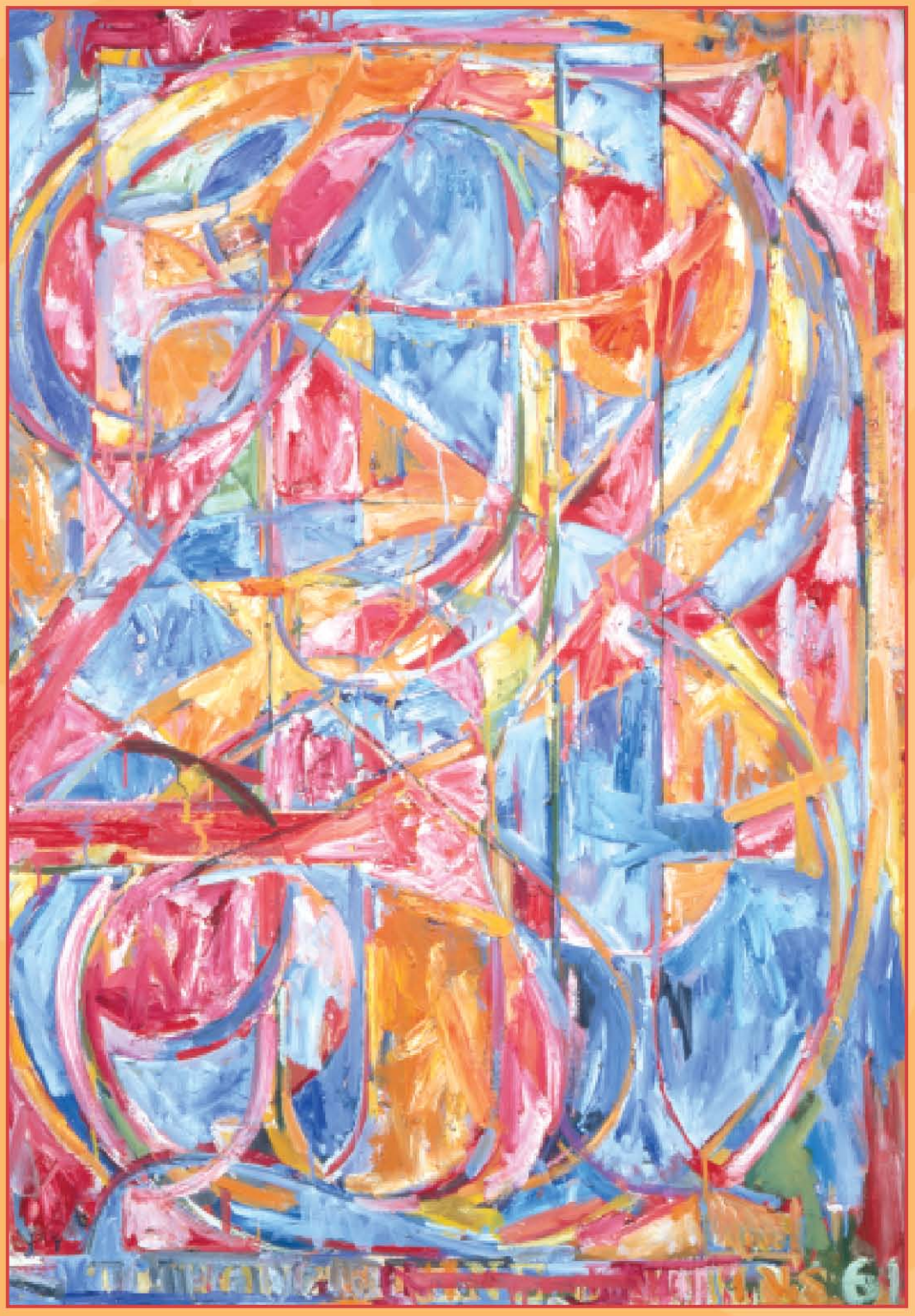
INTRODUCTION TO THE PRACTICE OF STATISTICS



SIXTH EDITION

MOORE
McCABE
CRAIG

Authors' note about the cover

Introduction to the Practice of Statistics emphasizes the use of graphical and numerical summaries to understand data. The front cover shows a painting entitled *0 to 9,* by the American artist Jasper Johns in 1961. In this work, the structure of the painting is determined by number sequence, just as our graphical summaries are determined by the numerical calculations that we perform when we analyze data. Can you find all of the digits in the painting?

Introduction to the Practice of Statistics

Senior Publisher:	CRAIG BLEYER
Publisher:	RUTH BARUTH
Development Editors:	SHONA BURKE, ANNE SCANLAN-ROHRER
Senior Media Editor:	ROLAND CHEYNEY
Assistant Editor:	BRIAN TEDESCO
Editorial Assistant:	KATRINA WILHELM
Marketing Coordinator:	DAVE QUINN
Photo Editor:	CECILIA VARAS
Photo Researcher:	ELYSE RIEDER
Cover and Text Designer:	VICKI TOMASELLI
Senior Project Editor:	MARY LOUISE BYRD
Illustrations:	INTEGRE TECHNICAL PUBLISHING CO.
Production Manager:	JULIA DE ROSA
Composition:	INTEGRE TECHNICAL PUBLISHING CO.
Printing and Binding:	RR DONNELLEY

Library of Congress Control Number: 2007938575

ISBN-13: 978-1-4292-1623-4
ISBN-10: 1-4292-1623-9 (Extended Version, Casebound)

ISBN-13: 978-1-4292-1622-7
ISBN-10: 1-4292-1622-0 (Casebound)

ISBN-13: 978-1-4292-1621-0
ISBN-10: 1-4292-1621-2 (Paperback)

Printed in the United States of America

First printing

W. H. Freeman and Company
41 Madison Avenue
New York, NY 10010
Houndmills, Basingstoke RG21 6XS, England
www.whfreeman.com

Introduction to the Practice of Statistics

SIXTH EDITION

DAVID S. MOORE

GEORGE P. McCABE

BRUCE A. CRAIG

Purdue University

W. H. Freeman and Company
New York

a 95% confidence interval, we know that the probability that the interval we compute will cover the parameter is 0.95. That's the meaning of 95% confidence. If we use several such intervals, however, our confidence that *all* of them give correct results is less than 95%. Suppose we take independent samples each month for five months and report a 95% confidence interval for each set of data.

