement based on a two-sided test?
12. a. How is a one-sided test performed based on a confidence interval?
    b. How is a one-sided test performed based on the *t*-statistic?
13. Suppose you have an estimator and would like to test whether or not the population mean value equals 0. What do you need in addition to the estimated value?
14. What standard error would you use to test whether a new observation came from the same population as a sample? (Give both its name and the formula.)
15. a. What is a paired *t*-test?
    b. Identify the two hypotheses involved in a paired *t*-test.
    c. What is the "pairing" requirement? Give a concrete example.
    d. How is a paired *t*-test similar to and different from an ordinary *t*-test for just one sample?
16. a. What is an unpaired *t*-test?
    b. Identify the two hypotheses involved in an unpaired *t*-test.
    c. What is the "independence" requirement? Give a concrete example.
    d. How is an unpaired *t*-test similar to and different from an ordinary *t*-test for just one sample?
    e. When is each standard error appropriate? (Answer both in words and using a formula.)
    f. What new assumption is needed for the unpaired *t*-test to be valid for small samples? What can you do if this assumption is grossly violated?
17. a. Describe the general process of constructing confidence intervals and performing hypothesis tests using the rule of thumb when you have an estimator and its standard error.
    b. If you also know the number of degrees of freedom, how would your answer change to be more exact?

Problems

Problems marked with an asterisk (*) are solved in the Self-Test in Appendix C.

1.* To help your restaurant marketing campaign target the right age levels, you want to find out if there is a statistically significant difference, on the average, between the age of your customers and the age of the general population in town, 43.1 years. A random sample of 50 customers shows an average age of 33.6 years with a standard deviation of 16.2 years.
    a. Identify the null and research hypotheses for a two-sided test using both words and mathematical symbols.
    b. Perform a two-sided test at the 5% significance level and describe the result.
2.* a. Perform a two-sided test at the 1% significance level for the previous problem and describe the result.
    b. State the *p*-value as either $p > 0.05$, $p < 0.05$, $p < 0.01$, or $p < 0.001$.
    c. Find the *p*-value using statistical software.
3. Part of the assembly line will need adjusting if the consistency of the injected plastic becomes either too viscous or not viscous enough as compared with a value (56.00) your engineers consider reasonable. You will decide to adjust only if you are convinced that the system is "not in control," that is, there is a real need for

adjustment. The average viscosity for 13 recent measurements was 51.22 with a standard error of 3.18.

a. Identify the null and research hypotheses for a two-sided test, using both words and mathematical symbols.
b. Perform a two-sided test at the 5% significance level and describe the result.
c. Perform a two-sided test at the 1% significance level and describe the result.
d. State the p-value as either $p>0.05$, $p<0.05$, $p<0.01$, or $p<0.001$.
e. Find the p-value using statistical software.

4. a. Why is a two-sided test appropriate for the previous problem?
b. State the one-sided result of the two-sided test at the 5% level, if appropriate.

5. Some of your advertisements seem to get no reaction, as though they are being ignored by the public. You have arranged for a study to measure the public's awareness of your brand before and after viewing a TV show that includes the advertisement in question. You wish to see if the ad has a statistically significant effect as compared with zero, representing no effect. Your brand awareness, measured on a scale from 1 to 5, was found to have increased an average of 0.22 point when 200 people were shown an advertisement and questioned before and after. The standard deviation of the increase was 1.39 points.
a. Identify the null and research hypotheses for a two-sided test, using both words and mathematical symbols.
b. Perform a two-sided test at the 5% significance level and describe the result.
c. Perform a two-sided test at the 1% significance level and describe the result.
d. State the p-value as either $p>0.05$, $p<0.05$, $p<0.01$, or $p<0.001$.
e. Find the p-value using statistical software.

6. a. Why is a two-sided test appropriate for the previous problem?
b. State the one-sided result of the two-sided test at the 5% level, if appropriate.

7. In a random sample of 725 selected for interview from your database of 13,916 customers, 113 said they are dissatisfied with your company's service.
a. Find the best estimate of the percentage of all customers in your entire database who are dissatisfied.
b. Find the standard error of your estimate of the percentage of all customers who are dissatisfied.
c. Find the best estimate of the overall number of dissatisfied customers within your database.
d. Find the 95% confidence interval for the percentage of dissatisfied customers.
e. The company's goal has been to keep the percentage of dissatisfied customers at or below 10%. Could this reasonably still be the case, or do you have convincing evidence that the percentage is larger than 10%? Justify your answer.

8. Your factory's inventory level was determined at 12 randomly selected times last year, with the following results: 313, 891, 153, 387, 584, 162, 742, 684, 277, 271, 285, 845
a. Find the typical inventory level throughout the whole year, using the standard statistical summary.
b. Identify the population.
c. Find the 95% confidence interval for the population mean inventory level.
d. Is the average of the measured inventory levels significantly different from 500, which is the number used for management budgeting purposes? Justify your answer.

