CHAPTER 3

Principles of Inference

CONTENTS

■ Example 3.1

The National Center for Education Statistics reports that the year 2007 reading scores for fourth graders had a national mean of 220.99 and a standard deviation of 35.73. (This data is from the National Assessment of Educational Progress administered to 191,000 children in fourth grade, and is for the reading average scale score.) You believe that your school district is doing a superlative job of teaching reading, and want to show that mean scores on this exam in your district would be higher than this national mean. You randomly select 50 children in fourth grade in your district and give the same exam. The mean in your sample is 230.2. This seems to vindicate your belief, but a critic points out that you simply may have been lucky in your sample. Since you could not afford to test every fourth grader in your school system, you only have sample data. Is it possible that if you tested all your fourth graders, the mean would be the same as the 220.99 observed nationally? Or can we eliminate sampling variability as an explanation for the high score in your data? This chapter presents methodology that can be used to help answer this question. This problem will be solved in Section 3.2. ■

Statistical Methods, DOI: 10.1016/B978-0-12-374970-3.00003-2

3.1 INTRODUCTION

As we have repeatedly noted, one of the primary objectives of a statistical analysis is to use data from a sample to make inferences about the population from which the sample was drawn. In this chapter we present the basic procedures for making such inferences.

As we will see, the sampling distributions discussed in Chapter 2 play a pivotal role in statistical inference. Because inference on an unknown population parameter is usually based solely on a statistic computed from a single sample, we rely on these distributions to determine how reliable this inference is. That is, a statistical inference is composed of two parts:

1. a *statement* about the value of that parameter, and
2. a measure of the *reliability* of that statement, usually expressed as a probability.

Traditionally statistical inference is done with one of two different but related objectives in mind.

1. We conduct tests of hypotheses, in which we hypothesize that one or more parameters have some specific values or relationships, and make our decision about the parameter(s) based on one or more sample statistic(s). In this type of inference, the reliability of the decision is the probability that the decision is incorrect.
2. We estimate one or more parameters using sample statistics. This estimation is usually done in the form of an interval, and the reliability of this inference is expressed as the level of confidence we have in the interval.

We usually refer to an incorrect decision in a hypothesis test as "making an error" of one kind or another. Making an error in a statistical inference is not the same as making a mistake; the term simply recognizes the fact that the possibility of making an incorrect inference is an inescapable fact of statistical inference. The best we can do is to try to evaluate the reliability of our inference. Fortunately, if the data used to perform a statistical inference are a random sample, we can use sampling distributions to calculate the probability of making an error and therefore quantify the reliability of our inference.

In this chapter we present the basic principles for making these inferences and see how they are related. As you go through this and the next two chapters, you will note that hypothesis testing is presented before estimation. The reason for this is that it is somewhat easier to introduce them in this order, and since they are closely related, once the concept of the hypothesis test is understood, the estimation principles are easily grasped. We want to emphasize that both are equally important and each should be used where appropriate. To avoid extraneous issues, in this chapter we use two extremely simple examples that have little practical application. Once we have learned these principles, we can apply them to more interesting and useful applications. That is the subject of the remainder of this book.

3.2 HYPOTHESIS TESTING

A hypothesis usually results from speculation concerning observed behavior, natural phenomena, or established theory. If the hypothesis is stated in terms of population parameters such as the mean and variance, the hypothesis is called a **statistical hypothesis**. Data from a sample (which may be an experiment) are used to test the validity of the hypothesis. A procedure that enables us to agree or disagree with the statistical hypothesis using data from a sample is called a **test** of the hypothesis. Some examples of hypothesis tests are:

- A consumer-testing organization determining whether a type of appliance is of standard quality (say, an average lifetime of a prescribed length) would base their test on the examination of a sample of prototypes of the appliance. The result of the test may be that the appliance is not of acceptable quality and the organization will recommend against its purchase.
- A test of the effect of a diet pill on weight loss would be based on observed weight losses of a sample of healthy adults. If the test concludes the pill is effective, the manufacturer can safely advertise to that effect.
- To determine whether a teaching procedure enhances student performance, a sample of students would be tested before and after exposure to the procedure and the differences in test scores subjected to a statistical hypothesis test. If the test concludes that the method is not effective, it will not be used.

3.2.1 General Considerations

To illustrate the general principles of hypothesis testing, consider the following two simple examples:

■ Example 3.2

There are two identically appearing bowls of jelly beans. Bowl 1 contains 60 red and 40 black jelly beans, and bowl 2 contains 40 red and 60 black jelly beans. Therefore, the proportion of red jelly beans, p, for the two bowls are

$$\text{Bowl 1} : p = 0.6,$$
$$\text{Bowl 2} : p = 0.4.$$

One of the bowls is sitting on the table, but you do not know which one it is (you cannot see inside it). You suspect that it is bowl 2, but you are not sure. To test your hypothesis that bowl 2 is on the table you sample five jelly beans.[1] The data from this sample, specifically the number of red jelly beans, is the sample statistic

[1] To make the necessary probability calculations easier, you replace each jelly bean before selecting a new one; this is called sampling with replacement and allows the use of the binomial probability distribution presented in Section 2.3.

that will be used to test the hypothesis that bowl 2 is on the table. That is, based on this sample, you will decide whether bowl 2 is the one on the table. ■

■ Example 3.3

A company that packages salted peanuts in 8-oz. jars is interested in maintaining control on the amount of peanuts put in jars by one of its machines. Control is defined as averaging 8 oz. per jar and not consistently over- or underfilling the jars. To monitor this control, a sample of 16 jars is taken from the line at random time intervals and their contents weighed. The mean weight of peanuts in these 16 jars will be used to test the hypothesis that the machine is indeed working properly. If it is deemed not to be doing so, a costly adjustment will be needed.[2] ■

These two examples will be used to illustrate the procedures presented in this chapter.

3.2.2 The Hypotheses

Statistical hypothesis testing starts by making a set of two statements about the parameter or parameters in question. These are usually in the form of simple mathematical relationships involving the parameters. The two statements are exclusive and exhaustive, which means that one or the other statement must be true, but they cannot both be true. The first statement is called the *null* hypothesis and is denoted by H_0, and the second is called the *alternative* hypothesis and is denoted by H_1.

The two hypotheses will not be treated equally. The null hypothesis, which represents the status quo, or the statement of "no effect," gets the benefit of the doubt. The alternative hypothesis, which is the statement that we are trying to establish, requires positive evidence before we can conclude it is correct. This is done by showing that the data is inconsistent with the null hypothesis. Since we rule out the null hypothesis as an explanation, we are left with the alternative hypothesis. In cases where we cannot rule out the null hypothesis, it does not mean we regard H_0 as true. We simply reserve judgment, possibly until additional data is gathered. The distinction between the null and alternative hypothesis is fundamental to understanding everything in the remainder of this text.

Definition 3.1 *The* ***null hypothesis*** *is a statement about the values of one or more parameters. This hypothesis represents the status quo and is usually not rejected unless the sample results strongly imply that it is false.*

For Example 3.2, the null hypothesis is

Bowl 2 is on the table.

[2] Note the difference between this problem and Example 2.13, the control chart example. In this case, a decision to adjust the machine is to be made on one sample only, while in Example 2.13 it is made by an examination of its performance over time.

In bowl 2, since 40 of the 100 jelly beans are red, the statistical hypothesis is stated in terms of a population parameter, $p =$ the proportion of red jelly beans in bowl 2. Thus the null hypothesis is

$$H_0 \colon p = 0.4.$$

Definition 3.2 *The **alternative hypothesis** is a statement that contradicts the null hypothesis. This hypothesis is accepted if the null hypothesis is rejected. The alternative hypothesis is often called the research hypothesis because it usually implies that some action is to be performed, some money spent, or some established theory overturned.*

In Example 3.2 the alternative hypothesis is

Bowl 1 is on the table,

for which the statistical hypothesis is

$$H_1 \colon p = 0.6,$$

since 60 of the 100 jelly beans in bowl 1 are red. Because there are no other choices, the two statements form a set of two exclusive and exhaustive hypotheses. That is, the two statements specify all possible values of parameter p.

For Example 3.3, the hypothesis statements are given in terms of the population parameter μ, the mean weight of peanuts per jar. The null hypothesis is

$$H_0 : \mu = 8,$$

which is the specification for the machine to be functioning correctly. The alternative hypothesis is

$$H_1 : \mu \neq 8,$$

which means the machine is malfunctioning. These statements also form a set of two exclusive and exhaustive hypotheses, even though the alternative hypothesis does not specify a single value as it did for Example 3.2.

3.2.3 Rules for Making Decisions

After stating the hypotheses we specify what sample results will lead to the rejection of the null hypothesis. Intuitively, sample results (summarized as sample statistics) that lead to rejection of the null hypothesis should reflect an apparent contradiction to the null hypothesis. In other words, if the sample statistics have values that are unlikely to occur if the null hypothesis is true, then we decide the null hypothesis is false. The statistical hypothesis testing procedure consists of defining sample results that appear to sufficiently contradict the null hypothesis to justify rejecting it.

In Section 2.5 we showed that a sampling distribution can be used to calculate the probability of getting values of a sample statistic from a given population. If we now define "unlikely" as some small probability, we can use the sampling distribution to determine a range of values of a sample statistic that is unlikely to occur if the null hypothesis is true. The occurrence of values in that range may then be considered grounds for rejecting that hypothesis. Statistical hypothesis testing consists of appropriately defining that region of values.

Definition 3.3 *The **rejection region** (also called the **critical region**) is the range of values of a sample statistic that will lead to rejection of the null hypothesis.*

In Example 3.2, the null hypothesis specifies the bowl having the lower proportion of red jelly beans; hence observing a large proportion of red jelly beans would tend to contradict the null hypothesis. For now, we will arbitrarily decide that having a sample with all red jelly beans provides sufficient evidence to reject the null hypothesis. If we let Y be the number of red jelly beans, the rejection region is defined as $y = 5$.

In Example 3.3, any sample mean weight $\bar{Y}$ not equal to 8 oz. would seem to contradict the null hypothesis. However, since some variation is expected, we would probably not want to reject the null hypothesis for values reasonably close to 8 oz. For the time being we will arbitrarily decide that a mean weight of below 7.9 or above 8.1 oz. is not "reasonably close," and we will therefore reject the null hypothesis if the mean weight of our sample occurs in this region. Thus, the rejection region for this example contains the values of $\bar{y} < 7.9$ or $\bar{y} > 8.1$.

If the value of the sample statistic falls in the rejection region, we know what decision to make. If it does not fall in the rejection region, we have a choice of decisions. First, we could accept the null hypothesis as being true. As we will see, this decision may not be the best choice. Our other choice would be to "fail to reject" the null hypothesis. As we will see, this is not necessarily the same as accepting the null hypothesis.

3.2.4 Possible Errors in Hypothesis Testing

In Section 3.1 we emphasized that statistical inferences based on sample data may be subject to what we called errors. Actually, it turns out that results of a hypothesis test may be subject to two distinctly different errors, which are called type I and type II errors. These errors are defined in Definitions 3.4 and 3.5 and illustrated in Table 3.1.

Definition 3.4 *A **type I error** occurs when we incorrectly reject H_0, that is, when H_0 is actually true and our sample-based inference procedure rejects it.*

Definition 3.5 *A **type II error** occurs when we incorrectly fail to reject H_0, that is, when H_0 is actually not true, and our inference procedure fails to detect this fact.*

Table 3.1 Results of a Hypothesis Test

	IN THE POPULATION	
The Decision	**H_0 is True**	**H_0 is not True**
H_0 is not rejected	Decision is correct	A type II error has been committed
H_0 is rejected	A type I error has been committed	Decision is correct

In Example 3.2 the rejection region consisted of finding all five jelly beans in the sample to be red. Hence, the type I error occurs if all five sample jelly beans are red, the null hypothesis is rejected, and we proclaim the bowl to be bowl 1 but, in fact, bowl 2 is actually on the table. Alternatively, a type II error will occur if our sample has four or fewer red jelly beans (or one or more black jelly beans), in which case H_0 is not rejected, and we therefore proclaim that it is bowl 2, but, in fact, bowl 1 is on the table.

