6.1 Estimating with Confidence

The unbiasedness of an estimator concerns the center of its sampling distribution, but questions about variation are answered by looking at its spread. The central limit theorem says that if the entire population of SATM scores has mean μ and standard deviation σ, then in repeated SRSs of size 500, the sample mean $\bar{x}$ is approximately $N\left(\mu, \sigma/\sqrt{500}\right)$. Let us suppose that we know that the standard deviation σ of SATM scores in our California population is $\sigma = 100$. (We will see in the next chapter how to proceed when σ is not known. For now, we are more interested in statistical reasoning than in details of realistic methods.) This means that in repeated sampling the sample mean $\bar{x}$ has an approximately Normal distribution centered at the unknown population mean μ and a standard deviation of

$$\sigma_{\bar{x}} = \frac{100}{\sqrt{500}} = 4.5$$

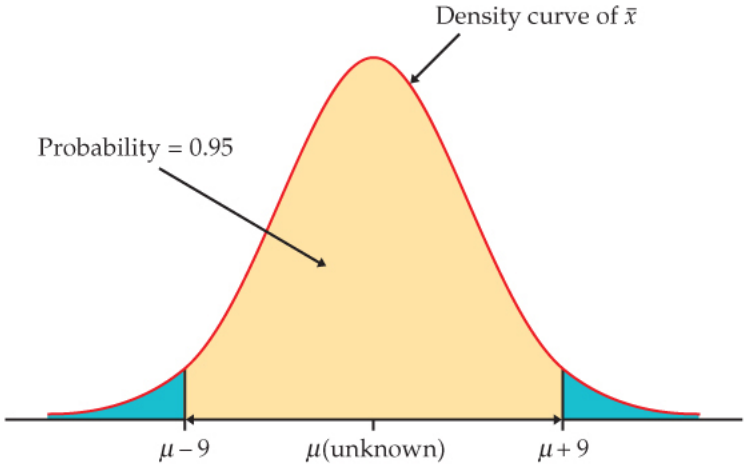


FIGURE 6.2 Distribution of the sample mean, Example 6.3. $\bar{x}$ lies within ±9 points of μ in 95% of all samples. This also means that μ is within ±9 points of $\bar{x}$ in those samples.

Now we are ready to proceed. Consider this line of thought, which is illustrated in Figure 6.2:

- The 68–95–99.7 rule says that the probability is about 0.95 that $\bar{x}$ will be within 9 points (that is, two standard deviations of $\bar{x}$) of the population mean score μ.
- To say that $\bar{x}$ lies within 9 points of μ is the same as saying that μ is within 9 points of $\bar{x}$.
- So about 95% of all samples will contain the true μ in the interval from $\bar{x} - 9$ to $\bar{x} + 9$.

We have simply restated a fact about the sampling distribution of $\bar{x}$. *The language of statistical inference uses this fact about what would happen in the long run to express our confidence in the results of any one sample.* Our sample gave $\bar{x} = 495$. We say that we are *95% confident* that the unknown mean score for all California seniors lies between

$$\bar{x} - 9 = 495 - 9 = 486$$

and

$$\bar{x} + 9 = 495 + 9 = 504$$

Be sure you understand the grounds for our confidence. There are only two possibilities for our SRS:

1. The interval between 486 and 504 contains the true μ.
2. The interval between 486 and 504 does not contain the true μ.

We cannot know whether our sample is one of the 95% for which the interval $\bar{x} \pm 9$ contains μ or one of the unlucky 5% for which it does not contain μ. The statement that we are 95% confident is shorthand for saying, "We arrived at these numbers by a method that gives correct results 95% of the time."

Confidence intervals

In the setting of Example 6.3, the interval of numbers between the values $\bar{x} \pm 9$ is called a *95% confidence interval* for μ. Like most confidence intervals we will discuss, this one has the form

$$\text{estimate} \pm \text{margin of error}$$

margin of error, p. 287

The estimate ($\bar{x} = 495$ in this case) is our guess for the value of the unknown parameter. The margin of error (9 here) reflects how accurate we believe our guess is, based on the variability of the estimate, and how confident we are that the procedure will produce an interval that will contain the true population mean μ.

Figure 6.3 illustrates the behavior of 95% confidence intervals in repeated sampling from a Normal distribution with mean μ. The center of each interval (marked by a dot) is at $\bar{x}$ and varies from sample to sample. The sampling distribution of $\bar{x}$ (also Normal) appears at the top of the figure to show the long-term pattern of this variation.

The 95% confidence intervals, $\bar{x} \pm$ margin of error, from 25 SRSs appear below the sampling distribution. The arrows on either side of the

dot ($\bar{x}$) span the confidence interval. All except one of the 25 intervals contain the true value of μ. In those intervals that contain μ, sometimes μ is near the middle of the interval and sometimes it is closer to one of the ends. This again reflects the variation of $\bar{x}$. In practice, we don't know the value of μ, but we have a method such that, in a very large number of samples, 95% of the confidence intervals will contain μ.