(a) What is the probability that all five intervals cover the true means? This probability (expressed as a percent) is our overall confidence level for the five simultaneous statements.

(b) What is the probability that at least four of the five intervals cover the true means?

6.34 Telemarketing wages. An advertisement in the student newspaper asks you to consider working for a telemarketing company. The ad states, "Earn between $500 and $1000 per week." Do you think that the ad is describing a confidence interval? Explain your answer.

6.35 Like your job? A Gallup Poll asked working adults about their job satisfaction. One question was "All in all, which best describes how you feel about your job?" The possible answers were "love job," "like job," "dislike job," and "hate job." Fifty-nine percent of the sample responded that they liked their job. Material provided with the results of the poll noted:

Results are based on telephone interviews with 1,001 national adults, aged 18 and older, conducted Aug. 8–11, 2005. For results based on the total sample of national adults, one can say with 95% confidence that the maximum margin of sampling error is ±3 percentage points.[11]

The Gallup Poll uses a complex multistage sample design, but the sample percent has approximately a Normal sampling distribution.

(a) The announced poll result was 59% ± 3%. Can we be certain that the true population percent falls in this interval?

(b) Explain to someone who knows no statistics what the announced result 59% ± 3% means.

(c) This confidence interval has the same form we have met earlier:

$$\text{estimate} \pm z^*\sigma_{\text{estimate}}$$

What is the standard deviation σ_{estimate} of the estimated percent?

(d) Does the announced margin of error include errors due to practical problems such as undercoverage and nonresponse?

6.2 Tests of Significance

The confidence interval is appropriate when our goal is to estimate population parameters. The second common type of inference is directed at a quite different goal: to assess the evidence provided by the data in favor of some claim about the population parameters.

The reasoning of significance tests

A significance test is a formal procedure for comparing observed data with a hypothesis whose truth we want to assess. The hypothesis is a statement about the population parameters. The results of a test are expressed in terms of a probability that measures how well the data and the hypothesis agree. We use the following examples to illustrate these concepts.

EXAMPLE

6.8 Debt levels of private and public college borrowers. One purpose of the National Student Loan Survey described in Example 6.4 (page 361) is to compare the debt of different subgroups of students. For example, the 525 borrowers who last attended a private four-year college had a mean debt of $21,200, while those who last attended a public four-year college had a mean debt of $17,100. The difference of $4100 is fairly large, but we know that these

numbers are estimates of the true means. If we took different samples, we would get different estimates. Can we conclude from these data that the average debt of borrowers who attended a private college is different than the average debt of borrowers who attended a public college?

One way to answer this question is to compute the probability of obtaining a difference as large or larger than the observed $4100 assuming that, in fact, there is no difference in the true means. This probability is 0.17. Because this probability is not particularly small, we conclude that observing a difference of $4100 is not very surprising when the true means are equal. The data do not provide evidence for us to conclude that the mean debts for private four-year borrowers and public four-year borrowers are different.

Here is an example with a different conclusion.

EXAMPLE

6.9 Change in average debt levels between 1997 and 2002. Another purpose of the National Student Loan Survey is to look for changes over time. For example, in 1997, the survey found that the mean debt for undergraduate study was $11,400. How does this compare with the value of $18,900 in the 2002 study? The difference is $7500. As we learned in the previous example, an observed difference in means is not necessarily sufficient for us to conclude that the true means are different. Do the data provide evidence that there is an increase in borrowing? Again, we answer this question with a probability calculated under the assumption that there is *no difference in the true means*. The probability is 0.00004 of observing an increase in mean debt that is $7500 or more when there really is no difference. Because this probability is so small, we have sufficient evidence in the data to conclude that there has been a change in borrowing between 1997 and 2002.

What are the key steps in these examples?

- We started each with a question about the difference between two mean debts. In Example 6.8, we compare private four-year borrowers with public four-year borrowers. In Example 6.9, we compare borrowers in 2002 with borrowers in 1997. In both cases, we ask whether or not the data are compatible with no difference, that is, a difference of $0.
- Next we compared the data, $4100 in the first case and $7500 in the second, with the value that comes from the question, $0.
- The results of the comparisons are probabilities, 0.17 in the first case and 0.00004 in the second.

The 0.17 probability is not particularly small, so we have no evidence to question the possibility that the true difference is zero. In the second case, however, the probability is quite small. Something that happens with probability 0.00004 occurs only about 4 times out of 100,000. In this case we have two possible explanations:

1. we have observed something that is very unusual, or
2. the assumption that underlies the calculation, no difference in mean debt, is not true.

Because this probability is so small, we prefer the second conclusion: there has been a change in the mean debt between 1997 and 2002.

The probabilities in Examples 6.8 and 6.9 are measures of the compatibility of the data (a difference in means of \$4100 and \$7500) with the *null hypothesis* that there is no difference in the true means. Figures 6.7 and 6.8 compare the two results graphically. For each a Normal curve centered at 0 is the sampling distribution. You can see that we are not particularly surprised to observe the difference \$4100 in Figure 6.7, but the difference \$7500 in Figure 6.8 is clearly an unusual observation. We will now consider some of the formal aspects of significance testing.

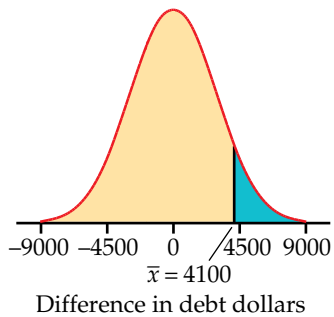


FIGURE 6.7 Comparison of the sample mean in Example 6.8 relative to the null hypothesized value 0.