In Example 3.3, a type I error will occur if the machine is indeed working properly, but our sample yields a mean weight of over 8.1 or under 7.9 oz., leading to rejection of the null hypothesis and therefore an unnecessary adjustment to the machine. Alternatively, a type II error will occur if the machine is malfunctioning but the sample mean weight falls between 7.9 and 8.1 oz. In this case we fail to reject H_0 and do nothing when the machine really needs to be adjusted.

Obviously we cannot make both types of errors simultaneously, and in fact we may not make either, but the possibility does exist. In fact, we will usually never know whether any error has been committed. The only way to avoid any chance of error is not to make a decision at all, hardly a satisfactory alternative.

3.2.5 Probabilities of Making Errors

If we assume that we have the results of a random sample, we can use the characteristics of sampling distributions presented in Chapter 2 to calculate the probabilities of making either a type I or type II error for any specified decision rule.

Definition 3.6

α: denotes the probability of making a type I error
β: denotes the probability of making a type II error

The ability to provide these probabilities is a key element in statistical inference, because they measure the reliability of our decisions. We will now show how to calculate these probabilities for our examples.

Calculating α for Example 3.2

The null hypothesis specifies that the probability of drawing a red jelly bean is 0.4 (bowl 2), and the null hypothesis is to be rejected with the occurrence of five red jelly beans. Then the probability of making a type I error is the probability of getting five red jelly beans in a sample of five from bowl 2. If we let Y be the number of red jelly beans in our sample of five, then

$$\alpha = P(Y = 5 \text{ when } p = 0.4).$$

The use of the binomial probability distribution (Section 2.3) provides the result $\alpha = (0.4)^5 = 0.01024$. Thus the probability of incorrectly rejecting a true null hypothesis in this case is 0.01024; that is, there is approximately a 1 in 100 chance that bowl 2 will be mislabeled bowl 1 using the described decision rule.

Calculating α for Example 3.3

For this example, the null hypothesis was to be rejected if the mean weight was less than 7.9 or greater than 8.1 oz. If $\bar{Y}$ is the sample mean weight of 16 jars, the probability of a type I error is

$$\alpha = P(\bar{Y} < 7.9 \text{ or } \bar{Y} > 8.1 \text{ when } \mu = 8).$$

Assume for now that we know[3] that σ, the standard deviation of the population of weights, is 0.2 and that the distribution of weights is approximately normal. If the null hypothesis is true, the sampling distribution of the mean of 16 jars is normal with $\mu = 8$ and $\sigma = 0.2/\sqrt{16} = 0.05$ (see discussion on the normal distribution in Section 2.5). The probability of a type I error corresponds to the shaded area in Fig. 3.1.

Using the tables of the normal distribution we compute the area for each portion of the rejection region

$$P(\bar{Y} < 7.9) = P\left[Z < \frac{7.9 - 8}{(0.2/\sqrt{16})}\right] = P(Z < -2.0) = 0.0228$$

and

$$P(\bar{Y} > 8.1) = P\left(Z > \frac{8.1 - 8}{0.2/\sqrt{16}}\right) = P(Z > 2.0) = 0.0228.$$

Hence

$$\alpha = 0.0228 + 0.0228 = 0.0456.$$

Thus the probability of adjusting the machine when it does not need it (using the described decision rule) is slightly less than 0.05 (or 5%).

[3] This is an assumption made here to simplify matters. In Chapter 4 we present the method required if we calculate the standard deviation from the sample data.

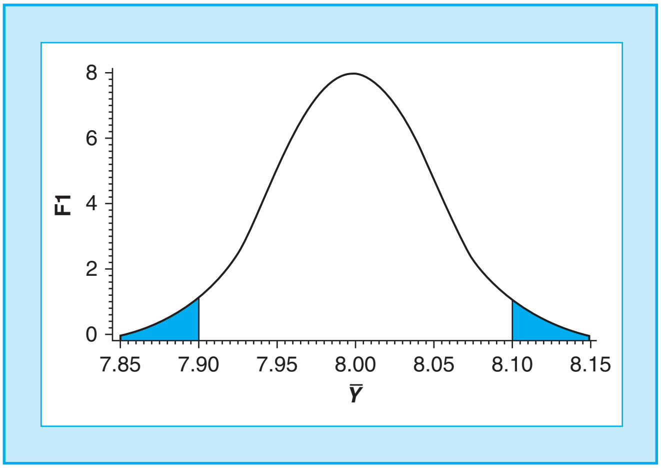

FIGURE 3.1
Rejection Region for Sample Mean.

Calculating β for Example 3.2

Having determined α for a specified decision rule, it is of interest to determine β. This probability can be readily calculated for Example 3.2. Recall that the type II error occurs if we fail to reject the null hypothesis when it is not true. For this example, this occurs if bowl 1 is on the table but we did not get the five red jelly beans required to reject the null hypothesis that bowl 2 is on the table. The probability of a type II error, which is denoted by β, is then the probability of getting four or fewer red jelly beans in a sample of five from bowl 1. If we let Y be the number of red jelly beans in the sample, then

$$\beta = P(Y \leq 4 \text{ when } p = 0.6).$$

Using the probability rules from Section 2.2, we know that

$$P(Y \leq 4) + P(Y = 5) = 1.$$

Since $(Y = 5)$ is the complement of $(Y \leq 4)$,

$$P(Y \leq 4) = 1 - P(Y = 5).$$

Now

$$P(Y = 5) = (0.6)^5,$$

and therefore

$$\beta = 1 - (0.6)^5 = 1 - 0.07776 = 0.92224.$$

That is, the probability of making a type II error in Example 3.2 is over 92%. This value of β is unacceptably large. If bowl 1 is truly on the table, the probability we will be unable to detect it is 0.92!

Calculating β for Example 3.3

For Example 3.3, H_1 does not specify a single value for μ but instead includes all values of $\mu \neq 8$. Therefore, calculating the probability of the type II error requires that we examine the probability of the sample mean being outside the rejection region for every value of $\mu \neq 8$. These calculations and further discussion of β are presented later in this section where we discuss type II errors.

3.2.6 Choosing between α and β

The probability of making a type II error can be decreased by making rejection easier, which is accomplished by making the rejection region larger. For example, suppose we decide to reject H_0 if either four or five of the jelly beans are red. In this case,

$$\alpha = P(Y \geq 4 \text{ when } p = 0.4) = 0.087$$

and

$$\beta = P(Y < 4 \text{ when } p = 0.6) = 0.663.$$

Note that by changing the rejection region we succeeded in lowering β but we increased α. This will always happen if the sample size is unchanged. In fact, if by changing the rejection region α becomes unacceptably large, no satisfactory testing procedure is available for a sample of five jelly beans, a condition that often occurs when sample sizes are small (see Section 3.4). This relationship between the two types of errors prevents us from constructing a hypothesis test that has a probability of 0 for either error. In fact, the only way to ensure that $\alpha = 0$ is to never reject a hypothesis, while to ensure that $\beta = 0$ the hypothesis should always be rejected, regardless of any sample results.

3.2.7 Five-Step Procedure for Hypothesis Testing

In the above presentation we have shown how to determine the probability of making a type I error for some arbitrarily chosen rejection region. The more frequently used method is to specify an acceptable maximum value for α and then delineate a rejection region for a sample statistic that satisfies this value. A hypothesis test can be formally summarized as a five-step process. Briefly these steps are as follows:

Step 1: Specify H_0, H_1, and an acceptable level of α.

Step 2: Define a sample-based test statistic and the rejection region for the specified H_0.

Step 3: Collect the sample data and calculate the test statistic.

Step 4: Make a decision to either reject or fail to reject H_0. This decision will normally result in a recommendation for action.

Step 5: Interpret the results in the language of the problem. It is imperative that the results be usable by the practitioner. Since H_1 is of primary interest, this conclusion should be stated in terms of whether there was or was not evidence for the alternative hypothesis.

We now discuss various aspects of these steps.

Step 1 consists of specifying H_0 and H_1 and a choice of a maximum acceptable value of α. This value is based on the seriousness or cost of making a type I error in the problem being considered.

Definition 3.7 *The* ***significance level*** *of a hypothesis test is the maximum acceptable probability of rejecting a true null hypothesis.*[4]

The reason for specifying α (rather than β) for a hypothesis test is based on the premise that the type I error is of prime concern. For this reason the hypothesis statement must be set up in such a manner that the type I error is indeed the more costly. The significance level is then chosen considering the cost of making that error.

In Example 3.2, H_0 was the assertion that the bowl on the table was bowl 2. In this example interchanging H_0 and H_1 would probably not cause any major changes unless there was some extra penalty for one of the errors. Thus, we could just as easily have hypothesized that the bowl was really 1, which would have made $H_0\colon p = 0.6$ instead of $H_0\colon p = 0.4$.

In Example 3.3 we stated that the null hypothesis is $\mu = 8$. In this example the choice of the appropriate H_0 is clear: There is a definite cost if we make a type I error since this error may cause an unnecessary adjustment on a properly working machine. Of course, making a type II error is not without cost, but since we have not accepted H_0, we are free to repeat the sampling at another time, and if the machine is indeed malfunctioning, the null hypothesis will eventually be rejected.

3.2.8 Why Do We Focus on the Type I Error?

In general, the null hypothesis is usually constructed to be that of the status quo; that is, it is the hypothesis requiring no action to be taken, no money to be spent, or in general nothing changed. This is the reason for denoting this as the null or nothing hypothesis. Since it is usually costlier to incorrectly reject the status quo than it is to do the reverse, this characterization of the null hypothesis does indeed cause the type I error to be of greater concern. In statistical hypothesis testing, the null hypothesis will invariably be stated in terms of an “equal” condition existing.

[4] Because the selection and use of the significance level is fundamental to this procedure, it is often referred to as a significance test. Although some statisticians make a minor distinction between hypothesis and significance testing, we use the two labels interchangeably.

On the other hand, the alternative hypothesis describes conditions for which something will be done. It is the action or research hypothesis. In an experimental or research setting, the alternative hypothesis is that an established (status quo) hypothesis is to be replaced with a new one. Thus, the research hypothesis is the one we actually want to support, which is accomplished by rejecting the null hypothesis with a sufficiently low level of α such that it is unlikely that the new hypothesis will be erroneously pronounced as true. The significance level represents a standard of evidence. The smaller the value of α, the stronger the evidence needed to establish H_1.

In Example 3.2, we thought the bowl was 2 (the status quo), and would only change our mind if the sample showed significant evidence that we were wrong. In Example 3.3 the status quo is that the machine is performing correctly; hence the machine would be left alone unless the sample showed so many or so few peanuts so as to provide sufficient evidence to reject H_0.

We can now see that it is quite important to specify an appropriate significance level. Because making the type I error is likely to have the more serious consequences, the value of α is usually chosen to be a relatively small number, and smaller in some cases than in others. That is, α must be selected so that an acceptable level of risk exists that the test will incorrectly reject the null hypothesis. Historically and traditionally, α has been chosen to have values of 0.10, 0.05, or 0.01, with 0.05 being most frequently used. These values are not sacred but do represent convenient numbers and allow the publication of statistical tables for use in hypothesis testing. We shall use these values often throughout the text. (See, however, the discussion of p values later in this section.)

