Danita Delimont/Getty Images

CHAPTER 6

Introduction to Inference

Introduction

Statistical inference draws conclusions about a population or process from sample data. It also provides a statement of how much confidence we can place in our conclusions. Although there are numerous methods for inference, there are only a few general types of statistical inference. This chapter introduces the two most common types: *confidence intervals* and *tests of significance*.

Because the underlying reasoning for these two types of inference remains the same across different settings, this chapter considers just one simple setting that is closely related to our study of the sampling distributions of $\bar{x}$ in Section 5.2 (page 293): inference about the mean of a large population whose standard deviation is known. This setting, although unrealistic, allows us to focus on the underlying rationale of statistical inference rather than the calculations.

Later chapters present inference methods to use in most of the settings we met in learning to explore data. In fact, there are libraries—both of books and of computer software—full of more elaborate statistical techniques. Informed use of any of these methods, however, requires a firm understanding of the underlying reasoning. That is the goal of this chapter. A computer or calculator will do the arithmetic, but *you must exercise sound judgment based on understanding*.

Overview of Inference

The purpose of statistical inference is to draw conclusions from data. Formal inference emphasizes substantiating our conclusions via probability calculations. Probability allows us to take chance variation into account. Here is an example.

EXAMPLE 6.1

WADE

Clustering of trees in a forest. The Wade Tract in Thomas County, Georgia, is an old-growth forest of longleaf pine trees (*Pinus palustris*) that has survived in a relatively undisturbed state since before the settlement of the area by Europeans. Foresters who study these trees are interested in how the trees are distributed in the forest. Is there some sort of clustering, resulting in regions of the forest with more trees than others? Or are the tree locations random, resulting in no particular patterns? Figure 6.1 gives a plot of the locations of all 584 longleaf pine trees in a 200-meter by 200-meter region in the Wade Tract.[1]

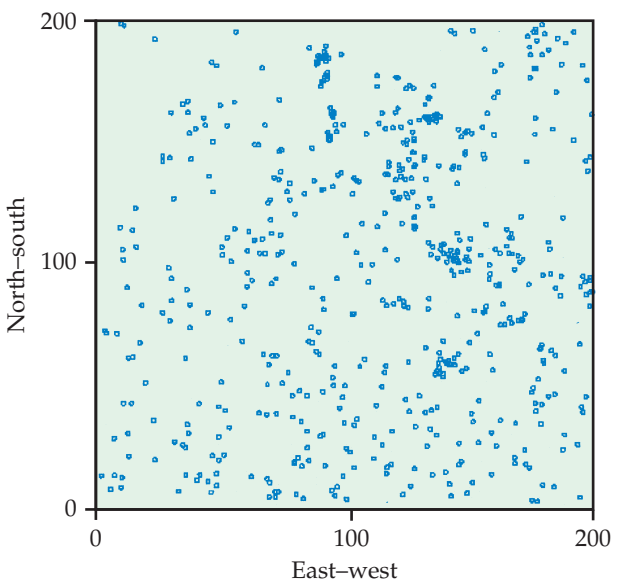


FIGURE 6.1 The distribution of longleaf pine trees, Example 6.1.

Do the locations appear to be random, or do there appear to be clusters of trees? One approach to the analysis of these data indicates that a pattern as clustered as, or more clustered than, the one in Figure 6.1 would occur only 4% of the time if, in fact, the locations of longleaf pine trees in the Wade Tract are random. Because this chance is fairly small, we conclude that there is some clustering of these trees.

This probability calculation helps us to distinguish between patterns that are consistent or inconsistent with the random location scenario. Here is an example assessing a new oral antibiotic for acne—with a different conclusion.

EXAMPLE 6.2

Effectiveness of a new oral antibiotic. Researchers want to know if a new oral antibiotic is more effective in relieving acne than a popular topical (on the skin) antibiotic. Twenty patients are randomly assigned to receive the oral medication, and another 20 receive the topical medication. Fifteen (75%) of those taking the oral medication find satisfactory symptom relief versus only 11 (55%) of the topical medication patients.

Our unaided judgment suggests that the oral medication is better, 75% to 55%. However, probability calculations tell us that a difference this large or larger between the results in the two groups of 20 patients would occur about one time in five simply because of chance variation. In this case, it is better to conclude that the data fail to establish a real difference between the two treatments. This probability (nearly 0.19) is too large to ignore.

In this chapter, we introduce the two most frequently used types of statistical inference. Section 6.1 concerns *confidence intervals* for estimating the value of a population parameter. Section 6.2 presents *tests of significance*, which assess the evidence for a claim, such as those in Examples 6.1 and 6.2.

sampling distribution, p. 286

Both types of inference are based on the sampling distributions of statistics. That is, both report probabilities that state *what would happen if we used the inference method many times*. This kind of probability statement is characteristic of standard statistical inference. Users of statistics must understand the nature of this reasoning and the meaning of the probability statements that appear, for example, online and in journal articles and statistical software output.

Because the methods of formal inference are based on sampling distributions, they require a probability model for the data. Trustworthy probability models can arise in many ways, but the model is most secure and inference is most reliable when the data are produced by a properly randomized design.

When you use statistical inference, you are acting as if the data come from a random sample or a randomized experiment. If this is not true, your conclusions may be open to challenge. Do not be overly impressed by the complex details of formal inference. This elaborate machinery cannot remedy basic flaws in producing the data such as voluntary response samples and confounded experiments. Use the common sense developed in your study of the first three chapters of this book, and proceed to detailed formal inference only when you are satisfied that the data deserve such analysis.

6.1 Estimating with Confidence

When you complete this section, you will be able to:

AU: Please check.
PR edit here okay?

- Describe a level C confidence interval for a population parameter in terms of an estimate and its margin of error.
- Construct a level C confidence interval for μ from a simple random sample (SRS) of size n from a large population having known standard deviation σ.
- Explain how the margin of error changes with a change in the confidence level C.
- Determine the sample size needed to obtain a specified margin of error for a level C confidence interval for μ.
- Identify situations where inference about μ based on the confidence interval $\bar{x} \pm z^*\sigma/\sqrt{n}$ may be suspect.

The SAT is a widely used measure of readiness for college study. It consists of three sections, one for mathematical reasoning ability (SATM), one for verbal reasoning ability (SATV), and one for writing ability (SATW). Possible scores on each section range from 200 to 800, for a total range of 600 to 2400. Since 1995, section scores have been *recentered* so that the mean is approximately 500 with a standard deviation of 100 in a large "standardized group." This scale has been maintained so that scores have a constant interpretation.

LOOK BACK
linear transformations, p. 44

EXAMPLE 6.3


Peter Cade/The Image Bank/Getty Images

Estimating the mean SATM score for seniors in California. Suppose that you want to estimate the mean SATM score for the 485,264 high school seniors in California.[2] You know better than to trust data from the students who choose to take the SAT. Only about 38% of California students typically take the SAT. These self-selected students are planning to attend college and are not representative of all California seniors. At considerable effort and expense, you give the test to a simple random sample (SRS) of 500 California high school seniors. The mean score for your sample is $\bar{x} = 495$. What can you say about the mean score μ in the population of all 485,264 seniors?

LOOK BACK
unbiased estimator, p. 287
law of large numbers, p. 250

The sample mean $\bar{x}$ is the natural estimator of the unknown population mean μ. We know that $\bar{x}$ is an unbiased estimator of μ. More important, the law of large numbers says that the sample mean must approach the population mean as the size of the sample grows. The value $\bar{x} = 495$, therefore, appears to be a reasonable estimate of the mean score μ that all 485,264 students would achieve if they took the test.

But how reliable is this estimate? A second sample of 500 students would surely not give a sample mean of 495 again. Unbiasedness says only that there is no systematic tendency to underestimate or overestimate the truth. Could we plausibly get a sample mean of 485 and a sample mean of 520 in repeated samples? *An estimate without an indication of its variability is of little value.*

Statistical confidence

LOOK BACK
central limit theorem, p. 298

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 $N(\mu, \sigma/\sqrt{500})$. 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

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

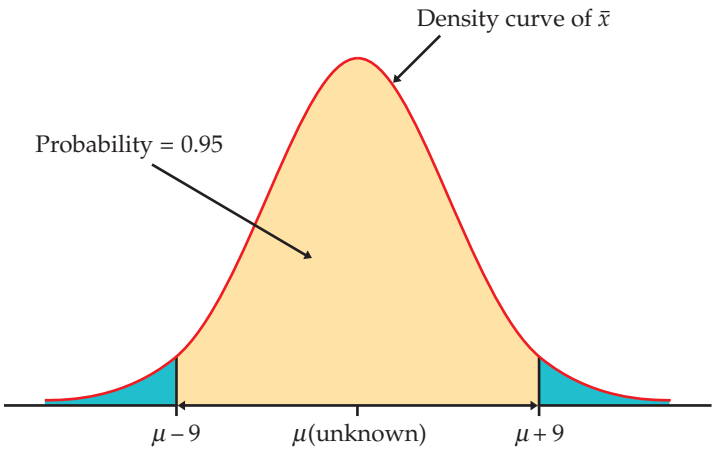


FIGURE 6.2 Distribution of the sample mean, Example 6.3. $\bar{x}$ lies within ± 9 points of μ in 95% of all samples. This also means that μ is within ± 9 points of $\bar{x}$ in those samples.

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."

USE YOUR KNOWLEDGE

6.1 How much do you spend on lunch? The average amount you spend on a lunch during the week is not known. Based on past experience, you are willing to assume that the standard deviation is \$2.10. If you take a random sample of 28 lunches, what is the value of the standard deviation of $\bar{x}$?

6.2 Applying the 68–95–99.7 rule. In the setting of the previous exercise, the 68–95–99.7 rule says that the probability is about 0.95 that $\bar{x}$ is within \$_____ of the population mean μ. Fill in the blank.

6.3 Constructing a 95% confidence interval. In the setting of the previous two exercises, about 95% of all samples will capture the true mean in the interval $\bar{x}$ plus or minus \$_____. Fill in the blank.

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

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

LOOK BACK
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

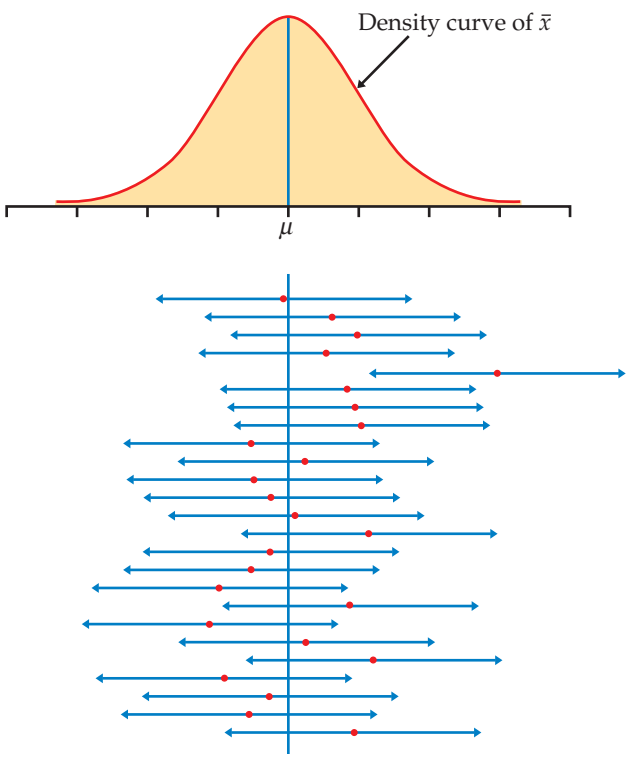


FIGURE 6.3 Twenty-five samples from the same population gave these 95% confidence intervals. In the long run, 95% of all samples give an interval that covers μ. The sampling distribution of $\bar{x}$ is shown at the top.

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.

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.
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.

With the *Confidence Interval* applet, you can construct diagrams similar to the one displayed in Figure 6.3. The only difference is that the applet displays the Normal population distribution at the top rather than the Normal sampling distribution of $\bar{x}$. You choose the confidence level C, the sample size n, and whether you want to generate 1 or 25 samples at a time. A running total (and percent) of the number of intervals that contain μ is displayed so you can consider a larger number of samples.

When generating single samples, the data for the latest SRS are shown below the confidence interval. The spread in these data reflects the spread of the population distribution. This spread is assumed known, and it does not change with sample size. What does change, as you vary n, is the margin of error, since it reflects the uncertainty in the estimate of μ. As you increase n, you'll find that the span of the interval gets smaller.

USE YOUR KNOWLEDGE

6.4 Generating a single confidence interval. Using the default settings in the *Confidence Interval* applet (95% confidence level and $n = 20$), click "Sample" to choose an SRS and display its confidence interval.

(a) Is the spread in the data, shown as yellow dots below the confidence interval, larger than the span of the confidence interval? Explain why this would typically be the case.

(b) For the same data set, you can compare the span of the confidence interval for different values of C by sliding the confidence level to a

new value. For the SRS you generated in part (a), what happens to the span of the interval when you move C to 99%? What about 90%? Describe the relationship you find between the confidence level C and the span of the confidence interval.

6.5 80% confidence intervals. The idea of an 80% confidence interval is that the interval captures the true parameter value in 80% of all samples. That's not high enough confidence for practical use, but 80% hits and 20% misses make it easy to see how a confidence interval behaves in repeated samples from the same population.

(a) Set the confidence level in the *Confidence Interval* applet to 80%. Click "Sample 25" to choose 25 SRSs and display their confidence intervals. How many of the 25 intervals contain the true mean μ? What proportion contain the true mean?

(b) We can't determine whether a new SRS will result in an interval that contains μ or not. The confidence level only tells us what percent will contain μ in the long run. Click "Sample 25" again to get the confidence intervals from 50 SRSs. What proportion hit? Keep clicking "Sample 25" and record the proportion of hits among 100, 200, 300, 400, and 500 SRSs. As the number of samples increases, we expect the percent of captures to get closer to the confidence level, 80%. Do you find this pattern in your results?

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 $N(\mu, \sigma/\sqrt{n})$ 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.

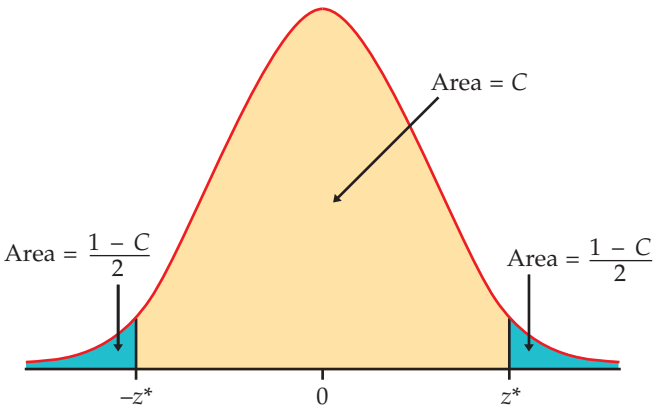


FIGURE 6.4 To construct a level C confidence interval, we must find the number z^*. The area between $-z^*$ and z^* under the standard Normal curve is C.

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.

Starting in 2008, Sallie Mae, a major provider of education loans and savings programs, has conducted an annual study titled "How America Pays for College." In the 2015 survey, 1600 randomly selected individuals (800 parents of undergraduate students and 800 undergraduate students) were surveyed by telephone.[3]

Many of the survey questions focus on the composition of funding sources used to pay for college, so the undergraduates in the survey are often responding for their parents. For example, each participant is asked to report how much of the parent's current income is used to pay for college. Do you think it is wise to combine responses across the parents and undergraduates? Are you fully aware of how much money your parents are spending and borrowing for college? The authors report overall averages and percents in their report. We will also consider this a sample from one population but this is certainly debatable.

EXAMPLE 6.4

Average college savings fund contribution. One survey question asked how much money from a college savings fund, such as a 529 plan, is used to pay for college. Of the 1600 who were surveyed, $n = 1593$ provided an answer. *Nonresponse should always be considered as a source of bias.* In this case, the nonresponse is very low, so we'll proceed by treating the $n = 1593$ sample as if it were an unbiased sample.

The average amount is $1768. It's very likely that this distribution is highly skewed to the right with many small amounts and a few very large amounts. Nevertheless, because the sample size is quite large, we can rely on the central limit theorem to assure us that the confidence interval based on the Normal distribution will be a good approximation.