9. Your bakery produces loaves of bread with "1 pound" written on the label. Here are weights of randomly sampled loaves from today's production:
1.02, 0.97, 0.98, 1.10, 1.00, 1.02, 0.98, 1.03, 1.03, 1.05, 1.02, 1.06
a. Find the 95% confidence interval for the mean weight of all loaves produced today.
b. Find the reference value for testing the average of the actual weights against the claim on the label.
c. Find the hypotheses, H_0 and H_1.
d. Perform the hypothesis test (two-sided, level 0.05) and report the result.
e. What error, if any, might you have committed?

10. View the 20,000 people in the donations database on the companion site as a random sample from a much larger group of potential donors. Determine whether or not the amount donated in response to the current mailing (named "Donation" in the database), on average, is enough to cover the per-person cost (assumed to be 38 cents) of preparing materials and mailing them. In particular, can you conclude that it was significantly worthwhile to solicit a donation from this group?

11. Suppose that the target response rate was 4% when the current mailing was sent to the 20,000 people in the donations database on the companion site.
a. Find the actual response rate represented by the 989 donations received in response to this mailing to 20,000 people.
b. How does the actual response rate compare to the target? Give a statement that includes information about statistical significance (or lack of significance).

12. If the list price of the Eureka 4750A Bagged Upright Vacuum cleaner is \$79.99, is the average price, based on the data from Table 9.6.1, significantly different from a 10% discount?

13. At a recent meeting, it was decided to go ahead with the introduction of a new product if "interested consumers would be willing, on average, to pay \$20.00 for the product." A study was conducted, with 315 random interested consumers indicating that they would pay an average of \$18.14 for the product. The standard deviation was \$2.98.
a. Identify the reference value for testing the mean for all interested consumers.

b. Identify the null and research hypotheses for a two-sided test using both words and mathematical symbols.
c. Perform a two-sided test at the 5% significance level and describe the result.
d. Perform a two-sided test at the 1% significance level and describe the result.
e. State the p-value as either $p>0.05$, $p<0.05$, $p<0.01$, or $p<0.001$.
f. Find the p-value using statistical software.

14. a. Why might a one-sided test be appropriate for the preceding problem?
b. Identify the null and research hypotheses for a one-sided test, using both words and mathematical symbols.
c. Perform a one-sided test at the 5% significance level and describe the result.

15. The p-value is 0.0371. What conclusions can you reach and what error might have been made?

16. Do initial public offerings (IPOs) of stock significantly increase in value, on average, in the short term? Test using the data from Table 4.3.7 that show performance of initial offerings as percent increases from the offer price, with most newly traded companies increasing in value while some lost money. Please give the p-value (as either $p>0.05$, $p<0.05$, $p<0.01$, or $p<0.001$) as part of your answer.

17. A recent poll of 809 randomly selected registered voters revealed that 426 plan to vote for your candidate in the coming election.
a. Is the observed percentage more than 50%?
b. Is the observed percentage significantly more than 50%? How do you know? Base your answer on a two-sided test.

18. Test whether or not the population percentage could reasonably be 20%, based on the observed 18.4% who like your products, from a random sample of 500 consumers.

19. As part of a decision regarding a new product launch, you want to test whether or not a large enough percentage (10% or more) of the community would be interested in purchasing it. You will launch the product only if you find convincing evidence of such demand. A survey of 400 randomly selected people in the community finds that 13.0% are willing to try your proposed new product.
a. Why is a one-sided test appropriate here?
b. Identify the null and research hypotheses for a one-sided test using both words and mathematical symbols.
c. Perform the test at the 5% significance level and describe the result.
d. Perform the test at the 1% significance level and describe the result.
e. State the p-value as either $p>0.05$, $p<0.05$, $p<0.01$, or $p<0.001$.

20. You are considering a new delivery system and wish to test whether delivery times are significantly different, on average, than your current system. It is well established that the mean delivery time of the current system is 2.38 days. A test of the new system shows that, with 48 observations, the average delivery time is 1.91 days with a standard deviation of 0.43 day.
a. Identify the null and research hypotheses for a two-sided test, using both words and mathematical symbols.
b. Perform a two-sided test at the 5% significance level and describe the result.
c. Perform a two-sided test at the 1% significance level and describe the result.
d. State the p-value as either $p>0.05$, $p<0.05$, $p<0.01$, or $p<0.001$.
e. Summarize the results in a brief memo to management.

21. You work for a company that prepares and distributes frozen foods. The package claims a net weight of 14.5 ounces. A random sample of today's production was weighed, producing the following data set:
14.43, 14.37, 14.38, 14.29, 14.60, 14.45, 14.16, 14.52, 14.19, 14.04, 14.31
A sample was also selected from yesterday's production. The average was 14.46 and the standard deviation was 0.31.
a. Estimate the mean weight you would have found had you been able to weigh all packages produced today.
b. For a typical individual package produced yesterday, approximately how different was the actual weight from yesterday's average?
c. Find the 95% confidence interval for the mean weight for all packages produced today.
d. Identify the hypotheses you would work with to test whether or not your claimed weight is correct, on average, today.
e. Is there a significant difference between claimed and actual mean weight today? Justify your answer.