3.2.9 Choosing α

As we saw in Example 3.2, α and β are inversely related. Unless the sample size is increased, we can reduce α only at the price of increasing β. In Example 3.2 there was little difference in the consequences of a type I or type II error; hence, the hypothesis test would probably be designed to have approximately equal levels of α and β. In Example 3.3 making the type I error will cause a costly adjustment to be made to a properly working machine, while if the type II error is committed we do not adjust the machine when needed. This error also entails some cost such as wasted peanuts or unsatisfied customers. Unless the cost of adjusting the machine is extremely high, a reasonable choice here would be to use the "standard" value of 0.05.

Some examples of problems for which one or the other type of error is more serious include the following:

- An industrial plant emits a pollutant that the state environmental agency requires have a mean less than a threshold T. If the benefit of the doubt goes to the industry, so that the agency has to prove a violation exists, then $H_0: \mu = T$

and $H_1: \mu > T$.[5] A type I error occurs when the plant is actually operating in compliance, but sampling data leads the agency to conclude a violation exists. A type II error occurs when the plant is actually noncompliant, but the agency is not able to show the violation exists. Bearing in mind that the cost of controlling the pollutant is likely to be expensive, the choice of α is likely to depend on the toxicity of the pollutant. If extremely dangerous, we will want to set α high (perhaps even 10%), so that we detect a violation with only moderate levels of evidence.

- When a drug company tests a new drug, there are two considerations that must be tested: (1) the toxicity (side effects) and (2) the effectiveness. For (1), the null hypothesis would be that the drug is toxic. This is because we would want to "prove" that it is not. For this test we would want a very small α, because a type I error would have extremely serious consequences (a significance level of 0.0001 would not be uncommon). For (2), the null hypothesis would be that the drug is not effective and a type I error would result in the drug being put on the market when it is not effective. The ramifications of this error would depend on the existing competitive drug market and the cost to both the company and society of marketing an ineffective drug.

Definition 3.8 *The* ***test statistic*** *is a sample statistic whose sampling distribution can be specified for both the null and alternative hypothesis case (although the sampling distribution when the alternative hypothesis is true may often be quite complex). After specifying the appropriate significance level of α, the sampling distribution of this statistic is used to define the rejection region.*

Definition 3.9 *The* ***rejection region*** *comprises the values of the test statistic for which (1) the probability when the null hypothesis is true is less than or equal to the specified α and (2) probabilities when H_1 is true are greater than they are under H_0.*

In **Step 2** we define the **test statistic** and the **rejection region**.

For Example 3.3 the appropriate test statistic is the sample mean. The sampling distribution of this statistic has already been used to show that the initially proposed rejection region of $\bar{y} < 7.9$ and $\bar{y} > 8.1$ produces a value of 0.0456 for α. If we had wanted α to be 0.05, this rejection region would appear to have been a very lucky guess! However, in most hypothesis tests it is necessary to specify α first and then use this value to delineate the rejection region. In the discussion of the significance level for Example 3.3 an appropriate level of α was chosen to be 0.05.

Remember, α is defined as

$$P(\bar{Y} \text{ falls in the rejection region when } H_0 \text{ is true}).$$

[5] An alternative hypothesis that specifies values in only one direction from the null hypothesis is called a one-sided or one-tailed alternative and requires some modifications in the testing procedure. One-tailed hypothesis tests are discussed later in this section.

We define the rejection region by a set of boundary values, often called critical values, that are denoted by $C1$ and $C2$. The probability α is then defined as

$$P(\bar{Y} < C1 \text{ when } \mu = 8) + P(\bar{Y} > C2 \text{ when } \mu = 8).$$

We want to find values of $C1$ and $C2$ so that this probability is 0.05. This is obtained by finding the $C1$ and $C2$ that satisfy the expression

$$\alpha = P\left[Z < \frac{C1 - 8}{0.2/\sqrt{16}}\right] + P\left[Z > \frac{C2 - 8}{0.2/\sqrt{16}}\right] = 0.05,$$

where Z is the standard normal variable. Because of the symmetry of the normal distribution, exactly half of the rejection region is in each tail; hence,

$$P\left[Z < \frac{C1 - 8}{0.05}\right] = P\left[Z > \frac{C2 - 8}{0.05}\right] = 0.025.$$

The values of $C1$ and $C2$ that satisfy this probability statement are found by using the standard normal table, where we find that the values of $z = -1.96$ and $z = +1.96$ satisfy our probability criteria. We use these values to solve for $C1$ and $C2$ in the equations $[(C1 - 8)/0.05] = -1.96$ and $[(C2 - 8)/0.05] = 1.96$. The solution yields $C1 = 7.902$ and $C2 = 8.098$; hence, the rejection region is

$$\bar{y} < 7.902 \quad \text{or} \quad \bar{y} > 8.098,$$

as seen in Fig. 3.2. The rejection region of Fig. 3.2 is given in terms of the test statistic $\bar{Y}$, the sample mean.

It is computationally more convenient to express the rejection region in terms of a test statistic that can be compared directly to a table, such as that of the normal

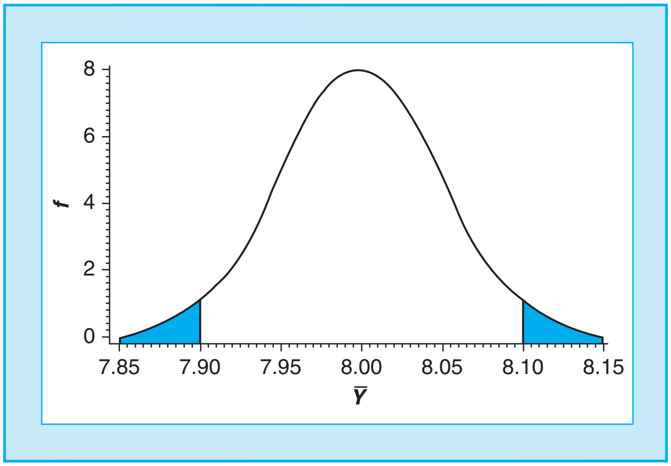

FIGURE 3.2
Rejection Region for 0.05 Significance.

distribution. In this case the test statistic is

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$$

$$= \frac{\bar{Y} - 8}{0.05},$$

which has the standard normal distribution and can be compared directly with the values read from the table. Then the rejection region for this statistic is

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

which can be more compactly written as $|z| > 1.96$. In other words we reject the null hypothesis if the value we calculate for Z has an absolute value (value ignoring sign) larger than 1.96.

Step 3 of the hypothesis test is to collect the sample data and compute the test statistic. (While this strict order may not be explicitly followed in practice, the sample data should not be used until the first two steps have been completed!) In Example 3.3, suppose our sample of 16 peanut jars yielded a sample mean value $\bar{y} = 7.89$. Then

$$z = (7.89 - 8)/0.05 = -2.20, \quad \text{or} \quad |z| = 2.20.$$

Step 4 compares the value of the test statistic to the rejection region to make the decision. In this case we have observed that the value 2.20 is larger than 1.96 so our decision is to reject H_0. This is often referred to as a "statistically significant" result, which means that the difference between the hypothesized value of $\mu = 8$ and the observed value of $\bar{y} = 7.89$ is large enough to be statistically significant.

In **Step** 5 we then conclude that the mean weight of nuts being put into jars is not the desired 8 oz. and the machine should be adjusted.

3.2.10 The Five Steps for Example 3.3

The hypothesis for Example 3.3 is summarized as follows:

Step 1:

$$H_0\colon \mu = 8,$$
$$H_1\colon \mu \neq 8,$$
$$\alpha = 0.05.$$

Step 2: The test statistic is

$$Z = \frac{\bar{Y} - 8}{0.2/\sqrt{16}}$$

whose sampling distribution is the standard normal. We specify $\alpha = 0.05$; hence we will reject H_0 if $|z| > 1.96$.

Step 3: Sample results: $n = 16, \bar{y} = 7.89, \sigma = 0.2$ (assumed);

$$z = (7.89 - 8)/[0.2/\sqrt{16}] = -2.20, \quad \text{hence } |z| = 2.20.$$

Step 4: $|z| > 1.96$; hence we reject H_0.

Step 5: We conclude $\mu \neq 8$ and recommend that the machine be adjusted.

Suppose that in our initial setup of the hypothesis test we had chosen α to be 0.01 instead of 0.05. What changes? This test is summarized as follows:

Step 1:

$$H_0\colon \mu = 8,$$
$$H_1\colon \mu \neq 8,$$
$$\alpha = 0.01.$$

Step 2: Reject H_0 if $|z| > 2.576$.

Step 3: Sample results: $n = 16, \sigma = 0.2, \bar{y} = 7.89$;

$$z = (7.89 - 8)/0.05 = -2.20.$$

Step 4: $|z| < 2.576$; hence we fail to reject $H_0\colon \mu = 8$.

Step 5: We do not recommend that the machine be readjusted.

We now have a problem. We have failed to reject the null hypothesis and do nothing. However, remember that we have not proved that the machine is working perfectly. In other words, *failing to reject the null hypothesis does not mean the null hypothesis was accepted*. Instead, we are simply saying that this particular test (or experiment) does not provide sufficient evidence to have the machine adjusted at this time. In fact, in a continuing quality control program, the test will be repeated in due time.

3.2.11 *p* Values

Having to specify a significance level before making a hypothesis test seems unnecessarily restrictive because many users do not have a fixed or definite idea of what constitutes an appropriate value for α. Also it is quite difficult to do when using computers because the user would have to specify an alpha for every test being requested. Another problem with using a specified significance level is that the ultimate conclusion may be affected by very minor changes in sample statistics.

As an illustration, we observed that in Example 3.3 the sample value of 7.89 leads to rejection with $\alpha = 0.05$. However, if the sample mean had been 7.91, certainly a very similar result, the test statistic would be -1.8, and we would not reject H_0. In other words, the decision of whether to reject may depend on minute differences in sample results.

We also noted that with a sample mean of 7.89 we would reject H_0 with $\alpha = 0.05$ but not with $\alpha = 0.01$. The logical question then is this: What about $\alpha = 0.02$, or $\alpha = 0.03$, or . . . ? This question leads to a method of reporting the results of a significance test without having to choose an exact level of significance, but instead leaves that decision to the individual who will actually act on the conclusion of the test. This method of reporting results is referred to as reporting the p value.

Definition 3.10 *The* ***p value*** *is the probability of committing a type I error if the actual sample value of the statistic is used as the boundary of the rejection region. It is therefore the smallest level of significance for which we would reject the null hypothesis with that sample. Consequently, the p value is often called the "attained" or the "observed" significance level. It is also interpreted as an indicator of the weight of evidence against the null hypothesis.*

In Example 3.3, the use of the normal table allows us to calculate the p value accurate to about four decimal places. For the sample $\bar{y} = 7.89$, this value is $P(|Z| > 2.20)$. Remembering the symmetry of the normal distribution, this is easily calculated to be $2P(Z > 2.20) = 0.0278$. This means that the management of the peanut-packing establishment can now evaluate the results of this experiment. They would reject the null hypothesis with a level of significance of 0.0278 or higher, and fail to reject it at anything lower.

Using the p value approach, Example 3.3 is summarized as follows:

Step 1:

$$H_0\colon \mu = 8,$$
$$H_1\colon \mu \neq 8.$$

Step 2: Sample results: $n = 16, \sigma = 0.2, \bar{y} = 7.89$;

$$z = (7.89 - 8)/0.05 = -2.20.$$

Step 3: $p = P(|Z| > 2.20) = 0.0278$; hence the p value is 0.0278. Therefore, we can say that the probability of observing a test statistic at least this extreme if the null hypothesis is true is 0.0278.

