Chapter 10

One- and Two-Sample Tests of Hypotheses

10.1 Statistical Hypotheses: General Concepts

Often, the problem confronting the scientist or engineer is not so much the estimation of a population parameter, as discussed in Chapter 9, but rather the formation of a data-based decision procedure that can produce a conclusion about some scientific system. For example, a medical researcher may decide on the basis of experimental evidence whether coffee drinking increases the risk of cancer in humans; an engineer might have to decide on the basis of sample data whether there is a difference between the accuracy of two kinds of gauges; or a sociologist might wish to collect appropriate data to enable him or her to decide whether a person's blood type and eye color are independent variables. In each of these cases, the scientist or engineer *postulates* or *conjectures* something about a system. In addition, each must make use of experimental data and make a decision based on the data. In each case, the conjecture can be put in the form of a statistical hypothesis. Procedures that lead to the acceptance or rejection of statistical hypotheses such as these comprise a major area of statistical inference. First, let us define precisely what we mean by a **statistical hypothesis**.

Definition 10.1: A **statistical hypothesis** is an assertion or conjecture concerning one or more populations.

The truth or falsity of a statistical hypothesis is never known with absolute certainty unless we examine the entire population. This, of course, would be impractical in most situations. Instead, we take a random sample from the population of interest and use the data contained in this sample to provide evidence that either supports or does not support the hypothesis. Evidence from the sample that is inconsistent with the stated hypothesis leads to a rejection of the hypothesis.

The Role of Probability in Hypothesis Testing

It should be made clear to the reader that the decision procedure must include an awareness of the *probability of a wrong conclusion*. For example, suppose that the hypothesis postulated by the engineer is that the fraction defective p in a certain process is 0.10. The experiment is to observe a random sample of the product in question. Suppose that 100 items are tested and 12 items are found defective. It is reasonable to conclude that this evidence does not refute the condition that the binomial parameter $p = 0.10$, and thus it may lead one not to reject the hypothesis. However, it also does not refute $p = 0.12$ or perhaps even $p = 0.15$. As a result, the reader must be accustomed to understanding that **rejection of a hypothesis implies that the sample evidence refutes it**. Put another way, **rejection means that there is a small probability of obtaining the sample information observed when, in fact, the hypothesis is true**. For example, for our proportion-defective hypothesis, a sample of 100 revealing 20 defective items is certainly evidence for rejection. Why? If, indeed, $p = 0.10$, the probability of obtaining 20 or more defectives is approximately 0.002. With the resulting small risk of a wrong conclusion, it would seem safe to **reject the hypothesis** that $p = 0.10$. In other words, rejection of a hypothesis tends to all but "rule out" the hypothesis. On the other hand, it is very important to emphasize that acceptance or, rather, failure to reject does not rule out other possibilities. As a result, the *firm conclusion is established by the data analyst when a hypothesis is rejected.*

The formal statement of a hypothesis is often influenced by the structure of the probability of a wrong conclusion. If the scientist is interested in *strongly supporting* a contention, he or she hopes to arrive at the contention in the form of rejection of a hypothesis. If the medical researcher wishes to show strong evidence in favor of the contention that coffee drinking increases the risk of cancer, the hypothesis tested should be of the form "there is no increase in cancer risk produced by drinking coffee." As a result, the contention is reached via a rejection. Similarly, to support the claim that one kind of gauge is more accurate than another, the engineer tests the hypothesis that there is no difference in the accuracy of the two kinds of gauges.

The foregoing implies that when the data analyst formalizes experimental evidence on the basis of hypothesis testing, the formal **statement of the hypothesis** is very important.

The Null and Alternative Hypotheses

The structure of hypothesis testing will be formulated with the use of the term **null hypothesis**, which refers to any hypothesis we wish to test and is denoted by H_0. The rejection of H_0 leads to the acceptance of an **alternative hypothesis**, denoted by H_1. An understanding of the different roles played by the null hypothesis (H_0) and the alternative hypothesis (H_1) is crucial to one's understanding of the rudiments of hypothesis testing. The alternative hypothesis H_1 usually represents the *question to be answered or the theory to be tested,* and thus its specification is crucial. The null hypothesis H_0 *nullifies or opposes* H_1 and is often the logical complement to H_1. As the reader gains more understanding of hypothesis testing, he or she should note that the analyst arrives at one of the two following

conclusions:

***reject* H_0** in favor of H_1 because of sufficient evidence in the data or

***fail to reject* H_0** because of insufficient evidence in the data.

Note that the *conclusions do not involve a formal and literal "accept* H_0*."* The statement of H_0 often represents the "status quo" in opposition to the new idea, conjecture, and so on, stated in H_1, while failure to reject H_0 represents the proper conclusion. In our binomial example, the practical issue may be a concern that the historical defective probability of 0.10 no longer is true. Indeed, the conjecture may be that p exceeds 0.10. We may then state

$$H_0\text{: } p = 0.10,$$
$$H_1\text{: } p > 0.10.$$

Now 12 defective items out of 100 does not refute $p = 0.10$, so the conclusion is "fail to reject H_0." However, if the data produce 20 out of 100 defective items, then the conclusion is "reject H_0" in favor of H_1: $p > 0.10$.