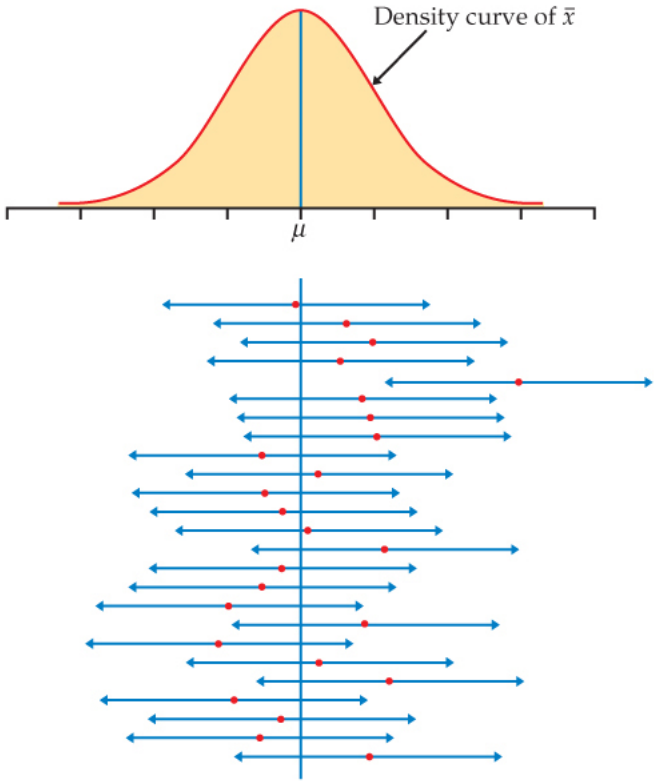


FIGURE 6.3 Twenty-five samples from the same population gave these 95% confidence intervals. In the long run, 95% of all samples give an interval that covers μ. The sampling distribution of $\bar{x}$ is shown at the top.

We can construct confidence intervals for many different parameters based on a variety of designs for data collection. We will learn the details of a number of these in later chapters. Two important things about a confidence interval are common to all settings:

1. It is an interval of the form (a, b), where a and b are numbers computed from the sample data.

confidence level

2. It has a property called a **confidence level** that gives the probability of producing an interval that contains the unknown parameter.

Users can choose the confidence level, but 95% is the standard for most situations. Occasionally, 90% or 99% is used. We use C to stand for the confidence level in decimal form. For example, a 95% confidence level corresponds to $C = 0.95$.

CONFIDENCE INTERVAL

A level C **confidence interval** for a parameter is an interval computed from sample data by a method that has probability C of producing an interval containing the true value of the parameter.

Confidence interval for a population mean

LOOK BACK

central limit theorem, p. 298

We now construct a level C confidence interval for the mean μ of a population when the data are an SRS of size n. The construction is based on the sampling distribution of the sample mean $\bar{x}$. This distribution is exactly $N\left(\mu, \sigma/\sqrt{n}\right)$ when the population has the $N(\mu, \sigma)$ distribution. The central limit theorem says that this same sampling distribution is approximately correct for large samples whenever the population mean and standard deviation are μ and σ. For now, we will assume we are in one of these two situations. We discuss what we mean by "large sample" after we briefly study these intervals.

Our construction of a 95% confidence interval for the mean SATM score began by noting that any Normal distribution has probability about 0.95 within ±2 standard deviations of its mean. To construct a level C confidence interval we first catch the central C area under a Normal curve. That is, we must find the number z^* such that any Normal distribution has probability C within $\pm z^*$ standard deviations of its mean.

Because all Normal distributions have the same standardized form, we can obtain everything we need from the standard Normal curve. Figure 6.4 shows how C and z^* are related. Values of z^* for many choices of C appear in the row labeled z^* at the bottom of Table D. Here are the most important entries from that row:

z^*	1.645	1.960	2.576
C	90%	95%	99%

Notice that for 95% confidence the value 2 obtained from the 68–95–99.7 rule is replaced with the more precise 1.96.

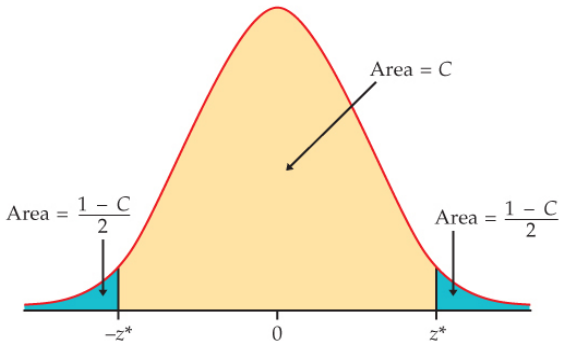


FIGURE 6.4 To construct a level C confidence interval, we must find the number z^*. The area between $-z^*$ and z^* under the standard Normal curve is C.

z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

As Figure 6.4 reminds us, any Normal curve has probability C between the point z^* standard deviations below the mean and the point z^* standard deviations above the mean. The sample mean $\bar{x}$ has the Normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ so there is probability C that $\bar{x}$ lies between

$$\mu - z * \frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \mu + z * \frac{\sigma}{\sqrt{n}}$$

This is exactly the same as saying that the unknown population mean μ lies between

$$\bar{x} - z * \frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \bar{x} + z * \frac{\sigma}{\sqrt{n}}$$

That is, there is probability C that the interval $\bar{x} \pm z * \sigma/\sqrt{n}$ contains μ. This is our confidence interval. The estimate of the unknown μ is $\bar{x}$, and the margin of error is $z * \sigma/\sqrt{n}$.

CONFIDENCE INTERVAL FOR A POPULATION MEAN

Choose an SRS of size n from a population having unknown mean μ and known standard deviation σ. The **margin of error** for a level C confidence interval for μ is

$$m = z * \frac{\sigma}{\sqrt{n}}$$

Here, z^* is the value on the standard Normal curve with area C between the critical points $-z^*$ and z^*. The level C **confidence interval** for μ is

$$\bar{x} \pm m$$

The confidence level of this interval is exactly C when the population distribution is Normal and is approximately C when n is large in other cases.