Let's compute an approximate 95% confidence interval for the true mean amount contributed from a college savings fund among all undergraduates. We'll assume that the standard deviation for the population of college savings fund contributions is $1483. For 95% confidence, we see from Table D that $z^* = 1.960$. The margin of error for the 95% confidence interval for μ is, therefore,

$$\begin{aligned} m &= z^*\frac{\sigma}{\sqrt{n}} \\ &= 1.960\,\frac{1483}{\sqrt{1593}} \\ &= 37.16 \end{aligned}$$

We have computed the margin of error with more digits than we really need. Our mean is rounded to the nearest $1, so we will do the same for the margin of error. Keeping additional digits would provide no additional useful information. Therefore, we will use $m = 37$. The approximate 95% confidence interval is

$$\begin{aligned} \bar{x} \pm m &= 1768 \pm 37 \\ &= (1731, 1805) \end{aligned}$$

We are 95% confident that the mean amount contributed from a college savings fund among all undergraduates is between $1731 and $1805.

Suppose that the researchers who designed this study had used a different sample size. How would this affect the confidence interval? We can answer this question by changing the sample size in our calculations and assuming that the sample mean is the same.

EXAMPLE 6.5

How sample size affects the confidence interval. As in Example 6.4, the sample mean of the college savings fund contribution is \$1768 and the population standard deviation is \$1483. Suppose that the sample size is only 177 but still large enough for us to rely on the central limit theorem. In this case, the margin of error for 95% confidence is

$$m = z^* \frac{\sigma}{\sqrt{n}}$$
$$= 1.960 \frac{1483}{\sqrt{177}}$$
$$= 111.47$$

and the approximate 95% confidence interval is

$$\bar{x} \pm m = 1768 \pm 111$$
$$= (1657, 1879)$$

Notice that the margin of error for this example is three times as large as the margin of error that we computed in Example 6.4. The only change that we made was to assume that the sample size is 177 rather than 1593. This sample size is one-ninth of the original 1593. Thus, we triple the margin of error when we reduce the sample size to one-ninth of the original value. Figure 6.5 illustrates the effect in terms of the intervals.

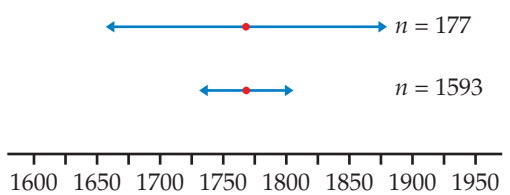


FIGURE 6.5 Confidence intervals for $n = 1593$ and $n = 177$, Examples 6.4 and 6.5. A sample size nine times as large results in a confidence interval that is one-third as wide.

USE YOUR KNOWLEDGE

6.6 Average amount paid for college. Refer to Example 6.4 (page 350). The average annual amount the $n = 1593$ families paid for college was \$24,164.[4] If the population standard deviation is \$8500, give the 95% confidence interval for μ, the average annual amount a family pays for a college undergraduate.

6.7 Changing the sample size. In the setting of the previous exercise, would the margin of error for 95% confidence be roughly doubled or halved if the sample size were raised to $n = 6375$? Verify your answer by performing the calculations.

6.8 Changing the confidence level. In the setting of Exercise 6.7, would the margin of error for 99% confidence be larger or smaller? Verify your answer by performing the calculations.

The argument leading to the form of confidence intervals for the population mean μ rested on the fact that the statistic $\bar{x}$ used to estimate μ has a Normal distribution. Because many sample estimates have Normal distributions

(at least approximately), it is useful to notice that the confidence interval has the form

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

The estimate based on the sample is the center of the confidence interval. The margin of error is $z^*\sigma_{\text{estimate}}$. The desired confidence level determines z^* from Table D. The standard deviation of the estimate is found from knowledge of the sampling distribution in a particular case. When the estimate is $\bar{x}$ from an SRS, the standard deviation of the estimate is $\sigma_{\text{estimate}} = \sigma/\sqrt{n}$. We return to this general form numerous times in the following chapters.

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.

AU: add page x-ref?

EXAMPLE 6.6

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.

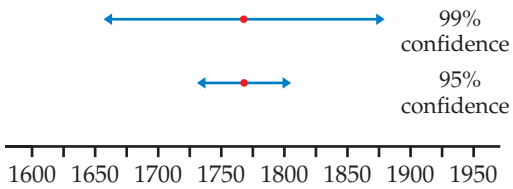


FIGURE 6.6 Confidence intervals, Examples 6.4 and 6.6. The larger the value of C, the wider the interval.

AU: digits OK?

Similarly, choosing a larger sample size n reduces the margin of error for any fixed confidence level. The square root in the formula implies that we must multiply the number of observations by 4 in order to cut the margin of error in half. Likewise, if we want to reduce the standard deviation of $\bar{x}$ by a factor of 4, we must take a sample 16 times as large.

The standard deviation σ measures the variation in the population. You can think of the variation among individuals in the population as noise that obscures the average value μ. It is harder to pin down the mean μ of a highly variable population; that is why the margin of error of a confidence interval increases with σ.

In practice, we can sometimes reduce σ by carefully controlling the measurement process. We also might change the mean of interest by restricting our attention to only part of a large population. Focusing in a subpopulation will often result in a smaller\sigma. This is why many medical studies only use healthy male subjects. The tradeoff, however, is less generalizable results.

AU/DE/PE: Caution icon?

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

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

This formula does not account for collection costs. In practice, taking observations costs time and money. The required sample size may be impossibly expensive. In those situations, you might consider a larger margin of error and/or a lower confidence level to find a workable sample size.

EXAMPLE 6.7

How many undergraduates should we survey? Suppose that we are planning a survey similar to the one described in Example 6.4 (page 350). If we want the margin of error for the average amount contributed from a college savings plan to be \$30 with 95% confidence, what sample size n do we need? For 95% confidence, Table D gives $z^* = 1.960$. For σ we will use the value from the previous study, \$1483. If the margin of error is \$30, we have

$$n = \left(\frac{z^*\sigma}{m}\right)^2 = \left(\frac{1.96 \times 1483}{30}\right)^2 = 9387.54$$

Because 9387 measurements will give a slightly wider interval than desired and 9388 measurements a slightly narrower interval, we should choose $n = 9388$. We need information from 9388 undergraduates to determine an estimate of mean college savings fund contribution with the desired margin of error.

It is always safe to round *up* to the next higher whole number when finding n because this will give us a smaller margin of error. The purpose of this calculation is to determine a sample size that is sufficient to provide useful results, but the determination of what is useful is a matter of judgment.

Would we need a much larger sample size to obtain a margin of error of \$25? Here is the calculation:

$$n = \left(\frac{z^*\sigma}{m}\right)^2 = \left(\frac{1.96 \times 1483}{25}\right)^2 = 13{,}518.06$$

A sample of $n = 13{,}519$ is much larger, and the costs of such a large sample may be prohibitive.

Unfortunately, the actual number of usable observations is often less than what we plan at the beginning of a study. This is particularly true of data collected in surveys but is an important consideration in most studies. Careful study designers often assume a nonresponse rate or dropout rate that specifies what proportion of the originally planned sample will fail to provide data. We use this information to calculate the sample size to be used at the start of the study.

For example, if in Example 6.7 we expect only 50% of those contacted to respond, we would need to start with a sample size of $2 \times 9388 = 18{,}776$ to obtain usable information from 9388 undergraduates and parents of undergraduates.

USE YOUR KNOWLEDGE

6.9 Starting salaries. You are planning a survey of starting salaries for recent computer science majors. In the latest survey by the National Association of Colleges and Employers, the average starting salary was reported to be \$61,287.[5] If you assume that the standard deviation is \$3850, what sample size do you need to have a margin of error equal to \$500 with 95% confidence?

6.10 Changes in sample size. Suppose that in the setting of the previous exercise you have the resources to contact 300 recent graduates. If all respond, will your margin of error be larger or smaller than \$500? What if only 50% respond? Verify your answers by performing the calculations.

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.

- 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.

- 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 s 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.

LOOK BACK
resistant measure, p. 30

LOOK BACK
standard deviation s, p. 38

AU/DE/PE: Caution icons?

AU: add page xrefs here?

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.

AU/DE/PE: Caution icon?

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.

AU: add page x-ref

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.

USE YOUR KNOWLEDGE

6.11 Nonresponse in a survey. In earlier versions of the Sallie Mae survey of Example 6.4 (page 350), participants were asked to report the undergraduate's outstanding credit card balance. Only about a third reported this amount. Provide a couple of reasons why a survey respondent might not provide an amount. Based on these reasons, do you think the sample mean using just the reported amounts is biased? Is the margin of error based just on the reported amounts a good measure of precision? Explain your answers.

SECTION 6.1 SUMMARY

- The purpose of a **confidence interval** is to estimate an unknown parameter with an indication of how accurate the estimate is and of how confident we are that the result is correct.
- Any confidence interval has two parts: an interval computed from the data and a confidence level. The interval often has the form

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

- The **confidence level** states the probability that the method will give a correct answer. That is, if you use 95% confidence intervals, in the long run 95% of your intervals will contain the true parameter value. When you apply the method once (that is, to a single sample), you do not know if your interval gave a correct answer (this happens 95% of the time) or not (this happens 5% of the time).

• The **margin of error** for a level C confidence interval for the mean μ of a Normal population with known standard deviation σ, based on an SRS of size n, is given by

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

Here z^* is obtained from the row labeled z^* at the bottom of Table D. The probability is C that a standard Normal random variable takes a value between $-z^*$ and z^*. The confidence interval is

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

If the population is not Normal and n is large, the confidence level of this interval is approximately correct.

• Other things being equal, the margin of error of a confidence interval decreases as

- the confidence level C decreases,
- the sample size n increases, and
- the population standard deviation σ decreases.

• The sample size n required to obtain a confidence interval of specified margin of error m for a population mean is

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

where z^* is the critical point for the desired level of confidence.

• A specific confidence interval formula is correct only under specific conditions. The most important conditions concern the method used to produce the data. Other factors such as the form of the population distribution may also be important. These conditions should be investigated *prior* to any calculations.

SECTION 6.1 EXERCISES

For Exercises 6.1 through 6.3, see pages 345–346; for Exercises 6.4 and 6.5, see pages 347–348; for Exercises 6.6 through 6.8, see page 351; for Exercises 6.9 and 6.10, see page 354; and for Exercise 6.11, see page 356.

6.12 Margin of error and the confidence interval. A study of stress on the campus of your university reported a mean stress level of 78 (on a 0 to 100 scale with a higher score indicating more stress) with a margin of error of 5 for 95% confidence. The study was based on a random sample of 64 undergraduates.

(a) Give the 95% confidence interval.

(b) If you wanted 99% confidence for the same study, would your margin of error be greater than, equal to, or less than 5? Explain your answer.

6.13 Changing the sample size. Consider the setting of the previous exercise. Suppose that the sample mean is again 78 and the population standard deviation is 20. Make a diagram similar to Figure 6.5 (page 351) that illustrates the effect of sample size on the width of a 95% interval. Use the following sample sizes: 9, 25, 81, and 100. Summarize what the diagram shows.

6.14 Changing the confidence level. Consider the setting of the previous two exercises. Suppose that the sample mean is still 78, the sample size is 64, and the population standard deviation is 20. Make a diagram similar to Figure 6.6 (page 353) that illustrates the effect of the confidence level on the width of the interval. Use 80%, 90%, 95%, and 99%. Summarize what the diagram shows.

6.15 Confidence interval mistakes and misunderstandings. Suppose that 500 randomly selected alumni of the University of Okoboji were asked to rate the university's academic advising services on a 1 to 10 scale. The sample mean $\bar{x}$ was found to be 8.6. Assume that the population standard deviation is known to be $\sigma = 2.2$.

(a) Ima Bitlost computes the 95% confidence interval for the average satisfaction score as $8.6 \pm 1.96(2.2)$. What is her mistake?

(b) After correcting her mistake in part (a), she states, "I am 95% confident that the sample mean falls between 8.4 and 8.8." What is wrong with this statement?

(c) She quickly realizes her mistake in part (b) and instead states, "The probability that the true mean is between 8.4 and 8.8 is 0.95." What misinterpretation is she making now?

(d) Finally, in her defense for using the Normal distribution to determine the confidence interval she says, "Because the sample size is quite large, the population of alumni ratings will be approximately Normal." Explain to Ima her misunderstanding and correct this statement.

6.16 More confidence interval mistakes and misunderstandings. Suppose that 100 randomly selected members of the Karaoke Channel were asked how much time they typically spend on the site during the week.[6] The sample mean $\bar{x}$ was found to be 3.8 hours. Assume that the population standard deviation is known to be $\sigma = 2.9$.

(a) Cary Oakey computes the 95% confidence interval for the average time on the site as $3.8 \pm 1.96(2.9/100)$. What is his mistake?

(b) He corrects this mistake and then states that "95% of the members spend between 3.23 and 4.37 hours a week on the site." What is wrong with his interpretation of this interval?

(c) The margin of error is slightly larger than half an hour. To reduce this to roughly 15 minutes, Cary says that the sample size needs to be doubled to 200. What is wrong with this statement?

6.17 The state of stress in the United States. Since 2007, the American Psychological Association has supported an annual nationwide survey to examine stress across the United States.[7] This year, a total of 720 millennials (18- to 33-year-olds) were asked to indicate their average stress level (on a 10-point scale) during the past month. The mean score was 5.5. Assume that the population standard deviation is 2.8.

(a) Give the margin of error and find the 95% confidence interval for this sample.

(b) Repeat these calculations for a 99% confidence interval. How do the results compare with those in part (a)?

6.18 Inference based on integer values. Refer to Exercise 6.17. The data for this study are integer values between 1 and 10. Explain why the confidence interval based on the Normal distribution should be a good approximation.

6.19 Mean TRAP in young women. For many important processes that occur in the body, direct measurement of characteristics of the process is not possible. In many cases, however, we can measure a *biomarker,* a biochemical substance that is relatively easy to measure and is associated with the process of interest. Bone turnover is the net effect of two processes: the breaking down of old bone, called resorption, and the building of new bone, called formation. One biochemical measure of bone resorption is tartrate-resistant acid phosphatase (TRAP), which can be measured in blood. In a study of bone turnover in young women, serum TRAP was measured in 31 subjects.[8] The mean was 13.2 units per liter (U/l). Assume that the standard deviation is known to be 6.5 U/l. Give the margin of error and find a 95% confidence interval for the mean TRAP amount in young women represented by this sample.

6.20 Mean OC in young women. Refer to the previous exercise. A biomarker for bone formation measured in the same study was osteocalcin (OC), measured in the blood. For the 31 subjects in the study, the mean was 33.4 nanograms per milliliter (ng/ml). Assume that the standard deviation is known to be 19.6 ng/ml. Report the 95% confidence interval.

6.21 Populations sampled and margins of error. Consider the following two scenarios. (A) Take a simple random sample of 200 freshman students at your college or university. (B) Take a simple random sample of 200 students at your college or university. For each of these samples, you will record the amount spent on textbooks used for classes during the fall semester. Which sample should have the smaller margin of error? Explain your answer.

6.22 Average starting salary. The National Association of Colleges and Employers (NACE) Spring Salary Survey shows that the current class of college graduates received an average starting-salary offer of $48,127.[9] Your institution collected an SRS ($n = 300$) of its recent graduates and obtained a 95% confidence interval of ($46,382, $48,008). What can we conclude about the *difference* between the average starting salary of recent graduates at your institution and the overall NACE average? Write a short summary.

6.23 Consumption of sweet snacks. A recent study reported that the U.S. per capita consumption of sweet snacks among healthy weight children aged 12 to 19 years is 251.2 kilocalories per day (kcal/d).[10] This was

based on 24-hour dietary recall records of $n = 2265$ adolescents.

(a) Suppose that the population distribution is heavily skewed, with a standard deviation equal to 540 kcal/d. What is the margin of error for a 95% confidence interval of the per capita consumption of sweet snacks?

(b) A future study is being planned and the goal is to have the margin of error no more than 15 kcal/d. Based on your answer to part (a), will this study require an examination of more or fewer recall records? Explain your answer without calculations.

(c) Compute the sample size necessary for the study described in part (b).