22. Although your product, a word game, has a list price of $12.95, each store is free to set the price as it wishes. You have just completed a quick survey, and the marked prices at a random sample of stores that sell the product were as follows:
$12.95, 9.95, 8.95, 12.95, 12.95, 9.95, 9.95, 9.98, 13.00, 9.95
a. Estimate the mean selling price you would have found had you been able to survey all stores selling your product.
b. For a typical store, approximately how different is the actual selling price from the average?
c. Find the 95% confidence interval for the mean selling price for all stores selling your product.
d. Your marketing department believes that games generally sell at a mean discount of 12% from the list price. Identify the hypotheses you would work with to test the population mean selling price against this belief.
e. Test the hypotheses from part d.

23. Some frozen food dinners were randomly selected from this week's production and destroyed in order to

measure their actual calorie content. The claimed calorie content is 200. Here are the calorie counts for each dinner:

221, 198, 203, 223, 196, 202, 219, 189, 208, 215, 218, 207

a. Estimate the mean calorie content you would have found had you been able to measure all packages produced this week.

b. Approximately how different is the average calorie content (for the sample) from the mean value for all dinners produced this week?

c. Find the 99% confidence interval for the mean calorie content for all packages produced this week.

d. Is there a significant difference between claimed and measured calorie content? Justify your answer.

24. Consider the dollar value (in thousands) of gifts returned to each of your department stores after the holiday season (Table 10.7.2):

a. Compute the standard deviation.

b. Interpret the standard deviation as a measure of the variation from one store to another.

c. Compute the standard error of the average and briefly describe its meaning.

d. Find the two-sided 95% confidence interval for the mean value of returned merchandise for all downtown stores.

e. The Association of Downtown Merchants had been expecting an average value of $10,000 of returned merchandise per store, since this has been typical in the past. Test to see if this year's average differs significantly from their expectation.

25. Here are the satisfaction scores given by 12 randomly selected customers:

89, 98, 96, 65, 99, 81, 76, 51, 82, 90, 96, 76

Does the observed average score differ significantly from the target score of 80? Justify your answer.

26. Regulations require that your factory provide convincing evidence that it discharges less than 25 milligrams of a certain pollutant each week, on average, over the long run. A recent sample shows weekly amounts of 13, 12, 10, 8, 22, 14, 10, 15, 9, 10, 6, and 12 milligrams released.

TABLE 10.7.2 Dollar Value of Returned Gifts

Store	Returned
A	13
B	8
C	36
D	18
E	6
F	21

a. Have you complied with the regulations? Explain your answer based on a one-sided hypothesis test at the 5% level.

b. Report the p-value as either $p>0.05$, $p<0.05$, $p<0.01$, or $p<0.001$. In particular, is the result highly significant?

c. Identify the underlying hypotheses and assumptions involved in these tests.

d. All else equal, would the use of a two-sided test, instead of a one-sided test, result in more or fewer instances of "out-of-compliance" findings? Explain.

27. A manufacturing process is considered to be "in control" if the long run mean weight of components produced is 0.20 kilograms, even though individual components may vary from this mean. Here are weights of a random sample of recently produced components:

0.253, 0.240, 0.247, 0.183, 0.247, 0.223, 0.252, 0.195, 0.235, 0.241, 0.251, 0.261, 0.194, 0.236, 0.256, and 0.241

Does this process seem to be in control? Justify your answer.

28. Production yields vary and can be high or low on a given day. If they are high, you want to find out why so that yields could be similarly increased on other days. If they are low, you want to fix the problem. You have just learned that today's production yields seem to be lower than usual. Should you use a one-sided test or a two-sided test to investigate? Why?

29. A recent poll of 1,423 randomly sampled likely voters shows your favorite candidate ahead, with 51.93% in favor. There are two candidates. Use hypothesis testing to infer to the larger group of all likely voters to see whether or not this indicates that your candidate is ahead in the larger population.

a. Carefully identify the two-sided hypotheses.

b. Perform the hypothesis test at level 0.05 and give the result.

c. Make a careful, exact statement summarizing the result of the test and what it means.

d. Repeat parts b and c assuming that the percentage is 56.64% instead of 51.93%.

e. Explain why a one-sided test would be inappropriate here by showing that each of the three possible outcomes of a two-sided test would be of interest.