How confidence intervals behave

The margin of error $z^*\sigma/\sqrt{n}$ for the mean of a Normal population illustrates several important properties that are shared by all confidence intervals in common use. The user chooses the confidence level, and the margin of error follows from this choice.

Both high confidence and a small margin of error are desirable characteristics of a confidence interval. High confidence says that our method almost always gives correct answers. A small margin of error says that we have pinned down the parameter quite precisely.

Suppose that in planning a study you calculate the margin of error and decide that it is too large. Here are your choices to reduce it:

- Use a lower level of confidence (smaller C).
- Choose a larger sample size (larger n).
- Reduce σ.

For most problems, you would choose a confidence level of 90%, 95%, or 99%, so z^* will be 1.645, 1.960, or 2.576, respectively. Figure 6.4 shows that z^* will be smaller for lower confidence (smaller C). The bottom row of Table D also shows this. If n and σ are unchanged, a smaller z^* leads to a smaller margin of error.

How the confidence level affects the confidence interval. Suppose that for the college saving fund contribution data in Example 6.4 (page 350), we wanted 99% confidence. Table D tells us that for 99% confidence, $z^* = 2.576$. The margin of error for 99% confidence based on 1593 observations is

$$\begin{aligned} m &= z^* \frac{\sigma}{\sqrt{n}} \\ &= 2.576 \frac{1483}{\sqrt{1593}} \\ &= 95.71 \end{aligned}$$

and the 99% confidence interval is

$$\begin{aligned} \bar{x} \pm m &= 1768 \pm 96 \\ &= (1672, 1864) \end{aligned}$$

Requiring 99%, rather than 95%, confidence has increased the margin of error from 37 to 96. Figure 6.6 compares the two intervals.

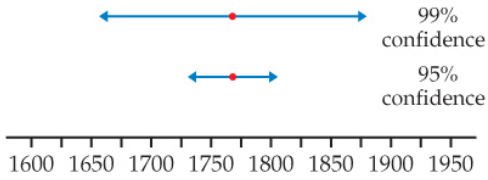


FIGURE 6.6 Confidence intervals, Examples 6.4 and 6.6. The larger the value of C, the wider the interval.

Choosing the sample size

A wise user of statistics never plans data collection without, at the same time, planning the inference. You can arrange to have both high confidence and a small margin of error. The margin of error of the confidence interval for a population mean is

$$m = z^* \frac{\sigma}{\sqrt{n}}$$

Notice once again that it is the size of the *sample* that determines the margin of error. The size of the *population* (as long as the population is much larger than the sample) does not influence the sample size we need.

To obtain a desired margin of error m, plug in the value of σ and the value of z^* for your desired confidence level, and solve for the sample size n. Here is the result.

SAMPLE SIZE FOR DESIRED MARGIN OF ERROR

The confidence interval for a population mean will have a specified margin of error m when the sample size is

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

Some cautions

We have already seen that small margins of error and high confidence can require large numbers of observations. You should also be keenly aware that *any formula for inference is correct only in specific circumstances*. If the government required statistical procedures to carry warning labels like those on drugs, most inference methods would have long labels. Our formula $\bar{x} \pm z^*\sigma/\sqrt{n}$ for estimating a population mean comes with the following list of warnings for the user:

- The data should be an SRS from the population. We are completely safe if we actually did a randomization and drew an SRS. We are not in great danger if the data can plausibly be thought of as independent observations from a population. That is the case in Examples 6.4 through 6.7, provided the undergraduates and parents can be considered one population.
- The formula is not correct for probability sampling designs more complex than an SRS. Correct methods for other designs are available. We will not discuss confidence intervals based on multistage or stratified samples (page 195). If you plan such samples, be sure that you (or your statistical consultant) know how to carry out the inference you desire.

- There is no correct method for inference from data haphazardly collected with bias of unknown size. Fancy formulas cannot rescue badly produced data.

resistant measure, p. 30

- Because $\bar{x}$ is not a resistant measure, outliers can have a large effect on the confidence interval. *You should search for outliers and try to correct them or justify their removal before computing the interval.* If the outliers cannot be removed, ask your statistical consultant about procedures that are not sensitive to outliers.
- If the sample size is small and the population is not Normal, the true confidence level will be different from the value C used in computing the interval. *Prior to any calculations, examine your data carefully for skewness and other signs of non-Normality.* Remember though that the interval relies only on the distribution of $\bar{x}$, which even for quite small sample sizes is much closer to Normal than is the distribution of the individual observations. When $n \geq 15$, the confidence level is not greatly disturbed by non-Normal populations unless extreme outliers or quite strong skewness are present. Our college fund contribution data in Example 6.4 are very likely skewed, but because of the large sample size, we are confident that the distribution of the sample mean will be approximately Normal.

standard deviation s, p. 38

- The interval $\bar{x} \pm z^* \sigma/\sqrt{n}$ assumes that the standard deviation σ of the population is known. This unrealistic requirement renders the interval of little use in statistical practice. We will learn in the next chapter what to do when σ is unknown. If, however, the sample is large, the sample standard deviation σ will be close to the unknown σ. The interval $\bar{x} \pm z^* s/\sqrt{n}$ is then an approximate confidence interval for μ.

The most important caution concerning confidence intervals is a consequence of the first of these warnings. *The margin of error in a confidence interval covers only random sampling errors.* The margin of error is obtained from the sampling distribution and indicates how much error can be expected because of chance variation in randomized data production.

Practical difficulties such as undercoverage and nonresponse in a sample survey cause additional errors. These errors can be larger than the random sampling error. This often happens when the sample size is large (so that $\sigma/\sqrt{n}$ is small). Remember this unpleasant fact when reading the results of an opinion poll or other sample survey. The practical conduct of the survey influences the trustworthiness of its results in ways that are not included in the announced margin of error.