6.24 Total sleep time of college students. In Example 5.4 (page 293), the total sleep time per night among college students was approximately Normally distributed with mean $\mu = 6.78$ hours and standard deviation $\sigma = 1.24$ hours. You initially plan to take an SRS of size $n = 175$ and compute the average total sleep time.

(a) What is the standard deviation for the average time in hours? in minutes?

(b) Use the 95 part of the 68–95–99.7 rule to describe the variability of this sample mean.

(c) What is the probability that your average will be below 6.9 hours?

6.25 Determining sample size. Refer to the previous exercise. You really want to use a sample size such that about 95% of the averages fall within ± 5 minutes of the true mean $\mu = 6.78$.

(a) Based on your answer to part (b) in Exercise 6.24, should the sample size be larger or smaller than 175? Explain.

(b) What standard deviation of $\bar{x}$ do you need such that 95% of all samples will have a mean within 5 minutes of μ?

(c) Using the standard deviation you calculated in part (b), determine the number of students you need to sample.

6.26 Inference based on skewed data. The mean OC for the 31 subjects in Exercise 6.20 was 33.4 ng/ml. In our calculations, we assumed that the standard deviation was known to be 19.6 ng/ml. Use the 68–95–99.7 rule from Chapter 1 (page 57) to find the approximate bounds on the values of OC that would include these percents of the population. If the assumed standard deviation is correct, this distribution may be highly skewed. Why? (*Hint:* The measured values for a variable such as this are all positive.) Do you think that this skewness will invalidate the use of the Normal confidence interval in this case? Explain your answer.

6.27 Average hours per week listening to the radio. The *Student Monitor* surveys 1200 undergraduates from four-year colleges and universities throughout the United States semiannually to understand trends among college students.[11] Recently, the *Student Monitor* reported that the average amount of time listening to the radio per week was 11.5 hours. Of the 1200 students surveyed, 83% said that they listened to the radio, so this collection of listening times has around 204 (17% $\times$ 1200) zeros. Assume that the standard deviation is 8.3 hours.

(a) Give a 95% confidence interval for the mean time spent per week listening to the radio.

(b) Is it true that 95% of the 1200 students reported weekly times that lie in the interval you found in part (a)? Explain your answer.

(c) It appears that the population distribution has many zeros and is skewed to the right. Explain why the confidence interval based on the Normal distribution should nevertheless be a good approximation.

6.28 Average minutes per week listening to the radio. Refer to the previous exercise.

(a) Give the mean and standard deviation in minutes.

(b) Calculate the 95% confidence interval in minutes from your answer to part (a).

(c) Explain how you could have directly calculated this interval from the 95% interval that you calculated in the previous exercise.

6.29 Outlook on life. Since 2008, the Gallup-Healthways Well-Being Index tracks how people feel about their daily lives. In 2014, 54.1% of the respondents were classified as "thriving." This classification is based on how a respondent rates his or her current and future lives. This is the highest percent of respondents in this category since the index started. Material provided with the results noted:

> *Results are based on telephone interviews . . . with a random sample of 176,903 adults, living in all 50 U.S. states and the District of Columbia. For results based on the total sample of national adults, the margin of sampling error is ±1 percentage points at the 95% confidence level.*[12]

AU: Please check PR correction here.

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

(a) The announced poll result was 54.1% ± 1%. Can we be certain that the true population percent falls in this interval? Explain your answer.

(b) Explain to someone who knows no statistics what the announced result 54.1% ± 1% 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 nonresponse? Explain your answer.

6.30 Fuel efficiency. Computers in some vehicles calculate various quantities related to performance. One of these is the fuel efficiency, or gas mileage, usually expressed as miles per gallon (mpg). For one vehicle equipped in this way, the miles per gallon were recorded each time the gas tank was filled, and the computer was then reset.[13] Here are the mpg values for a random sample of 20 of these records: MPG

41.5	50.7	36.6	37.3	34.2	45.0	48.0	43.2	47.7	42.2
43.2	44.6	48.4	46.4	46.8	39.2	37.3	43.5	44.3	43.3

Suppose that the standard deviation is known to be $\sigma = 3.5$ mpg.

(a) What is $\sigma_{\bar{x}}$, the standard deviation of $\bar{x}$?

(b) Examine the data for skewness and other signs of non-Normality. Show your plots and numerical summaries. Do you think it is reasonable to construct a confidence interval based on the Normal distribution? Explain your answer.

(c) Give a 95% confidence interval for μ, the mean miles per gallon for this vehicle.

6.31 Fuel efficiency in metric units. In the previous exercise, you found an estimate with a margin of error for the average miles per gallon. Convert your estimate and margin of error to the metric units kilometers per liter (kpl). To change mpg to kpl, use the fact that 1 mile = 1.609 kilometers and 1 gallon = 3.785 liters.

6.32 How many "hits"? The *Confidence Interval* applet lets you simulate large numbers of confidence intervals quickly. Select 95% confidence and then sample 50 intervals. Record the number of intervals that cover the true value (this appears in the "Hit" box in the applet). Press the "Reset" button and repeat 30 times. Make a stemplot of the results and find the mean. Describe the results. If you repeated this experiment very many times, what would you expect the average number of hits to be?

6.33 Required sample size for specified margin of error. A new bone study is being planned that will measure the biomarker TRAP described in Exercise 6.19. Using the value of σ given there, 6.5 U/l, find the sample size required to provide an estimate of the mean TRAP with a margin of error of 1.5 U/l for 95% confidence.

6.34 Adjusting required sample size for dropouts. Refer to the previous exercise. In similar previous studies, about 20% of the subjects drop out before the study is completed. Adjust your sample size requirement so that you will have enough subjects at the end of the study to meet the margin of error criterion.

6.35 Radio poll. A national public radio (NPR) station invites listeners to enter a dispute about a proposed "pay as you throw" waste collection program. The station asks listeners to call in and state how much each 10 gallon bag of trash should cost. A total of 179 listeners call in. The station calculates the 95% confidence interval for the average fee to be \$0.53 to \$1.39. Is this result trustworthy? Explain your answer.

6.36 Accuracy of a laboratory scale. To assess the accuracy of a laboratory scale, a standard weight known to weigh 10 grams is weighed repeatedly. The scale readings are Normally distributed with unknown mean (this mean is 10 grams if the scale has no bias). The standard deviation of the scale readings is known to be 0.0002 gram.

(a) The weight is measured six times. The mean result is 10.0023 grams. Give a 99% confidence interval for the mean of repeated measurements of the weight.

(b) Based on the interval in part (a), do you think the scale is accurate? Explain your answer.

(c) How many measurements must be averaged to get a margin of error of ± 0.0001 with 99% confidence?

6.37 More than one confidence interval. As we prepare to take a sample and compute 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 plan to use several such intervals, however, our confidence that *all* of them will give correct results is less than 95%. Suppose that we plan to 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 will cover the true means? This probability (expressed as a percent) is our overall confidence level for the five simultaneous statements.

(b) Suppose we instead considered individual 99% confidence intervals. Now, what is the overall confidence level for the five simultaneous statements?

(c) Based on the results of parts (a) and (b), how could you keep the overall confidence level near 95% if you were considering 10 simultaneous intervals?

6.2 Tests of Significance

When you complete this section, you will be able to:

- Outline the four steps common to all tests of significance.
- Formulate the null and alternative hypotheses of a significance test.
- Describe a common form for the test statistic in terms of the parameter estimate, its standard deviation, and the hypothesized value.
- Define what a *P*-value is and explain whether a small *P*-value provides evidence for or against the null hypothesis.
- Draw a conclusion from a test of significance based on the test's *P*-value and significance level α.
- Describe the relationship between a level α two-sided significance test for μ and the $1 - \alpha$ confidence interval.

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

Scholarship amount by borrower status. One purpose of Sallie Mae's annual study described in Example 6.4 (page 350) is to allow comparisons of different subgroups. For example, in the latest report, 980 of the 1593 participants (61.5%) did not borrow any money to pay for college. The average scholarship amount among these participants was $3925. The average scholarship among those who did borrow was $4350. The difference of $425 is fairly large, but we know that these numbers are estimates of the population means. If we took different samples, we would get different estimates.

Can we conclude from these data that the average scholarship amounts in these two groups are different? One way to answer this question is to compute the probability of obtaining a difference as large or larger than the observed $425 assuming that, in fact, there is no difference in the population means. This probability is 0.23. Because this probability is not particularly small, we conclude that observing a difference of $425 is not very surprising when the population means are equal. The data do not provide enough evidence for us to conclude that the average scholarship amount for borrowers and non-borrowers differ.

Here is an example with a different conclusion.

EXAMPLE 6.9

Parent income contribution by school type. Sallie Mae's study also reports that the parents' current income contribution among undergraduates going to a four-year public or four-year private college. The parents' contribution averages $4444 among undergraduates at public colleges, while it is $6083 among undergraduates at private schools. Do parents pay more of their current income for undergraduates going to private schools? The observed difference is $1639, but as we learned in the previous example, an observed difference in means is not necessarily sufficient for us to conclude that the population means are different.

Again, we answer this question with a probability calculated under the assumption that there is *no difference in the population means*. The probability is 0.0003 of observing a difference in mean contributions that is $1639 or more when there really is no difference. Because this probability is so small, we have sufficient evidence in the data to conclude that the average current income contribution of parents is higher for undergraduates going to a private school than undergraduates going to a public school.

What are the key steps in these examples?

- We started each with a question about the difference between two means. In Example 6.8, we compare borrowers with nonborrowers. In Example 6.9, we compare undergraduates attending private and public four-year colleges. 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 difference given by the data, $425 in the first case and $1639 in the second, with the value assumed in the question, $0.
- The results of the comparisons are probabilities, 0.23 in the first case and 0.0003 in the second.

The 0.23 probability is not particularly small, so we have limited evidence to question the possibility that the true difference is zero. In the second case, however, the probability is very small. Something that happens with probability 0.0003 occurs only about 3 times out of 10,000. In this case we have two possible explanations:

1. We have observed something that is very unusual.
2. The assumption that underlies the calculation, no difference in mean balance, is not true.

Because this probability is so small, we prefer the second conclusion: the average current income contribution from parents for undergraduates attending public colleges and for undergraduates attending private colleges is different, with the private school group contribution higher than that of the public school group.

The probabilities in Examples 6.8 and 6.9 are measures of the compatibility of the data (a difference in means of $425 and $1639) 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 from Figure 6.7 that we should not be particularly surprised to observe the difference $425, but the difference $1639 in Figure 6.8 is clearly an unusual observation. We will now consider some of the formal aspects of significance testing.

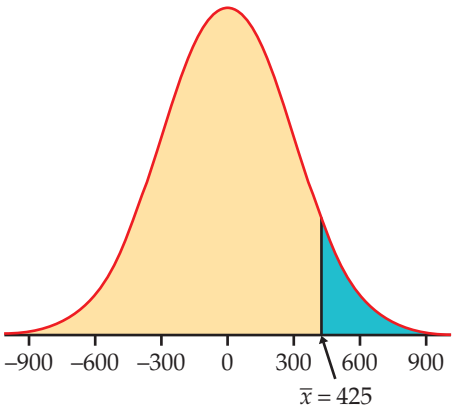


FIGURE 6.7 Comparison of the sample mean in Example 6.8 with the null hypothesized value 0.

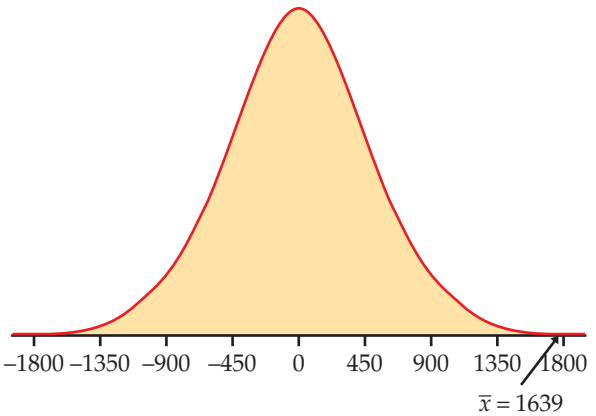


FIGURE 6.8 Comparison of the sample mean in Example 6.9 with the null hypothesized value 0.

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 population 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. Often, 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 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.

Because H_a expresses the effect that we hope to find evidence *for,* we will sometimes begin with H_a and then set up H_0 as the statement that the hoped-for effect is not present. Stating H_a, however, 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.

one-sided or two-sided alternatives

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

6.38 Dining court survey. The dining court closest to your university residence has been redesigned. A survey is planned to assess whether or not students think that the new design is an improvement. It will contain eight questions; a seven-point scale will be used for the answers, with scores less than 4 favoring the previous design and scores greater than 4 favoring the new design (to varying degrees). The average of these eight questions will be used as the student's response. State the null and alternative hypotheses you would use for examining whether or not the new design is viewed more favorably.

6.39 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. One researcher believes that her scanner is not giving accurate readings. To assess this, the researcher uses an object called a "phantom" that has known mineral density $\mu = 1.4$ grams per square centimeter. The researcher 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. 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

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

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.

Let's return to our comparison of the scholarship amount among borrowers and nonborrowers and specify the hypotheses as well as calculate the test statistic.

EXAMPLE 6.10

Average scholarship amount of borrowers and nonborrowers: The hypotheses. In Example 6.8, the hypotheses are stated in terms of the difference in the average scholarship amount between borrowers and nonborrowers:

H_0: there is no difference in the population means

H_a: there is a difference in the population means

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

AU: add page x-refs?

We can also state the null hypothesis as H_0: the true mean difference is 0. This statement makes it more clear that the hypothesized value for this comparison of average scholarship amounts is 0.

EXAMPLE 6.11

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.

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

LOOK BACK

Normal distribution, p. 56

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 6.44 on page 370), and this is the criterion used in most applications involving the law. But because not all test statistics are Normal, we use the language of probability to express the meaning of a test statistic.

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.

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.

AU: add page x-ref?

In Example 6.8, we want to know if the average scholarship amount for borrowers differs from the average scholarship amount for non-borrowers. The difference we calculated based on our sample is \$425, which corresponds to 1.20 standard deviations away from zero—that is, $z = 1.20$. 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 1.20.

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

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

versus

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

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

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

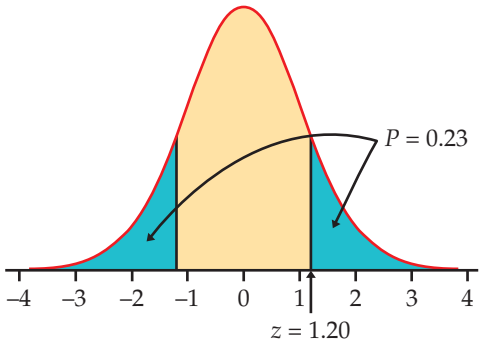


FIGURE 6.9 The *P*-value, Example 6.12. The *P*-value is the probability (when H_0 is true) that $\bar{x}$ takes a value as extreme or more extreme than the actual observed value, $z = 1.20$. Because the alternative hypothesis is two-sided, we use both tails of the distribution.

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.

USE YOUR KNOWLEDGE

6.40 Normal curve and the *P*-value. A test statistic for a two-sided significance test for a population mean is $z = 2.47$. 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.

6.41 More on the Normal curve and the *P*-value. A test statistic for a two-sided significance test for a population mean is $z = -1.57$. 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 is called the **significance level**. It is commonly denoted by α (the Greek letter alpha). 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 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 $\alpha = 0.05$ level but is not significant at the $\alpha = 0.01$ level. We discuss this in more detail at the end of this section.

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 ($z = 1.20$, $P = 0.23$)."

This statement does not mean that we conclude that the null hypothesis is true, only that the level of evidence we require to reject the null hypothesis is not met. Our criminal court system follows a similar procedure in which a defendant is presumed innocent (H_0) until proven guilty. If the level of evidence presented is not strong enough for the jury to find the defendant guilty beyond a reasonable doubt, the defendant is acquitted. Acquittal does not imply innocence, only that the degree of evidence was not strong enough to prove guilt.

If the P-value is small, we reject the null hypothesis. Here is the conclusion for our second example.

AU: add page x-ref?

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

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

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 ($z = 3.81$, $P = 0.0001$)."