Though the applications of hypothesis testing are quite abundant in scientific and engineering work, perhaps the best illustration for a novice lies in the predicament encountered in a jury trial. The null and alternative hypotheses are

$$H_0\text{: defendant is innocent,}$$
$$H_1\text{: defendant is guilty.}$$

The indictment comes because of suspicion of guilt. The hypothesis H_0 (the status quo) stands in opposition to H_1 and is maintained unless H_1 is supported by evidence "beyond a reasonable doubt." However, "failure to reject H_0" in this case does not imply innocence, but merely that the evidence was insufficient to convict. So the jury does not necessarily *accept* H_0 but *fails to reject* H_0.

10.2 Testing a Statistical Hypothesis

To illustrate the concepts used in testing a statistical hypothesis about a population, we present the following example. A certain type of cold vaccine is known to be only 25% effective after a period of 2 years. To determine if a new and somewhat more expensive vaccine is superior in providing protection against the same virus for a longer period of time, suppose that 20 people are chosen at random and inoculated. (In an actual study of this type, the participants receiving the new vaccine might number several thousand. The number 20 is being used here only to demonstrate the basic steps in carrying out a statistical test.) If more than 8 of those receiving the new vaccine surpass the 2-year period without contracting the virus, the new vaccine will be considered superior to the one presently in use. The requirement that the number exceed 8 is somewhat arbitrary but appears reasonable in that it represents a modest gain over the 5 people who could be expected to receive protection if the 20 people had been inoculated with the vaccine already in use. We are essentially testing the null hypothesis that the new vaccine is equally effective after a period of 2 years as the one now commonly used. The alternative

hypothesis is that the new vaccine is in fact superior. This is equivalent to testing the hypothesis that the binomial parameter for the probability of a success on a given trial is $p = 1/4$ against the alternative that $p > 1/4$. This is usually written as follows:

$$H_0: \ p = 0.25,$$
$$H_1: \ p > 0.25.$$

The Test Statistic

The **test statistic** on which we base our decision is X, the number of individuals in our test group who receive protection from the new vaccine for a period of at least 2 years. The possible values of X, from 0 to 20, are divided into two groups: those numbers less than or equal to 8 and those greater than 8. All possible scores greater than 8 constitute the **critical region**. The last number that we observe in passing into the critical region is called the **critical value**. In our illustration, the critical value is the number 8. Therefore, if $x > 8$, we reject H_0 in favor of the alternative hypothesis H_1. If $x \leq 8$, we fail to reject H_0. This decision criterion is illustrated in Figure 10.1.

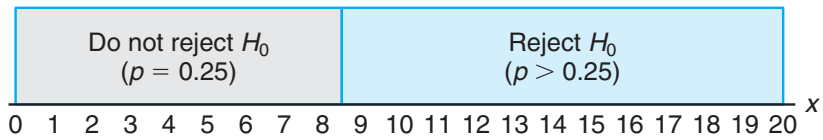

Figure 10.1: Decision criterion for testing $p = 0.25$ versus $p > 0.25$.

The Probability of a Type I Error

The decision procedure just described could lead to either of two wrong conclusions. For instance, the new vaccine may be no better than the one now in use (H_0 true) and yet, in this particular randomly selected group of individuals, more than 8 surpass the 2-year period without contracting the virus. We would be committing an error by rejecting H_0 in favor of H_1 when, in fact, H_0 is true. Such an error is called a **type I error**.

Definition 10.2: Rejection of the null hypothesis when it is true is called a **type I error**.

A second kind of error is committed if 8 or fewer of the group surpass the 2-year period successfully and we are unable to conclude that the vaccine is better when it actually is better (H_1 true). Thus, in this case, we fail to reject H_0 when in fact H_0 is false. This is called a **type II error**.

Definition 10.3: Nonrejection of the null hypothesis when it is false is called a **type II error**.

In testing any statistical hypothesis, there are four possible situations that determine whether our decision is correct or in error. These four situations are

summarized in Table 10.1.

Table 10.1: Possible Situations for Testing a Statistical Hypothesis

	H_0 **is true**	H_0 **is false**
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

The probability of committing a type I error, also called the **level of significance**, is denoted by the Greek letter α. In our illustration, a type I error will occur when more than 8 individuals inoculated with the new vaccine surpass the 2-year period without contracting the virus and researchers conclude that the new vaccine is better when it is actually equivalent to the one in use. Hence, if X is the number of individuals who remain free of the virus for at least 2 years,

$$\alpha = P(\text{type I error}) = P\left(X > 8 \text{ when } p = \frac{1}{4}\right) = \sum_{x=9}^{20} b\left(x; 20, \frac{1}{4}\right)$$

$$= 1 - \sum_{x=0}^{8} b\left(x; 20, \frac{1}{4}\right) = 1 - 0.9591 = 0.0409.$$

We say that the null hypothesis, $p = 1/4$, is being tested at the $\alpha = 0.0409$ level of significance. Sometimes the level of significance is called the **size of the test**. A critical region of size 0.0409 is very small, and therefore it is unlikely that a type I error will be committed. Consequently, it would be most unusual for more than 8 individuals to remain immune to a virus for a 2-year period using a new vaccine that is essentially equivalent to the one now on the market.