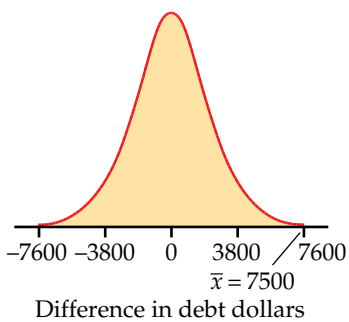


FIGURE 6.8 Comparison of the sample mean in Example 6.9 relative to the null hypothesized value 0.

Stating hypotheses

In Examples 6.8 and 6.9, we asked whether the difference in the observed means is reasonable if, in fact, there is no difference in the true means. To answer this, we begin by supposing that the statement following the "if" in the previous sentence is true. In other words, we suppose that the true difference is \$0. We then ask whether the data provide evidence against the supposition we have made. If so, we have evidence in favor of an effect (the means are different) we are seeking. The first step in a test of significance is to state a claim that we will try to find evidence *against*.

NULL HYPOTHESIS

The statement being tested in a test of significance is called the **null hypothesis.** The test of significance is designed to assess the strength of the evidence against the null hypothesis. Usually the null hypothesis is a statement of "no effect" or "no difference."

We abbreviate "null hypothesis" as H_0. A null hypothesis is a statement about the population parameters. For example, our null hypothesis for Example 6.8 is

H_0: there is no difference in the true means

Note that the null hypothesis refers to the *true* means for all borrowers from either a four-year private or public college, including those for whom we do not have data.

alternative hypothesis

It is convenient also to give a name to the statement we hope or suspect is true instead of H_0. This is called the **alternative hypothesis** and is abbreviated as H_a. In Example 6.8, the alternative hypothesis states that the means are different. We write this as

H_a: the true means are not the same

Hypotheses always refer to some populations or a model, not to a particular outcome. For this reason, we must state H_0 and H_a in terms of population parameters.

one-sided or two-sided alternatives

Because H_a expresses the effect that we hope to find evidence *for,* we often begin with H_a and then set up H_0 as the statement that the hoped-for effect is not present. Stating H_a is often the more difficult task. It is not always clear, in particular, whether H_a should be **one-sided** or **two-sided,** which refers to whether a parameter differs from its null hypothesis value in a specific direction or in either direction.

The alternative hypothesis should express the hopes or suspicions we bring to the data. *It is cheating to first look at the data and then frame H_a to fit what the data show.* If you do not have a specific direction firmly in mind in advance, you must use a two-sided alternative. Moreover, some users of statistics argue that we should always use a two-sided alternative.

USE YOUR KNOWLEDGE

6.36 Food court survey. The food court at your dormitory has been redesigned. A survey is planned to determine whether or not students think that the new design is an improvement. Sampled students will respond on a seven-point scale with scores less than 4 favoring the old design and scores greater than 4 favoring the new design (to varying degrees). State the null and alternative hypotheses that provide a framework for examining whether or not the new design is an improvement.

6.37 DXA scanners. A dual-energy X-ray absorptiometry (DXA) scanner is used to measure bone mineral density for people who may be at risk

for osteoporosis. To ensure its accuracy, the company uses an object called a "phantom" that has known mineral density $\mu = 1.4$ grams per square centimeter. Once installed, the company scans the phantom 10 times and compares the sample mean reading $\bar{x}$ with the theoretical mean μ using a significance test. State the null and alternative hypotheses for this test.

Test statistics

We will learn the form of significance tests in a number of common situations. Here are some principles that apply to most tests and that help in understanding these tests:

- The test is based on a statistic that estimates the parameter that appears in the hypotheses. Usually this is the same estimate we would use in a confidence interval for the parameter. When H_0 is true, we expect the estimate to take a value near the parameter value specified by H_0.
- Values of the estimate far from the parameter value specified by H_0 give evidence against H_0. The alternative hypothesis determines which directions count against H_0.
- To assess how far the estimate is from the parameter, standardize the estimate. In many common situations the test statistic has the form

$$z = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard deviation of the estimate}}$$

Let's return to our student loan example.

EXAMPLE

6.10 Debt levels of private and public college borrowers: the hypotheses. In Example 6.8, the hypotheses are stated in terms of the difference in debt between borrowers who attended a private college and those who attended a public college:

H_0: there is no difference in the true means

H_a: there is a difference in the true means

Because H_a is two-sided, large values of both positive and negative differences count as evidence against the null hypothesis.

test statistic

A **test statistic** measures compatibility between the null hypothesis and the data. We use it for the probability calculation that we need for our test of significance. It is a random variable with a distribution that we know.

EXAMPLE

6.11 Debt levels of private and public college borrowers: the test statistic. In Example 6.8, we can state the null hypothesis as H_0: the true mean difference is 0. The estimate of the difference is \$4100. Using methods that we will discuss in detail later, we can determine that the standard deviation of the estimate is \$3000. For this problem the test statistic is

$$z = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard deviation of the estimate}}$$

For our data,

$$z = \frac{4100 - 0}{3000} = 1.37$$

We have observed a sample estimate that is about one and a third standard deviations away from the hypothesized value of the parameter. Because the sample sizes are sufficiently large for us to conclude that the distribution of the sample estimate is approximately Normal, the standardized test statistic z will have approximately the $N(0, 1)$ distribution.

LOOK BACK
Normal distribution, page 58

We will use facts about the Normal distribution in what follows.

P-values

If all test statistics were Normal, we could base our conclusions on the value of the z test statistic. In fact, the Supreme Court of the United States has said that "two or three standard deviations" ($z = 2$ or 3) is its criterion for rejecting H_0 (see Exercise 6.42 on page 381), and this is the criterion used in most applications involving the law. Because not all test statistics are Normal, we translate the value of test statistics into a common language, the language of probability.

A test of significance finds the probability of getting an outcome *as extreme or more extreme than the actually observed outcome.* "Extreme" means "far from what we would expect if H_0 were true." The direction or directions that count as "far from what we would expect" are determined by H_a and H_0.