30. Managers perceived employee stock ownership as having a significant positive effect on product quality. As part of that same study, managers were also asked to rate the effect of employee stock ownership on unit labor cost.[24] This effect, on a scale from -2 (large negative effect) to 2 (large positive effect), was 0.12 with a standard error of 0.11, based on a sample of 343 managers.

a. Find the 95% confidence interval and state carefully what this represents. Keep in mind that these are opinions of randomly selected managers.

b. Is there a significant relationship between employee stock ownership and the unit cost of labor as perceived by managers? Why or why not?

c. Identify the null and research hypotheses.
d. Which hypothesis has been accepted? Is this a weak or a strong conclusion?
e. Has the accepted hypothesis been absolutely proven? If not, what type of error may have been made?

31. The goal of your marketing campaign is for more than 25% of supermarket shoppers to recognize your brand name. A recent survey of 150 random shoppers found that 21.3% recognized your brand name.
a. It might be argued that the burden of proof is to show that more than 25% of shoppers recognize your brand name. Identify the appropriate one-sided hypotheses in this case and perform the test at level 0.05.
b. On the other hand, it might be argued that you would be interested in knowing about all three possibilities: significantly more than 25% (indicating success), significantly less than 25% (indicating failure), and not significantly different from 25% (indicating that there is not enough information to say for sure). Identify the appropriate two-sided hypotheses in this case and perform the test at level 0.05.
c. For the two-sided test, write a brief paragraph describing the result, the error that might have been made, and its implications for your marketing strategy.

32. You are supervising an audit to decide whether or not any errors in the recording of account transactions are "material errors." Each account has a reported balance, whose accuracy can be verified only by careful and costly investigation; the account's error is defined as the difference between the reported balance and the actual balance. Note that the error is zero for any account that is correctly reported. In practical terms, for this situation involving 12,000 accounts, the total error is material only if it is at least $5,000. The average error amount for 250 randomly selected accounts was found to be $0.25, and the standard deviation of the error amount was $193.05. You may assume that your reputation as an auditor is on the line, so you want to be fairly certain before declaring that the total error is not material.
a. Find the estimated total error based on your sample and compare it to the material amount.
b. Identify the null and research hypotheses for a one-sided test of the population mean error per account and explain why a one-sided test is appropriate here.
c. Find the appropriate one-sided 95% confidence interval statement for the population mean error per account.
d. Find the *t*-statistic.
e. Which hypothesis is accepted as a result of a one-sided test at the 5% level?
f. Write a brief paragraph explaining the results of this audit.

33. Dishwasher detergent is packaged in containers that claim a weight of 24 ounces. Although there is some variation from one package to another, your policy is to ensure that the mean weight for each day's production is slightly over 24 ounces. A random sample of 100 packages from today's production indicates an average of 24.23 ounces with a standard deviation of 0.15 ounce.
a. Find the *p*-value (as either $p>0.05$, $p<0.05$, $p<0.01$, or $p<0.001$) for a one-sided hypothesis test to check if the population mean weight is above the claimed weight.
b. Write a brief paragraph summarizing your test and its results.
c. Is your conclusion a strong one or a weak one? Why?

34. Do employees take more sick leave in the year before retirement? They may well have an incentive to do so if their accumulated paid sick leave (the number of days they are entitled to be away with full pay) is about to expire. Indeed, this appears to happen with government workers. One evaluation of this issue looked at statistics gathered by the U.S. General Accounting Office (GAO).[25] The study concluded,

[What if] the bulge in sick days was just an aberration in the GAO sample rather than a real symptom of goofing off? In zeroing in on this question, we note that the 714 retirees in the GAO sample averaged 30 sick days in their last year instead of the "expected" 14 days. So in a work year of 251 days (average for federal employees), the retirees were finding themselves indisposed 12.0% of the time instead of 5.6%. Could that happen by chance? The science of statistics tells us that the probability of any such swing in so large a sample is low. To be precise, one in 200,000.

a. Identify the population and the sample.
b. Identify the hypotheses being tested, in terms of percent of time indisposed.
c. Identify the *p*-value here.
d. Which hypothesis (if any) has been rejected? Which has been accepted?
e. How significant (statistically) is the result?

35. Selected mutual funds that practice socially aware investing, with year-to-date rates of return, are shown in Table 10.7.3. On average, these funds lost value in the first half of 2010, in the sense that their average rate of return was negative. However, the Standard & Poor's 500 stock market index lost 9.03% of its value during the same period, so this was a difficult time for the market in general.
a. On average, as a group, did socially aware mutual funds lose significantly more than the market index? Please use the market index as the reference value.
b. Find the *p*-value for this test (as either $p>0.05$, $p<0.05$, $p<0.01$, or $p<0.001$). In particular, is it highly significant?
c. Identify the underlying hypotheses and assumptions involved in part a.
d. Under these assumptions, the hypothesis test makes a clear and correct statement. However, are the