The Probability of a Type II Error

The probability of committing a type II error, denoted by β, is impossible to compute unless we have a specific alternative hypothesis. If we test the null hypothesis that $p = 1/4$ against the alternative hypothesis that $p = 1/2$, then we are able to compute the probability of not rejecting H_0 when it is false. We simply find the probability of obtaining 8 or fewer in the group that surpass the 2-year period when $p = 1/2$. In this case,

$$\beta = P(\text{type II error}) = P\left(X \leq 8 \text{ when } p = \frac{1}{2}\right)$$

$$= \sum_{x=0}^{8} b\left(x; 20, \frac{1}{2}\right) = 0.2517.$$

This is a rather high probability, indicating a test procedure in which it is quite likely that we shall reject the new vaccine when, in fact, it is superior to what is now in use. Ideally, we like to use a test procedure for which the type I and type II error probabilities are both small.

It is possible that the director of the testing program is willing to make a type II error if the more expensive vaccine is not significantly superior. In fact, the only

time he wishes to guard against the type II error is when the true value of p is at least 0.7. If $p = 0.7$, this test procedure gives

$$\beta = P(\text{type II error}) = P(X \leq 8 \text{ when } p = 0.7)$$
$$= \sum_{x=0}^{8} b(x; 20, 0.7) = 0.0051.$$

With such a small probability of committing a type II error, it is extremely unlikely that the new vaccine would be rejected when it was 70% effective after a period of 2 years. As the alternative hypothesis approaches unity, the value of β diminishes to zero.

The Role of α, β, and Sample Size

Let us assume that the director of the testing program is unwilling to commit a type II error when the alternative hypothesis $p = 1/2$ is true, even though we have found the probability of such an error to be $\beta = 0.2517$. It is always possible to reduce β by increasing the size of the critical region. For example, consider what happens to the values of α and β when we change our critical value to 7 so that all scores greater than 7 fall in the critical region and those less than or equal to 7 fall in the nonrejection region. Now, in testing $p = 1/4$ against the alternative hypothesis that $p = 1/2$, we find that

$$\alpha = \sum_{x=8}^{20} b\left(x; 20, \frac{1}{4}\right) = 1 - \sum_{x=0}^{7} b\left(x; 20, \frac{1}{4}\right) = 1 - 0.8982 = 0.1018$$

and

$$\beta = \sum_{x=0}^{7} b\left(x; 20, \frac{1}{2}\right) = 0.1316.$$

By adopting a new decision procedure, we have reduced the probability of committing a type II error at the expense of increasing the probability of committing a type I error. For a fixed sample size, a decrease in the probability of one error will usually result in an increase in the probability of the other error. Fortunately, **the probability of committing both types of error can be reduced by increasing the sample size**. Consider the same problem using a random sample of 100 individuals. If more than 36 of the group surpass the 2-year period, we reject the null hypothesis that $p = 1/4$ and accept the alternative hypothesis that $p > 1/4$. The critical value is now 36. All possible scores above 36 constitute the critical region, and all possible scores less than or equal to 36 fall in the acceptance region.

To determine the probability of committing a type I error, we shall use the normal curve approximation with

$$\mu = np = (100)\left(\frac{1}{4}\right) = 25 \quad \text{and} \quad \sigma = \sqrt{npq} = \sqrt{(100)(1/4)(3/4)} = 4.33.$$

Referring to Figure 10.2, we need the area under the normal curve to the right of $x = 36.5$. The corresponding z-value is

$$z = \frac{36.5 - 25}{4.33} = 2.66.$$

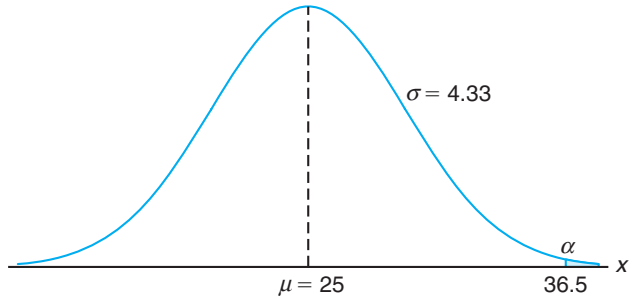

Figure 10.2: Probability of a type I error.

From Table A.3 we find that

$$\alpha = P(\text{type I error}) = P\left(X > 36 \text{ when } p = \frac{1}{4}\right) \approx P(Z > 2.66)$$
$$= 1 - P(Z < 2.66) = 1 - 0.9961 = 0.0039.$$

If H_0 is false and the true value of H_1 is $p = 1/2$, we can determine the probability of a type II error using the normal curve approximation with

$$\mu = np = (100)(1/2) = 50 \quad \text{and} \quad \sigma = \sqrt{npq} = \sqrt{(100)(1/2)(1/2)} = 5.$$

The probability of a value falling in the nonrejection region when H_0 is true is given by the area of the shaded region to the left of $x = 36.5$ in Figure 10.3. The z-value corresponding to $x = 36.5$ is

$$z = \frac{36.5 - 50}{5} = -2.7.$$

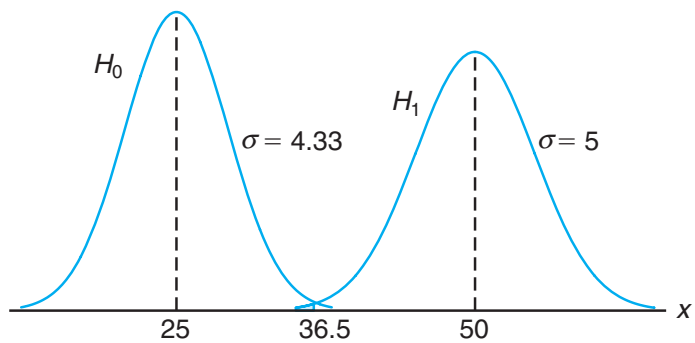

Figure 10.3: Probability of a type II error.