Every inference procedure that we will meet has its own list of warnings. Because many of the warnings are similar to those we have mentioned, we will not print the full warning label each time. It is easy to state (from the mathematics of probability) conditions under which a method of inference is exactly correct. These conditions are *never* fully met in practice.

For example, no population is exactly Normal. *Deciding when a statistical procedure should be used in practice often requires judgment assisted by exploratory analysis of the data.* Mathematical facts are, therefore, only a part of statistics. The difference between statistics and mathematics can be stated thusly: mathematical theorems are true; statistical methods are often effective when used with skill.

Finally, you should understand what statistical confidence does not say. Based on our SRS in Example 6.3, we are 95% confident that the mean SATM score for the California students lies between 486 and 504. This says that this interval was calculated by a method that gives correct results in 95% of all possible samples. It does *not* say that the probability is 0.95 that the true mean falls between 486 and 504. *No randomness remains after we draw a particular sample and compute the interval.* The true mean either is or is not between 486 and 504. The probability calculations of standard statistical inference describe how often the *method,* not a particular sample, gives correct answers.

6.2 Tests of Significance

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 population means

or equivalently,

H_0: the difference in population means is zero

Note that the null hypothesis refers to the *population* means for all undergraduates, 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 population means are not the same

or equivalently,

H_a: the difference in population means is not zero

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.

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. We call this specified value the hypothesized value.
- Values of the estimate far from the hypothesized value give evidence against H_0. The alternative hypothesis determines which directions count against H_0.
- To assess how far the estimate is from the hypothesized value, standardize the estimate. In many common situations the test statistic has the form

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

Average scholarship amount of borrowers and nonborrowers: The test statistic. In Example 6.8, the estimate of the difference is \$425. Using methods that we will discuss in detail later, we can determine that the standard deviation of the estimate is \$353. For this problem the test statistic is

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

For our data,

$$z = \frac{425 - 0}{353} = 1.20$$

We have observed a sample estimate that is one and one-fifth standard deviations away from the hypothesized value of the parameter.

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.

EXAMPLE 6.12

Average scholarship amount of borrowers and nonborrowers: The *P*-value. In Example 6.11, we found that the test statistic for testing

H_0: the true mean difference is 0

versus

H_a: there is a difference in the population means

is

$$z = \frac{425 - 0}{353} = 1.20$$

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, $z = 1.20$. From Table A, our table of standard Normal probabilities, we find

$$P(Z \geq 1.20) = 1 - 0.8849 = 0.1151$$

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

$$P(Z \leq -1.20) = 0.1151$$

So the P-value is

$$P = 2P(Z \geq 1.20) = 2(0.1151) = 0.2302$$

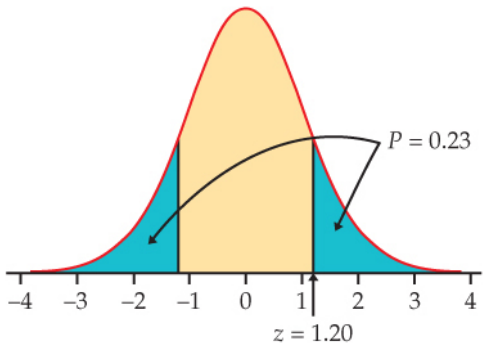


FIGURE 6.9 The *P*-value, Example 6.12. The P-value is the probability (when H_0 is true) that $\bar{x}$ takes a value as extreme or more extreme than the actual observed value, $z = 1.20$. Because the alternative hypothesis is two-sided, we use both tails of the distribution.

This is the value that we reported on page 361. There is a 23% chance of observing a difference as extreme as the $425 in our sample if the true population difference is zero. This P-value tells us that our outcome is not particularly extreme. In other words, the data do not provide substantial evidence for us to doubt the validity of the null hypothesis.

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 has reached the standard set by α. For example, 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 α = 0.05 level but is not significant at the α = 0.01 level. We discuss this in more detail at the end of this section.

EXAMPLE 6.14

Parent income contribution by school type: The conclusion. In Example 6.9, we found that the difference in the average parent current income contribution between undergraduates going to a private college versus public college was $1639. Because the cost of tuition at a private college is typically higher than the cost at a public college,[14] we had a prior expectation that the parental current income contribution would be higher for undergraduates going to a private college. It is appropriate to use a one-sided alternative in this situation. So, our hypotheses are

H_0: the true mean difference is 0

versus

H_a: the difference between the average parent income contribution of undergraduates at a private college and public college is positive

The standard deviation is $428 (again, we defer details regarding this calculation), and the test statistic is

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

$$z = \frac{1639 - 0}{430}$$

$$= 3.81$$

Because only positive differences in parental contributions 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. In Table A, the largest z is 3.49. This means that for z = 3.81, $P < 0.0002$. Using software, we can be more precise. The P-value is

$$P = P(Z \geq 3.81)$$

$$= 0.0001$$

The calculation is illustrated in Figure 6.10. There is about a 1-in-10,000 chance of observing a difference as large or larger than the $1639 in our sample if the true population difference is zero. This *P*-value tells us that our outcome is extremely rare. We conclude that the null hypothesis must be false. Because the observed difference is positive, here is one way to report the result: "The data clearly show that the average parent income contribution for undergraduates at a private college is larger than the average parent income contribution for undergraduates at a public college ($z = 3.81$, $P = 0.0001$)."