One feature of this approach is that the significance level need not be specified by the statistical analyst. In situations where the statistical analyst is not the same person who makes decisions, the analyst provides the p value and the decision maker determines the significance level based on the costs of making the type I error. For these reasons, many research journals now require that the results of such tests be published in this manner.

It is, in fact, actually easier for a computer program to provide p values, which are often given to three or more decimal places. However, when tests are calculated manually we must use tables. And because many tables provide for only a limited set of probabilities, p values can only be approximately determined. For example, we may only be able to state that the p value for the peanut jar example is between 0.01 and 0.05.

Note that the five steps of a significance test require that the significance level α be specified before conducting the test, while the p value is determined after the data have been collected and analyzed. Thus the use of a p value and a significance test are similar, but not strictly identical. It is, however, possible to use the p value in a significance test by specifying α in Step 1 and then altering Step 3 to read: Compute the p value and compare with the desired α. If the p value is smaller than α, reject the null hypothesis; otherwise fail to reject.

Alternate Definition 3.10 *A* ***p value*** *is the probability of observing a value of the test statistic that is at least as contradictory to the null hypothesis as that computed from the sample data.*

Thus the p value measures the extent to which the test statistic disagrees with the null hypothesis.

■ Example 3.4

An aptitude test has been used to test the ability of fourth graders to reason quantitatively. The test is constructed so that the scores are normally distributed with a mean of 50 and standard deviation of 10. It is suspected that, with increasing exposure to computer-assisted learning, the test has become obsolete. That is, it is suspected that the mean score is no longer 50, although σ remains the same. This suspicion may be tested based on a sample of students who have been exposed to a certain amount of computer-assisted learning.

Solution

The test is summarized as follows:

1.

$$H_0\colon \mu = 50,$$
$$H_1\colon \mu \neq 50.$$

2. The test is administered to a random sample of 500 fourth graders. The test statistic is

$$Z = \frac{\bar{Y} - 50}{10/\sqrt{500}}.$$

The sample yields a mean of 51.07. The test statistic has a value of

$$z = \frac{51.07 - 50}{10/\sqrt{500}} = 2.39.$$

3. The p value is computed as $2P(Z > 2.39) = 0.0168$. Because the construction of a new test is quite expensive, it may be determined that the level of significance should be less than 0.01, in which case the null hypothesis will not be rejected. However, the p value of 0.0168 may be considered sufficiently small to justify further investigation, say, by performing another experiment. ■

3.2.12 The Probability of a Type II Error

In presenting the procedures for hypothesis and significance tests we have concentrated exclusively on the control over α, the probability of making the type I error. However, just because that error is the more serious one, we cannot completely ignore the type II error. There are many reasons for ascertaining the probability of that error, for example:

- The probability of making a type II error may be so large that the test may not be useful. This was the case for Example 3.2.
- Because of the trade-off between α and β, we may find that we may need to increase α in order to have a reasonable value for β.
- Sometimes we have a choice of testing procedures where we may get different values of β for a given α.

Unfortunately, calculating β is not always straightforward. Consider Example 3.3. The alternative hypothesis, H_1: $\mu \neq 8$, encompasses all values of μ not equal to 8. Hence there is a sampling distribution of the test statistic for each unique value of μ, each producing a different value for β. Therefore β must be evaluated for all values of μ contained in the alternative hypothesis, that is, all values of μ not equal to 8.

This is not really necessary. For practical purposes it is sufficient to calculate β for a few representative values of μ and use these values to plot a function representing β for all values of μ not equal to 8. A graph of β versus μ is called an "operating characteristic curve" or simply an OC curve.

To construct the OC curve for Example 3.3, we first select a few values of μ and calculate the probability of a type II error at these values. For example, consider $\mu = 7.80, 7.90, 7.95, 8.05, 8.10$, and 8.20. Recall that for $\alpha = 0.05$ the rejection region is $\bar{y} < 7.902$ or $\bar{y} > 8.098$. The probability of a type II error is then the probability that $\bar{Y}$ does not fall in the rejection region, that is, $P(7.902 \leq \bar{Y} \leq 8.098)$, which is to be calculated for each of the specific values of μ given above.

Figure 3.3 shows the sampling distribution for the mean if the population mean is 7.95 as well as the rejection region (nonshaded area) for testing the null hypothesis

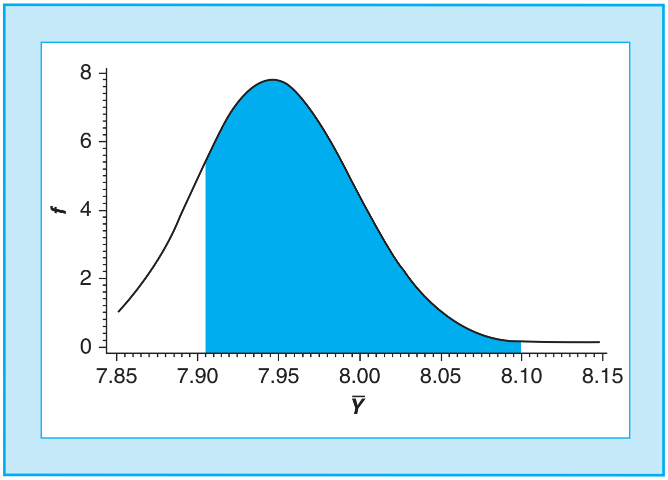

FIGURE 3.3
Probability of a Type II Error When the Mean Is 7.95.

that $\mu = 8$. The type II error occurs when the sample mean is not in the rejection region. Therefore, as seen in the figure, the probability of a type II error when the true value of μ is 7.95 is

$$\begin{aligned}\beta &= P(7.902 \leq \bar{Y} \leq 8.098 \text{ when } \mu = 7.95)\\ &= P\{[(7.902 - 7.95)/0.05] \leq Z \leq [(8.098 - 7.95)/0.05]\}\\ &= P(-0.96 \leq Z \leq 2.96) = 0.8300,\end{aligned}$$

obtained by using the table of the normal distribution. This probability corresponds to the shaded area in Fig. 3.3.

Similarly, the probability of a type II error when $\mu = 8.05$ is

$$\begin{aligned}\beta &= P(7.902 \leq \bar{Y} \leq 8.098 \text{ when } \mu = 8.05)\\ &= P\{[(7.902 - 8.05)/0.05] \leq Z \leq [(8.098 - 8.05)/0.05]\}\\ &= P(-2.96 \leq Z \leq 0.96) = 0.8300.\end{aligned}$$

These two values of β are the same because of the symmetry of the normal distribution and also because in both cases μ is 0.05 units from the null hypothesis value. The probability of a type II error when $\mu = 7.90$, which is the same as that for $\mu = 8.10$, is calculated as

$$\begin{aligned}\beta &= P(7.902 \leq \bar{Y} \leq 8.098 \text{ when } \mu = 7.90)\\ &= P(0.04 \leq Z \leq 3.96) = 0.4840.\end{aligned}$$

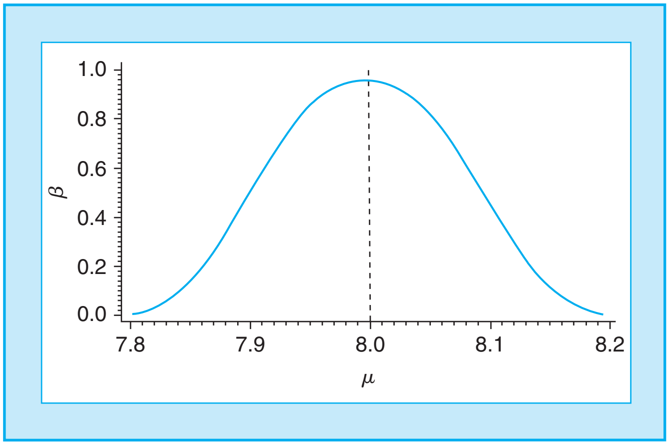

FIGURE 3.4
The OC Curve for Example 3.3.

In a similar manner we can obtain β for $\mu = 7.80$ and $\mu = 8.20$, which has the value 0.0207.

While it is impossible to make a type II error when the true mean is equal to the value specified in the null hypothesis, β approaches $(1 - \alpha)$ as the true value of the parameter approaches that specified in the null hypothesis. The OC curve can now be constructed using these values. Figure 3.4 gives the OC curve for Example 3.3. Note that the curve is indeed symmetric and continuous. Its maximum value is $(1 - \alpha) = 0.95$ at $\mu = 8$, and it approaches zero as the true mean moves further from the H_0 value. From this OC curve we may read (at least approximately) the value of β for any value of μ we desire.

The OC curve shows the logic behind the hypothesis testing procedure as follows:

- We have controlled the probability of making the more serious type I error.
- The OC curve shows that the probability of making the type II error is larger when the difference between the true value of the mean is close to the null hypothesis value, but decreases as that difference becomes greater. In other words, the higher probabilities of failing to reject the null hypothesis occur when the null hypothesis is "almost" true, in which case the type II error may not have serious consequences.

For example, in the peanut jar problem, failing to reject simply means that we continue using the machine but also continue the sampling inspection plan. If the machine is only slightly off, continuing the operation is not likely to have very serious consequences, but since sampling inspection continues, we will have the larger probability of rejection if the machine strays very far from its target.

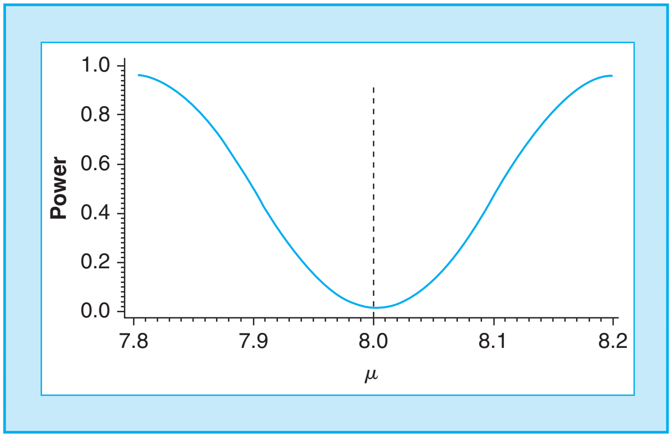

FIGURE 3.5
Power Curve
for Example 3.3.

3.2.13 Power

As a practical matter we are usually more interested in the probability of not making a type II error, that is, the probability of correctly rejecting the null hypothesis when it is false.

Definition 3.11 *The **power** of a test is the probability of correctly rejecting the null hypothesis when it is false.*

The power of a test is $(1 - \beta)$ and depends on the true value of the parameter μ. The graph of power versus all values of μ is called a **power curve**. The power curve for Example 3.3 is given in Fig. 3.5. Some features of a power curve are as follows:

- The power of the test increases and approaches unity as the true mean gets further from the null hypothesis value. This feature simply confirms that it is easier to deny a hypothesis as it gets further from the truth.
- As the true value of the population parameter approaches that of the null hypothesis, the power approaches α.
- Decreasing α while keeping the sample size fixed will produce a power curve that is everywhere lower. That is, decreasing α decreases the power.
- Increasing the sample size will produce a power curve that has a sharper "trough"; hence (except at the null hypothesis value) the power is higher everywhere. That is, increasing the sample size increases the power.

3.2.14 Uniformly Most Powerful Tests

Obviously high power is a desirable property of a test. If a choice of tests is available, the test with the largest power should be chosen. In certain cases, theory leads us to a test that has the largest possible power for any specified alternative hypothesis, sample size, and level of significance. Such a test is considered to be the best possible test for the hypothesis and is called a "uniformly most powerful" test. The test discussed in Example 3.3 is a uniformly most powerful test for the conditions specified in the example.