Therefore,

$$\beta = P(\text{type II error}) = P\left(X \leq 36 \text{ when } p = \frac{1}{2}\right) \approx P(Z < -2.7) = 0.0035.$$

Obviously, the type I and type II errors will rarely occur if the experiment consists of 100 individuals.

The illustration above underscores the strategy of the scientist in hypothesis testing. After the null and alternative hypotheses are stated, it is important to consider the sensitivity of the test procedure. By this we mean that there should be a determination, for a fixed α, of a reasonable value for the probability of wrongly accepting H_0 (i.e., the value of β) when the true situation represents some *important deviation from* H_0. A value for the sample size can usually be determined for which there is a reasonable balance between the values of α and β computed in this fashion. The vaccine problem provides an illustration.

Illustration with a Continuous Random Variable

The concepts discussed here for a discrete population can be applied equally well to continuous random variables. Consider the null hypothesis that the average weight of male students in a certain college is 68 kilograms against the alternative hypothesis that it is unequal to 68. That is, we wish to test

$$H_0: \ \mu = 68,$$
$$H_1: \ \mu \neq 68.$$

The alternative hypothesis allows for the possibility that $\mu < 68$ or $\mu > 68$.

A sample mean that falls close to the hypothesized value of 68 would be considered evidence in favor of H_0. On the other hand, a sample mean that is considerably less than or more than 68 would be evidence inconsistent with H_0 and therefore favoring H_1. The sample mean is the test statistic in this case. A critical region for the test statistic might arbitrarily be chosen to be the two intervals $\bar{x} < 67$ and $\bar{x} > 69$. The nonrejection region will then be the interval $67 \leq \bar{x} \leq 69$. This decision criterion is illustrated in Figure 10.4.

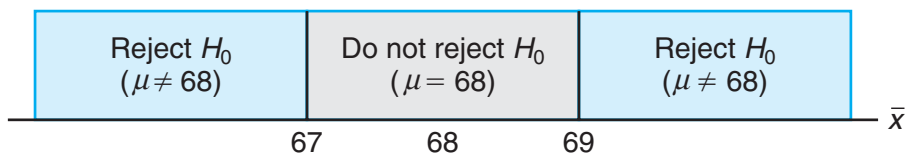

Figure 10.4: Critical region (in blue).

Let us now use the decision criterion of Figure 10.4 to calculate the probabilities of committing type I and type II errors when testing the null hypothesis that $\mu = 68$ kilograms against the alternative that $\mu \neq 68$ kilograms.

Assume the standard deviation of the population of weights to be $\sigma = 3.6$. For large samples, we may substitute s for σ if no other estimate of σ is available. Our decision statistic, based on a random sample of size $n = 36$, will be $\bar{X}$, the most efficient estimator of μ. From the Central Limit Theorem, we know that the sampling distribution of $\bar{X}$ is approximately normal with standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 3.6/6 = 0.6$.

The probability of committing a type I error, or the level of significance of our test, is equal to the sum of the areas that have been shaded in each tail of the distribution in Figure 10.5. Therefore,

$$\alpha = P(\bar{X} < 67 \text{ when } \mu = 68) + P(\bar{X} > 69 \text{ when } \mu = 68).$$

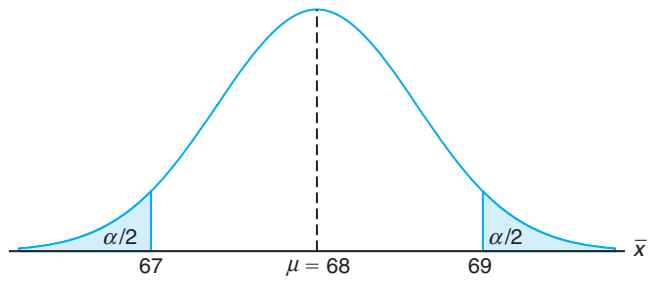

Figure 10.5: Critical region for testing $\mu = 68$ versus $\mu \neq 68$.