In Example 6.8 we want to know if the debt of private college borrowers is different from the debt of public college borrowers. The difference we calculated based on our sample is \$4100, which corresponds to 1.37 standard deviations away from zero—that is, $z = 1.37$. Because we are using a two-sided alternative for this problem, the evidence against H_0 is measured by the probability that we observe a value of Z as extreme or more extreme than 1.37. More formally, this probability is

$$P(Z \le -1.37 \text{ or } Z \ge 1.37)$$

where Z has the standard Normal distribution $N(0, 1)$.

P-VALUE

The probability, assuming H_0 is true, that the test statistic would take a value as extreme or more extreme than that actually observed is called the ***P*-value** of the test. The smaller the *P*-value, the stronger the evidence against H_0 provided by the data.

The key to calculating the *P*-value is the sampling distribution of the test statistic. For the problems we consider in this chapter, we need only the standard Normal distribution for the test statistic z.

EXAMPLE

6.12 Debt levels of private and public college borrowers: the *P*-value. In Example 6.11 we found that the test statistic for testing

$$H_0: \text{the true mean difference is } 0$$

versus

$$H_a: \text{there is a difference in the true means}$$

is

$$z = \frac{4100 - 0}{3000} = 1.37$$

If H_0 is true, then z is a single observation from the standard Normal, $N(0, 1)$, distribution. Figure 6.9 illustrates this calculation. The *P*-value is the probability of observing a value of Z at least as extreme as the one that we observed, $z = 1.37$. From Table A, our table of standard Normal probabilities, we find

$$P(Z \geq 1.37) = 1 - 0.9147 = 0.0853$$

The probability for being extreme in the negative direction is the same:

$$P(Z \leq -1.37) = 0.0853$$

So the *P*-value is

$$P = 2P(Z \geq 1.37) = 2(0.0853) = 0.1706$$

This is the value that was reported on page 373. There is a 17% chance of observing a difference as extreme as the $4100 in our sample if the true population difference is zero. The *P*-value tells us that our outcome is not particularly extreme, so we conclude that the data do not provide evidence that would cause us to doubt the validity of the null hypothesis.

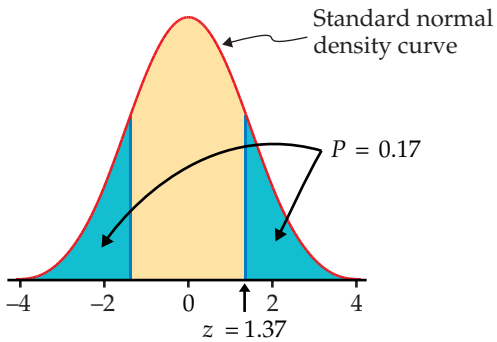


FIGURE 6.9 The *P*-value for Example 6.12. The *P*-value here is the probability (when H_0 is true) that $\bar{x}$ takes a value as extreme or more extreme than the actual observed value.

USE YOUR KNOWLEDGE

6.38 Normal curve and the *P*-value. A test statistic for a two-sided significance test for a population mean is $z = 2.7$. Sketch a standard Normal curve and mark this value of z on it. Find the *P*-value and shade the appropriate areas under the curve to illustrate your calculations.

6.39 More on the Normal curve and the *P*-value. A test statistic for a two-sided significance test for a population mean is $z = -1.2$. Sketch a standard Normal curve and mark this value of z on it. Find the *P*-value and shade the appropriate areas under the curve to illustrate your calculations.

Statistical significance

We started our discussion of the reasoning of significance tests with the statement of null and alternative hypotheses. We then learned that a test statistic is the tool used to examine the compatibility of the observed data with the null hypothesis. Finally, we translated the test statistic into a *P*-value to quantify the evidence against H_0. One important final step is needed: to state our conclusion.

We can compare the *P*-value we calculated with a fixed value that we regard as decisive. This amounts to announcing in advance how much evidence against H_0 we will require to reject H_0. The decisive value of P is called the **significance level.** It is commonly denoted by α. If we choose $\alpha = 0.05$, we are requiring that the data give evidence against H_0 so strong that it would happen no more than 5% of the time (1 time in 20) when H_0 is true. If we choose $\alpha = 0.01$, we are insisting on stronger evidence against H_0, evidence so strong that it would appear only 1% of the time (1 time in 100) if H_0 is in fact true.

significance level

STATISTICAL SIGNIFICANCE

If the *P*-value is as small or smaller than α, we say that the data are **statistically significant at level α.**

"Significant" in the statistical sense does not mean "important." The original meaning of the word is "signifying something." In statistics the term is used to indicate only that the evidence against the null hypothesis reached the standard set by α. Significance at level 0.01 is often expressed by the statement "The results were significant ($P < 0.01$)." Here P stands for the *P*-value. The *P*-value is more informative than a statement of significance because we can then assess significance at any level we choose. For example, a result with $P = 0.03$ is significant at the $\alpha = 0.05$ level but is not significant at the $\alpha = 0.01$ level.

A test of significance is a process for assessing the significance of the evidence provided by data against a null hypothesis. The four steps common to all tests of significance are as follows:

1. State the *null hypothesis* H_0 and the *alternative hypothesis* H_a. The test is designed to assess the strength of the evidence against H_0; H_a is the statement that we will accept if the evidence enables us to reject H_0.
2. Calculate the value of the *test statistic* on which the test will be based. This statistic usually measures how far the data are from H_0.
3. Find the *P-value* for the observed data. This is the probability, calculated assuming that H_0 is true, that the test statistic will weigh against H_0 at least as strongly as it does for these data.

4. State a conclusion. One way to do this is to choose a *significance level* α, how much evidence against H_0 you regard as decisive. If the P-value is less than or equal to α, you conclude that the alternative hypothesis is true; if it is greater than α, you conclude that the data do not provide sufficient evidence to reject the null hypothesis. Your conclusion is a sentence that summarizes what you have found by using a test of significance.