The computations involved in the construction of a power curve are not simple, and they become increasingly difficult for the applications in subsequent chapters. Fortunately, the performance of such computations often is not necessary because virtually all of the procedures we will be using provide uniformly most powerful tests, assuming that basic assumptions are met. We discuss these assumptions in subsequent chapters and provide some information on what the consequences may be of nonfulfillment of assumptions.

Power calculations for more complex applications can be made easier through the use of computer programs. While there is no single program that calculates power for all hypothesis tests, some programs either have the option of calculating power for specific situations or can be adapted to do so. One example using the SAS System can be found in Wright and O'Brien (1988).

3.2.15 One-Tailed Hypothesis Tests

In Examples 3.3 and 3.4 the alternative hypothesis simply stated that μ was not equal to the specified null hypothesis value. That is, the null hypothesis was to be rejected if the evidence showed that the population mean was either larger or smaller than that specified by the null hypothesis. For some applications we may want to reject the null hypothesis only if the value of the parameter is larger or smaller than that specified by the null hypothesis.

Solution to Example 3.1

In the example that introduced this chapter, we wished to know if our sample constituted evidence that the mean reading score among all fourth graders in our district (μ) is higher than the national mean of 220.99, that is,

$$H_0\colon \mu = 220.99 \quad \textit{versus} \quad H_1\colon \mu > 220.99.$$

The alternative hypothesis is now "greater than."[6] We would decide we had evidence for H_1 only if $\bar{X}$ is large; that is, our rejection region has all α of the probability in the upper tail. If we use $\alpha = 0.05$, our rejection rule is "reject H_0 if $z > 1.645$." Assuming that we may use the national standard deviation (35.73) as an estimate for σ, we get $z = (230.2 - 220.99)/(35.73/\sqrt{50}) = 1.82$. Hence we reject the null hypothesis. There is significant evidence that the mean in our district is higher than the national mean.

[6] To be consistent with the specification that the two hypotheses must be exhaustive, some authors will specify the null hypothesis as $\mu \le 220.99$ for this situation. We will stay with the single-valued null hypothesis statement whether we have a one- or two-tailed alternative. We maintain the exclusive and exhaustive nature of the two hypothesis statements by stating that we do not concern ourselves with values of the parameter in the "other" tail.

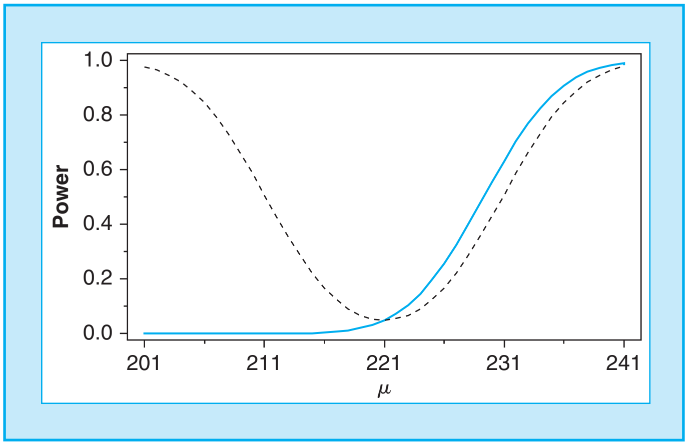

FIGURE 3.6
Power Curve for One- and Two-Tailed Tests.

We could also calculate the p value for our result as $\mathrm{P}(Z > 1.82) = .0344$, reaching the same conclusion.

Notice that the conclusion is about the mean among all fourth graders in our district. On the basis of a limited sample of only 50, we are reaching a conclusion about this much larger group.

This is an example of a one-tailed alternative hypothesis. It is important to try a different version of this problem, where you look for evidence that the mean among all fourth graders in our district *differs* from the national mean of 220.99. Now H_1: $\mu \neq 220.99$ and the rejection rule is "reject H_0 if $|z| > 1.96$." You would fail to reject the null hypothesis, even though the data has not changed!

The advantage of a one-tailed test over a two-tailed test is that for a given level of significance, the one-tailed test generally has a better chance of establishing H_1. Figure 3.6 shows how the power curve for the one-tailed test is slightly better when the true value of μ really does exceed 220.99. On the other hand, if the actual mean in our school district is really less than 220.99, the one-tailed test will not catch this, no matter how much sample data we have available. Since the one-tailed rejection region is only looking for large values of the test statistic, small values will not raise any alarm.

Generally, p values from one-tailed tests are smaller than from a two-tailed test. This raises the possibility for abuse, as researchers might decide (after examining their results) that they would achieve significance if they switch from two-tailed to one-tailed hypotheses.

For this reason, statisticians look on one-tailed tests as valid only if there are clear reasons, specified in advance, for choosing one particular direction for the hypothesis.

The decision on whether to perform a one- or two-tailed test is determined entirely by the problem statement. A one-tailed test is indicated by the alternative or research hypothesis, stating that only larger (or smaller) values of the parameter are of interest. In the absence of such specification, a two-tailed test should be employed.

3.3 ESTIMATION

In many cases we do not necessarily have a hypothesized value for the parameter that we want to test; instead we simply want to make a statement about the value of the parameter. For example, a large business may want to know the mean income of families in a target population near a proposed retail sales outlet. A chemical company may want to know the average amount of a chemical produced in a certain reaction. An animal scientist may want to know the mean yield of marketable meat of animals fed a certain ration. In each of these examples we use data from a sample to estimate the value of a parameter of the population. These are all examples of the inferential procedure called **estimation**.

As we will see, estimation and testing share some common characteristics and are often used in conjunction. For example, assume that we had rejected the hypothesis that the peanut-filling machine was putting 8 oz. of peanuts in the jars. It is then logical to ask, how much is the machine putting in the jars? The answer to this question could be useful in the effort to fix it.

The most obvious estimate of a population parameter is the corresponding sample statistic. This single value is known as a **point estimate**. For example, for estimating the parameter μ, the best point estimate is the sample mean $\bar{y}$. For estimating the parameter p in a binomial experiment, the best point estimate is the sample proportion $\hat{p} = y/n$.

For Example 3.3, the best point estimate of the mean weight of peanuts is the sample mean, which we found to be 7.89. We know that a point estimate will vary among samples from the same population. In fact, the probability that any point estimate exactly equals the true population parameter value is essentially zero for any continuous distribution. This means that if we make an unqualified statement of the form "μ is $\bar{y}$," that statement has almost no probability of being correct.

Thus a point estimate appears to be precise, but the precision is illusory because we have no confidence that the estimate is correct. In other words, it provides no information on the reliability of the estimate. A common practice for avoiding this dilemma is to "hedge," that is, to make a statement of the form "μ is almost certainly between 7.8 and 8." This is an **interval estimate**, and is the idea behind the statistical inference procedure known as the **confidence interval**. Admittedly a confidence interval does not seem as precise as a point estimate, but it has the advantage of having a known (and hopefully high) reliability.

Definition 3.12 *A **confidence interval** consists of a range of values together with a percentage that specifies how confident we are that the parameter lies in the interval.*

Estimation of parameters with intervals uses the sampling distribution of the point estimate. For example, to construct an interval estimate of μ we use the already established sampling distribution of $\bar{Y}$ (see Section 2.5). Using the characteristics of this distribution we can make the statement

$$P[(\mu - 1.96\sigma/\sqrt{n}) < \bar{Y} < (\mu + 1.96\sigma/\sqrt{n})] = 0.95.$$

An exercise in algebra provides a rearrangement of the inequality inside the parentheses without affecting the probability statement:

$$P[(\bar{Y} - 1.96\sigma/\sqrt{n}) < \mu < (\bar{Y} + 1.96\sigma/\sqrt{n})] = 0.95.$$

In general, using the notation of Chapter 2 we can write the probability statement as

$$P[(\bar{Y} - z_{\alpha/2}\sigma/\sqrt{n}) < \mu < (\bar{Y} + z_{\alpha/2}\sigma/\sqrt{n})] = (1 - \alpha).$$

Then, our interval estimate of μ is

$$(\bar{y} - z_{\alpha/2}\sigma/\sqrt{n}) \text{ to } (\bar{y} + z_{\alpha/2}\sigma/\sqrt{n}).$$

This interval estimate is called a **confidence interval**, and the lower and upper boundary values of the interval are known as **confidence limits**. The probability used to construct the interval is called the **level of confidence** or confidence coefficient. This confidence level is the equivalent of the "almost certainly" alluded to in the preceding introduction. We thus say that we are $(1 - \alpha)$ confident that this interval contains the population mean. The confidence coefficient is often given as a percentage, for example, a 95% confidence interval.

For Example 3.3, a 0.95 confidence interval (or 95% confidence interval) lies between the values

$$7.89 - 1.96(0.2)/\sqrt{16} \quad \text{and} \quad 7.89 + 1.96(0.2)/\sqrt{16}$$

or

$$7.89 \pm 1.96(0.05), \quad \text{or} \quad 7.89 \pm 0.098.$$

Hence, we say that we are 95% confident that the true mean weight of peanuts is between 7.792 and 7.988 oz. per jar.

3.3.1 Interpreting the Confidence Coefficient

We must emphasize that the confidence interval statement is not a standard probability statement. That is, we cannot say that with 0.95 probability μ lies between

7.792 and 7.988. Remember that μ is a fixed number, which by definition has no distribution. This true value of the parameter either is or is not in a particular interval, and we will likely never know which event has occurred for a particular sample. We can, however, state that 95% of the intervals constructed in this manner will contain the true value of μ.

Definition 3.13 *The* ***maximum error of estimation****, also called the margin of error, is an indicator of the precision of an estimate and is defined as one-half the width of a confidence interval.*

We can write the formula for the confidence limits on μ as $\bar{y} \pm E$, where

$$E = z_{\alpha/2}\sigma/\sqrt{n}$$

is one-half of the width of the $(1 - \alpha)$ confidence interval. The quantity E can also be described as the farthest that μ may be from $\bar{y}$ and still be in the confidence interval. This value is a measure of how "close" our estimate may be to the true value of the parameter. This bound on the error of estimation, E, is most often associated with a 95% confidence interval, but other confidence coefficients may be used. Incidentally, the "margin of error" often quoted in association with opinion polls is indeed E with an unstated 0.95 confidence level.

The formula for E illustrates for us the following relationships among E, α, n, and σ:

1. If the confidence coefficient is increased (α decreased) and the sample size remains constant, the maximum error of estimation will increase (the confidence interval will be wider). In other words, the more confidence we require, the less precise a statement we can make, and vice versa.
2. If the sample size is increased and the confidence coefficient remains constant, the maximum error of estimation will be decreased (the confidence interval will be narrower). In other words, by increasing the sample size we can increase precision without loss of confidence, or vice versa.
3. Decreasing σ has the same effect as increasing the sample size. This may seem a useless statement, but it turns out that proper experimental design (Chapter 10) can often reduce the standard deviation.

Thus there are trade-offs in interval estimation just as there are in hypothesis testing. In this case we trade precision (narrower interval) for higher confidence. The only way to have more confidence without increasing the width (or vice versa) is to have a larger sample size.

■ Example 3.5

Suppose that a population mean is to be estimated from a sample of size 25 from a normal population with $\sigma = 5.0$. Find the maximum error of estimation with confidence coefficients 0.95 and 0.99. What changes if n is increased to 100 while the confidence coefficient remains at 0.95?