The z-values corresponding to $\bar{x}_1 = 67$ and $\bar{x}_2 = 69$ when H_0 is true are

$$z_1 = \frac{67 - 68}{0.6} = -1.67 \quad \text{and} \quad z_2 = \frac{69 - 68}{0.6} = 1.67.$$

Therefore,

$$\alpha = P(Z < -1.67) + P(Z > 1.67) = 2P(Z < -1.67) = 0.0950.$$

Thus, 9.5% of all samples of size 36 would lead us to reject $\mu = 68$ kilograms when, in fact, it is true. To reduce α, we have a choice of increasing the sample size or widening the fail-to-reject region. Suppose that we increase the sample size to $n = 64$. Then $\sigma_{\bar{X}} = 3.6/8 = 0.45$. Now

$$z_1 = \frac{67 - 68}{0.45} = -2.22 \quad \text{and} \quad z_2 = \frac{69 - 68}{0.45} = 2.22.$$

Hence,

$$\alpha = P(Z < -2.22) + P(Z > 2.22) = 2P(Z < -2.22) = 0.0264.$$

The reduction in α is not sufficient by itself to guarantee a good testing procedure. We must also evaluate β for various alternative hypotheses. If it is important to reject H_0 when the true mean is some value $\mu \geq 70$ or $\mu \leq 66$, then the probability of committing a type II error should be computed and examined for the alternatives $\mu = 66$ and $\mu = 70$. Because of symmetry, it is only necessary to consider the probability of not rejecting the null hypothesis that $\mu = 68$ when the alternative $\mu = 70$ is true. A type II error will result when the sample mean $\bar{x}$ falls between 67 and 69 when H_1 is true. Therefore, referring to Figure 10.6, we find that

$$\beta = P(67 \leq \bar{X} \leq 69 \text{ when } \mu = 70).$$

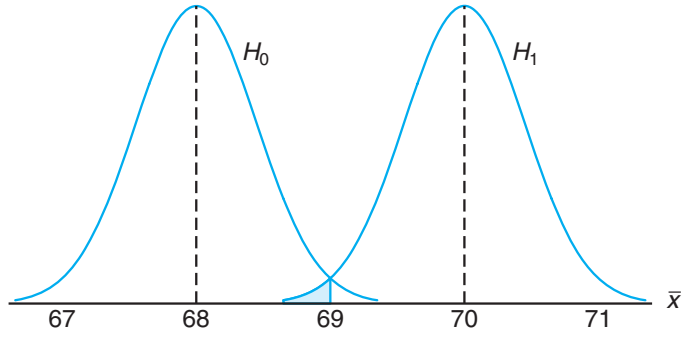

Figure 10.6: Probability of type II error for testing $\mu = 68$ versus $\mu = 70$.

The z-values corresponding to $\bar{x}_1 = 67$ and $\bar{x}_2 = 69$ when H_1 is true are

$$z_1 = \frac{67 - 70}{0.45} = -6.67 \quad \text{and} \quad z_2 = \frac{69 - 70}{0.45} = -2.22.$$

Therefore,

$$\begin{aligned}\beta = P(-6.67 < Z < -2.22) &= P(Z < -2.22) - P(Z < -6.67) \\ &= 0.0132 - 0.0000 = 0.0132.\end{aligned}$$

If the true value of μ is the alternative $\mu = 66$, the value of β will again be 0.0132. For all possible values of $\mu < 66$ or $\mu > 70$, the value of β will be even smaller when $n = 64$, and consequently there would be little chance of not rejecting H_0 when it is false.

The probability of committing a type II error increases rapidly when the true value of μ approaches, but is not equal to, the hypothesized value. Of course, this is usually the situation where we do not mind making a type II error. For example, if the alternative hypothesis $\mu = 68.5$ is true, we do not mind committing a type II error by concluding that the true answer is $\mu = 68$. The probability of making such an error will be high when $n = 64$. Referring to Figure 10.7, we have

$$\beta = P(67 \leq \bar{X} \leq 69 \text{ when } \mu = 68.5).$$

The z-values corresponding to $\bar{x}_1 = 67$ and $\bar{x}_2 = 69$ when $\mu = 68.5$ are

$$z_1 = \frac{67 - 68.5}{0.45} = -3.33 \quad \text{and} \quad z_2 = \frac{69 - 68.5}{0.45} = 1.11.$$

Therefore,

$$\begin{aligned}\beta = P(-3.33 < Z < 1.11) &= P(Z < 1.11) - P(Z < -3.33) \\ &= 0.8665 - 0.0004 = 0.8661.\end{aligned}$$

The preceding examples illustrate the following important properties:

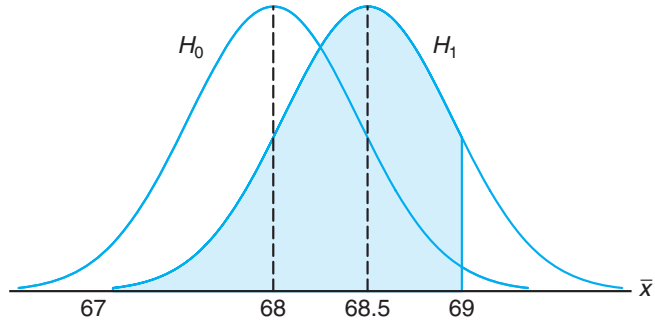

Figure 10.7: Type II error for testing $\mu = 68$ versus $\mu = 68.5$.

Important Properties of a Test of Hypothesis

1. The type I error and type II error are related. A decrease in the probability of one generally results in an increase in the probability of the other.
2. The size of the critical region, and therefore the probability of committing a type I error, can always be reduced by adjusting the critical value(s).
3. An increase in the sample size n will reduce α and β simultaneously.
4. If the null hypothesis is false, β is a maximum when the true value of a parameter approaches the hypothesized value. The greater the distance between the true value and the hypothesized value, the smaller β will be.