We will learn the details of many tests of significance in the following chapters. The proper test statistic is determined by the hypotheses and the data collection design. We use computer software or a calculator to find its numerical value and the P-value. The computer will not formulate your hypotheses for you, however. Nor will it decide if significance testing is appropriate or help you to interpret the P-value that it presents to you. The most difficult and important step is the last one: stating a conclusion.

EXAMPLE

6.13 Debt levels of private and public college borrowers: significance. In Example 6.12 we found that the P-value is 0.1706. There is a 17% chance of observing a difference as extreme as the \$4100 in our sample if the true population difference is zero. The P-value tells us that our outcome is not particularly extreme. We could report the result as "the data do not provide evidence that would cause us to conclude that there is a difference in student loan debt between private college borrowers and public college borrowers ($z = 1.37$, $P = 0.17$)."

If the P-value is small, we reject the null hypothesis. Here is an example.

EXAMPLE

6.14 Change in mean debt levels: significance. In Example 6.9 we found that the average debt has risen by \$7500 from 1997 to 2002. Since we would have a prior expectation that the debt would increase over this period because of rising costs of a college education, it is appropriate to use a one-sided alternative in this situation. So, our hypotheses are

$$H_0: \text{the true mean difference is } 0$$

versus

$$H_a: \text{the mean debt has increased between 1997 and 2002}$$

The standard deviation is \$1900 (again we defer details regarding this calculation), and the test statistic is

$$z = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard deviation of the estimate}}$$

$$z = \frac{7500 - 0}{1900}$$

$$= 3.95$$

Because only increases in debt count against the null hypothesis, the one-sided alternative leads to the calculation of the P-value using the upper tail

of the Normal distribution. The P-value is

$$P = P(Z \geq 3.95)$$
$$= 0.00004$$

The calculation is illustrated in Figure 6.10. There is about a 4 in 100,000 chance of observing a difference as large or larger than the $7500 in our sample if the true population difference is zero. This P-value tells us that our outcome is extremely rare. We conclude that the null hypothesis must be false. Here is one way to report the result: "The data clearly show that the mean debt for college loans has increased between 1997 and 2002 ($z = 3.95$, $P < 0.001$)."

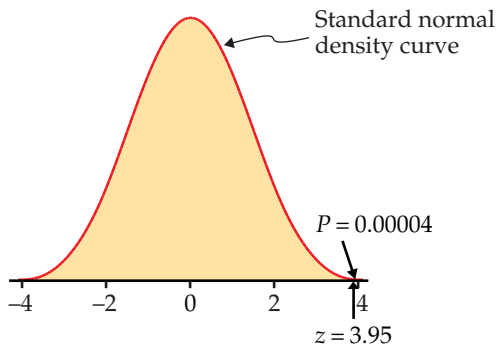


FIGURE 6.10 The P-value for Example 6.14. The P-value here is the probability (when H_0 is true) that $\bar{x}$ takes a value as large or larger than the actual observed value.

Note that the calculated P-value for this example is 0.00004 but we reported the result as $P < 0.001$. The value 0.001, 1 in 1000, is sufficiently small to force a clear rejection of H_0. Standard practice is to report very small P-values as simply less than 0.001.

USE YOUR KNOWLEDGE

6.40 Finding significant z-scores. Consider a significance test of the true mean based on an SRS of 30 observations from a Normal population. The alternative hypothesis is that the true mean is different from 1000. What values of the z statistic are statistically significant at the $\alpha = 0.05$ level?

6.41 More on finding significant z-scores. Consider a significance test of the true mean based on an SRS of 30 observations from a Normal population. The alternative hypothesis is that the true mean is larger than 1000. What values of the z statistic are statistically significant at the $\alpha = 0.05$ level?

6.42 The Supreme Court speaks. The Supreme Court has said that z-scores beyond $z^* = 2$ or 3 are generally convincing statistical evidence. For a two-sided test, what significance level corresponds to $z^* = 2$? To $z^* = 3$?

Tests for a population mean

Our discussion has focused on the reasoning of statistical tests, and we have outlined the key ideas for one type of procedure. Here is a summary. We want to test the hypothesis that a parameter has a specified value. This is the null hypothesis. For a test of a population mean μ, the null hypothesis is

$$H_0: \text{the true population mean is equal to } \mu_0$$

which often is expressed as

$$H_0: \mu = \mu_0$$

where μ_0 is the specified value of μ that we would like to examine.

The test is based on data summarized as an estimate of the parameter. For a population mean this is the sample mean $\overline{x}$. Our test statistic measures the difference between the sample estimate and the hypothesized parameter in terms of standard deviations of the test statistic:

$$z = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard deviation of the estimate}}$$

LOOK BACK

distribution of sample mean, page 339

central limit theorem, page 339

Recall from Chapter 5 that the standard deviation of $\overline{x}$ is $\sigma/\sqrt{n}$. Therefore, the test statistic is

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

Again recall from Chapter 5 that, if the population is Normal, then $\overline{x}$ will be Normal and z will have the standard Normal distribution when H_0 is true. By the central limit theorem both distributions will be approximately Normal when the sample size is large even if the population is not Normal.

Suppose we have calculated a test statistic $z = 1.7$. If the alternative is one-sided on the high side, then the P-value is the probability that a standard Normal random variable Z takes a value as large or larger than the observed 1.7. That is,

$$\begin{aligned} P &= P(Z \geq 1.7) \\ &= 1 - P(Z < 1.7) \\ &= 1 - 0.9554 \\ &= 0.0446 \end{aligned}$$

Similar reasoning applies when the alternative hypothesis states that the true μ lies below the hypothesized μ_0 (one-sided). When H_a states that μ is simply unequal to μ_0 (two-sided), values of z away from zero in either direction count against the null hypothesis. The P-value is the probability that a standard Normal Z is at least as far from zero as the observed z. Again, if the test statistic is $z = 1.7$, the two-sided P-value is the probability that $Z \leq -1.7$ or $Z \geq 1.7$. Because the standard Normal distribution is symmetric, we calculate this probability by finding $P(Z \geq 1.7)$ and *doubling* it:

$$\begin{aligned} P(Z \leq -1.7 \text{ or } Z \geq 1.7) &= 2P(Z \geq 1.7) \\ &= 2(1 - 0.9554) = 0.0892 \end{aligned}$$

We would make exactly the same calculation if we observed $z = -1.7$. It is the absolute value $|z|$ that matters, not whether z is positive or negative. Here is a statement of the test in general terms.