Solution

1. The maximum error of estimation of μ with confidence coefficient 0.95 is

$$E = 1.96(5/\sqrt{25}) = 1.96.$$

2. The maximum error of estimation of μ with confidence coefficient 0.99 is

$$E = 2.576(5/\sqrt{25}) = 2.576.$$

3. If $n = 100$ then the maximum error of estimation of μ with confidence coefficient 0.95 is

$$E = 1.96(5/\sqrt{100}) = 0.98.$$

Note that increasing n fourfold only halved E. The relationship of sample size to confidence intervals is discussed further in Section 3.4. ■

3.3.2 Relationship between Hypothesis Testing and Confidence Intervals

As noted previously there is a direct relationship between hypothesis testing and confidence interval estimation. A confidence interval on μ gives all acceptable values for that parameter with confidence $(1 - \alpha)$. This means that any value of μ not in the interval is not an "acceptable" value for the parameter. The probability of being incorrect in making this statement is, of course, α. Therefore,

A hypothesis test for $H_0\colon \mu = \mu_0$ against $H_1\colon \mu \neq \mu_0$ will be rejected at a significance level of α if μ_0 is not in the $(1 - \alpha)$ confidence interval for μ.

Conversely,

Any value of μ inside the $(1 - \alpha)$ confidence interval will not be rejected by an α-level significance test.

For Example 3.3, the 95% confidence interval is 7.792 to 7.988. The hypothesized value of 8 is not contained in the interval; therefore we would reject the hypothesis $H_0: \mu = 8$ at the 0.05 level of significance. For Example 3.4, a 99% confidence interval on μ is 49.92 to 52.22. The hypothesis $H_0: \mu = 50$ would not be rejected with $\alpha = 0.01$ because the value 50 does lie within the interval. These results are, of course, consistent with results obtained from the hypothesis tests presented previously.

As in hypothesis testing, one-sided confidence intervals can be constructed. In Example 3.1 we used a one-sided alternative hypothesis, $H_1: \mu > 220.99$. This corresponds to finding the lower confidence limit so that the confidence statement will indicate that the mean score is at least that amount or higher. For this example, then, the lower $(1-\alpha)$ limit is

$$\bar{y} - z_a(\sigma/\sqrt{n}),$$

which results in the lower 0.90 confidence limit

$$230.2 - 1.645(35.73/\sqrt{50}) = 221.89.$$

Since the lower limit of the set of "feasible" μ lies above the national mean of 220.99, this is consistent with the result of our earlier one-sided hypothesis test.

3.4 SAMPLE SIZE

We have noted that in both hypothesis testing and interval estimation, a definite relationship exists between sample size and the precision of our results. In fact, the best possible sample appears to be the one that contains the largest number of observations. This is not necessarily the case. The cost and effort of obtaining the sample and processing and analyzing the data may offset the added precision of the results. Remember that costs often increase linearly with sample size, while precision, in terms of E, decreases only with the square root of the sample size. It is therefore not surprising that the question of sample size is of major concern. Because of the relationship of sample size to the precision of statistical inference, we can answer the question of optimal sample size.

Consider the problem of estimating μ using a sample from a normal population with known standard deviation, σ. We want to find the required sample size, n, for a specified maximum value of E. Using the formula for E,

$$E = \frac{z_{\alpha/2}\sigma}{\sqrt{n}},$$

we can solve for n, resulting in

$$n = \frac{z_{\alpha/2}^2\sigma^2}{E^2}.$$

Thus, given values for σ and α and a specified maximum E, we can determine the required sample size for the desired precision. For example, suppose that in Example 3.3 we wanted a 99% confidence interval for the mean weight to be no wider than 0.10 oz. This means that $E = 0.05$. The required sample size is

$$\begin{aligned} n &= (2.576)^2(0.2)^2/(0.05)^2 \\ &= 106.2. \end{aligned}$$

We round up to the nearest integer, so the required sample size is 107. This is a large sample, but both the confidence coefficient and the required precision were both quite strict. This example illustrates an often encountered problem: Requirements are often made so strict that unreasonably large sample sizes are required.

Sample size determination must satisfy two prespecified criteria:

1. the value of E, the maximum error of estimation (or, equivalently, half the width of the confidence interval), and
2. the required level of confidence (the confidence coefficient, $1 - \alpha$).

In other words, it is not only sufficient to require a certain degree of precision, but it is also necessary to state the degree of confidence. Since the degree of confidence is so often assumed to be 0.95, it is usually not stated, which may give the incorrect impression of 100% confidence! It is, of course, also necessary to have an estimated value for σ^2 if we are estimating μ. In many cases, we have to use rough approximations of the variance. One such approximation can be obtained from the empirical rule discussed in Chapter 1. If we can determine the expected range of values of the results of the experiment, we can use the empirical rule to obtain an estimate of the standard deviation. That is, we could use the range divided by 4 to estimate the standard deviation. This is because the empirical rule states that 95% of the values of a distribution will be plus or minus 2σ from the mean. Thus, 95% of the values will be in the 4σ range.

■ Example 3.6

In a study of the effect of a certain drug on the behavior of laboratory animals, a research psychologist needed to determine the appropriate sample size. The study was to estimate the time necessary for the animal to travel through a maze under the influence of this drug.

Solution

Since no previous studies had been conducted on this drug, no independent estimate for the variation of times was available. Using the conventional confidence level of 95%, a bound on the error of estimation of 5 seconds, and an anticipated range of times of 15 to 60 seconds, what sample size would the psychologist need?

1. First, an estimate of the standard deviation was obtained from the range by dividing by 4:

$$EST(\sigma) = (60 - 15)/4 = 11.25.$$

2. The sample size was determined as $n = [(1.96)^2(11.25)^2]/5^2 = 19.4$.
3. Round up to $n = 20$, so the researcher needs 20 animals in the study.

The formula for the required sample size clearly indicates the trade-off between the interval width (the value of E) and the degree of confidence. In Example 3.6, narrowing the width to 1 would give

$$n = (1.96)^2(11.25)^2/(1)^2 = 487.$$

Requirements for being able to detect a specified difference between the null and alternate hypotheses with a given degree of significance can be converted to the desired width of a confidence interval by remembering the equivalence of the two procedures.

In Example 3.4 we may want to be able to detect, at the 0.01 level of significance, a change of one unit in the average test score. According to the equivalence, this requires a 99% confidence interval of plus or minus one unit, hence $E = 1$. The required sample size is

$$n = (2.576)^2(10)^2/(1)^2 = 664.$$

This, of course, may not always be possible, or may not be the best way to approach the problem. What we need is a way to compute directly the required sample size for conducting a hypothesis test, using the constraints usually developed in the process of testing a hypothesis. For example, we might be interested in determining how big a sample we need to have reasonable power against a specified value of μ, say μ_{a}, in the hypothesis

$$H_0\colon \mu = \mu_0 \quad \textit{versus} \quad H_1\colon \mu > \mu_0.$$

That is, we want to determine what sample size will give us adequate protection against mean values in the alternative (values of μ_{a} greater than μ_0) that have some negative impact on the process under scrutiny. In this case, however, several prespecified criteria must be considered. We need to satisfy:

1. the required level of significance (α),
2. the difference, called δ (delta), between the hypothesized value and the specified value ($\delta = \mu_{\mathrm{a}} - \mu_0$), and
3. the probability of a type II error (β) when the real mean is at this specified value (or one larger than the specified value).

The value of n that satisfies these criteria can be obtained using the formula

$$n = \frac{\sigma^2(z_\alpha + z_\beta)}{\delta^2},$$

where all the components of this formula have been defined.

Suppose that in Example 3.6 we wanted to test the following set of hypotheses:

$$H_0\colon \mu = 35\,\text{s} \quad \textit{versus} \quad H_1\colon \mu > 35\,\text{s}.$$

We use a level of significance $\alpha = 0.05$, and we decide that we are willing to risk making a type II error of $\beta = 0.10$ if the actual mean time is 37 s. This means that the power of the test at $\mu = 37$ s will be 0.90. The difference between the hypothesized value of the mean and the specified value of the mean is $\delta = 37 - 35 = 2$. In Example 3.6 we estimated the value of the standard deviation as 11.25. We can substitute this value for σ in the formula, obtain the necessary values from Appendix Table A.1A, and calculate n as

$$n = (11.25)^2 \frac{(1.64485 + 1.28155)^2}{(2)^2} = 271.$$

Therefore, if we take a sample of size $n = 271$ we can expect to reject the hypothesis that $\mu = 35$ if the real mean value is 37 or higher with probability 0.90.

The procedure for a hypothesis test with a one-sided alternative in the other direction is almost identical. The only difference is that μ_{a} will be less than μ_0. To use a two-sided alternative, we use the following formula to calculate the required sample size:

$$n = \frac{\sigma^2 (z_{\alpha/2} + z_\beta)^2}{\delta^2},$$

where $\delta = |\mu_{\text{a}} - \mu_0|$.

In Example 3.4 we might want to be more rigorous in our definition of the problem, and rather than saying that we simply want to detect a difference of one unit, say instead that we want to reject the null hypothesis if the deviation from the hypothesized value is one unit or more with probability 99%. That is, we would reject the null hypothesis if it were less than 49 or greater than 51 with power of 0.99. Using the values of $\sigma = 10, \alpha = 0.01, \beta = 0.01$, and $\delta = 1$, we get

$$n = (10)^2 \frac{(2.57583 + 2.32635)^2}{(1)^2} = 2404.$$

Note that this is larger than the value we obtained using the confidence interval approach; this is because we imposed more rigorous criteria.

These examples of sample size determination are relatively straightforward because of the simplicity of the methods used. If we did not know the standard deviation in a hypothesis test on the mean, or if we were using any of the hypothesis testing procedures discussed in subsequent chapters, we would not have such simple formulas for calculating n. There are, however, tables and charts that enable sample size determination to be done for most hypotheses tests. See, for example, Neter *et al.* (1996).

3.5 ASSUMPTIONS

In this chapter we have considered inferences on the population mean in situations where it can be assumed that the sampling distribution of the mean is reasonably close to normal. Inference procedures based on the assumption of a normally distributed sample statistic are referred to as normal theory methods.

In Section 2.5 we pointed out that the sampling distribution of the sample mean is normal if the population itself is normal, or if the sample size is large enough to satisfy the central limit theorem. However, normality of the sampling distribution of the mean is not always assured for relatively small samples, especially those from highly skewed distributions or where the observations may be dominated by a few extreme values. In addition, as noted in Chapter 1, some data may be obtained as ordinal values such as ranks, or nominal values such as categorical data. Such data are not readily amenable to analysis by the methods designed for interval data.

When the assumption of normality does not hold, use of methods requiring this assumption may produce misleading inferences. That is, the significance level of a hypothesis test or the confidence level of an estimate may not be as specified by the procedure. For instance, the use of the normal distribution for a test statistic may indicate rejection at the 0.05 significance level, but due to nonfulfillment of the assumptions, the true protection against making a type I error may be as high as 0.10. (Refer to Section 4.5 for ways to determine whether the normality assumption is valid.)

Unfortunately, we cannot know the true value of α in such cases. For this reason alternate procedures have been developed for situations in which normal theory methods are not applicable. Such methods are often described as “robust” methods, because they provide the specified α for virtually all situations. However, this added protection is not free: Most of these robust methods have wider confidence intervals and/or have power curves generally lower than those provided by normal theory methods when the assumption of normality is indeed satisfied.

Various principles are used to develop robust methods. Two often used principles are as follows:

1. Trimming, which consists of discarding a small prespecified portion of the most extreme observations and making appropriate adjustments to the test statistics.
2. Nonparametric methods, which avoid dependence on the sampling distribution by making strictly probabilistic arguments (often referred to as distribution-free methods).