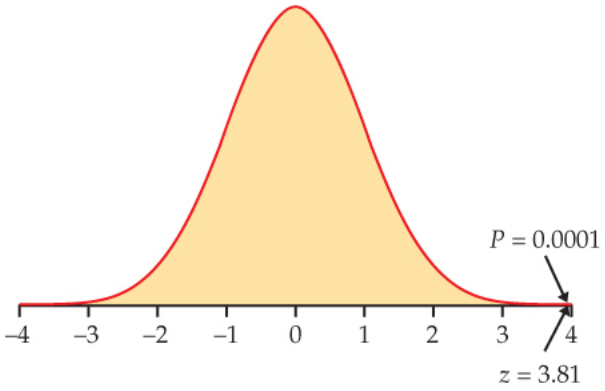


FIGURE 6.10 The *P*-value, Example 6.14. The *P*-value is the probability (when H_0 is true) that $\bar{x}$ takes a value as extreme or more extreme than the actual observed value, $z = 3.81$. We look at only the right tail because we are considering the one-sided (>) alternative.

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 or two that summarizes what you have found by using a test of significance.

EXAMPLE 6.13

Average scholarship amount of borrowers and nonborrowers: The conclusion. In Example 6.12, we found that the P-value is

$$P = 2P(Z \geq 1.20) = 2(0.1151) = 0.2302$$

There is an 23% chance of observing a difference as extreme as the \$425 in our sample if the true population difference is zero. Because this P-value is larger than the $\alpha = 0.05$ significance level, we conclude that our test result is not significant. We could report the result as "the data fail to provide evidence that would cause us to conclude that there is a difference in average scholarship amount between borrowers and nonborrowers ($z = 1.20$, $P = 0.23$)."

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. Our examples focused on the comparison of two population means. Here is a summary for a test about one population mean.

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 hypothesized 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 $\bar{x}$. Our test statistic measures the difference between the sample estimate and the hypothesized parameter in terms of standard deviations of the test statistic:

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

Suppose that 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**

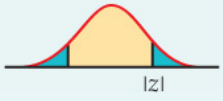
$$z = \frac{x - \mu_0}{\sigma/\sqrt{n}}$$

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.

Energy intake from sugar-sweetened beverages. Consumption of sugar-sweetened beverages (SSBs) has been positively associated with weight gain and obesity and negatively associated with the intake of important micronutrients. One study used data from the National Health and Nutrition Examination Survey (NHANES) to estimate SSB consumption among adolescents (aged 12 to 19 years). More than 2400 individuals provided data for this study.[15] The mean consumption was 298 calories per day.

You survey 100 students at your large university and find the average consumption of SSBs per day to be 262 calories. Is there evidence that the average calories per day from SSBs at your university differs from this large U.S. survey average?

The null hypothesis is "no difference" from the published mean $\mu_0 = 298$. The alternative is two-sided because you did not have a particular direction in mind before examining the data. So the hypotheses about the unknown mean μ of the students at your university are

$$H_0: \mu = 298$$
$$H_a: \mu \neq 298$$

As usual in this chapter, we make the unrealistic assumption that the population standard deviation is known. In this case, we'll use the standard deviation from the large national study, $\sigma = 435$ calories.

We compute the test statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{262 - 298}{435/\sqrt{100}} = -0.83$$

Figure 6.11 illustrates the P-value, which is the probability that a standard Normal variable Z takes a value at least 0.83 away from zero. From Table A, we find that this probability is

$$P = 2P(Z \geq 0.83) = 2(1 - 0.7967) = 0.4066$$

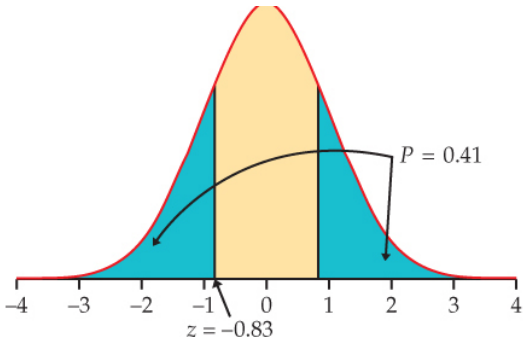


FIGURE 6.11 Sketch of the P-value calculation for the two-sided test, Example 6.15. The test statistic is $z = -0.83$.

That is, if the population mean were 298, more than 40% of the time an SRS of size 100 from the students at your university would have a mean consumption from SSBs at least as far from 298 as that of this sample. The observed $\bar{x} = 262$ is, therefore, not strong evidence that the student population mean at your university differs from that of the large population of adolescents.

Significance test of the mean SATM score. In a discussion of SAT Mathematics (SATM) scores, someone comments: "Because only a select 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 485." You do not agree with this claim and decide to use the SRS of 500 seniors from Example 6.3 (page 344) to assess the degree of evidence against it. Those 500 seniors had a mean SATM score of $\bar{x} = 495$. Is this strong enough evidence to conclude that this person's claim is wrong?

Because the claim states that the mean is "no more than 485," the alternative hypothesis is one-sided. The hypotheses are

$$H_0: \mu = 485$$

$$H_a: \mu > 485$$

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

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{495 - 485}{100/\sqrt{500}} = 2.24$$

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

$$P = P(Z \geq 2.24) = 1 - 0.9875 = 0.0125$$

Figure 6.12 illustrates this P-value. A mean score as large as that observed would occur roughly 12 times in 1000 samples if the population mean were 485. This is convincing evidence that the mean SATM score for all California high school seniors is higher than 485. You can confidently tell this person that his or her claim is incorrect.