TABLE 10.7.3 Performance of Socially Aware Investment Funds

Fund	Rate of Return
Calvert Global Alternative Energy Fund A	−26.99%
Calvert Global Water Fund	−9.28%
Calvert New Vision Small Cap A	−5.42%
Calvert Social Investment Balanced A	−2.16%
Calvert World Values International A	−11.07%
Domini Social Equity A	4.38%
Gabelli SRI Green Fund Inc A	−16.32%
Green Century Balanced	−4.00%
Legg Mason Prt Social Awareness Fund A	−4.27%
Neuberger Berman Socially Resp Inv	−0.53%
Pax World Global Green Fund—Individual Investor	−10.35%
Sentinel Sustainable Core Opportunities Fund	−7.38%
TIAA-CREF Social Choice Eq Retail	−5.98%
Walden Social Balanced Fund	−2.75%
Winslow Green Growth Fund	−15.59%

Source: From Social Investment Forum, accessed at http://www.social-invest.org/resources/mfpc/ on Jul. 14, 2010. Their source is Bloomberg.

assumptions realistic? Be sure to address independence (note that some of these funds are part of the same group).

e. Why is a two-sided test appropriate in this case? (*Hint:* You may wish to consider how the situation would have appeared if these funds had performed better than the market, on average.)

36. World investments markets were highly volatile in 1998. Table 10.7.4 shows one-year rates of return on closed-end mutual funds that specialize in income from international sources.

a. Do the rates of return of these closed-end world income funds, as a group, differ significantly on average from the 2.59% overall performance representing all world mutual funds over the same time period? If so, were these closed-end funds significantly better or significantly worse? In your calculations, you may assume that the overall performance is measured without randomness.

b. Do the rates of return of these closed-end world income funds, as a group, differ significantly on average from the −26.83% overall performance representing all emerging markets' mutual funds over the same time period? If so, were these closed-end funds significantly better or significantly worse? In your calculations, you may assume that the overall performance is measured without randomness.

TABLE 10.7.4 Performance of Closed-End World Income Funds: One-Year Market Return

Fund	Return (%)
ACM Mgd $-x	−27.7
Alliance Wld $	−17.1
Alliance Wld $ 2	−27.0
BlckRk North Am -x	3.9
Dreyfus Str Govt	4.0
Emer Mkts Float	−19.7
Emer Mkts Inc -x	−18.4
Emer Mkts Inc II -x	−16.9
First Aust Prime -x	−5.3
First Commonwlth -x	−3.5
Global HI Inc $	−10.7
Global Income Fund -x	−17.3
Global Partners -x	−16.7
Kleinwort Aust	−5.4
Morg St Em Debt -x	−24.9
Morgan St Glbl -x	−27.3
Salomon SBG -x	−0.6
Salomon SBW -x	−18.9
Scudder Glbl High Inc -x	−53.8
Strategic GI Inc	5.8
Templeton Em Inc	−12.8
Templtn GI Govt	−1.1
Templtn Glbl Inc	2.2
Worldwide $Vest -x	−48.2

Source: From "Quarterly Closed-End Funds Review," *Wall Street Journal*, Jan. 7, 1999, p. R14. Overall performance measures are from "Mutual-Fund Performance Yardsticks," p. R3.

37. Your broker achieved a rate of return of 18.3% on your portfolio last year. For a sample of 25 other brokers in the area, according to a recent news article, the average rate of return was 15.2% with a standard deviation of 3.2% (as percentage points).

a. To test whether your broker significantly outperformed this group, identify the idealized population and the hypotheses being tested. In particular, are you testing against a mean or against a new observation?

b. Find the standard error for prediction.

c. Find the two-sided 95% prediction interval for a new observation.
d. Did your broker outperform this group?
e. Did your broker significantly outperform this group?
f. Find the t-value and the p-value (as either $p>0.05$, $p<0.05$, $p<0.01$, or $p<0.001$) for this two-sided test.

38. Last year you received an average of 129.2 complaints (i e, individual items to be fixed under warranty) per new car sold, with a standard deviation of 42.1 complaints based on 3,834 new cars sold. This year you have set up a quality assurance program to fix some of these problems before the car is delivered. So far this year, you have had an average of just 93.4 complaints per new car sold with a standard deviation of 37.7, based on 74 cars sold so far.
a. To see if your new quality assurance program is working, what hypothesis testing method would you use?
b. Identify the populations, samples, and hypotheses.
c. Perform a two-sided test at the 5% level and report the results.