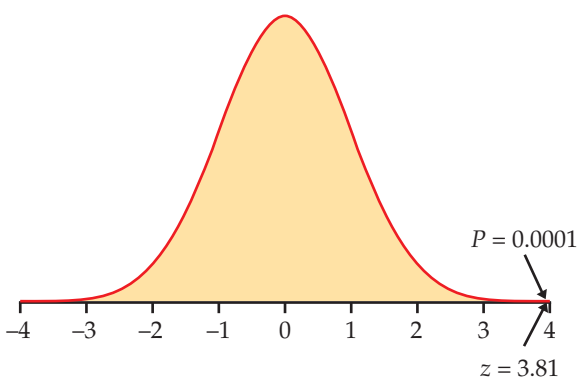


FIGURE 6.10 The P-value, Example 6.14. The P-value is the probability (when H_0 is true) that $\bar{x}$ takes a value as extreme or more extreme than the actual observed value, $z = 3.81$. We look at only the right tail because we are considering the one-sided ($>$) alternative.

USE YOUR KNOWLEDGE

6.42 Finding significant *z*-scores. Consider a two-sided significance test for a population mean.

(a) Sketch a Normal curve similar to that shown in Figure 6.9 (page 367), but find the value z such that $P = 0.05$.

(b) Based on your curve from part (a), what values of the z statistic are statistically significant at the $\alpha = 0.05$ level?

6.43 More on finding significant *z*-scores. Consider a one-sided significance test for a population mean, where the alternative is "greater than."

(a) Sketch a Normal curve similar to that shown in Figure 6.10, but find the value z such that $P = 0.05$.

(b) Based on your curve from part (a), what values of the z statistic are statistically significant at the $\alpha = 0.05$ level?

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

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.

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. These steps require judgment based on a sound understanding of this type of inference.

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

$$H_0\text{: } \mu = \mu_0$$

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:

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

distribution of sample mean, p. 298

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

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

central limit theorem, p. 298

Again recall from Chapter 5 that, if the population is Normal, then $\bar{x}$ 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. We'll assume that we're in one of these two settings for now.

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:

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

AU: Please check. I think you just wanted the one arrow moved down with both items follow it. Please confirm.

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.

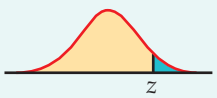
z TEST FOR A POPULATION MEAN

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**

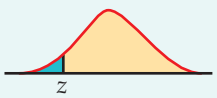
$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

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

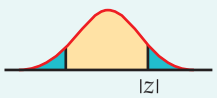
H_a: $\mu > \mu_0$ is $P(Z \geq z)$



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

Joe Raedle/Getty Images

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

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

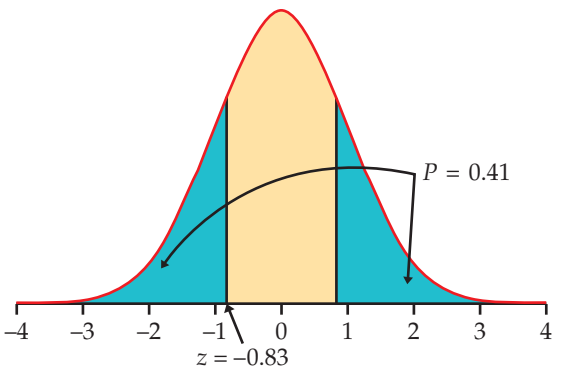


FIGURE 6.11 Sketch of the *P*-value calculation for the two-sided test, Example 6.15. The test statistic is $z = -0.83$.

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.

This z test requires that the 100 students in the sample are an SRS from the population of students at your university. We will assume that the students in the sample were selected in a proper random manner. We'll also assume that $n = 100$ is sufficiently large that we can rely on the central limit theorem to assure us that the *P*-value based on the Normal distribution will be a good approximation.

The data in Example 6.15 do *not* establish that the mean consumption μ for the students at your university is 298 calories. We sought evidence that μ differed from 298 and failed to find convincing evidence. That is all we can say. No doubt the mean amount at your university is not exactly equal to 298 calories. 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.*

AU/DE/PE: Caution icon?

EXAMPLE 6.16

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.

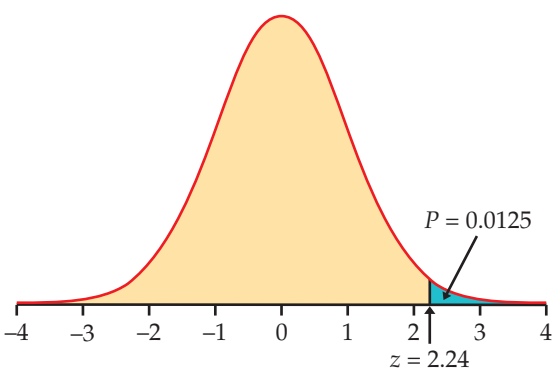


FIGURE 6.12 Sketch of the P-value calculation for the one-sided test, Example 6.16. The test statistic is $z = 2.24$.

USE YOUR KNOWLEDGE

6.45 Computing the test statistic and *P*-value. You will perform a significance test of $H_0: \mu = 30$ based on an SRS of $n = 49$. Assume that $\sigma = 14$.

(a) If $\bar{x} = 33.5$, what is the test statistic z?

(b) What is the P-value if $H_a: \mu > 30$?

(c) What is the P-value if $H_a: \mu \neq 30$?

6.46 Testing a random number generator. Statistical software often has a "random number generator" that is supposed to produce numbers uniformly distributed between 0 and 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 $\bar{x} = 0.469$ and $s = 0.286$. Because the sample is reasonably large, take the population standard deviation also to be $\sigma = 0.286$. 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 Section 6.1. We constructed an interval that would include the true value of μ with a specified probability C. Suppose that we use a 95% confidence interval ($C = 0.95$). Then the values of μ_0 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

Voisin/ Phanie/Science Source

PBTEST

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

$$15.84 \quad 15.33 \quad 15.58$$

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.

We can compute a 99% confidence interval for the same data to get a likely range for the actual mean concentration μ of this specimen.

EXAMPLE 6.18

PBTEST

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

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

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.

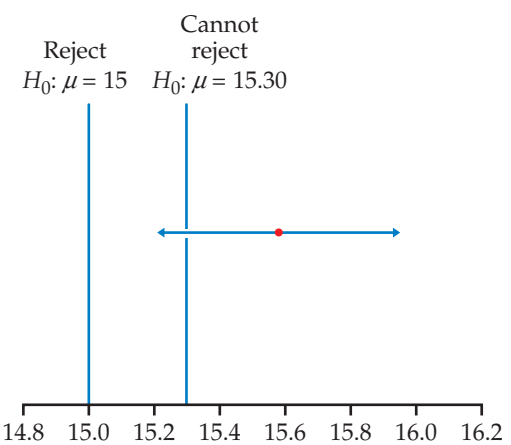


FIGURE 6.13 The link between two-sided significance tests and confidence intervals. For the study described in Examples 6.17 and 6.18, values of μ falling outside a 99% confidence interval can be rejected at the 1% significance level; values falling inside the interval cannot be rejected. This holds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

USE YOUR KNOWLEDGE

6.47 Two-sided significance tests and confidence intervals. The P-value for a two-sided test of the null hypothesis H_0: $\mu = 30$ is 0.037.

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

(b) Does the 99% confidence interval include the value 30? Explain.

6.48 More on two-sided tests and confidence intervals. A 95% confidence interval for a population mean is (29, 58).

(a) Can you reject the null hypothesis that $\mu = 50$ against the two-sided alternative at the 5% significance level? Explain.

(b) Can you reject the null hypothesis that $\mu = 60$ against the two-sided alternative at the 5% significance level? Explain.

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

$$H_0: \mu = 485$$
$$H_a: \mu > 485$$

AU: add page x-ref?

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.

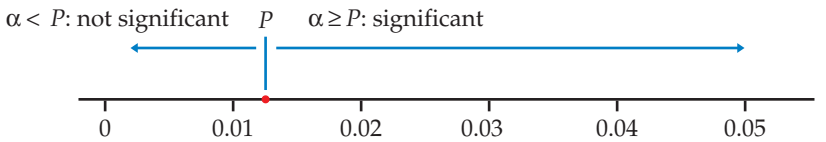


FIGURE 6.14 Link between the *P*-value and the significance level α. An outcome with *P*-value *P* is significant at all levels α at or above *P* and is not significant at smaller levels α.

critical value

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 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 or a calculator 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.

EXAMPLE 6.20

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.

USE YOUR KNOWLEDGE

6.49 *P*-value and the significance level. The *P*-value for a significance test is 0.033.

(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 how you determined your answers to parts (a) and (b).

6.50 More on *P*-value and the significance level. The *P*-value for a significance test is 0.069.

(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 how you determined your answers to parts (a) and (b).

6.51 One-sided and two-sided *P*-values. The *P*-value for a two-sided significance test is 0.076.

(a) State the *P*-values for the two 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**:

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

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 6.38 and 6.39, see page 364; for Exercises 6.40 and 6.41, see page 367; for Exercises 6.42 through 6.44, see page 370; for Exercises 6.45 and 6.46, see pages 374–375; for Exercises 6.47 and 6.48, see page 377; and for Exercises 6.49 through 6.51, see page 378.

6.52 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 researcher tests the following null hypothesis: H_0: $\bar{x} = 23$.

(b) A random sample of size 30 is taken from a population that is assumed to have a standard deviation of 5. The standard deviation of the sample mean is 5/30.

(c) A study with $\bar{x} = 45$ reports statistical significance for H_a: $\mu > 50$.

(d) A researcher tests the hypothesis H_0: $\mu = 350$ and concludes that the population mean is equal to 350.

6.53 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 500.

(b) A test preparation company wants to test that the average score of its students on the ACT is better than the national average score of 21.2. The company states its null hypothesis to be H_0: $\mu > 21.2$.

(c) A study summary says that the results are statistically significant and the P-value is 0.98.

(d) The z test statistic is equal to 0.018. Because this is less than $\alpha = 0.05$, the null hypothesis was rejected.

6.54 Determining hypotheses. State the appropriate null hypothesis H_0 and alternative hypothesis H_a in each of the following cases.

(a) A 2015 study reported that 96% of students owned a cell phone. You plan to take an SRS of students to see if the percent has increased.

(b) The examinations in a large freshman chemistry class are scaled after grading so that the mean score is 75. The professor thinks that students who attend early-morning recitation sections will have a higher mean score than the class as a whole. Her students in these sections this semester can be considered a sample from the population of all students who might attend an early-morning section, so she compares their mean score with 75.

(c) The student newspaper at your college recently changed the format of its opinion page. You want to test whether students find the change an improvement. You take a random sample of students and select those who regularly read the newspaper. They are asked to indicate their opinions on the changes using a five-point scale: -2 if the new format is much worse than the old, -1 if the new format is somewhat worse than the old, 0 if the new format is the same as the old, $+1$ if the new format is somewhat better than the old, and $+2$ if the new format is much better than the old.

6.55 More on determining hypotheses. State the null hypothesis H_0 and the alternative hypothesis H_a in each case. Be sure to identify the parameters that you use to state the hypotheses.

(a) A university gives credit in first-year calculus to students who pass a placement test. The mathematics department wants to know if students who get credit in this way differ in their success with second-year calculus. Scores in second-year calculus are scaled so the average each year is equivalent to a 77. This year, 21 students who took second-year calculus passed the placement test.

(b) Experiments on learning in animals sometimes measure how long it takes a mouse to find its way through a maze. The mean time is 20 seconds for one particular maze. A researcher thinks that playing rap music will cause the mice to complete the maze more slowly. She measures how long each of 12 mice takes with the rap music as a stimulus.

(c) The average square footage of one-bedroom apartments in a new student-housing development is advertised to be 880 square feet. A student group thinks that the apartments are smaller than advertised. They hire an engineer to measure a sample of apartments to test their suspicion.

6.56 Even more on determining hypotheses. In each of the following situations, state an appropriate null hypothesis H_0 and alternative hypothesis H_a. Be sure to identify the parameters that you use to state the hypotheses. (We have not yet learned how to test these hypotheses.)

(a) A sociologist asks a large sample of high school students which television channel they like best. She suspects that a higher percent of males than of females will name MTV as their favorite channel.

(b) An education researcher randomly divides sixth-grade students into two groups for physical education class. He teaches both groups basketball skills, using the same methods of instruction in both classes. He encourages Group A with compliments and other positive behavior but acts cool and neutral toward Group B. He hopes to show that positive teacher attitudes result in a higher mean score on a test of basketball skills than do neutral attitudes.

(c) An education researcher believes that, among college students, there is a negative correlation between time spent at social network sites and self-esteem, measured on a 0 to 100 scale. To test this, she gathers social-networking information and self-esteem data from a sample of students at your college.

6.57 Translating research questions into hypotheses. Translate each of the following research questions into appropriate H_0 and H_a.

(a) U.S. Census Bureau data show that the mean household income in the area served by a shopping mall is $42,800 per year. A market research firm questions shoppers at the mall to find out whether the mean household income of mall shoppers is higher than that of the general population.

(b) Last year, your online registration technicians took an average of 0.4 hour to respond to trouble calls from

students trying to register. Do this year's data show a different average response time?

6.58 Computing the *P*-value. A test of the null hypothesis H_0: $\mu = \mu_0$ gives test statistic $z = 1.89$.

(a) What is the *P*-value if the alternative is H_a: $\mu > \mu_0$?

(b) What is the *P*-value if the alternative is H_a: $\mu < \mu_0$?

(c) What is the *P*-value if the alternative is H_a: $\mu \neq \mu_0$?

6.59 More on computing the *P*-value. A test of the null hypothesis H_0: $\mu = \mu_0$ gives test statistic $z = -1.33$.

(a) What is the *P*-value if the alternative is H_a: $\mu > \mu_0$?

(b) What is the *P*-value if the alternative is H_a: $\mu < \mu_0$?

(c) What is the *P*-value if the alternative is H_a: $\mu \neq \mu_0$?

6.60 Timing of food intake and weight loss. A study found that a large group of late lunch eaters lost less weight over a 20-week observation period than a large group of early lunch eaters ($P = 0.002$).[16] Explain what this $P = 0.002$ means in a way that could be understood by someone who has not studied statistics.

6.61 Average starting salary. Refer to Exercise 6.22 (page 358). Use the information presented in the exercise to test that the average income of graduates from your institution is different from the national average ($\alpha = 0.01$). Write a short paragraph summarizing your conclusions.

6.62 Change in consumption of sweet snacks? Refer to Exercise 6.23 (page 358). A similar study performed four years earlier reported the average consumption of sweet snacks among healthy weight children aged 12 to 19 years to be 369.4 kilocalaries per day (kcal/d). Does this current study suggest a change in the average consumption? Perform a significance test using the 5% significance level. Write a short paragraph summarizing the results.

6.63 Peer pressure and choice of major. A study followed a cohort of students entering a business/economics program.[17] All students followed a common track during the first three semesters and then chose to specialize in either business or economics. Through a series of surveys, the researchers were able to classify roughly 50% of the students as either peer driven (ignored abilities and chose major to follow peers) or ability driven (ignored peers and chose major based on ability). When looking at entry wages after graduation, the researchers conclude that a peer-driven student can expect an average wage that is 13% less than that of an ability-driven student. The report states that the significance level is $P = 0.09$. Can you be confident of the researchers' conclusion statement regarding the wage decrease? Explain your answer.

6.64 Symbol of wealth in ancient China? Every society has its own symbols of wealth and prestige. In ancient China, it appears that owning pigs was such a symbol. Evidence comes from examining burial sites. If the skulls of sacrificed pigs tend to appear along with expensive ornaments, that suggests that the pigs, like the ornaments, signal the wealth and prestige of the person buried. A study of burials from around 3500 B.C. concluded that "there are striking differences in grave goods between burials with pig skulls and burials without them... A test indicates that the two samples of total artifacts are significantly different at the 0.01 level."[18] Explain clearly why "significantly different at the 0.01 level" gives good reason to think that there really is a systematic difference between burials that contain pig skulls and those that lack them.

6.65 Alcohol awareness among college students. A study of alcohol awareness among college students reported a higher awareness for students enrolled in a health and safety class than for those enrolled in a statistics class.[19] The difference is described as being statistically significant. Explain what this means in simple terms and offer an explanation for why the health and safety students had a higher mean score.