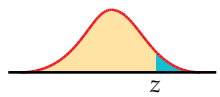
z TEST FOR A POPULATION MEAN

To test the hypothesis $H_0: \mu = \mu_0$ based on an SRS of size n from a population with unknown mean μ and known standard deviation σ, compute the test statistic

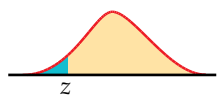
$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

In terms of a standard Normal random variable Z, the P-value for a test of H_0 against

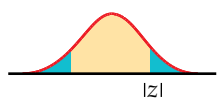
$H_a: \mu > \mu_0$ is $P(Z \geq z)$



$H_a: \mu < \mu_0$ is $P(Z \leq z)$



$H_a: \mu \neq \mu_0$ is $2P(Z \geq |z|)$



These P-values are exact if the population distribution is Normal and are approximately correct for large n in other cases.

EXAMPLE

6.15 Cholesterol level of sedentary female undergraduates. In 1999, it was reported that the mean serum cholesterol level for female undergraduates was 168 mg/dl with a standard deviation of 27 mg/dl. A recent study at Baylor University investigated the lipid levels in a cohort of sedentary university students.[12] The mean total cholesterol level among $n = 71$ females was $\bar{x} = 173.7$. Is this evidence that cholesterol levels of sedentary students differ from the previously reported average?

The null hypothesis is "no difference" from the published mean $\mu_0 = 168$. The alternative is two-sided because the researcher did not have a particular direction in mind before examining the data. So the hypotheses about the unknown mean μ of the sedentary population are

$$H_0: \mu = 168$$
$$H_a: \mu \neq 168$$

As usual in this chapter, we make the unrealistic assumption that the population standard deviation is known, in this case that sedentary female students have the same $\sigma = 27$ as the general population of female undergraduates. The z test requires that the 71 students in the sample are an SRS from the population of all sedentary female students. We check this assumption by asking how the data were produced. In this case, all participants were enrolled in a health class at Baylor, so there may be some concerns about whether the sample is an SRS. We will press on for now.

We compute the test statistic:

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{173.7 - 168}{27/\sqrt{71}} = 1.78$$

Figure 6.11 illustrates the P-value, which is the probability that a standard Normal variable Z takes a value at least 1.78 away from zero. From Table A we find that this probability is

$$P = 2P(Z \geq 1.78) = 2(1 - 0.9625) = 0.075$$

That is, more than 7% of the time an SRS of size 71 from the general undergraduate female population would have a mean cholesterol level at least as far from 168 as that of the sedentary sample. The observed $\overline{x} = 173.7$ is therefore not strong evidence that the sedentary female undergraduate population differs from the general female undergraduate population.

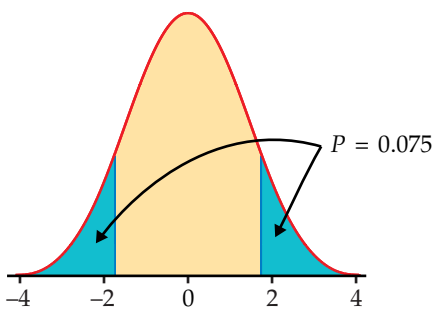


FIGURE 6.11 The P-value for the two-sided test in Example 6.15.

The data in Example 6.15 do *not* establish that the mean cholesterol level μ for the sedentary population is 168. We sought evidence that μ differed from 168 and failed to find convincing evidence. That is all we can say. No doubt the mean cholesterol level of the entire sedentary population is not exactly equal to 168. A large enough sample would give evidence of the difference, even if it is very small. Tests of significance assess the evidence *against* H_0. If the evidence is strong, we can confidently reject H_0 in favor of the alternative. Failing to find evidence against H_0 means only that the data are consistent with H_0, not that we have clear evidence that H_0 is true.

EXAMPLE

6.16 Significance test of the mean SATM score. In a discussion of SAT Mathematics (SATM) scores, someone comments: "Because only a minority of California high school students take the test, the scores overestimate the ability of typical high school seniors. I think that if all seniors took the test, the mean score would be no more than 450." You decided to test this claim (H_0) and gave the SAT to an SRS of 500 seniors from California (Example 6.3). These students had a mean SATM score of $\bar{x} = 461$. Is this good evidence against this claim? Because the claim states the mean is "no more than 450," the alternative hypothesis is one-sided. The hypotheses are

$$H_0: \mu = 450$$

$$H_a: \mu > 450$$

As we did in the discussion following Example 6.3, we assume that $\sigma = 100$. The z statistic is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{461 - 450}{100/\sqrt{500}} = 2.46$$

Because H_a is one-sided on the high side, large values of z count against H_0. From Table A, we find that the P-value is

$$P = P(Z \geq 2.46) = 1 - 0.9931 = 0.0069$$

Figure 6.12 illustrates this P-value. A mean score as large as that observed would occur fewer than seven times in 1000 samples if the population mean were 450. This is convincing evidence that the mean SATM score for all California high school seniors is higher than 450.

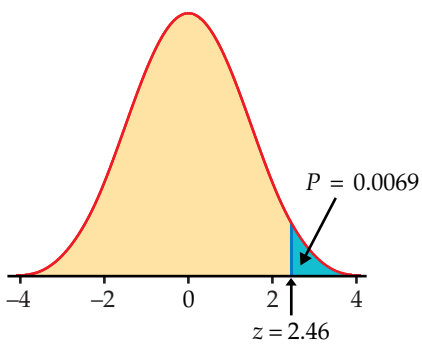


FIGURE 6.12 The P-value for the one-sided test in Example 6.16.

USE YOUR KNOWLEDGE

6.43 Computing the test statistic and *P*-value. You will perform a significance test of $H_0: \mu = 25$ based on an SRS of $n = 25$. Assume $\sigma = 5$.