39. Why do firms change ownership? One possible reason for acquisitions is that the new owners expect to be able to manage the operations more efficiently than the current management. This theory leads to testable hypotheses. For example, it predicts that productivity should increase following a takeover and also that firms changing ownership should have lower productivity than firms in general. A study of this situation examined the productivity year by year for some firms that changed ownership and other firms that did not change owners.[26] In particular, they reported

These numbers display a very clear pattern. Plants that changed owners … tended to be less efficient … than nonchangers.… But the differences … [after the change] were declining in magnitude.… This signifies that the productivity of … changers relative to that of … nonchangers was both low and declining before the ownership change, and increasing (albeit still low) after the ownership change. With one exception, all of the productivity differences are highly statistically significant.

a. In the last line of the preceding quote, explain what is implied by "highly statistically significant."
b. Consider the comparison of average productivity of firms that changed ownership (at the time of the change) to average productivity of firms that did not change ownership. Identify all elements of this hypothesis testing situation, in particular: the hypotheses, the sample data, the type of test used, and the assumptions being made.
c. One result they reported was "at the time of ownership change, productivity level was 3.9% lower as compared to plants that did not change ownership. The t-statistic is 9.10." Perform a hypothesis test based on this information and state your conclusion. You may assume reasonably large samples.
d. Why have they gone to the trouble of doing statistical hypothesis tests? What have they gained over and above simply observing and describing the productivity differences in their data?

40.* Stress levels were recorded during a true answer and a false answer given by each of six people in a study of lie-detecting equipment, based on the idea that the stress involved in telling a lie can be measured. The results are shown in Table 10.7.5.
a. Was everyone's stress level higher during a false answer than during a true answer?
b. Find the average stress levels for true and for false answers. Find the average change in stress level (false minus true).
c. Find the appropriate standard error for the average difference. In particular, is this a paired or an unpaired situation?
d. Find the 95% two-sided confidence interval for the mean difference in stress level.
e. Test to see if the average stress levels are significantly different. If they are significantly different, are they significantly higher or lower when a false answer is given?
f. Write a paragraph interpreting the results of this test. In particular, is this a conclusion about these six people or about some other group? Also, how can you find a significant difference when some individuals had higher stress and some had lower stress for the false answer?

41. A group of experts has rated your winery's two best varietals. Ratings are on a scale from 1 to 20, with higher numbers being better. The results are shown in Table 10.7.6.
a. Is this a paired or unpaired situation? Why?
b. Find the average rating for each varietal and the average difference in ratings (Chardonnay minus Cabernet Sauvignon).
c. Find the appropriate standard error for the average difference.
d. Find the 95% two-sided confidence interval for the mean difference in rating.

TABLE 10.7.5 Vocal Stress Level

Person	True Answer	False Answer
1	12.8	13.1
2	8.5	9.6
3	3.4	4.8
4	5.0	4.6
5	10.1	11.0
6	11.2	12.1

TABLE 10.7.6 Wine-Tasting Scores

Expert	Chardonnay	Cabernet Sauvignon	Expert	Chardonnay	Cabernet Sauvignon
1	17.8	16.6	6	19.9	18.8
2	18.6	19.9	7	17.1	18.9
3	19.5	17.2	8	17.3	19.5
4	18.3	19.0	9	18.0	16.2
5	19.8	19.7	10	19.8	18.6

TABLE 10.7.7 Days Until Failure

Your Products	Competitor's
1.0	0.2
8.9	2.8
1.2	1.7
10.3	7.2
4.9	2.2
1.8	2.5
3.1	2.6
3.6	2.0
2.1	0.5
2.9	2.3
8.6	1.9
5.3	1.2
	6.6
	0.5
	1.2

e. Test to see if the average ratings are significantly different. If they are significantly different, which varietal is superior?

f. Write a paragraph interpreting the results of this test.

42*. To understand your competitive position, you have examined the reliability of your product as well as the reliability of your closest competitor's product. You have subjected each product to abuse that represents about a year's worth of wear-and-tear per day. Table 10.7.7 shows the data indicating how long each item lasted.

a. Find the average time to failure for your and your competitor's products. Find the average difference (yours minus your competitor's).

b. Find the appropriate standard error for this average difference. In particular, is this a paired or an unpaired situation? Why?

c. Find the two-sided 99% confidence interval for the mean difference in reliability.

d. Test at the 1% level if there is a significant difference in reliability between your products and your competitor's at this test level.

e. Find the p-value for the difference in reliability (as either $p>0.05$, $p<0.05$, $p<0.01$, or $p<0.001$) and state whether or not the result is significant at the conventional test level.

f. Write a brief paragraph, complete with footnote(s) that might be used in an advertising brochure showing off your products.

43. Child care is one of life's necessities for working parents. Monthly rates per child at a sample of family day care centers in the North Seattle area are shown in Table 10.7.8. The Laurelhurst area is considered to be a highly desirable neighborhood, and real estate prices are higher in this area. Perform a one-sided hypothesis test at the 5% level to see if day care prices are also higher in the Laurelhurst area.

44. An advertising study interviewed six randomly selected people in each of two cities, recording each person's level of preference for a new product (Table 10.7.9).

a. Is this a paired or an unpaired two-sample problem?

b. Find the average preference level for each city.

c. Find the standard error of the difference between these average preference levels. (Note that these are small samples.)

d. Find the 95% two-sided confidence interval for the mean difference in preference between these two cities (Green Bay minus Milwaukee).

e. Test whether the apparent difference in preference is significant at the 5% test level.