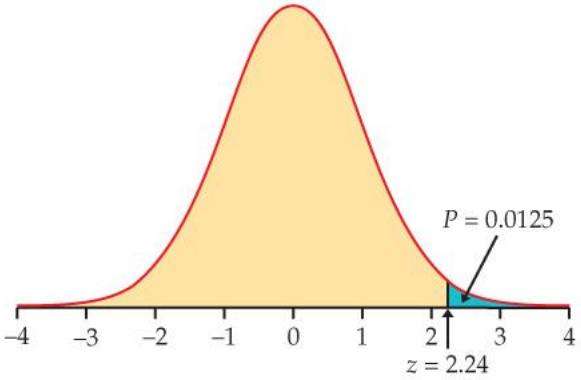


FIGURE 6.12 Sketch of the P-value calculation for the one-sided test, Example 6.16. The test statistic is $z = 2.24$.

Water quality testing. The Deely Laboratory is a drinking-water testing and analysis service. One of the common contaminants it tests for is lead. Lead enters drinking water through corrosion of plumbing materials, such as lead pipes, fixtures, and solder. The service knows that their analysis procedure is unbiased but not perfectly precise, so the laboratory analyzes each water sample three times and reports the mean result. The repeated measurements follow a Normal distribution quite closely. The standard deviation of this distribution is a property of the analytic procedure and is known to be $\sigma = 0.25$ parts per billion (ppb).

The Deely Laboratory has been asked by a university to evaluate a claim that the drinking water in the Student Union has a lead concentration above the Environmental Protection Agency's (EPA) action level of 15 ppb. Because the true concentration of the sample is the mean μ of the population of repeated analyses, the hypotheses are

$$H_0: \mu = 15$$

$$H_a: \mu \neq 15$$

We use the two-sided alternative here because there is no prior evidence to substantiate a one-sided alternative. The lab chooses the 1% level of significance, $\alpha = 0.01$.

Three analyses of one specimen give concentrations

15.84 15.33 15.58

The sample mean of these readings is

$$\bar{x} = \frac{15.84 + 15.33 + 15.58}{3} = 15.58$$

The test statistic is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{15.58 - 15.00}{0.25/\sqrt{3}} = 4.02$$

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

$$P = 2P(Z \geq 4.02)$$

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 $2P(Z \geq 3.49) = 2(1 - 0.9998) = 0.0004$. Software or a calculator could be used to give an accurate value of the P-value. However, because the P-value is clearly less than the lab's standard of 1%, we reject H_0. Because $\bar{x}$ is larger than 15.00, we can conclude that the true concentration level of lead in this one specimen is higher than the EPA's action level.

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

$$\begin{aligned} \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} &= 15.58 \pm 2.576 \left(0.25/\sqrt{3}\right) \\ &= 15.58 \pm 0.37 \\ &= (15.21, 15.95) \end{aligned}$$

The hypothesized value $\mu_0 = 15.00$ in Example 6.17 falls outside the confidence interval we computed in Example 6.18. In other words, it is in the region we are 99% confident that μ is *not* in. Thus, we can reject

$$H_0: \mu = 15.00$$

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

$$H_0: \mu = 15.30$$

at the 1% level in favor of the two-sided alternative $H_a: \mu \neq 15.30$, because 15.30 lies inside the 99% confidence interval for μ. Figure 6.13 illustrates both cases.

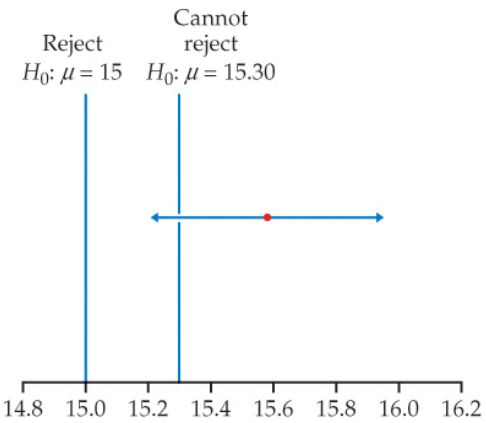


FIGURE 6.13 The link between two-sided significance tests and confidence intervals. For the study described in Examples 6.17 and 6.18, values of μ falling outside a 99% confidence interval can be rejected at the 1% significance level; values falling inside the interval cannot be rejected. This holds for any significance level α and $1 - \alpha$ confidence interval.

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 $C = 1 - \alpha$.

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.

The *P*-value versus a statement of significance

The observed result in Example 6.17 was $z = 4.02$. 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 actual P-value

$$2P(Z \geq 4.02) = 0.000058$$

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

$$H_0: \mu = 485$$

$$H_a: \mu \geq 485$$

concerning the mean SAT Mathematics score μ of California high school seniors. The test had the P-value $P = 0.0125$. This result is significant at the $\alpha = 0.05$ level because $0.0125 \leq 0.05$. It is not significant at the $\alpha = 0.01$ level, because the P-value is larger than 0.01. See Figure 6.14.

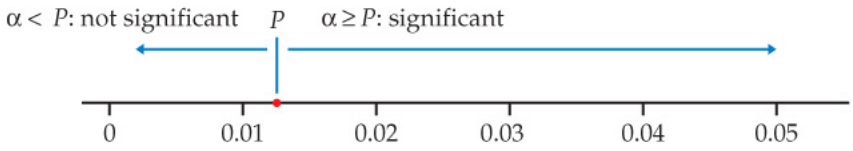


FIGURE 6.14 Link between the P-value and the significance level α. An outcome with P-value P is significant at all levels a at or above P and is not significant at smaller levels α.