(a) If $\bar{x} = 27$, what is the test statistic z?

(b) What is the P-value if H_a: $\mu > 25$?

(c) What is the P-value if H_a: $\mu \neq 25$?

6.44 Testing a random number generator. Statistical software has a "random number generator" that is supposed to produce numbers uniformly distributed between 0 to 1. If this is true, the numbers generated come from a population with $\mu = 0.5$. A command to generate 100 random numbers gives outcomes with mean $\bar{x} = 0.522$ and $s = 0.316$. Because the sample is reasonably large, take the population standard deviation also to be $\sigma = 0.316$. Do we have evidence that the mean of all numbers produced by this software is not 0.5?

Two-sided significance tests and confidence intervals

Recall the basic idea of a confidence interval, discussed in the first section of this chapter. We constructed an interval that would include the true value of μ with a specified probability C. Suppose we use a 95% confidence interval ($C = 0.95$). Then the values of μ that are not in our interval would seem to be incompatible with the data. This sounds like a significance test with $\alpha = 0.05$ (or 5%) as our standard for drawing a conclusion. The following examples demonstrate that this is correct.

EXAMPLE

6.17 Testing a pharmaceutical product. The Deely Laboratory analyzes specimens of a pharmaceutical product to determine the concentration of the active ingredient. Such chemical analyses are not perfectly precise. Repeated measurements on the same specimen will give slightly different results. The results of repeated measurements follow a Normal distribution quite closely. The analysis procedure has no bias, so that the mean μ of the population of all measurements is the true concentration in the specimen. The standard deviation of this distribution is a property of the analytical procedure and is known to be $\sigma = 0.0068$ grams per liter. The laboratory analyzes each specimen three times and reports the mean result.

The Deely Laboratory has been asked to evaluate the claim that the concentration of the active ingredient in a specimen is 0.86 grams per liter. The true concentration is the mean μ of the population of repeated analyses. The hypotheses are

$$H_0: \mu = 0.86$$

$$H_a: \mu \neq 0.86$$

The lab chooses the 1% level of significance, $\alpha = 0.01$.

Three analyses of one specimen give concentrations

$$0.8403 \quad 0.8363 \quad 0.8447$$

The sample mean of these readings is

$$\bar{x} = \frac{0.8403 + 0.8363 + 0.8447}{3} = 0.8404$$

The test statistic is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{0.8404 - 0.86}{0.0068/\sqrt{3}} = -4.99 \text{ standard deviations}$$

Because the alternative is two-sided, the P-value is

$$P = 2P(Z \geq |-4.99|) = 2P(Z \geq 4.99)$$

We cannot find this probability in Table A. The largest value of z in that table is 3.49. All that we can say from Table A is that P is less than $2P(Z \geq 3.49) = 2(1 - 0.9998) = 0.0004$. If we use the bottom row of Table D, we find that the largest value of z^* is 3.291, corresponding to a P-value of $1 - 0.999 = 0.001$. Software could be used to give an accurate value of the P-value. However, because the P-value is clearly less than the company's standard of 1%, we reject H_0.

Suppose we compute a 99% confidence interval for the same data.

EXAMPLE

6.18 99% confidence interval for the mean concentration. The 99% confidence interval for μ in Example 6.17 is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 0.8404 \pm 0.0101$$
$$= (0.8303, 0.8505)$$

The hypothesized value $\mu_0 = 0.86$ in Example 6.17 falls outside the confidence interval we computed in Example 6.18. We are therefore 99% confident that μ is *not* equal to 0.86, so we can reject

$$H_0: \mu = 0.86$$

at the 1% significance level. On the other hand, we cannot reject

$$H_0: \mu = 0.85$$

at the 1% level in favor of the two-sided alternative $H_a: \mu \neq 0.85$, because 0.85 lies inside the 99% confidence interval for μ. Figure 6.13 illustrates both cases.

The calculation in Example 6.17 for a 1% significance test is very similar to the calculation for a 99% confidence interval. In fact, a two-sided test at significance level α can be carried out directly from a confidence interval with confidence level $C = 1 - \alpha$.

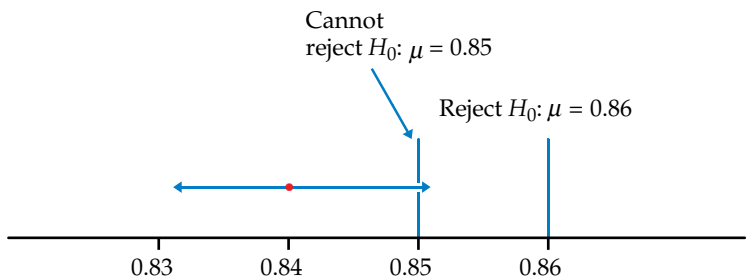


FIGURE 6.13 Values of μ falling outside a 99% confidence interval can be rejected at the 1% significance level; values falling inside the interval cannot be rejected.

TWO-SIDED SIGNIFICANCE TESTS AND CONFIDENCE INTERVALS

A level α two-sided significance test rejects a hypothesis $H_0: \mu = \mu_0$ exactly when the value μ_0 falls outside a level $1 - \alpha$ confidence interval for μ.

USE YOUR KNOWLEDGE

6.45 Two-sided significance tests and confidence intervals. The P-value for a two-sided test of the null hypothesis $H_0: \mu = 30$ is 0.08.

(a) Does the 95% confidence interval include the value 30? Explain.

(b) Does the 90% confidence interval include the value 30? Explain.

6.46 More on two-sided tests and confidence intervals. A 95% confidence interval for a population mean is (57, 65).

(a) Can you reject the null hypothesis that $\mu = 68$ at the 5% significance level? Explain.

(b) Can you reject the null hypothesis that $\mu = 62$ at the 5% significance level? Explain.