One very important concept that relates to error probabilities is the notion of the **power** of a test.

Definition 10.4: The **power** of a test is the probability of rejecting H_0 given that a specific alternative is true.

The power of a test can be computed as $1 - \beta$. Often **different types of tests are compared by contrasting power properties**. Consider the previous illustration, in which we were testing H_0: $\mu = 68$ and H_1: $\mu \neq 68$. As before, suppose we are interested in assessing the sensitivity of the test. The test is governed by the rule that we do not reject H_0 if $67 \leq \bar{x} \leq 69$. We seek the capability of the test to properly reject H_0 when indeed $\mu = 68.5$. We have seen that the probability of a type II error is given by $\beta = 0.8661$. Thus, the **power** of the test is $1 - 0.8661 = 0.1339$. In a sense, the power is a more succinct measure of how sensitive the test is for detecting differences between a mean of 68 and a mean of 68.5. In this case, if μ is truly 68.5, the test as described will *properly reject H_0 only 13.39% of the time*. As a result, the test would not be a good one if it was important that the analyst have a reasonable chance of truly distinguishing between a mean of 68.0 (specified by H_0) and a mean of 68.5. From the foregoing, it is clear that to produce a desirable power (say, greater than 0.8), one must either increase α or increase the sample size.

So far in this chapter, much of the discussion of hypothesis testing has focused on foundations and definitions. In the sections that follow, we get more specific

and put hypotheses in categories as well as discuss tests of hypotheses on various parameters of interest. We begin by drawing the distinction between a one-sided and a two-sided hypothesis.

One- and Two-Tailed Tests

A test of any statistical hypothesis where the alternative is **one sided**, such as

$$H_0\colon\ \theta = \theta_0,$$
$$H_1\colon\ \theta > \theta_0$$

or perhaps

$$H_0\colon\ \theta = \theta_0,$$
$$H_1\colon\ \theta < \theta_0,$$

is called a **one-tailed test**. Earlier in this section, we referred to the **test statistic** for a hypothesis. Generally, the critical region for the alternative hypothesis $\theta > \theta_0$ lies in the right tail of the distribution of the test statistic, while the critical region for the alternative hypothesis $\theta < \theta_0$ lies entirely in the left tail. (In a sense, the inequality symbol points in the direction of the critical region.) A one-tailed test was used in the vaccine experiment to test the hypothesis $p = 1/4$ against the one-sided alternative $p > 1/4$ for the binomial distribution. The one-tailed critical region is usually obvious; the reader should visualize the behavior of the test statistic and notice the obvious *signal* that would produce evidence supporting the alternative hypothesis.

A test of any statistical hypothesis where the alternative is **two sided**, such as

$$H_0\colon\ \theta = \theta_0,$$
$$H_1\colon\ \theta \neq \theta_0,$$

is called a **two-tailed test**, since the critical region is split into two parts, often having equal probabilities, in each tail of the distribution of the test statistic. The alternative hypothesis $\theta \neq \theta_0$ states that either $\theta < \theta_0$ or $\theta > \theta_0$. A two-tailed test was used to test the null hypothesis that $\mu = 68$ kilograms against the two-sided alternative $\mu \neq 68$ kilograms in the example of the continuous population of student weights.

How Are the Null and Alternative Hypotheses Chosen?

The null hypothesis H_0 will often be stated using the *equality sign*. With this approach, it is clear how the probability of type I error is controlled. However, there are situations in which "do not reject H_0" implies that the parameter θ might be any value defined by the natural complement to the alternative hypothesis. For example, in the vaccine example, where the alternative hypothesis is $H_1\colon p > 1/4$, it is quite possible that nonrejection of H_0 cannot rule out a value of p less than $1/4$. Clearly though, in the case of one-tailed tests, the statement of the alternative is the most important consideration.

Whether one sets up a one-tailed or a two-tailed test will depend on the conclusion to be drawn if H_0 is rejected. The location of the critical region can be determined only after H_1 has been stated. For example, in testing a new drug, one sets up the hypothesis that it is no better than similar drugs now on the market and tests this against the alternative hypothesis that the new drug is superior. Such an alternative hypothesis will result in a one-tailed test with the critical region in the right tail. However, if we wish to compare a new teaching technique with the conventional classroom procedure, the alternative hypothesis should allow for the new approach to be either inferior or superior to the conventional procedure. Hence, the test is two-tailed with the critical region divided equally so as to fall in the extreme left and right tails of the distribution of our statistic.

Example 10.1: A manufacturer of a certain brand of rice cereal claims that the average saturated fat content does not exceed 1.5 grams per serving. State the null and alternative hypotheses to be used in testing this claim and determine where the critical region is located.

***Solution*:** The manufacturer's claim should be rejected only if μ is greater than 1.5 milligrams and should not be rejected if μ is less than or equal to 1.5 milligrams. We test

$$H_0: \ \mu = 1.5,$$
$$H_1: \ \mu > 1.5.$$

Nonrejection of H_0 does not rule out values less than 1.5 milligrams. Since we have a one-tailed test, the greater than symbol indicates that the critical region lies entirely in the right tail of the distribution of our test statistic $\bar{X}$. ■

Example 10.2: A real estate agent claims that 60% of all private residences being built today are 3-bedroom homes. To test this claim, a large sample of new residences is inspected; the proportion of these homes with 3 bedrooms is recorded and used as the test statistic. State the null and alternative hypotheses to be used in this test and determine the location of the critical region.