6.66 Change in eighth-grade average mathematics score. A report based on the 2015 National Assessment of Educational Progress (NAEP)[20] states that the average score on their mathematics test for eighth-grade students attending public schools is significantly higher than in 2011. The report also states that the average score for eighth-grade students attending private schools is not significantly different from the average score in 2011. A footnote states that comparisons are determined by two-sided statistical tests with 0.05 as the level of significance. Explain what this footnote means in language understandable to someone who knows no statistics. Do not use the word "significance" in your answer.

6.67 More on change in eighth-grade average mathematics score. Refer to the previous exercise. On the basis of the NAEP study, a friend who works for the school newspaper wants to report that between 2011 and 2013 the average mathematics score improved for students attending public schools but stayed the same for students attending private schools. Do you agree with this statement? Explain your answer.

6.68 Background television in homes of U.S. children. In one study, U.S. parents were surveyed to determine the amount of background television

their children were exposed to. A total of $n = 1454$ families with one child between the ages of 8 months and 8 years participated.[21] For those families in which the caregiver had a high school degree or less, the child was exposed to an average of 313.0 minutes of background television per day. For those families in which the caregiver had some college or a college degree, the child was exposed to an average of 218.8 minutes per day. These average times were reported to be significantly different with $P < 0.05$. The actual P-value is 0.003. Explain why the actual P-value is more informative than the statement of significance at the 0.05 level.

6.69 Sleep quality and elevated blood pressure. A study looked at $n = 238$ adolescents, all free of severe illness.[22] Subjects wore a wrist actigraph, which allowed the researchers to estimate sleep patterns. Those subjects classified as having low sleep efficiency had an average systolic blood pressure that was 5.8 millimeters of mercury (mm Hg) higher than that of other adolescents. The standard deviation of this difference is 1.4 mm Hg. Based on these results, test whether this difference is significant at the 0.01 level.

6.70 Are the pine trees randomly distributed from north to south? In Example 6.1 (page 342), we looked at the distribution of longleaf pine trees in the Wade Tract. One way to formulate hypotheses about whether or not the trees are randomly distributed in the tract is to examine the average location in the north–south direction. The values range from 0 to 200, so if the trees are uniformly distributed in this direction, any difference from the middle value (100) should be due to chance variation. The sample mean for the 584 trees in the tract is 99.74. A theoretical calculation based on the assumption that the trees are uniformly distributed gives a standard deviation of 58. Carefully state the null and alternative hypotheses in terms of this variable. Note that this requires that you translate the research question about the random distribution of the trees into specific statements about the mean of a probability distribution. Test your hypotheses, report your results, and write a short summary of what you have found.

6.71 Are the pine trees randomly distributed from east to west? Answer the questions in the previous exercise for the east–west direction, for which the sample mean is 113.8.

6.72 Who is the author? Statistics can help decide the authorship of literary works. Sonnets by a certain Elizabethan poet are known to contain an average of $\mu = 8.9$ new words (words not used in the poet's other works). The standard deviation of the number of new words is $\sigma = 2.5$. Now a manuscript with six new sonnets has come to light, and scholars are debating whether it is the poet's work. The new sonnets contain an average of $\bar{x} = 10.2$ words not used in the poet's known works. We expect poems by another author to contain more new words, so to see if we have evidence that the new sonnets are not by our poet we test

$$H_0: \mu = 8.9$$
$$H_a: \mu > 8.9$$

Give the z test statistic and its P-value. What do you conclude about the authorship of the new poems?

6.73 Attitudes toward school. The Survey of Study Habits and Attitudes (SSHA) is a psychological test that measures the motivation, attitude toward school, and study habits of students. Scores range from 0 to 200. The mean score for U.S. college students is about 95, and the standard deviation is about 20. A teacher who suspects that older students have better attitudes toward school gives the SSHA to 25 students who are at least 30 years of age. Their mean score is $\bar{x} = 103.3$.

(a) Assuming that $\sigma = 30$ for the population of older students, carry out a test of

$$H_0: \mu = 95$$
$$H_a: \mu > 95$$

Report the P-value of your test, and state your conclusion clearly.

(b) Your test in part (a) required two important assumptions in addition to the assumption that the value of σ is known. What are they? Which of these assumptions is most important to the validity of your conclusion in part (a)?

6.74 Nutritional intake among Canadian high-performance athletes. Since previous studies have reported that elite athletes are often deficient in their nutritional intake (for example, total calories, carbohydrates, protein), a group of researchers decided to evaluate Canadian high-performance athletes.[23] A total of $n = 324$ athletes from eight Canadian sports centers participated in the study. One reported finding was that the average caloric intake among the $n = 201$ women was 2403.7 kilocalories per day (kcal/d). The recommended amount is 2811.5 kcal/d. Is there evidence that female Canadian athletes are deficient in caloric intake?

(a) State the appropriate H_0 and H_a to test this.

(b) Assuming a standard deviation of 880 kcal/d, carry out the test. Give the P-value, and then interpret the result in plain language.

6.75 Are the measurements similar? Refer to Exercise 6.30 (page 360). In addition to the computer's calculations of miles per gallon, the driver also recorded the miles per gallon by dividing the miles driven by the number of gallons at each fill-up. The following data are the differences between the computer's and the driver's calculations for that random sample of 20 records. The driver wants to determine if these calculations are different. Assume that the standard deviation of a difference is $\sigma = 3.0$. MPGDIFF

5.0	6.5	−0.6	1.7	3.7	4.5	8.0	2.2	4.9	3.0
4.4	0.1	3.0	1.1	1.1	5.0	2.1	3.7	−0.6	−4.2

(a) State the appropriate H_0 and H_a to test this suspicion.

(b) Carry out the test. Give the P-value, and then interpret the result in plain language.

6.76 Impact of $\bar{x}$ on significance. The *Statistical Significance* applet illustrates statistical tests with a fixed level of significance for Normally distributed data with known standard deviation. Open the applet and keep the default settings for the null ($\mu = 0$) and the alternative ($\mu > 0$) hypotheses, the sample size ($n = 10$), the standard deviation ($\sigma = 1$), and the significance level ($\alpha = 0.05$). In the "I have data, and the observed $\bar{x}$ is $\bar{x} =$" box, enter the value 1. Is the difference between $\bar{x}$ and μ_0 significant at the 5% level? Repeat for $\bar{x}$ equal to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. Make a table giving $\bar{x}$ and the results of the significance tests. What do you conclude?

6.77 Effect of changing α on significance. Repeat the previous exercise with significance level $\alpha = 0.01$. How does the choice of α affect which values of $\bar{x}$ are far enough away from μ_0 to be statistically significant?

6.78 Changing to a two-sided alternative. Repeat the previous exercise but with the two-sided alternative hypothesis. How does this change affect which values of $\bar{x}$ are far enough away from μ_0 to be statistically significant at the 0.01 level?

6.79 Changing the sample size. Refer to Exercise 6.76. Suppose that you increase the sample size n from 10 to 50. Again, make a table giving $\bar{x}$ and the results of the significance tests at the 0.05 significance level. What do you conclude?

6.80 Impact of $\bar{x}$ on the *P*-value. We can also study the P-value using the *Statistical Significance* applet. Reset the applet to the default settings for the null ($\mu = 0$) and the alternative ($\mu > 0$) hypotheses, the sample size ($n = 10$), the standard deviation ($\sigma = 1$), and the significance level ($\alpha = 0.05$). In the "I have data, and the observed $\bar{x}$ is $\bar{x} =$" box, enter the value 1. What is the P-value? It is shown at the top of the blue vertical line. Repeat for $\bar{x}$ equal to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. Make a table giving $\bar{x}$ and P-values. How does the P-value change as $\bar{x}$ moves farther away from μ_0?

6.81 Changing to a two-sided alternative, continued. Repeat the previous exercise but with the two-sided alternative hypothesis. How does this change affect the P-values associated with each $\bar{x}$? Explain why the P-values change in this way.

6.82 Other changes and the *P*-value. Refer to the previous exercise.

(a) What happens to the P-values when you change the significance level α to 0.01? Explain the result.

(b) What happens to the P-values when you change the sample size n from 10 to 50? Explain the result.

6.83 Understanding levels of significance. Explain in plain language why a significance test that is significant at the 1% level must always be significant at the 5% level.

6.84 More on understanding levels of significance. You are told that a significance test is significant at the 5% level. From this information, can you determine whether or not it is significant at the 1% level? Explain your answer.

6.85 Test statistic and levels of significance. Consider a significance test for a null hypothesis versus a two-sided alternative. Give a value of z that will give a result significant at the 1% level but not at the 0.5% level.

6.86 Using Table D to find a *P*-value. You have performed a two-sided test of significance and obtained a value of $z = 2.08$. Use Table D to find the approximate P-value for this test.

6.87 More on using Table D to find a *P*-value. You have performed a one-sided test of significance and obtained a value of $z = 1.03$. Use Table D to find the approximate P-value for this test when the alternative is greater than.

6.88 Using Table A and Table D to find a *P*-value. Consider a significance test for a null hypothesis versus a two-sided alternative. Between what values from Table D does the P-value for an outcome $z = 1.88$ lie? Calculate the P-value using Table A and verify that it lies between the values you found from Table D.

6.89 More on using Table A and Table D to find a *P*-value. Refer to the previous exercise. Find the P-value for $z = -1.88$.

6.3 Use and Abuse of Tests

When you complete this section, you will be able to:

- Explain why it is important to report the *P*-value and not just report whether the result is statistically significant or not.
- Discriminate between practical (or scientific) significance and statistical significance.
- Identify poorly designed studies where formal statistical inference is suspect.
- Understand the consequences of searching solely for statistical significance, whether through the investigation of multiple tests or by identifying and testing using the same data set.

Carrying out a test of significance is often quite simple, especially if the *P*-value is given effortlessly by a computer. Using tests wisely is not so simple. Each test is valid only in certain circumstances, with properly produced data being particularly important.

The z test, for example, should bear the same warning label that was attached in Section 6.1 to the corresponding confidence interval (page 355). Similar warnings accompany the other tests that we will learn. There are additional caveats that concern tests more than confidence intervals, enough to warrant this separate section. Some hesitation about the unthinking use of significance tests is a sign of statistical maturity.

The reasoning of significance tests has appealed to researchers in many fields, so that tests are widely used to report research results. In this setting H_a is a "research hypothesis" asserting that some effect or difference is present. The null hypothesis H_0 says that there is no effect or no difference. A low *P*-value represents good evidence that the research hypothesis is true. Here are some comments on the use of significance tests, with emphasis on their use in reporting scientific research.

Choosing a level of significance

The intention of a test of significance is to give a clear statement of the degree of evidence provided by the sample against the null hypothesis. The *P*-value does this. It is common practice to report *P*-values and to describe results as statistically significant whenever $P \leq 0.05$. *However, there is no sharp border between "significant" and "not significant," only increasingly strong evidence as the P-value decreases.* Having the *P*-value with a description of the effect that we have found allows us to draw better conclusions from our data.

EXAMPLE 6.21

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.

Here is another example where the *P*-value provides useful information beyond that provided by the statement that we reject or fail to reject the null hypothesis.

EXAMPLE 6.22

More on information provided by the *P*-value. We have a test statistic of $z = -4.66$ for a two-sided significance test on a population mean. Software tells us that the *P*-value is 0.000003. This means that there are 3 chances in 1,000,000 of observing a sample mean this far or farther away from the null hypothesized value of μ. This kind of event is virtually impossible if the null hypothesis is true. There is no ambiguity in the result; we can clearly reject the null hypothesis.

We frequently report small *P*-values such as that in the previous example as $P < 0.001$. This corresponds to a chance of 1 in 1000 and is sufficiently small to lead us to a clear rejection of the null hypothesis.

One reason for the common use of $\alpha = 0.05$ is the great influence of Sir R. A. Fisher, the inventor of formal statistical methods for analyzing experimental data. Here is his opinion on choosing a level of significance: "A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance."[24]

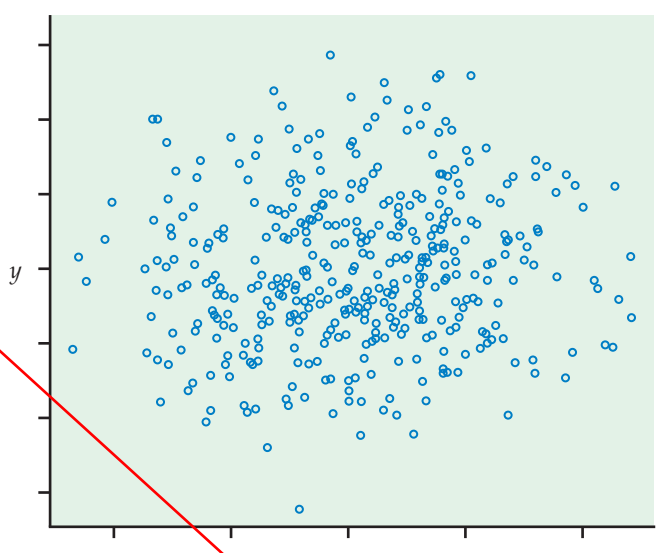
What statistical significance does not mean

When a null hypothesis ("no effect" or "no difference") can be rejected at the usual level $\alpha = 0.05$, there is good evidence that an effect is present. That effect, however, can be extremely small. *When large samples are available, even tiny deviations from the null hypothesis will be statistically significant.*

EXAMPLE 6.23

It's significant but is it important? Suppose that we are testing the null hypothesis of no correlation between two variables. With 400 observations, an observed correlation of only $r = 0.1$ is significant evidence at the $\alpha = 0.05$ level that the correlation in the population is not zero. Figure 6.15 is an

FIGURE 6.15 Scatterplot of $n = 400$ observations with an observed correlation of 0.10. There is not a strong association between the two variables even though there is significant evidence ($P < 0.05$) that the population correlation is not zero.



AU: add x-ref to Example 6.23?

example of 400 (x, y) pairs that have an observed correlation of 0.10. The low significance level does *not* mean that there is a strong association, only that there is strong evidence of some association. The proportion of the variability in one of the variables explained by the other is $r^2 = 0.01$, or 1%.

For practical purposes, we might well decide to ignore this association. *Statistical significance is not the same as practical significance.* Statistical significance rarely tells us about the importance of the experimental results. This depends on the context of the experiment.

The remedy for attaching too much importance to statistical significance is to pay attention to the actual experimental results as well as to the *P*-value. Plot your data and examine them carefully. Beware of outliers. *The user of statistics who feeds the data to a computer without exploratory analysis will often be embarrassed.* It is usually wise to give a confidence interval for the parameter in which you are interested. Confidence intervals are not used as often as they should be, while tests of significance are overused.

USE YOUR KNOWLEDGE

6.90 Is it significant? More than 200,000 people worldwide take the GMAT examination each year when they apply for MBA programs. Their scores vary Normally with mean $\mu = 540$ and standard deviation $\sigma = 100$. One hundred students go through a rigorous training program designed to raise their GMAT scores. Test the following hypotheses about the training program

$$H_0: \mu = 540$$
$$H_a: \mu > 540$$

in each of the following situations.

(a) The students' average score is $\bar{x} = 556.4$. Is this result significant at the 5% level?

(b) Now suppose that the average score is $\bar{x} = 556.5$. Is this result significant at the 5% level?

(c) Explain how you would reconcile this difference in significance, especially if any increase greater than 15 points is considered a success.

Don't ignore lack of significance

There is a tendency to conclude that there is no effect whenever a *P*-value fails to attain the usual 5% standard. A provocative editorial in the *British Medical Journal* entitled "Absence of Evidence Is Not Evidence of Absence" deals with this issue.[25] Here is one of the examples they cite.

EXAMPLE 6.24

Interventions to reduce HIV-1 transmission. A randomized trial of interventions for reducing transmission of HIV-1 reported an incident rate ratio of 1.00, meaning that the intervention group and the control group both had the same rate of HIV-1 infection. The 95% confidence interval was reported

as 0.63 to 1.58.[26] The editorial notes that a summary of these results that says the intervention has no effect on HIV-1 infection is misleading. The confidence interval indicates that the intervention may be capable of achieving a 37% decrease in infection; it might also be harmful and produce a 58% increase in infection. Clearly, more data are needed to distinguish between these possibilities.