45. There are two manufacturing processes, old and new, that produce the same product. The defect rate has been measured for a number of days for each process, resulting in the following summaries (Table 10.7.10).

a. By how much would we estimate that the defect rate would improve if we switched from the old to the new process?

TABLE 10.7.8 Monthly Day Care Rates in North Seattle[a]

Laurelhurst Area	Non-Laurelhurst Area
$400	$500
625	425
440	300
550	350
600	550
500	475
	325
	350
	350

[a]*I am grateful to Ms. Colleen Walker for providing this data set.*

TABLE 10.7.9 Preference Levels for Six Individuals in Each of Two Cities

Milwaukee	Green Bay
3	4
2	5
1	4
1	3
3	2
2	4

TABLE 10.7.10 Defect Rate Summaries for Two Manufacturing Processes

	Old	New
Average defect rate	0.047	0.023
Standard deviation	0.068	0.050
Sample size (days)	50	44

b. What is the standard error of your answer to part a?
c. Your firm is interested in switching to the new process only if it can be demonstrated convincingly that the new process improves quality. State the null and research hypotheses for this situation.
d. Find the appropriate one-sided 95% confidence interval for the (population) long-term reduction in the defect rate.

TABLE 10.7.11 Supplier Quality

Custom Cases Corp.	International Plastics, Inc.
54.3	93.6
58.8	69.7
77.8	87.7
81.1	96.0
54.2	82.2
78.3	

e. Is the improvement (as estimated in part a) statistically significant?

46. To help you decide which of your two current suppliers deserves the larger contract next year, you have rated a random sample of plastic cases from each one. The data are a composite of several measurements, with higher numbers indicating higher quality (Table 10.7.11).
a. Find the average quality for each supplier.
b. Find the standard deviation of quality for each supplier.
c. Find the average difference in quality (International minus Custom) and its standard error.
d. Find the two-sided 95% confidence interval for the quality difference.
e. Is there a significant difference in quality? How do you know?

47. Consider the weights for two samples of candy bars, before and after intervention, from Table 5.5.4.
a. Is this a paired or an unpaired situation?
b. Find the 95% confidence interval for the population mean difference in weight per candy bar (after minus before).
c. Did intervention produce a significant change in weight? How do you know?

48. Your Detroit division produced 135 defective parts out of the total production of 983 last week. The Kansas City division produced 104 defectives out of 1,085 produced during the same time period.
a. Find the percent defective for each division and compare them.
b. Find the difference between these two percentages (Detroit minus Kansas City) and interpret it.
c. Find the standard error for this difference using the large-sample formula.
d. Find the 95% confidence interval for the difference.
e. Test to see if these two divisions differ significantly in terms of quality of production, based on the defect rate.

49. You are analyzing the results of a consumer survey of a product, rated on a scale from 1 to 10. For the 130 consumers who described themselves as "outgoing," the average rating was 8.36, and the standard deviation

was 1.82. For the 218 "shy" consumers, the average was 8.78, and the standard deviation was 0.91.
   a. Test to see if there is a significant difference between the ratings of outgoing and shy consumers.
   b. Report the test results using *p*-value notation (as either $p>0.05$, $p<0.05$, $p<0.01$, or $p<0.001$).

50. Repeat the previous problem for a different product. For 142 outgoing consumers, the average rating was 7.28, and the standard deviation was 2.18. For 277 shy consumers, the average rating was 8.78, and the standard deviation was 1.32.
51. Repeat problem 49 for yet another product. For 158 outgoing consumers, the average rating was 7.93, and the standard deviation was 2.03. For 224 shy consumers, the average rating was 8.11, and the standard deviation was 1.55.
52. A cup of coffee is found to have only 72.8 milligrams of caffeine. Test (at the 5% level) whether the beans used could have come from the same population as those that generated the data in problem 47 of Chapter 9.

24. P. B. Voos, "Managerial Perceptions of the Economic Impact of Labor Relations Programs," *Industrial and Labor Relations Review* 40 (1987), pp. 195–208.
25. D. Seligman, "Sick in Washington," *Fortune*, March 28, 1988, p. 155.
26. F. Lichtenberg, "Productivity Improvements from Changes in Ownership," *Mergers & Acquisitions* 23 (1988), pp. 48–50.

Database Exercises

Refer to the employee database in Appendix A. View this data set as a random sample from a much larger population of employees.

1.* Is the average annual salary significantly different from $40,000?
2. You would like to claim that the population has significantly more than five years of experience, on average. Can you support this claim?
3. Test to see if the gender ratio differs significantly from 50%.
4. Test to see if the population mean annual salary for men differs from that for women.
5. Test to see if the population mean age for men differs from that for women.
6. Test to see if the average annual salary for training level A differs significantly from that for levels B and C combined.
7. Test to see if the population mean age for training level A differs from that for levels B and C combined.