P-values versus fixed α

The observed result in Example 6.17 was $z = -4.99$. The conclusion that this result is significant at the 1% level does not tell the whole story. The observed z is far beyond the z corresponding to 1%, and the evidence against H_0 is far stronger than 1% significance suggests. The P-value

$$2P(Z \geq 4.99) = 0.0000006$$

gives a better sense of how strong the evidence is. *The P-value is the smallest level α at which the data are significant.* Knowing the P-value allows us to assess significance at any level.

EXAMPLE

6.19 Test of the mean SATM score: significance. In Example 6.16, we tested the hypotheses

$$H_0: \mu = 450$$

$$H_a: \mu > 450$$

concerning the mean SAT Mathematics score μ of California high school seniors. The test had the P-value $P = 0.0069$. This result is significant at the $\alpha = 0.01$ level because $0.0069 \leq 0.01$. It is not significant at the $\alpha = 0.005$ level, because the P-value is larger than 0.005. See Figure 6.14.

A P-value is more informative than a reject-or-not finding at a fixed significance level. But assessing significance at a fixed level α is easier, because no

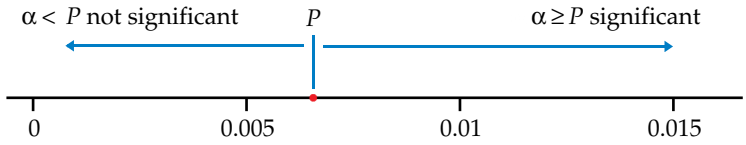


FIGURE 6.14 An outcome with *P*-value *P* is significant at all levels α at or above *P* and is not significant at smaller levels α.

probability calculation is required. You need only look up a number in a table. A value z^* with a specified area to its right under the standard Normal curve is called a **critical value** of the standard Normal distribution. Because the practice of statistics almost always employs computer software that calculates *P*-values automatically, the use of tables of critical values is becoming outdated. We include the usual tables of critical values (such as Table D) at the end of the book for learning purposes and to rescue students without good computing facilities. The tables can be used directly to carry out fixed α tests. They also allow us to approximate *P*-values quickly without a probability calculation. The following example illustrates the use of Table D to find an approximate *P*-value.

critical value

EXAMPLE

6.20 Debt levels of private and public college borrowers: assessing significance. In Example 6.11 we found the test statistic $z = 1.37$ for testing the null hypothesis that there was no difference in the mean debt between borrowers who attended a private college and those who attended a public college. The alternative was two-sided. Under the null hypothesis, z has a standard Normal distribution, and from the last row in Table D we can see that there is a 95% chance that z is between ± 1.96. Therefore, we reject H_0 in favor of H_a whenever z is outside this range. Since our calculated value is 1.37, we are within the range and we do not reject the null hypothesis at the 5% level of significance.

USE YOUR KNOWLEDGE

6.47 *P*-value and the significance level. The *P*-value for a significance test is 0.026.

(a) Do you reject the null hypothesis at level $\alpha = 0.05$?

(b) Do you reject the null hypothesis at level $\alpha = 0.01$?

(c) Explain your answers.

6.48 More on the *P*-value and the significance level. The *P*-value for a significance test is 0.074.

(a) Do you reject the null hypothesis at level $\alpha = 0.05$?

(b) Do you reject the null hypothesis at level $\alpha = 0.01$?

(c) Explain your answers.

6.49 One-sided and two-sided *P*-values. The *P*-value for a two-sided significance test is 0.06.

(a) State the *P*-values for the one-sided tests.

(b) What additional information do you need to properly assign these P-values to the $>$ and $<$ (one-sided) alternatives?

SECTION 6.2 Summary

A **test of significance** is intended to assess the evidence provided by data against a **null hypothesis H_0** in favor of an **alternative hypothesis H_a.**

The hypotheses are stated in terms of population parameters. Usually H_0 is a statement that no effect or no difference is present, and H_a says that there is an effect or difference, in a specific direction (**one-sided alternative**) or in either direction (**two-sided alternative**).

The test is based on a **test statistic.** The ***P*-value** is the probability, computed assuming that H_0 is true, that the test statistic will take a value at least as extreme as that actually observed. Small P-values indicate strong evidence against H_0. Calculating P-values requires knowledge of the sampling distribution of the test statistic when H_0 is true.

If the P-value is as small or smaller than a specified value α, the data are **statistically significant** at significance level α.

Significance tests for the hypothesis $H_0: \mu = \mu_0$ concerning the unknown mean μ of a population are based on the z **statistic:**

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

The z test assumes an SRS of size n, known population standard deviation σ, and either a Normal population or a large sample. P-values are computed from the Normal distribution (Table A). Fixed α tests use the table of **standard Normal critical values** (Table D).

SECTION 6.2 Exercises

For Exercises 6.36 and 6.37, see page 375; for Exercises 6.38 and 6.39, see pages 378 and 379; for Exercises 6.40 to 6.42, see page 381; for Exercises 6.43 and 6.44, see pages 385 and 386; for Exercises 6.45 and 6.46, see page 388; and for Exercises 6.47 to 6.49, see page 389.

6.50 What's wrong? Here are several situations where there is an incorrect application of the ideas presented in this section. Write a short paragraph explaining what is wrong in each situation and why it is wrong.

(a) A random sample of size 20 is taken from a population that is assumed to have a standard deviation of 12. The standard deviation of the sample mean is 12/20.

(b) A researcher tests the following null hypothesis: $H_0: \overline{x} = 10$.

(c) A study with $\overline{x} = 48$ reports statistical significance for $H_a: \mu > 54$.

(d) A researcher tests the hypothesis $H_0: \mu = 50$ and concludes that the population mean is equal to 50.

6.51 What's wrong? Here are several situations where there is an incorrect application of the ideas presented in this section. Write a short paragraph explaining what is wrong in each situation and why it is wrong.

(a) A significance test rejected the null hypothesis that the sample mean is equal to 1500.

(b) A change is made that should improve student satisfaction with the way grades are processed. The null hypothesis, that there is an improvement, is tested versus the alternative, that there is no change.

(c) A study summary says that the results are statistically significant and the P-value is 0.99.