The situation can be worse. Research in some fields has rarely been published unless significance at the 0.05 level is attained.

EXAMPLE 6.25

Journal survey of reported significance results. A survey of four journals published by the American Psychological Association showed that of 294 articles using statistical tests, only eight reported results that did not attain the 5% significance level.[27] It is very unlikely that these were the only eight studies of scientific merit that did not attain significance at the 0.05 level. Manuscripts describing other studies were likely rejected because of a lack of statistical significance or never submitted in the first place due to the expectation of rejection.

In some areas of research, small effects that are detectable only with large sample sizes can be of great practical significance. Data accumulated from a large number of patients taking a new drug may be needed before we can conclude that there are life-threatening consequences for a small number of people.

On the other hand, sometimes a meaningful result is not found significant.

EXAMPLE 6.26

A meaningful but statistically insignificant result. A sample of size 10 gave a correlation of $r = 0.5$ between two variables. The P-value is 0.102 for a two-sided significance test. In many situations, a correlation this large would be interesting and worthy of additional study. When it takes a lot of effort (say, in terms of time or money) to obtain samples, researchers often use small studies like these as pilot projects to gain interest from various funding sources. With financial support, a larger, more powerful study can then be run.

Another important aspect of planning a study is to verify that the test you plan to use does have high probability of detecting an effect of the size you hope to find. This probability is the *power* of the test. Power calculations are discussed in Section 6.4.

design of experiments, p. 171

Statistical inference is not valid for all sets of data

In Chapter 3, we learned that badly designed surveys or experiments often produce invalid results. *Formal statistical inference cannot correct basic flaws in the design.*

EXAMPLE 6.27

English vocabulary and studying a foreign language. There is no doubt that there is a significant difference in English vocabulary scores between high school seniors who have studied a foreign language and those who have not. But because the effect of actually studying a language is confounded with the differences between students who choose language study and those who do not, this statistical significance is hard to interpret. The most plausible explanation is that students who were already good at English chose to study another language. A randomized comparative experiment would isolate the actual effect of language study and so make significance meaningful. Do you think it would be ethical to do such a study?

confounding, p. 150

Tests of significance and confidence intervals are based on the laws of probability. Randomization in sampling or experimentation ensures that these laws apply. But we must often analyze data that do not arise from randomized samples or experiments. *To apply statistical inference to such data, we must have confidence in a probability model for the data.* The diameters of successive holes bored in auto engine blocks during production, for example, may behave like independent observations from a Normal distribution. We can check this probability model by examining the data. If the Normal distribution model appears approximately correct, we can apply the methods of this chapter to do inference about the process mean diameter μ.

USE YOUR KNOWLEDGE

6.91 Home security systems. A recent TV advertisement for home security systems said that homes without an alarm system are three times more likely to be broken into. Suppose that this conclusion was obtained by examining an SRS of police records of break-ins and determining whether the percent of homes with alarm systems was significantly smaller than 50%. Explain why the significance of this study is suspect and propose an alternative study that would help clarify the importance of an alarm system.

Beware of searching for significance

Statistical significance is an outcome much desired by researchers. It means (or ought to mean) that you have found an effect that you were looking for. *The reasoning behind statistical significance works well if you decide what effect you are seeking, design an experiment or sample to search for it, and use a test of significance to weigh the evidence you get.* But because a successful search for a new scientific phenomenon often ends with statistical significance, it is all too tempting to make significance itself the object of the search. There are several ways to do this, none of them acceptable in polite scientific society.

EXAMPLE 6.28

Genomics studies. In genomics experiments, it is common to assess the differences in expression for tens of thousands of genes. If each of these genes was examined separately and statistical significance declared for all that had *P*-values that pass the 0.05 standard, we would have quite a mess. In the absence of any real biological effects, we would expect that, by chance alone, approximately 5% of these tests will show statistical significance. Much research in genomics is directed toward appropriate ways to deal with this situation.[28]

We do not mean that searching data for suggestive patterns is not proper scientific work. It certainly is. Many important discoveries have been made by accident rather than by design. Exploratory analysis of data is an essential part of statistics. We do mean that the usual reasoning of statistical inference does not apply when the search for a pattern is successful. *You cannot legitimately test a hypothesis on the same data that first suggested that hypothesis.* The remedy is clear. Once you have a hypothesis, design a study to search specifically for the effect you now think is there. If the result of this study is statistically significant, you have real evidence.

SECTION 6.3 SUMMARY

- *P*-values are more informative than the reject-or-not result of a level α test. Beware of placing too much weight on traditional values of α, such as $\alpha = 0.05$.
- Very small effects can be highly significant (small *P*), especially when a test is based on a large sample. A statistically significant effect need not be practically important. Plot the data to display the effect you are seeking, and use confidence intervals to estimate the actual values of parameters.
- On the other hand, lack of significance does not imply that H_0 is true, especially when the test has a low probability of detecting an effect.
- Significance tests are not always valid. Faulty data collection, outliers in the data, and testing a hypothesis on the same data that suggested the hypothesis can invalidate a test. Many tests run at once will probably produce some significant results by chance alone, even if all the null hypotheses are true.

SECTION 6.3 EXERCISES

For Exercise 6.90, see page 386; and for Exercise 6.91, see page 388.

6.92 A role as a statistical consultant. You are the statistical expert for a graduate student planning her PhD research. After you carefully present the mechanics of significance testing, she suggests using $\alpha = 0.20$ for the study because she would be more likely to obtain statistically significant results and she *really* needs significant results to graduate. Explain in simple terms why this would not be a good use of statistical methods.

6.93 What do you know? A research report described two results that both achieved statistical significance at the 5% level. The *P*-value for the first is 0.048; for the second it is 0.0002. Do the *P*-values add any useful information beyond that conveyed by the statement that both results are statistically significant? Write a short paragraph explaining your views on this question.

6.94 Selective publication based on results. In addition to statistical significance, selective publication can also be due to the observed outcome. A recent review of 74 studies of antidepressant agents found 38 studies with positive results and 36 studies with negative or questionable results. All but one of the 38 positive studies were published. Of the remaining 36, 22 were not published and 11 were published in such a way as to convey a positive outcome.[29] Describe how this selective reporting can have adverse consequences on health care.

6.95 What a test of significance can answer. Explain whether a test of significance can answer each of the following questions.

(a) Is the sample or experiment properly designed?

(b) Is the observed effect compatible with the null hypothesis?

(c) Is the observed effect important?

6.96 Vitamin C and colds. In a study to investigate whether vitamin C will prevent colds, 400 subjects are assigned at random to one of two groups. The experimental group takes a vitamin C tablet daily, while

the control group takes a placebo. At the end of the experiment, the researchers calculate the difference between the percents of subjects in the two groups who were free of colds. This difference is statistically significant ($P = 0.03$) in favor of the vitamin C group. Can we conclude that vitamin C has a strong effect in preventing colds? Explain your answer.

6.97 How far do rich parents take us? How much education children get is strongly associated with the wealth and social status of their parents, termed "socioeconomic status," or SES. The SES of parents, however, has little influence on whether children who have graduated from college continue their education. One study looked at whether college graduates took the graduate admissions tests for business, law, and other graduate programs. The effects of the parents' SES on taking the LSAT test for law school were "both statistically insignificant and small."

(a) What does "statistically insignificant" mean?

(b) Why is it important that the effects were small in size as well as statistically insignificant?

6.98 Do you agree? State whether or not you agree with each of the following statements and provide a short summary of the reasons for your answers.

(a) If the *P*-value is larger than 0.05, the null hypothesis is true.

(b) Practical significance is not the same as statistical significance.

(c) We can perform a statistical analysis using any set of data.

(d) If you find an interesting pattern in a set of data, it is appropriate to then use a significance test to determine its significance.

(e) It's always better to use a significance level of $\alpha = 0.05$ than to use $\alpha = 0.01$ because it is easier to find statistical significance.

6.99 Practical significance and sample size. Every user of statistics should understand the distinction between statistical significance and practical importance. A sufficiently large sample will declare very small effects statistically significant. Consider the study of elite female Canadian athletes in Exercise 6.74 (page 382). Female athletes were consuming an average of 2403.7 kcal/d with a standard deviation of 880 kcal/d. Suppose that a nutritionist is brought in to implement a new health program for these athletes. This program should increase mean caloric intake but not change the standard deviation. Given the standard deviation and how calorie deficient these athletes are, a change in the mean of 50 kcal/d to 2453.7 is of little importance. However, with a large enough sample, this change can be significant. To see this, calculate the *P*-value for the test of

$$H_0: \mu = 2403.7$$
$$H_a: \mu > 2403.7$$

in each of the following situations:

(a) A sample of 100 athletes; their average caloric intake is $\bar{x} = 2453.7$.

(b) A sample of 500 athletes; their average caloric intake is $\bar{x} = 2453.7$.

(c) A sample of 2500 athletes; their average caloric intake is $\bar{x} = 2453.7$.

6.100 Statistical versus practical significance. A study with 7500 subjects reported a result that was statistically significant at the 5% level. Explain why this result might not be particularly important.

6.101 More on statistical versus practical significance. A study with 14 subjects reported a result that failed to achieve statistical significance at the 5% level. The *P*-value was 0.051. Write a short summary of how you would interpret these findings.

6.102 Find journal articles. Find two journal articles that report results with statistical analyses. For each article, summarize how the results are reported and write a critique of the presentation. Be sure to include details regarding use of significance testing at a particular level of significance, *P*-values, and confidence intervals.

6.103 Create an example of your own. For each of the following cases, provide an example and an explanation as to why it is appropriate.

(a) A set of data or an experiment for which statistical inference is not valid.

(b) A set of data or an experiment for which statistical inference is valid.

6.104 Predicting success of trainees. What distinguishes managerial trainees who eventually become executives from those who, after expensive training, don't succeed and leave the company? We have abundant data on past trainees—data on their personalities and goals, their college preparation and performance, even their family backgrounds and their hobbies. Statistical software makes it easy to perform dozens of significance tests on these dozens of variables to see which ones best predict later success. We find that future executives are significantly more likely than washouts to have an urban or suburban upbringing and an undergraduate degree in a technical field.

Explain clearly why using these "significant" variables to select future trainees is not wise. Then suggest a follow-up study using this year's trainees as subjects that

should clarify the importance of the variables identified by the first study.

6.105 Searching for significance. Give an example of a situation where searching for significance would lead to misleading conclusions.

6.106 More on searching for significance. You perform 1000 significance tests using $\alpha = 0.05$. Assuming that all null hypotheses are true, about how many of the test results would you expect to be statistically significant? Explain how you obtained your answer.

6.107 Interpreting a very small *P*-value. Assume that you are performing a large number of significance tests. Let n be the number of these tests. How large would n need to be for you to expect about one P-value to be 0.00001 or smaller? Use this information to write an explanation of how to interpret a result that has $P = 0.00001$ in this setting.

6.108 An adjustment for multiple tests. One way to deal with the problem of misleading P-values when performing more than one significance test is to adjust the criterion you use for statistical significance. The **Bonferroni procedure** does this in a simple way. If you perform two tests and want to use the $\alpha = 5\%$ significance level, you would require a P-value of $0.05/2 = 0.025$ to declare either one of the tests significant. In general, if you perform k tests and want protection at level α, use α/k as your cutoff for statistical significance. You perform six tests and obtain individual P-values of 0.075, 0.021, 0.285, 0.002, 0.015, and <0.001. Which of these are statistically significant using the Bonferroni procedure with $\alpha = 0.05$?

6.109 Significance using the Bonferroni procedure. Refer to the previous exercise. A researcher has performed 12 tests of significance and wants to apply the Bonferroni procedure with $\alpha = 0.05$. The calculated P-values are 0.141, 0.519, 0.186, 0.753, 0.001, 0.008, 0.646, 0.038, 0.898, 0.013, <0.002, and 0.538. Which of these tests reject their null hypotheses with this procedure?

6.4 Power and Inference as a Decision

When you complete this section, you will be able to:

- Define what is meant by the power of a test.
- Determine the power of a test to detect an alternative for a given sample size n.
- Describe the two types of possible errors when performing a test that focuses on deciding between two hypotheses.
- Relate the two errors to the significance level and power of the test.

Although we prefer to use P-values rather than the reject-or-not view of the level α significance test, the latter view is very important for planning studies and for understanding statistical decision theory. We will discuss these two topics in this section.

Power

Level α significance tests are closely related to confidence intervals—in fact, we saw that a two-sided test can be carried out directly from a confidence interval (pages 353–354). The significance level, like the confidence level, says how reliable the method is in repeated use. If we use 5% significance tests repeatedly when H_0 is, in fact, true, we will be wrong (the test will reject H_0) 5% of the time and right (the test will fail to reject H_0) 95% of the time.

The ability of a test to detect that H_0 is false is measured by the probability that the test will reject H_0 when an alternative is true. The higher this probability is, the more sensitive the test is.

POWER

The probability that a level α significance test will reject H_0 when a particular alternative value of the parameter is true is called the **power** of the test to detect that alternative.

EXAMPLE 6.29

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 μ 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

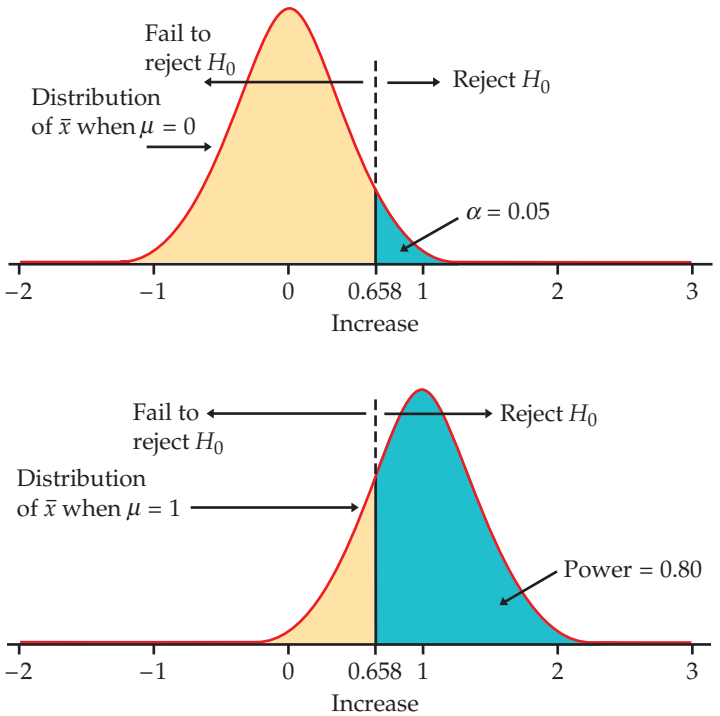


FIGURE 6.16 The sampling distributions of $\bar{x}$ when $\mu = 0$ and when $\mu = 1$. The power is the probability that the test rejects H_0 when the alternative is true.

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 illustrates the power with the sampling distribution of $\bar{x}$ when $\mu = 1$. This significance test rejects the null hypothesis that exercise has no effect on TBBMC 80% of the time if the true effect of exercise is a 1% increase in TBBMC. If the true effect of exercise is a greater percent increase, the test will have greater power; it will reject with a higher probability.

Here is another example of a power calculation, this time for a two-sided z test.

EXAMPLE 6.30

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

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

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.

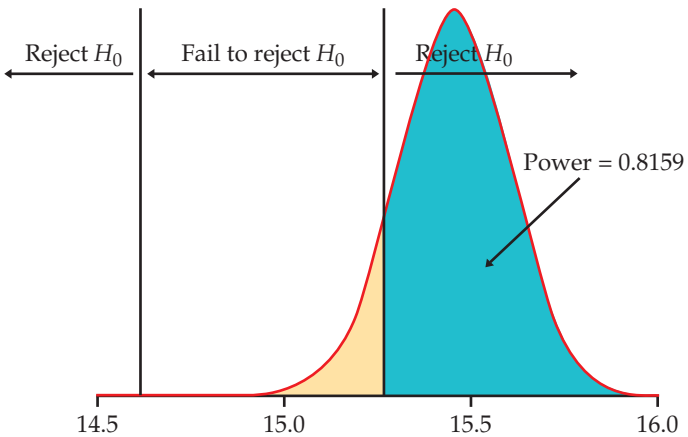


FIGURE 6.17 The power, Example 6.30. Unlike Figure 6.16, only the sampling distribution under the alternative is shown.