Projects

1. Identify a decision process within your work or business interests that could be resolved based on data.
   a. Describe the null and research hypotheses.
   b. Compute (or use an educated guess for) an appropriate estimate and its standard error.
   c. Find a confidence interval.
   d. Test the hypothesis.
   e. Interpret and explain your results.
2. Find a news item (from the Internet, a newspaper, a magazine, radio, or television) that reaches a conclusion based on data.
   a. Identify the null and research hypotheses.
   b. Identify the population and the sample, to the extent that you can from the information given. Was any important information omitted?
   c. What was the result of their hypothesis test?
   d. Is their conclusion a weak one or a strong one?
   e. Discuss and interpret their claims.

Case

So Many Ads, So Little Time

It is almost decision time, and the stakes are huge. With astronomical TV advertising costs per minute of airtime, it has been worthwhile to do some preliminary work so that nothing is wasted. In particular, you have been helping manage an effort to produce 22 ads for a personal hygiene product, even though only just a few will ever actually be shown to the general public. They have all been tested and ranked using the responses of representative consumers who were each randomly selected and assigned to view one ad, answering questions before and after. A composite score from 0 to 10 points, representing both recall and persuasion, has been produced for each consumer in the sample.

At your firm, the ads traditionally have been ranked using the average composite results, and the highest have run on nationwide TV. Recently, however, statistical hypothesis testing has been used to make sure that the ad or ads to be run are significantly better than a minimum score of 3.5 points.

Everything looks straightforward this time, with the two best ads scoring significantly above the minimum. The decision meeting should be straightforward, with Country Picnic the favorite for the most airtime and Coffee Break as an alternate. Following are the summaries, sorted in descending order by average composite score. The number of consumers viewing the ad is *n*. The *p*-values are from one-sided hypothesis tests against the reference value 3.5, computed separately for each ad.

Ad	*n*	*avg*	*stDev*	*stdErr*	*t*	*p*
Country Picnic	49	3.95	0.789	0.113	3.985	0.0001
Coffee Break	51	3.70	0.744	0.104	1.921	0.0302
Anniversary	51	3.66	0.934	0.131	1.214	0.1153
Ocean Breeze	49	3.63	0.729	0.104	1.255	0.1078
Friends at Play	56	3.62	0.896	0.120	0.969	0.1683
Tennis Match	56	3.60	0.734	0.098	1.037	0.1521
Walking Together	51	3.57	0.774	0.108	0.687	0.2476
Swimming Pool	52	3.56	0.833	0.116	0.532	0.2984
Shopping	49	3.54	0.884	0.126	0.355	0.3619
Jogging	47	3.54	0.690	0.101	0.423	0.3372
Family Scene	54	3.54	0.740	0.101	0.404	0.3438
Mountain Retreat	49	3.53	0.815	0.116	0.298	0.3836
Cool & Comfortable	52	3.52	0.780	0.108	0.195	0.4229

Coffee Together	53	3.52	0.836	0.115	0.148	0.4415
City Landscape	47	3.51	0.756	0.110	0.058	0.4770
Friends at Work	53	3.50	0.674	0.093	0.020	0.4919
Sailing	48	3.49	0.783	0.113	−0.055	0.5219
Desert Oasis	55	3.48	0.716	0.097	−0.226	0.5890
Birthday Party	50	3.48	0.886	0.125	−0.175	0.5693
Weekend Brunch	53	3.45	0.817	0.112	−0.437	0.6681
Home from Work	55	3.35	0.792	0.107	−1.430	0.9207
Windy	47	3.34	0.678	0.099	−1.593	0.9410

Thinking it over, you have some second thoughts. Because you want to really understand what the decision is based on, and because you remember material about errors in hypothesis testing from a course taken long ago, you wonder. The probability of a type I error is 0.05, so you expect to find about one ad in 20 to be significantly good even if it is not. That says that sometimes none would be significant, yet other times more than one could reasonably be significant.

Your speculation continues: Could it be that decisions are being made on the basis of pure randomness? Could it be that consumers, on average, rate these ads equally good? Could it be that all you have here is the randomness of the particular consumers who were chosen for each ad?

You decide to run a computer simulation model, setting the population mean score for all ads to exactly 3.5. Hitting the recalculation button on the spreadsheet 10 times, you observe that three times no ads are significant, 5 times one ad is significant, once two ads are, and once three ads are significant. Usually, the significant ads are different each time. Even more troubling, the random simulated results look a lot like the real ones that are about to be used to make real decisions.

Discussion Questions

1. Choose two ads, one that is significant and one that is not. Verify significance based on the average, standard error, and n, to make sure that they are correct. Is it appropriate to use one-sided tests here?
2. If the type I error is supposed to be controlled at 5%, how is it that in the computer simulation model, type I errors occurred 70% of the time?
3. Could it reasonably be that no ads are worthwhile, in a study for which 2 of 22 are significant?
4. What is your interpretation of the effectiveness of the ads in this study? What would you recommend in this situation?