In subsequent chapters we will give examples of situations in which assumptions are not fulfilled and briefly describe some results of alternative methods. A more complete presentation of nonparametric methods is found in Chapter 14. Trimming and other robust methods are not presented in this text (see Koopmans, 1987).

3.5.1 Statistical Significance versus Practical Significance

The use of statistical hypothesis testing provides a powerful tool for decision making. In fact, there really is no other way to determine whether two or more population means differ based solely on the results of one sample or one experiment. However, a statistically significant result cannot be interpreted simply by itself. In fact, we can have a statistically significant result that has no practical implications, or we may not have a statistically significant result, yet useful information may be obtained from the data. For example, a market research survey of potential customers might find that a potential market exists for a particular product. The next question to be answered is whether this market is such that a reasonable expectation exists for making profit if the product is marketed in the area. That is, does the mere existence of a potential market guarantee a profit? Probably not. Further investigation must be done before recommending marketing of the product, especially if the marketing is expensive. The following examples are illustrations of the difference between statistical significance and practical significance.

■ Example 3.7

This is an example of a statistically significant result that is not practically significant.

In the January/February 1992 *International Contact Lens Clinic* publication, there is an article that presented the results of a clinical trial designed to determine the effect of defective disposable contact lenses on ocular integrity (Efron and Veys, 1992). The study involved 29 subjects, each of whom wore a defective lens in one eye and a nondefective one in the other eye. The design of the study was such that neither the research officer nor the subject was informed of which eye wore the defective lens. In particular, the study indicated that a significantly greater ocular response was observed in eyes wearing defective lenses in the form of corneal epithelial microcysts (among other results). The test had a p value of 0.04. Using a level of significance of 0.05, the conclusion would be that the defective lenses resulted in more microcysts being measured. The study reported a mean number of microcysts for the eyes wearing defective lenses as 3.3 and the mean for eyes wearing the nondefective lenses as 1.6. In an invited commentary following the article, Dr. Michel Guillon makes an interesting observation concerning the presence of microcysts. The commentary points out that the observation of fewer than 50 microcysts per eye requires no clinical action other than regular patient follow-up. The commentary further states that it is logical to conclude that an incidence of microcysts so much lower than the established guideline for action is not clinically significant. Thus, we have an example of the case where statistical significance exists but where there is no practical significance. ■

■ Example 3.8

A major impetus for developing the statistical hypothesis test was to avoid jumping to conclusions simply on the basis of apparent results. Consequently, if some result is not statistically significant the story usually ends. However it is possible to have practical significance but not statistical significance. In a recent study of the effect of a certain diet on weight reduction, a random sample of 10 subjects was weighed, put on a diet for 2 weeks, and weighed again. The results are given in Table 3.2.

Table 3.2 Weight Difference (in pounds)

Subject	Weight Before	Weight After	Difference (Before – After)
1	120	119	+1
2	131	130	+1
3	190	188	+2
4	185	183	+2
5	201	188	+13
6	121	119	+2
7	115	114	+1
8	145	144	+1
9	220	243	−23
10	190	188	+2

Solution

A hypothesis test comparing the mean weight before with the mean weight after (see Section 5.4 for the exact procedure for this test) would result in a p value of 0.21. Using a level of significance of 0.05 there would not be sufficient evidence to reject the null hypothesis and the conclusion would be that there is no significant loss in weight due to the diet. However, note that 9 of the 10 subjects lost weight! This means that the diet is probably effective in reducing weight, but perhaps does not take a lot of it off. Obviously, the observation that almost all the subjects did in fact lose weight does not take into account the amount of weight lost, which is what the hypothesis test did. So in effect, the fact that 9 of the 10 subjects lost weight (90%) really means that the proportion of subjects losing weight is high rather than that the mean weight loss differs from 0.

We can evaluate this phenomenon by calculating the probability that the results we observed occurred strictly due to chance using the basic principles of probability of Chapter 2. That is, we can calculate the probability that 9 of the 10 differences in before and after weight are in fact positive if the diet does not affect the subjects' weight. If the sign of the difference is really due to chance, then the probability of an individual difference being positive would be 0.5 or 1/2. The probability of 9 of the 10 differences being positive would then be $10(0.5)(0.5)^9$ or 0.009765—a very small value. Thus, it is highly unlikely that we could get 9 of the 10 differences positive due to chance so there is something else causing the differences. That something must be the diet.

Note that although the results appear to be contradictory, we actually tested two different hypotheses. The first one was a test to compare the weight before and after. Thus, if there was a significant increase or decrease in the average weight we would have rejected this hypothesis. On the other hand, the second analysis was really a hypothesis test to determine whether the probability of losing weight is really 0.5 or 1/2. We discuss this type of a hypothesis test in the next chapter. ■

3.6 CHAPTER SUMMARY

We have discussed two types of statistical inference, hypothesis tests and confidence intervals, which are closely related but different in their purposes. Most of the discussion has focused on hypothesis testing, which is directed toward evaluating the strength of the evidence for a proposition. We will see many types of hypothesis tests in this text, but all will have the same three stages:

1. State the hypotheses and the significance level.
2. Collect data and compute test statistics.
3. Make a decision to confirm or deny hypothesis.

Confidence intervals are designed to estimate the value of a parameter, or size of an effect. They also have a set of formal stages:

1. Identify the parameter and the confidence level.
2. Collect data and compute the statistics for the confidence interval.
3. Interpret the interval in the context of the situation.

The statistical inference principles presented in this chapter, often referred to as the Neyman-Pearson principles, may seem awkward at first. This is especially true of the hypothesis testing procedures, with their distinction between null and alternative hypotheses. However, despite the jokes about statistics, statisticians, and liars, this cumbersome procedure is specifically devised to make it difficult to lie with statistics. It articulates a philosophy that says that evidence must be based on comparing the results in data sets to the predictions of a well-specified null hypothesis. It lays out a method for gauging the strength of the evidence, in the form of a probability calculation. The noncommittal sound to some of the conclusions (e.g., "there is no significant evidence that the medicine is effective") is an intentional reminder that the statistical results will always contain an element of uncertainty.

Both the potential and limitations of statistical inference can be illustrated by considering current public controversies. Consider two hypotheses: *human activity contributes to climate change*, and *human activity does not contribute to climate change*. The highly polarized debate can partly be understood as an argument as to which should be the alternative hypothesis. Considering the costs of changing human activity in the case we decided we were contributing to climate change, many argue that the first statement should be the alternative hypothesis. Others would say that the cost of failing to act, and then finding too late that we had caused climate change, implies that the latter should be the alternative hypothesis. In fact, Neyman-Pearson principles do not adapt well to H_1 of the form "does not contribute." A more sophisticated technique called sequential sampling (Wald, 1947) could be helpful here. Since that is beyond the scope of this text, perhaps a compromise would be to set

H_0: *human activity does not contribute to climate change*
H_1: *human activity does contribute to climate change*

and use a fairly high level of α, essentially agreeing to act as soon as moderately strong evidence is available.

This might provide a framework for the debate, but now the hard science of modeling and measuring human activities' contribution to climate change must proceed. Once that data arrives, a further statistical distinction will cause debate. In Neyman-Pearson theory, the dichotomy between *does not contribute* and *does contribute* is absolute, with no middle ground for *does contribute but at very small levels*. The inference is only concerned with whether a result is too large to be attributable to chance. After all, a result can be significant (i.e., not explainable by chance), but still such a small effect that the practical implications are nil. We can already read commentary in the popular press along these lines—*humans are probably contributing but only in very small ways and it would be too expensive to change the way we do things*.

Obviously, part of the problem is that the Neyman-Pearson framework only has the choice of α as a mechanism for comparing competing costs of incorrect decisions. A more elaborate framework for balancing costs is based on penalty or payoff functions. These assign a range of costs to different degrees of statistical error. This is the foundation of statistical decision theory, widely used in economics and finance. (See Pratt *et al*., 1995.)

It might seem, then, that the inferential procedures we have presented will be of little help in debating some of our thorniest controversies. In fact, it will be essential to the core problem of assessing the results from the science. In part, this is because scientists generally work with specific mathematical models of systems described by sets of parameters. Now Neyman-Pearson theory is wonderfully adapted to assessing statements about parameters, and so the scientific literature abounds with both confidence intervals and hypothesis tests derived using many of the statistical techniques we will cover in this text. These inferential techniques are applied in two ways. First, they are used in an exploratory mode, where large numbers of possible hypotheses are checked. Here p values cannot be interpreted precisely, but act as a useful sorting device to separate promising from unpromising hypotheses. Finally, inference is applied in confirmatory mode in follow-up experiments testing a focused set of statements, using precisely the steps outlined at the beginning of this section.

The concepts presented in this chapter therefore represent fundamental ideas for gauging evidence, whether it be in the behavioral, social, life, or physical sciences. In essence, we are presenting a formal framework for critical thinking in the presence of incomplete and variable data.

3.7 CHAPTER EXERCISES

Concept Questions

This section consists of some true/false questions regarding concepts of statistical inference. Indicate whether a statement is true or false and, if false, indicate what is required to make the statement true.

1. ________ In a hypothesis test, the p value is 0.043. This means that the null hypothesis would be rejected at $\alpha = 0.05$.

2. ________ If the null hypothesis is rejected by a one-tailed hypothesis test, then it will also be rejected by a two-tailed test.

3. ________ If a null hypothesis is rejected at the 0.01 level of significance, it will also be rejected at the 0.05 level of significance.

4. ________ If the test statistic falls in the rejection region, the null hypothesis has been proven to be true.

5. ________ The risk of a type II error is directly controlled in a hypothesis test by establishing a specific significance level.

6. ________ If the null hypothesis is true, increasing only the sample size will increase the probability of rejecting the null hypothesis.

7. ________ If the null hypothesis is false, increasing the level of significance (α) for a specified sample size will increase the probability of rejecting the null hypothesis.

8. ________ If we decrease the confidence coefficient for a fixed n, we decrease the width of the confidence interval.

9. ________ If a 95% confidence interval on μ was from 50.5 to 60.6, we would reject the null hypothesis that $\mu = 60$ at the 0.05 level of significance.

10. ________ If the sample size is increased and the level of confidence is decreased, the width of the confidence interval will increase.

11. ________ A research article reports that a 95% confidence interval for mean reaction time is from 0.25 to 0.29 seconds. About 95% of individuals will have reaction times in this interval.

Practice Exercises

The following exercises are designed to give the reader practice in doing statistical inferences through small examples. The solutions are given in the back of the text.

1. From extensive research it is known that the population of a particular species of fish has a mean length $\mu = 171\,\text{mm}$ and a standard deviation $\sigma = 44\,\text{mm}$. The lengths are known to have a normal distribution. A sample of 100 fish from such a population yielded a mean length $\bar{y} = 167$ mm. Compute the 0.95 confidence interval for the mean length of the sampled population. Assume the standard deviation of the population is also 44 mm.

2. Using the data in Exercise 1 and using a 0.05 level of significance, test the null hypothesis that the population sampled has a mean of $\mu = 171$. Use a two-tailed alternative.

3. What sample size is required for a maximum error of estimation of 10 for a population whose standard deviation is 40 using a confidence interval of 0.95? How much larger must the sample size be if the maximum error is to be 5?

4. The following sample was taken from a normally distributed population with a known standard deviation $\sigma = 4$. Test the hypothesis that the mean $\mu = 20$ using a level of significance of 0.05 and the alternative that $\mu > 20$:

23, 32, 22, 31, 27, 25, 21, 24, 20, 18.