Average scholarship amount of borrowers and nonborrowers: Assessing significance. In Example 6.11 (page 365), we found the test statistic $z = 1.20$ for testing the null hypothesis that there was no difference in the mean scholarship amount between borrowers and nonborrowers. 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. Because our calculated value is 1.20, we are within the range and we do not reject the null hypothesis at the 5% level of significance.

z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

6.3 Use and Abuse of Tests

Information provided by the *P*-value. Suppose that the test statistic for a two-sided significance test for a population mean is $z = 1.95$. From Table A we can calculate the P-value. It is

$$P = 2[1 - P(Z \leq 1.95)] = 2(1 - 0.9744) = 0.0512$$

We have failed to meet the standard of evidence for $\alpha = 0.05$. However, with the information provided by the P-value, we can see that the result just barely missed the standard. If the effect in question is interesting and potentially important, we might want to design another study with a larger sample to investigate it further.

6.4 Power and Inference as a Decision

The power of a TBBMC significance test. Can a six-month exercise program increase the total body bone mineral content (TBBMC) of young women? A team of researchers is planning a study to examine this question. Based on the results of a previous study, they are willing to assume that $\sigma = 2$ for the percent change in TBBMC over the six-month period. They also believe that a change in TBBMC of 1% is important, so they would like to have a reasonable chance of detecting a change this large or larger. Is 25 subjects a large enough sample for this project?

We will answer this question by calculating the power of the significance test that will be used to evaluate the data to be collected. The calculation consists of three steps:

1. State H_0, H_a (the particular alternative we want to detect), and the significance level α.
2. Find the values of $\bar{x}$ that will lead us to reject H_0.
3. Calculate the probability of observing these values of $\bar{x}$ when the alternative is true.

Step 1. The null hypothesis is that the exercise program has no effect on TBBMC. In other words, the mean percent change is zero. The alternative is that exercise is beneficial; that is, the mean change is positive. Formally, we have

$$H_0: \mu = 0$$

$$H_a: \mu > 0$$

The alternative of interest is $\mu = 1\%$ increase in TBBMC. A 5% test of significance will be used.

Step 2. The z test rejects H_0 at the $\alpha = 0.05$ level whenever

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 0}{2/\sqrt{25}} \geq 1.645$$

Be sure you understand why we use 1.645. Rewrite this in terms of $\bar{x}$:

$$\bar{x} \geq 1.645 \frac{2}{\sqrt{25}}$$

$$\bar{x} \geq 0.658$$

Because the significance level is $\alpha = 0.05$, this event has probability 0.05 of occurring *when the population mean m is 0.*

Step 3. The power to detect the alternative $\mu = 1\%$ is the probability that H_0 will be rejected *when in fact $\mu = 1\%$.* We calculate this probability by standardizing $\bar{x}$, using the value $\mu = 1$, the population standard deviation $\sigma = 2$, and the sample size $n = 25$. The power is

$$P(\bar{x} \geq 0.658 \text{ when } \mu = 1) = P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \geq \frac{0.658 - 1}{2/\sqrt{25}}\right)$$

$$= P(Z \geq -0.855) = 0.80$$

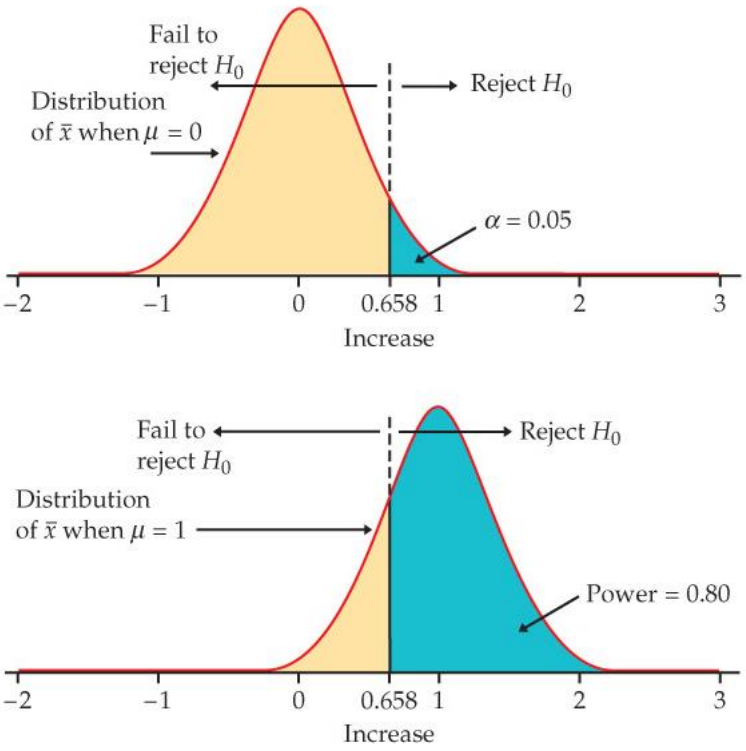


FIGURE 6.16 The sampling distributions of $\bar{x}$ when $\mu = 0$ and when $\mu = 1$, Example 6.29. The power is the probability that the test rejects H_0 when the alternative is true.

Power of the lead concentration test. Example 6.17 (page 375) presented a test of

$$H_0: \mu = 15.00$$

$$H_a: \mu \neq 15.00$$

at the 1% level of significance. What is the power of this test against the specific alternative $\mu = 15.50$?