***Solution*:** If the test statistic were substantially higher or lower than $p = 0.6$, we would reject the agent's claim. Hence, we should make the hypothesis

$$H_0: \ p = 0.6,$$
$$H_1: \ p \neq 0.6.$$

The alternative hypothesis implies a two-tailed test with the critical region divided equally in both tails of the distribution of $\hat{P}$, our test statistic. ■

10.3 The Use of *P*-Values for Decision Making in Testing Hypotheses

In testing hypotheses in which the test statistic is discrete, the critical region may be chosen arbitrarily and its size determined. If α is too large, it can be reduced by making an adjustment in the critical value. It may be necessary to increase the

sample size to offset the decrease that occurs automatically in the power of the test.

Over a number of generations of statistical analysis, it had become customary to choose an α of 0.05 or 0.01 and select the critical region accordingly. Then, of course, strict rejection or nonrejection of H_0 would depend on that critical region. For example, if the test is two tailed and α is set at the 0.05 level of significance and the test statistic involves, say, the standard normal distribution, then a z-value is observed from the data and the critical region is

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

where the value 1.96 is found as $z_{0.025}$ in Table A.3. A value of z in the critical region prompts the statement "The value of the test statistic is significant," which we can then translate into the user's language. For example, if the hypothesis is given by

$$H_0\colon\ \mu = 10,$$
$$H_1\colon\ \mu \neq 10,$$

one might say, "The mean differs significantly from the value 10."

Preselection of a Significance Level

This preselection of a significance level α has its roots in the philosophy that the maximum risk of making a type I error should be controlled. However, this approach does not account for values of test statistics that are "close" to the critical region. Suppose, for example, in the illustration with $H_0\colon \mu = 10$ versus $H_1\colon \mu \neq 10$, a value of $z = 1.87$ is observed; strictly speaking, with $\alpha = 0.05$, the value is not significant. But the risk of committing a type I error if one rejects H_0 in this case could hardly be considered severe. In fact, in a two-tailed scenario, one can quantify this risk as

$$P = 2P(Z > 1.87 \text{ when } \mu = 10) = 2(0.0307) = 0.0614.$$

As a result, 0.0614 is the probability of obtaining a value of z as large as or larger (in magnitude) than 1.87 when in fact $\mu = 10$. Although this evidence against H_0 is not as strong as that which would result from rejection at an $\alpha = 0.05$ level, it is important information to the user. Indeed, continued use of $\alpha = 0.05$ or 0.01 is only a result of what standards have been passed down through the generations. The ***P*-value approach has been adopted extensively by users of applied statistics**. The approach is designed to give the user an alternative (in terms of a probability) to a mere "reject" or "do not reject" conclusion. The P-value computation also gives the user important information when the z-value falls well *into the ordinary critical region*. For example, if z is 2.73, it is informative for the user to observe that

$$P = 2(0.0032) = 0.0064,$$

and thus the z-value is significant at a level considerably less than 0.05. It is important to know that under the condition of H_0, a value of $z = 2.73$ is an extremely rare event. That is, a value at least that large in magnitude would only occur 64 times in 10,000 experiments.

A Graphical Demonstration of a *P*-Value

One very simple way of explaining a P-value graphically is to consider two distinct samples. Suppose that two materials are being considered for coating a particular type of metal in order to inhibit corrosion. Specimens are obtained, and one collection is coated with material 1 and one collection coated with material 2. The sample sizes are $n_1 = n_2 = 10$, and corrosion is measured in percent of surface area affected. The hypothesis is that the samples came from common distributions with mean $\mu = 10$. Let us assume that the population variance is 1.0. Then we are testing

$$H_0\colon\ \mu_1 = \mu_2 = 10.$$

Let Figure 10.8 represent a point plot of the data; the data are placed on the distribution stated by the null hypothesis. Let us assume that the "×" data refer to material 1 and the "○" data refer to material 2. Now it seems clear that the data do refute the null hypothesis. But how can this be summarized in one number? **The *P*-value can be viewed as simply the probability of obtaining these data given that both samples come from the same distribution**. Clearly, this probability is quite small, say 0.00000001! Thus, the small P-value clearly refutes H_0, and the conclusion is that the population means are significantly different.

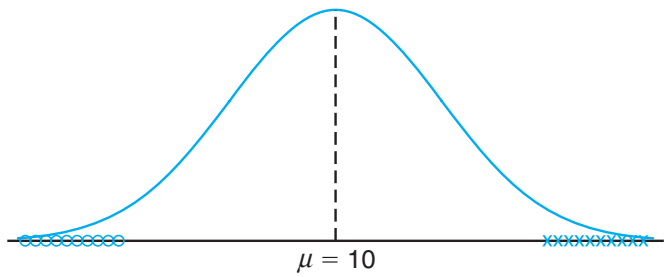

Figure 10.8: Data that are likely generated from populations having two different means.