Multiple Choice Questions

1. In testing the null hypothesis that $p = 0.3$ against the alternative that $p \neq 0.3$, the probability of a type II error is ________ when the true $p = 0.4$ than when $p = 0.6$.
   (1) the same
   (2) smaller
   (3) larger
   (4) none of the above

2. In a hypothesis test the p value is 0.043. This means that we can find statistical significance at:
   (1) both the 0.05 and 0.01 levels
   (2) the 0.05 but not at the 0.01 level
   (3) the 0.01 but not at the 0.05 level
   (4) neither the 0.05 or 0.01 levels
   (5) none of the above

3. A research report states: The differences between public and private school seventh graders' attitudes toward minority groups was statistically significant at the $\alpha = 0.05$ level. This means that:
   (1) It has been proven that the two groups are different.
   (2) There is a probability of 0.05 that the attitudes of the two groups are different.
   (3) There is a probability of 0.95 that the attitudes of the two groups are different.
   (4) If there is no difference between the groups, the difference observed in the sample would occur by chance with probability of no more than 0.05.
   (5) None of the above is correct.

4. If the null hypothesis is really false, which of these statements characterizes a situation where the value of the test statistic falls in the rejection region?
   (1) The decision is correct.
   (2) A type I error has been committed.
   (3) A type II error has been committed.
   (4) Insufficient information has been given to make a decision.
   (5) None of the above is correct.

5. If the null hypothesis is really false, which of these statements characterizes a situation where the value of the test statistic does not fall in the rejection region?
   (1) The decision is correct.
   (2) A type I error has been committed.
   (3) A type II error has been committed.
   (4) Insufficient information has been given to make a decision.
   (5) None of the above is correct.

6. If the value of any test statistic does not fall in the rejection region, the decision is:
   (1) Reject the null hypothesis.
   (2) Reject the alternative hypothesis.
   (3) Fail to reject the null hypothesis.
   (4) Fail to reject the alternative hypothesis.
   (5) There is insufficient information to make a decision.

7. For a particular sample, the 0.95 confidence interval for the population mean is from 11 to 17. You are asked to test the hypothesis that the population mean is 18 against a two-sided alternative. Your decision is:
   (1) Fail to reject the null hypothesis, $\alpha = 0.05$.
   (2) Reject the null hypothesis, $\alpha = 0.05$.
   (3) There is insufficient information to decide.

8. Failure to reject the null hypothesis means:
   (1) acceptance of the alternative hypothesis
   (2) rejection of the null hypothesis
   (3) rejection of the alternative hypothesis
   (4) absolute acceptance of the null hypothesis
   (5) none of the above

9. If we decrease the confidence level, the width of the confidence interval will:
   (1) increase
   (2) remain unchanged
   (3) decrease
   (4) double
   (5) none of the above

10. If the value of the test statistic falls in the rejection region, then:
   (1) We cannot commit a type I error.
   (2) We cannot commit a type II error.

(3) We have proven that the null hypothesis is true.
(4) We have proven that the null hypothesis is false.
(5) None of the above is correct.

11. You are reading a research article that states that there is no significant evidence that the median income in the two groups differs, at $\alpha = 0.05$. You are interested in this conclusion, but prefer to use $\alpha = 0.01$.
(1) You would also say there is no significant evidence that the medians differ.
(2) You would say there is significant evidence that the medians differ.
(3) You do not know whether there is significant evidence or not, until you know the p value.

Exercises

1. The following pose conceptual hypothesis test situations. For each situation define H_0 and H_1 so as to provide control of the more serious error. Justify your choice and comment on logical values for α.
(a) You are deciding whether you should take an umbrella to work.
(b) You are planning a proficiency testing procedure to determine whether some employees should be fired.
(c) Same as part (b) except you want to determine whether some employees deserve a special merit raise.
(d) A cigarette manufacturer is conducting a test of nicotine content in order to justify a new advertising claim.
(e) You are considering the procedure to decide guilt or innocence in a court of law.
(f) You are wondering whether you should buy a new battery for your calculator before the next statistics test.
(g) As a university administrator you are considering a policy to restrict student driving in order to improve scholastic achievement.

2. Suppose that in Example 3.3, σ was 0.15 instead of 0.2 and we decided to adjust the machine if a sample of 16 had a mean weight below 7.9 or above 8.1 (same as before).
(a) What is the probability of a type I error now?
(b) Draw the operating characteristic curve using the rejection region obtained in part (a).

3. Assume that a random sample of size 25 is to be taken from a normal population with $\mu = 10$ and $\sigma = 2$. The value of μ, however, is not known by the person taking the sample.
(a) Suppose that the person taking the sample tests H_0: $\mu = 10.4$ against H_1: $\mu \neq 10.4$. Although this null hypothesis is not true, it may not be rejected, and a type II error may therefore be committed. Compute β if $\alpha = 0.05$.
(b) Suppose the same hypothesis is to be tested as that of part (a) but $\alpha = 0.01$. Compute β.

(c) Suppose the person wanted to test $H_0: \mu = 11.2$ against $H_1: \mu \neq 11.2$. Compute β for $\alpha = 0.05$ and $\alpha = 0.01$.

(d) Suppose that the person decided to use $H_1: \mu < 11.2$. Calculate β for $\alpha = 0.05$ and $\alpha = 0.01$.

(e) What principles of hypothesis testing are illustrated by these exercises?

4. Repeat Exercise 3 using $n = 100$. What principles of hypothesis testing do these exercises illustrate?

5. A standardized test for a specific college course is constructed so that the distribution of grades should have $\mu = 100$ and $\sigma = 10$. A class of 30 students has a mean grade of 92.

   (a) Test the null hypothesis that the grades from this class are a random sample from the stated distribution. (Use $\alpha = 0.05$.)

   (b) What is the p value associated with this test?

   (c) Discuss the practical uses of the results of this statistical test.

6. The family incomes in a certain city in 1970 had a mean of \$14,200 with a standard deviation of \$2600. A random sample of 75 families taken in 1975 produced $\bar{y} = \$15{,}300$ (adjusted for inflation).

   (a) Assume σ has remained unchanged and test to see whether mean income has changed using a 0.05 level of significance.

   (b) Construct a 99% confidence interval on mean family income in 1975.

   (c) Construct the power curve for the test in part (a).

7. Suppose in Example 3.2 we were to reject H_0 if all the jelly beans in a sample of size four were red.

   (a) What is α?

   (b) What is β?

8. Suppose that for a given population with $\sigma = 7.2$ we want to test $H_0: \mu = 80$ against $H_1: \mu < 80$ based on a sample of $n = 100$.

   (a) If the null hypothesis is rejected when $\bar{y} < 76$, what is the probability of a type I error?

   (b) What would be the rejection region if we wanted to have a level of significance of exactly 0.05?

9. An experiment designed to estimate the mean reaction time of a certain chemical process has $\bar{y} = 79.6$ s, based on 144 observations. The standard deviation is $\sigma = 8$.

   (a) What is the maximum error of estimate at 0.95 confidence?

   (b) Construct a 0.95 confidence interval on μ.

   (c) How large a sample must be taken so that the 0.95 maximum error of estimate is 1 s or less?

10. A drug company is testing a drug intended to increase heart rate. A sample of 100 yielded a mean increase of 1.4 beats per minute, with a standard deviation known to be 3.6. Since the company wants to avoid marketing an ineffective

drug, it proposes a 0.001 significance level. Should it market the drug? (*Hint:* If the drug does not work, the mean increase will be zero.)

11. The manufacturer of auto windows discussed in Exercise 18 of Chapter 2 has developed a new plastic material that can be applied much thinner than the conventional material. To use this material, however, the production machinery must be adjusted. A trial adjustment was made on one of the 10 machines used in production, and a sample of 25 windshields measured. This sample had a mean thickness of 2.9 mm. Using the standard deviation of 0.25 mm, does this adjustment provide for a smaller thickness in the material than the old adjustment (4 mm)? (Use a hypothesis test and level of significance of 0.01. Assume the distribution of thickness is approximately normal.)

12. The manufacturer in Exercise 11 tried another, less expensive adjustment on another machine. A sample of 25 windshields was measured yielding a sample mean thickness of 3.4. Calculate the p value resulting from this mean using the same hypothesis and assumptions as in Exercise 11.

13. An experiment is conducted to determine whether a new computer program will speed up the processing of credit card billing at a large bank. The mean time to process billing using the present program is 12.3 min. with a standard deviation of 3.5 min. The new program is tested with 100 billings and yielded a sample mean of 10.9 min. Assuming the standard deviation of times in the new program is the same as the old, does the new program significantly reduce the time of processing? Use $\alpha = 0.05$.

14. Another bank is experimenting with programs to direct bill companies for commercial loans. They are particularly interested in the number of errors of a billing program. To examine a particular program, a simulation of 1000 typical loans is run through the program. The simulation yielded a mean of 4.6 errors with a standard deviation of 0.5. Construct a 95% confidence interval on the true mean error rate.

15. If the bank wanted to examine a program similar to that of Exercise 14 and wanted a maximum error of estimation of 0.01 with a level of confidence of 95%, how large a sample should be taken? (Assume that the standard deviation of the number of errors remains the same.)

16. In the United States, the probability a child will be born with a birth defect is thought to be 3%. In a certain community, there were 40 independent births during the last year, and three of those had birth defects. Using $\alpha = 0.05$, would this constitute evidence that this community had an elevated probability of birth defects?
    (a) State your hypotheses in terms of this community's true probability of birth defect, p.
    (b) Knowing that the number of birth defects out of 40 independent births follows a binomial distribution, calculate the p value.
    (c) Use your result from (b) to state the conclusion.

17. The public health official monitoring the community in exercise 16 uses the following rejection rule: decide there has been an increase in the probability of birth defects if there are four or more birth defects among 40 independent births.

(a) What α is the official using?

(b) What will β be, if the true probability of a birth defect has increased to 10%?

18. A large national survey of American dietary habits showed a mean calorie consumption of 2700 kcal and a standard deviation of 450 kcal among teenage boys. You are studying dietary habits of children in your county to see if they differ from the national norm.

(a) In a sample of 36 teenage boys, you find a mean consumption of 2620 kcal. At $\alpha = 0.05$, is this significant evidence that the mean in your county differs from the national mean? Assume that the standard deviation observed nationally can be used for σ.

(b) Using $\alpha = 0.05$ and a sample of size 36, what is the probability that you will actually be able to detect a situation where your county has a mean of only 2600 kcal? (That is, what is the power if $\mu = 2600$?)

19. Refer to the information in Problem 18.

(a) Give a 95% confidence interval for the mean consumption among teenage boys in your county.

(b) The confidence interval in (b) has a wide margin of error. What sample size would you suggest if you wanted a margin of error of only 75 kcal?

20. An insurance company randomly selects 100 claims from a chain of dialysis clinics and conducts an audit. The mean overpayment per claim in this sample is $21.32. The company is interested in extrapolating this information to the population of all claims from this chain. They want to make a statement of the form "with confidence level 95%, the mean overpayment per claim is at least ________."

(a) Based on past experience, the company assumes that $\sigma = 32.45$. Compute the appropriate confidence limit.

(b) What is likely to be true about the shape of the distribution for the individual overpayments? Why is the large sample size a critical part of this problem?

21. A hospital has observed that the number of nosocomial (hospital-acquired) pneumonias (NP) in its intensive care unit follows a Poisson distribution with a rate of 1.8 per month. The hospital's infection-control officer monitors the number of NP each month, and calls for an expensive additional equipment sterilization effort if four or more infections are reported. The alternative hypothesis is that the rate of infections has increased.

(a) Assuming that the rate of infections has not increased, what is the probability that the officer will call for the sterilization effort?

(b) The officer repeats the monitoring every month. Over a 12-month period in which the rate stays at 1.8, what is the probability the officer will call for the sterilization effort at least once? Assume each month's count of new infections is independent of the other months.