The test rejects H_0 when $|z| \geq 2.576$. The test statistic is

$$z = \frac{\bar{x} - 15.00}{0.25/\sqrt{3}}$$

Some arithmetic shows that the test rejects when either of the following is true:

$$z \geq 2.576 \quad \text{(in other words, } \bar{x} \geq 15.37\text{)}$$

$$z \leq -2.576 \quad \text{(in other words, } \bar{x} \leq 14.63\text{)}$$

These are disjoint events, so the power is the sum of their probabilities, *computed assuming that the alternative $\mu = 15.50$ is true.* We find that

$$\begin{aligned} P(\bar{x} \geq 15.37) &= P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \geq \frac{15.37 - 15.50}{0.25/\sqrt{3}}\right) \\ &= P(Z \geq -0.90) = 0.8159 \\ P(\bar{x} \leq 14.63) &= P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq \frac{14.63 - 15.50}{0.25/\sqrt{3}}\right) \\ &= P(Z \leq -6.03) = 0 \end{aligned}$$

Figure 6.17 illustrates this calculation. A power of about 0.82, we are quite confident that the test will reject H_0 when this alternative is true.

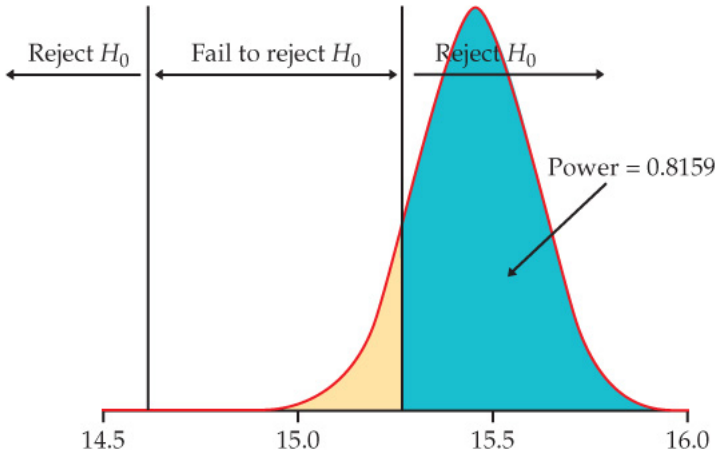


FIGURE 6.17 The power, Example 6.30. Unlike Figure 6.16, only the sampling distribution under the alternative is shown.

Outer diameter of a skateboard bearing. The mean outer diameter of a skateboard bearing is supposed to be 22.000 millimeters (mm). The outer diameters vary Normally with standard deviation $\sigma = 0.010$ mm. When a lot of the bearings arrives, the skateboard manufacturer takes an SRS of five bearings from the lot and measures their outer diameters. The manufacturer rejects the bearings if the sample mean diameter is significantly different from 22 mm at the 5% significance level.

This is a test of the hypotheses

$$H_0: \mu = 22$$

$$H_a: \mu \neq 22$$

To carry out the test, the manufacturer computes the z statistic:

$$z = \frac{\bar{x} - 22}{0.01/\sqrt{5}}$$

and rejects H_0 if

$$z < -1.96 \quad \text{or} \quad z > 1.96$$

A Type I error is to reject H_0 when in fact $\mu = 22$.

What about Type II errors? Because there are many values of μ in H_a, we will concentrate on one value. The producer and the manufacturer agree that a lot of bearings with mean 0.015 mm away from the desired mean 22.000 should be rejected. So a particular Type II error is to accept H_0 when in fact $\mu = 22.015$.

Figure 6.21 shows how the two probabilities of error are obtained from the two sampling distributions of $\bar{x}$, for $\mu = 22$ and for $\mu = 22.015$. When $\mu = 22$, H_0 is true and to reject H_0 is a Type I error. When $\mu = 22.015$, accepting H_0 is a Type II error. We will now calculate these error probabilities.

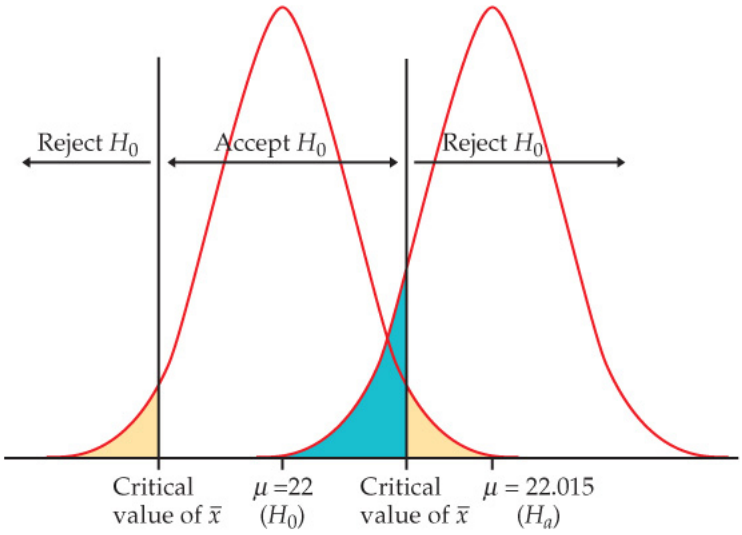


FIGURE 6.21 The two error probabilities, Example 6.33. The probability of a Type I error (yellow area) is the probability of rejecting H_0: $\mu = 22$ when, in fact, $\mu = 22$. The probability of a Type II error (blue area) is the probability of accepting H_0 when, in fact, $\mu = 22.015$.

The probability of a Type I error is the probability of rejecting H_0 when it is really true. In Example 6.33, this is the probability that $|z| \geq 1.96$ when $\mu = 22$. But this is exactly the significance level of the test. The critical value 1.96 was chosen to make this probability 0.05, so we do not have to compute it again. The definition of "significant at level 0.05" is that sample outcomes this extreme will occur with probability 0.05 when H_0 is true.