High power is desirable. Along with 95% confidence intervals and 5% significance tests, 80% power is becoming a standard. Many U.S. government agencies that provide research funds require that the sample size for the funded studies be sufficient to detect important results 80% of the time using a 5% test of significance.

EXAMPLE 6.31

Constructing a power curve. Example 6.30 considered one specific alternative, $\mu = 15.50$. Often, it is helpful to consider the power for a range of alternatives. Fortunately, most statistical software saves us from having to do these calculations manually. Figure 6.18 shows Minitab output for the power over the range 15.00 ppm to 15.80 ppm. The power calculation of Example 6.30 is represented by a dot on the curve at a difference of $15.50 - 15.00 = 0.50$. This curve is very informative. We see that with a sample size of three, the power is greater than 80% only for differences larger than about 0.48. If it is important to detect differences less than this, the Deely Laboratory needs to consider ways to increase the power.

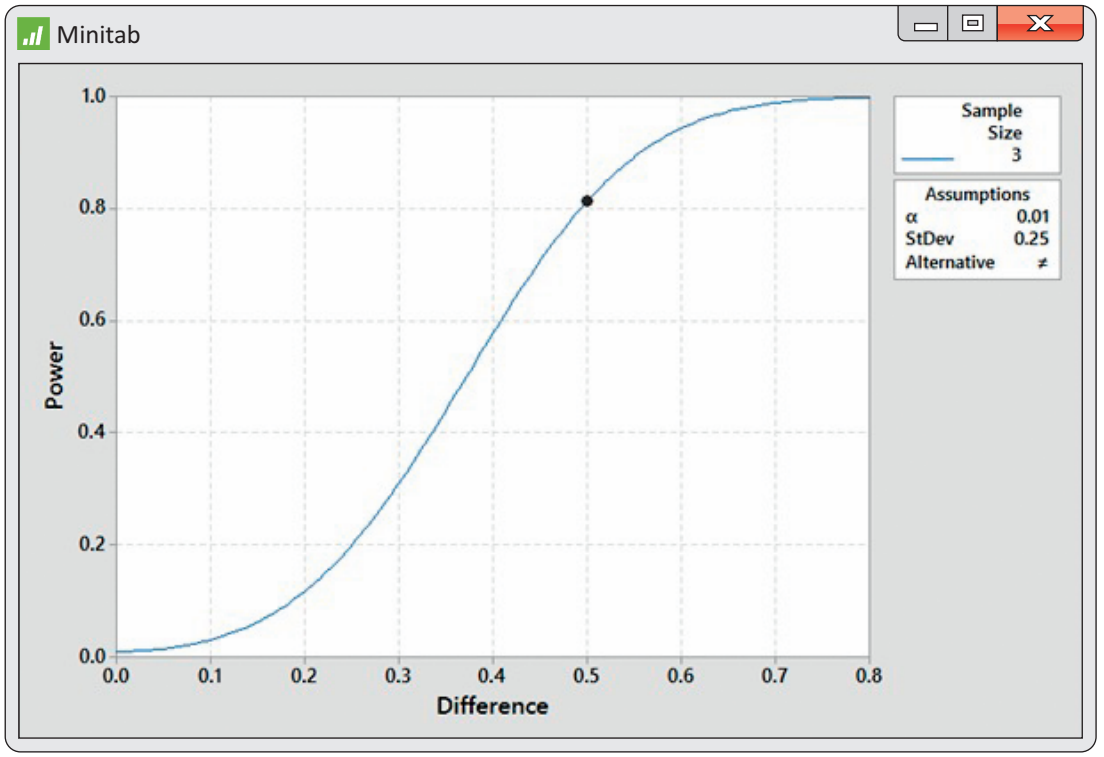


FIGURE 6.18 Minitab output (a power curve) for the one-sample power calculation, Example 6.31.

Increasing the power

Suppose that you have performed a power calculation and found that the power is too small. What can you do to increase it? Here are four ways. Note the similarity between these and the choices to reduce the margin of error (page 352).

- Increase α. A 5% test of significance will have a greater chance of rejecting the alternative than a 1% test because the strength of evidence required for rejection is less.
- Consider a particular alternative that is farther away from μ_0. Values of μ that are in H_a but lie close to the hypothesized value μ_0 are harder to detect (lower power) than values of μ that are far from μ_0.
- Increase the sample size. More data will provide more information about $\bar{x}$ so we have a better chance of distinguishing values of μ.
- Decrease σ. This has the same effect as increasing the sample size: more information about μ. Improving the measurement process and restricting attention to a subpopulation are possible ways to decrease σ.

Power calculations are important in planning studies. Using a significance test with low power makes it unlikely that you will find a significant effect even if the truth is far from the null hypothesis. A null hypothesis that is, in fact, false can become widely believed if repeated attempts to find evidence against it fail because of low power. The following example illustrates this point.

EXAMPLE 6.32

Are stock markets efficient? The "efficient market hypothesis" for the time series of stock prices says that future stock prices (when adjusted for inflation) show only random variation. No information available now will help us predict stock prices in the future because the efficient working of the market has already incorporated all available information in the present price. Many studies have tested the claim that one or another kind of information is helpful. In these studies, the efficient market hypothesis is H_0, and the claim that prediction is possible is H_a. Almost all the studies have failed to find good evidence against H_0. As a result, the efficient market hypothesis is quite popular. But an examination of the significance tests employed finds that the power is generally low. Failure to reject H_0 when using tests of low power is not evidence that H_0 is true. As one expert says, "The widespread impression that there is strong evidence for market efficiency may be due just to a lack of appreciation of the low power of many statistical tests."[30]

Inference as decision

We have presented tests of significance as methods for assessing the strength of evidence against the null hypothesis. This assessment is made by the P-value, which is a probability computed under the assumption that H_0 is true. The alternative hypothesis (the statement we seek evidence for) enters the test only to help us see what outcomes count against the null hypothesis.

There is another way to think about these issues. Sometimes, we are really concerned about making a decision or choosing an action based on our evaluation of the data. **Acceptance sampling** is one such circumstance. A producer of bearings and a skateboard manufacturer agree that each carload lot of bearings shall meet certain quality standards. When a carload arrives, the manufacturer chooses a sample of bearings to be inspected. On the basis of the sample outcome, the manufacturer will either accept or reject the carload. Let's examine how the idea of inference as a decision changes the reasoning used in tests of significance.

acceptance sampling

Two types of error

Tests of significance concentrate on H_0, the null hypothesis. If a decision is called for, however, there is no reason to single out H_0. There are simply two hypotheses, and we must accept one and reject the other. It is convenient to call the two hypotheses H_0 and H_a, but H_0 no longer has the special status (the statement we try to find evidence against) that it had in tests of significance. In the acceptance sampling problem, we must decide between

H_0: the lot of bearings meets standards

H_a: the lot does not meet standards

on the basis of a sample of bearings.

We hope that our decision will be correct, but sometimes it will be wrong. There are two types of incorrect decisions. We can accept a bad lot of bearings, or we can reject a good lot. Accepting a bad lot injures the consumer, while rejecting a good lot hurts the producer. To help distinguish these two types of error, we give them specific names.

		Truth about the population	
		H_0 true	H_a true
Decision based on sample	Reject H_0	Type I error	Correct decision
	Accept H_0	Correct decision	Type II error

FIGURE 6.19 The two types of error in testing hypotheses.

		Truth about the lot	
		Does meet standards	Does not meet standards
Decision based on sample	Reject the lot	Type I error	Correct decision
	Accept the lot	Correct decision	Type II error

FIGURE 6.20 The two types of error in the acceptance sampling setting.

TYPE I AND TYPE II ERRORS

If we reject H_0 (accept H_a) when in fact H_0 is true, this is a **Type I error**.
If we accept H_0 (reject H_a) when in fact H_a is true, this is a **Type II error**.

The possibilities are summed up in Figure 6.19. If H_0 is true, our decision either is correct (if we accept H_0) or is a Type I error. If H_a is true, our decision either is correct or is a Type II error. Only one error is possible at one time. Figure 6.20 applies these ideas to the acceptance sampling example.

Error probabilities

Any rule for making decisions is assessed in terms of the probabilities of the two types of error. This is in keeping with the idea that statistical inference is based on probability. We cannot (short of inspecting the whole lot) guarantee that good lots of bearings will never be rejected and bad lots never be accepted. But by random sampling and the laws of probability, we can say what the probabilities of both kinds of error are.

Significance tests with fixed level α give a rule for making decisions because the test either rejects H_0 or fails to reject it. If we adopt the decision-making way of thought, failing to reject H_0 means deciding that H_0 is true. We can then describe the performance of a test by the probabilities of Type I and Type II errors.

EXAMPLE 6.33

Photo by The Photo Works

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.

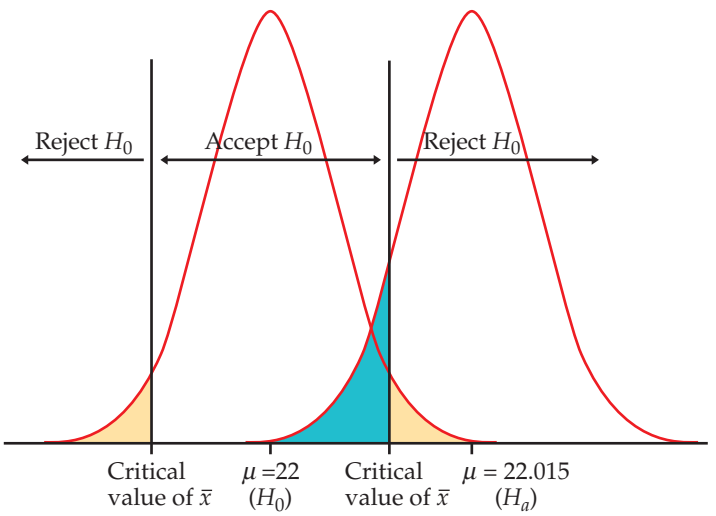


FIGURE 6.21 The two error probabilities, Example 6.33. The probability of a Type I error (yellow area) is the probability of rejecting H_0: $\mu = 22$ when, in fact, $\mu = 22$. The probability of a Type II error (blue area) is the probability of accepting H_0 when, in fact, $\mu = 22.015$.

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.

SIGNIFICANCE AND TYPE I ERROR

The significance level α of any fixed level test is the probability of a Type I error. That is, α is the probability that the test will reject the null hypothesis H_0 when H_0 is in fact true.

The probability of a Type II error for the particular alternative $\mu = 22.015$ in Example 6.33 is the probability that the test will fail to reject H_0 when μ has this alternative value. The *power* of the test to detect the alternative $\mu = 22.015$ is just the probability that the test *does* reject H_0. By following the method of Example 6.30, we can calculate that the power is about 0.92. The probability of a Type II error is therefore $1 - 0.92$, or 0.08.

POWER AND TYPE II ERROR

The power of a fixed level test to detect a particular alternative is 1 minus the probability of a Type II error for that alternative.

The two types of error and their probabilities give another interpretation of the significance level and power of a test. The distinction between tests of significance and tests as rules for deciding between two hypotheses does not lie in the calculations *but in the reasoning that motivates the calculations*. In a test of significance, we focus on a single hypothesis (H_0) and a single probability (the P-value). The goal is to measure the strength of the sample evidence against H_0. Calculations of power are done to check the sensitivity of the test. If we cannot reject H_0, we conclude only that there is not sufficient evidence against H_0, not that H_0 is actually true.

If the same inference problem is thought of as a decision problem, we focus on two hypotheses and give a rule for deciding between them based on the sample evidence. Therefore, we must focus equally on two probabilities, the probabilities of the two types of error. We must choose one hypothesis and cannot abstain on grounds of insufficient evidence.

The common practice of testing hypotheses

Such a clear distinction between the two ways of thinking is helpful for understanding. In practice, the two approaches often merge. We continued to call one of the hypotheses in a decision problem H_0. The common practice of *testing hypotheses* mixes the reasoning of significance tests and decision rules as follows:

1. State H_0 and H_a just as in a test of significance.
2. Think of the problem as a decision problem, so that the probabilities of Type I and Type II errors are relevant.
3. Because of Step 1, Type I errors are more serious. So choose an α (significance level) and consider only tests with probability of a Type I error no greater than α.
4. Among these tests, select one that makes the probability of a Type II error as small as possible (that is, power as large as possible). If this probability is too large, you will have to take a larger sample to reduce the chance of an error.

Testing hypotheses may seem to be a hybrid approach. It was, historically, the effective beginning of decision-oriented ideas in statistics. An impressive mathematical theory of hypothesis testing was developed between 1928 and 1938 by Jerzy Neyman and Egon Pearson. The decision-making approach

came later (1940s). Because decision theory in its pure form leaves you with two error probabilities and no simple rule on how to balance them, it has been used less often than either tests of significance or tests of hypotheses. Decision ideas have been applied in testing problems mainly by way of the Neyman-Pearson hypothesis-testing theory. That theory asks you first to choose α, and the influence of Fisher has often led users of hypothesis testing comfortably back to $\alpha = 0.05$ or $\alpha = 0.01$. Fisher, who was exceedingly argumentative, violently attacked the Neyman-Pearson decision-oriented ideas, and the argument still continues.

SECTION 6.4 SUMMARY

- The **power** of a significance test measures its ability to detect an alternative hypothesis. The power to detect a specific alternative is calculated as the probability that the test will reject H_0 when that alternative is true. This calculation requires knowledge of the sampling distribution of the test statistic under the alternative hypothesis. Increasing the size of the sample increases the power when the significance level remains fixed.
- An alternative to significance testing regards H_0 and H_a as two statements of equal status that we must decide between. This **decision theory** point of view regards statistical inference in general as giving rules for making decisions in the presence of uncertainty.
- In the case of testing H_0 versus H_a, decision analysis chooses a decision rule on the basis of the probabilities of two types of error. A **Type I error** occurs if H_0 is rejected when it is in fact true. A **Type II error** occurs if H_0 is accepted when in fact H_a is true.
- In a fixed level α significance test, the significance level α is the probability of a Type I error, and the power to detect a specific alternative is 1 minus the probability of a Type II error for that alternative.

SECTION 6.4 EXERCISES

6.110 Make a recommendation. Your manager has asked you to review a research proposal that includes a section on sample size justification. A careful reading of this section indicates that the power is 18% for detecting an effect that would be considered important. Write a short report for your manager explaining what this means and make a recommendation on whether or not this study should be run.

6.111 Explain power and sample size. Two studies are identical in all respects except for the sample sizes. Consider the power versus a particular sample size. Will the study with the larger sample size have more power or less power than the one with the smaller sample size? Explain your answer in terms that could be understood by someone with very little knowledge of statistics.

6.112 Power for a different alternative. The power for a two-sided test of the null hypothesis $\mu = 0$ versus the alternative $\mu = 6$ is 0.83. What is the power versus the alternative $\mu = -6$? Explain your answer.

6.113 More on the power for a different alternative. A one-sided test of the null hypothesis $\mu = 20$ versus the alternative $\mu = 30$ has power equal to 0.73. Will the power for the alternative $\mu = 35$ be higher or lower than 0.73? Draw a picture and use this to explain your answer.

6.114 Effect of changing the alternative μ on power. The *Statistical Power* applet illustrates the power calculation similar to that in Figure 6.16 (page 393). Open the applet and keep the default settings for the null ($\mu = 0$) and the alternative ($\mu > 0$) hypotheses, the sample size ($n = 10$), the standard deviation ($\sigma = 1$),

and the significance level ($\alpha = 0.05$). In the "alt $\mu =$" box, enter the value 1. What is the power? Repeat for alternative μ equal to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. Make a table giving μ and the power. What do you conclude?

6.115 Other changes and the effect on power. Refer to the previous exercise. For each of the following changes, explain what happens to the power for each alternative μ in the table.

(a) Change to the two-sided alternative.

(b) Decrease σ to 0.5.

(c) Increase n from 10 to 30.

6.116 Power of the random north–south distribution of trees test. In Exercise 6.70 (page 382), you performed a two-sided significance test of the null hypothesis that the average north–south location of the longleaf pine trees sampled in the Wade Tract was $\mu = 100$. There were 584 trees in the sample and the standard deviation was assumed to be 58. The sample mean in that analysis was $\bar{x} = 99.74$. Use the *Statistical Power* applet to compute the power for the alternative $\mu = 99$ using a two-sided test at the 5% level of significance.

6.117 Power of the random east–west distribution of trees test. Refer to the previous exercise. Note that in the east–west direction, the average location was 113.8. Use the *Statistical Power* applet to find the power for the alternative $\mu = 110$.

6.118 Planning another test to compare consumption. Example 6.15 (page 372) gives a test of a hypothesis about the mean consumption of sugar-sweetened beverages at your university based on a sample of size $n = 100$. The hypotheses are

$$H_0: \mu = 286$$
$$H_a: \mu \neq 286$$

While the result was not statistically significant, it did provide some evidence that the mean was smaller than 286. Thus, you plan to recruit another sample of students from your university, but this time use a one-sided alternative. You were thinking of surveying $n = 100$ students but now wonder if this sample size gives adequate power to detect a decrease of 15 calories per day to $\mu = 271$.

(a) Given $\alpha = 0.05$, for what values of z will you reject the null hypothesis?

(b) Using $\sigma = 155$ and $\mu = 286$, for what values of $\bar{x}$ will you reject H_0?

(c) Using $\sigma = 155$ and $\mu = 271$, what is the probability that $\bar{x}$ will fall in the region defined in part (b)?

(d) Will a sample size of $n = 100$ give you adequate power? Or do you need to find ways to increase the power? Explain your answer.

(e) Use the *Statistical Power* applet or other statistical software to determine the sample size n that gives you power near 0.80.

6.119 Planning the dining court survey. Exercise 6.38 (page 364) describes a survey to assess whether a newly designed dining court is viewed more favorably than the old design. The organizers are considering randomly surveying $n = 100$ student patrons but would like some statistical advice. The hypotheses are

$$H_0: \mu = 4$$
$$H_a: \mu > 4$$

and they've decided they want adequate power to detect a mean of at least 4.25.

(a) The organizers have no idea of σ. You suggest a small pilot study, which gives $s = 1.73$. Based on this result, you decide to use $\sigma = 2$. Provide an explanation for this choice to the organizers.

(b) Given $\alpha = 0.05$, for what values of $\bar{x}$ will you reject H_0?

(c) Using $\mu = 4.25$, what is the probability that $\bar{x}$ will fall in the region defined in part (b)?

(d) Will a sample size of $n = 100$ give you adequate power? Explain your answer.

(e) Use the *Statistical Power* applet or statistical software to determine the sample size n that gives you power near 0.80.

6.120 Choose the appropriate distribution. You must decide which of two discrete distributions a random variable X has. We will call the distributions p_0 and p_1. Here are the probabilities they assign to the values x of X:

x	0	1	2	3	4	5	6
p_0	0.1	0.1	0.2	0.3	0.1	0.1	0.1
p_1	0.1	0.3	0.2	0.1	0.1	0.1	0.1

You have a single observation on X and wish to test

$$H_0: p_0 \text{ is correct}$$
$$H_a: p_1 \text{ is correct}$$

One possible decision procedure is to reject H_0 only if $X \leq 1$.

(a) Find the probability of a Type I error, that is, the probability that you reject H_0 when p_0 is the correct distribution.

(b) Find the probability of a Type II error.

6.121 Power of the mean SATM score test. Example 6.16 (page 374) gives a test of a hypothesis about the SATM scores of California high school students based on an SRS of 500 students. The hypotheses are

$$H_0: \mu = 485$$
$$H_a: \mu > 485$$

Assume that the population standard deviation is $\sigma = 100$. The test rejects H_0 at the 1% level of significance when $z \geq 2.326$, where

$$z = \frac{\bar{x} - 485}{100/\sqrt{500}}$$

Is this test sufficiently sensitive to usually detect an increase of 14 points in the population mean SATM score? Answer this question by calculating the power of the test to detect the alternative $\mu = 499$.

6.122 More on choosing the appropriate distribution. Refer to Exercise 6.120. Suppose that instead of a single observation X, you obtained two observations and use the decision rule to reject when $\bar{x} \leq 1$.

(a) Under this scenario, would you expect the probabilities of a Type I and Type II errors to increase, decrease, or stay at the same values of Exercise 6.120? Explain your answer.

(b) Verify your answer to part (a) by computing the probabilities of a Type I and Type II error.

6.123 Computer-assisted career guidance systems. A wide variety of computer-assisted career guidance systems have been developed over the last decade. These programs use factors such as student interests, aptitude, skills, personality, and family history to recommend a career path. For simplicity, suppose that a program recommends a high school graduate either to go to college or to join the workforce.

(a) What are the two hypotheses and the two types of error that the program can make?

(b) The program can be adjusted to decrease one error probability at the cost of an increase in the other error probability. Which error probability would you choose to make smaller, and why? (This is a matter of judgment. There is no single correct answer.)

CHAPTER 6 EXERCISES

6.124 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.

6.125 Exercise and statistics exams. A study examined whether light exercise performed an hour before the final exam in statistics affects how students perform on the exam. The P-value was given as 0.13.

(a) State null and alternative hypotheses that could be used for this study. (*Note:* There is more than one correct answer.)

(b) Do you reject the null hypothesis? State your conclusion in plain language.

(c) What other facts about the study would you like to know for a proper interpretation of the results?

6.126 Roulette. A roulette wheel has 18 red slots among its 38 slots. You observe many spins and record the number of times that red occurs. Now you want to use these data to test whether the probability of a red has the value that is correct for a fair roulette wheel. State the hypotheses H_0 and H_a that you will test.

6.127 Food selection by children in school cafeterias. A group of researchers examined whether children's food selection in a school cafeteria met the standards set by the School Meals Initiative. They measured food selection and food intake of 2049 fourth- through sixth-grade students in 33 schools over a three-day period using digital photography. The following table summarizes some of the food intake measurements.[31]

	Boys $n = 852$		Girls $n = 1197$	
Food intake	**Mean**	**St. Dev.**	**Mean**	**St. Dev.**
Energy (kilojoules)	2448	717	2170	693
Protein (g)	24.5	7.5	22.1	7.7
Calcium (mg)	324.1	130.6	265.0	128.9

Given the large sample sizes, we can assume that the sample standard deviations are the population standard deviations.

(a) Compute 95% confidence intervals for all three intake measures for the boys.

(b) Compute 95% confidence intervals for all three intake measures for the girls.

(c) In the next chapter, we will describe the confidence interval for the difference between two means. For now, let's compare the boy and girl confidence intervals for each food intake measure. Do you think these pairs of intervals provide strong evidence against the null hypothesis that the boys and girls consume, on average, the same amount? Explain your answer.

6.128 Coverage percent of 95% confidence interval. For this exercise, you will use the *Confidence Interval* applet. Set the confidence level at 95% and click the "Sample" button 10 times to simulate 10 confidence intervals. Record the percent hit. Simulate another 10 intervals by clicking another 10 times (do not click the "Reset" button). Record the percent hit for your 20 intervals. Repeat the process of simulating 10 additional intervals and recording the results until you have a total of 200 intervals. Create a time plot of your results and write a summary of what you have found.

6.129 Coverage percent of 90% confidence interval. Refer to the previous exercise. Do the simulations and report the results for 90% confidence.

6.130 Effect of sample size on significance. You are testing the null hypothesis that $\mu = 0$ versus the alternative $\mu > 0$ using $\alpha = 0.05$. Assume that $\sigma = 16$. Suppose that $\bar{x} = 8$ and $n = 10$. Calculate the test statistic and its P-value. Repeat assuming the same value of $\bar{x}$ but with $n = 20$. Do the same for sample sizes of 30, 40, and 50. Plot the values of the test statistic versus the sample size. Do the same for the P-values. Summarize what this demonstration shows about the effect of the sample size on significance testing.

6.131 Survey response and margin of error. Suppose that a business conducts a marketing survey. As is often done, the survey is conducted by telephone. As it turns out, the business was only able to elicit responses from less than 10% of the randomly chosen customers. The low response rate is attributable to many factors, including caller ID screening. Undaunted, the marketing manager was pleased with the sample results because the margin of error was quite small, and thus the manager felt that the business had a good sense of the customers perceptions on various issues. Do you think the small margin of error is a good measure of the accuracy of the surveys results? Explain.

6.132 Reporting margins of error. A *U.S. News & World Report* article from July 17, 2014, reported Commerce Department *estimates* of changes in the construction industry:

> Construction fell 9.3 percent last month to a seasonally adjusted annual rate of 893,000 homes.

If we turn to the original Commerce Department report (released the same day), it states:

> Privately owned housing starts in June were at a seasonally adjusted annual rate of 893,000. This is 9.3 percent (10.3%) below the revised May estimate of 985,000.

(a) The 10.3% figure is the margin of error based on a 90% level of confidence. Given that fact, what is the 90% confidence interval for the percent change in housing starts from May to June?

(b) Explain why a credible media report should state:

> The Commerce Department has no evidence that privately owned housing starts rose or fell in June from the previous month.

6.133 Blood phosphorus level in dialysis patients. Patients with chronic kidney failure may be treated by dialysis, in which a machine removes toxic wastes from the blood, a function normally performed by the kidneys. Kidney failure and dialysis can cause other changes, such as retention of phosphorus, that must be corrected by changes in diet. A study of the nutrition of dialysis patients measured the level of phosphorus in the blood of several patients on six occasions. Here are the data for one patient (in milligrams of phosphorus per deciliter of blood):[32]

5.4 5.2 4.5 4.9 5.7 6.3

The measurements are separated in time and can be considered an SRS of the patient's blood phosphorus level. Assume that this level varies Normally with $\sigma = 0.9$ mg/dl. PMGDL

(a) Give a 95% confidence interval for the mean blood phosphorus level.

(b) The normal range of phosphorus in the blood is considered to be 2.6 to 4.8 mg/dl. Is there strong evidence that this patient has a mean phosphorus level that exceeds 4.8?

6.134 Cellulose content in alfalfa hay. An agronomist examines the cellulose content of a variety of alfalfa hay. Suppose that the cellulose content in the population has standard deviation $\sigma = 8$ milligrams per gram (mg/g). A sample of 15 cuttings has mean cellulose content $\bar{x} = 145$ mg/g.

(a) Give a 90% confidence interval for the mean cellulose content in the population.

(b) A previous study claimed that the mean cellulose content was $\mu = 140$ mg/g, but the agronomist believes that the mean is higher than that figure. State H_0 and H_a and carry out a significance test to see if the new data support this belief.

(c) The statistical procedures used in parts (a) and (b) are valid when several assumptions are met. What are these assumptions?

6.135 Odor threshold of future wine experts. Many food products contain small quantities of substances that would give an undesirable taste or smell if they are present in large amounts. An example is the "off-odors" caused by sulfur compounds in wine. Oenologists (wine experts) have determined the odor threshold, the lowest concentration of a compound that the human nose can detect. For example, the odor threshold for dimethyl sulfide (DMS) is given in the oenology literature as 25 micrograms per liter of wine (μg/l). Untrained noses may be less sensitive, however. Here are the DMS odor thresholds for 10 beginning students of oenology:

31 31 43 36 23 34 32 30 20 24

Assume (this is not realistic) that the standard deviation of the odor threshold for untrained noses is known to be $\sigma = 7$ μg/l. ODOR

(a) Make a stemplot to verify that the distribution is roughly symmetric with no outliers. (A Normal quantile plot confirms that there are no systematic departures from Normality.)

(b) Give a 95% confidence interval for the mean DMS odor threshold among all beginning oenology students.

(c) Are you convinced that the mean odor threshold for beginning students is higher than the published threshold, 25 μg/l? Carry out a significance test to justify your answer.

6.136 Where do you buy? Consumers can purchase nonprescription medications at food stores, mass merchandise stores such as Target and Walmart, or pharmacies. About 45% of consumers make such purchases at pharmacies. What accounts for the popularity of pharmacies, which often charge higher prices?

A study examined consumers' perceptions of overall performance of the three types of stores, using a long questionnaire that asked about such things as "neat and attractive store," "knowledgeable staff," and "assistance in choosing among various types of nonprescription medication." A performance score was based on 27 such questions. The subjects were 201 people chosen at random from the Indianapolis telephone directory. Here are the means and standard deviations of the performance scores for the sample:[33]

Store type	$\bar{x}$	s
Food stores	18.67	24.95
Mass merchandisers	32.38	33.37
Pharmacies	48.60	35.62

We do not know the population standard deviations, but a sample standard deviation s from so large a sample is usually close to σ. Use s in place of the unknown σ in this exercise.

(a) What population do you think the authors of the study want to draw conclusions about? What population are you certain they can draw conclusions about?

(b) Give 95% confidence intervals for the mean performance for each type of store.

(c) Based on these confidence intervals, are you convinced that consumers think that pharmacies offer higher performance than the other types of stores? (In Chapter 12, we will study a statistical method for comparing the means of several groups.)

6.137 CEO pay. A study of the pay of corporate chief executive officers (CEOs) examined the increase in cash compensation of the CEOs of 104 companies, adjusted for inflation, in a recent year. The mean increase in real compensation was $\bar{x} = 6.9$ %, and the standard deviation of the increases was $s = 55$ %. Is this good evidence that the mean real compensation μ of all CEOs increased that year? The hypotheses are

$$H_0: \mu = 0 \text{ (no increase)}$$
$$H_a: \mu > 0 \text{ (an increase)}$$

Because the sample size is large, the sample s is close to the population σ, so take $\sigma = 55$ %.

(a) Sketch the Normal curve for the sampling distribution of $\bar{x}$ when H_0 is true. Shade the area that represents the P-value for the observed outcome $\bar{x} = 6.9$ %.

(b) Calculate the P-value.

(c) Is the result significant at the $\alpha = 0.05$ level? Do you think the study gives strong evidence that the mean compensation of all CEOs went up?

6.138 Meaning of "statistically significant." When asked to explain the meaning of "statistically significant at the $\alpha = 0.01$ level," a student says, "This means there is only probability 0.01 that the null hypothesis is true." Is this an essentially correct explanation of statistical significance? Explain your answer.

6.139 More on the meaning of "statistically significant." Another student, when asked why statistical significance appears so often in research reports, says, "Because saying that results are significant tells us that

they cannot easily be explained by chance variation alone." Do you think that this statement is essentially correct? Explain your answer.

6.140 Increasing the power. Refer to Example 6.17 (page 375). Suppose the Deely Laboratory wants to make sure the power is at least 80% for $\mu = 15.30$ ppm. It cannot reduce σ, so the options are changing the significance level α and/or sample size n.

(a) With $\alpha = 0.01$, what sample size n is needed to have at least 80% power?

(b) With $\alpha = 0.05$, what sample size n is needed to have at least 80% power?

(c) Which of these options do you think the Deely Laboratory should choose? Explain your reasoning.

6.141 Simulation study of the confidence interval. Use a computer to generate $n = 15$ observations from a Normal distribution with mean 20 and standard deviation 5: $N(20, 5)$. Find the 95% confidence interval for μ. Repeat this process 100 times and then count the number of times that the confidence interval includes the value $\mu = 20$. Explain your results.

6.142 Simulation study of a test of significance. Use a computer to generate $n = 15$ observations from a Normal distribution with mean 20 and standard deviation 5: $N(20, 5)$. Test the null hypothesis that $\mu = 20$ using a two-sided significance test. Repeat this process 100 times and then count the number of times that you reject H_0. Explain your results.

6.143 Another simulation study of a test of significance. Use the same procedure for generating data as in the previous exercise. Now test the null hypothesis that $\mu = 24$. Explain your results.

6.144 Simulation study of power. Refer to the previous two exercises. What is the power of detecting a difference of four units (H_0: $\mu = 24$ versus $\mu = 20$) in this setting? Compare this power with the proportion of times you rejected H_0 in the previous exercise. Explain your results.

6.145 Find published studies with confidence intervals. Search the Internet or some journals that report research in your field and find two reports that provide an estimate with a margin of error or a confidence interval. For each report,

(a) describe the method used to collect the data.

(b) describe the variable being studied.

(c) give the estimate and the confidence interval.

(d) describe any practical difficulties that may have led to errors in addition to the sampling errors quantified by the margin of error.