Use of the P-value approach as an aid in decision-making is quite natural, and nearly all computer packages that provide hypothesis-testing computation print out P-values along with values of the appropriate test statistic. The following is a formal definition of a P-value.

Definition 10.5: A ***P*-value** is the lowest level (of significance) at which the observed value of the test statistic is significant.

How Does the Use of *P*-Values Differ from Classic Hypothesis Testing?

It is tempting at this point to summarize the procedures associated with testing, say, $H_0\colon \theta = \theta_0$. However, the student who is a novice in this area should understand that there are differences in approach and philosophy between the classic

fixed α approach that is climaxed with either a "reject H_0" or a "do not reject H_0" conclusion and the P-value approach. In the latter, no fixed α is determined and conclusions are drawn on the basis of the size of the P-value in harmony with the subjective judgment of the engineer or scientist. While modern computer software will output P-values, nevertheless it is important that readers understand both approaches in order to appreciate the totality of the concepts. Thus, we offer a brief list of procedural steps for both the classical and the P-value approach.

Approach to Hypothesis Testing with Fixed Probability of Type I Error

1. State the null and alternative hypotheses.
2. Choose a fixed significance level α.
3. Choose an appropriate test statistic and establish the critical region based on α.
4. Reject H_0 if the computed test statistic is in the critical region. Otherwise, do not reject.
5. Draw scientific or engineering conclusions.

Significance Testing (P-Value Approach)

1. State null and alternative hypotheses.
2. Choose an appropriate test statistic.
3. Compute the P-value based on the computed value of the test statistic.
4. Use judgment based on the P-value and knowledge of the scientific system.

In later sections of this chapter and chapters that follow, many examples and exercises emphasize the P-value approach to drawing scientific conclusions.

Exercises

10.1 Suppose that an allergist wishes to test the hypothesis that at least 30% of the public is allergic to some cheese products. Explain how the allergist could commit

(a) a type I error;

(b) a type II error.

10.2 A sociologist is concerned about the effectiveness of a training course designed to get more drivers to use seat belts in automobiles.

(a) What hypothesis is she testing if she commits a type I error by erroneously concluding that the training course is ineffective?

(b) What hypothesis is she testing if she commits a type II error by erroneously concluding that the training course is effective?

10.3 A large manufacturing firm is being charged with discrimination in its hiring practices.

(a) What hypothesis is being tested if a jury commits a type I error by finding the firm guilty?

(b) What hypothesis is being tested if a jury commits a type II error by finding the firm guilty?

10.4 A fabric manufacturer believes that the proportion of orders for raw material arriving late is $p = 0.6$. If a random sample of 10 orders shows that 3 or fewer arrived late, the hypothesis that $p = 0.6$ should be rejected in favor of the alternative $p < 0.6$. Use the binomial distribution.

(a) Find the probability of committing a type I error if the true proportion is $p = 0.6$.

(b) Find the probability of committing a type II error for the alternatives $p = 0.3$, $p = 0.4$, and $p = 0.5$.

10.5 Repeat Exercise 10.4 but assume that 50 orders are selected and the critical region is defined to be $x \leq 24$, where x is the number of orders in the sample that arrived late. Use the normal approximation.

10.6 The proportion of adults living in a small town who are college graduates is estimated to be $p = 0.6$. To test this hypothesis, a random sample of 15 adults is selected. If the number of college graduates in the sample is anywhere from 6 to 12, we shall not reject the null hypothesis that $p = 0.6$; otherwise, we shall conclude that $p \neq 0.6$.

(a) Evaluate α assuming that $p = 0.6$. Use the binomial distribution.

(b) Evaluate β for the alternatives $p = 0.5$ and $p = 0.7$.

(c) Is this a good test procedure?

10.7 Repeat Exercise 10.6 but assume that 200 adults are selected and the fail-to-reject region is defined to be $110 \leq x \leq 130$, where x is the number of college graduates in our sample. Use the normal approximation.

10.8 In *Relief from Arthritis* published by Thorsons Publishers, Ltd., John E. Croft claims that over 40% of those who suffer from osteoarthritis receive measurable relief from an ingredient produced by a particular species of mussel found off the coast of New Zealand. To test this claim, the mussel extract is to be given to a group of 7 osteoarthritic patients. If 3 or more of the patients receive relief, we shall not reject the null hypothesis that $p = 0.4$; otherwise, we conclude that $p < 0.4$.

(a) Evaluate α, assuming that $p = 0.4$.

(b) Evaluate β for the alternative $p = 0.3$.

10.9 A dry cleaning establishment claims that a new spot remover will remove more than 70% of the spots to which it is applied. To check this claim, the spot remover will be used on 12 spots chosen at random. If fewer than 11 of the spots are removed, we shall not reject the null hypothesis that $p = 0.7$; otherwise, we conclude that $p > 0.7$.

(a) Evaluate α, assuming that $p = 0.7$.

(b) Evaluate β for the alternative $p = 0.9$.

10.10 Repeat Exercise 10.9 but assume that 100 spots are treated and the critical region is defined to be $x > 82$, where x is the number of spots removed.

10.11 Repeat Exercise 10.8 but assume that 70 patients are given the mussel extract and the critical region is defined to be $x < 24$, where x is the number of osteoarthritic patients who receive relief.

10.12 A random sample of 400 voters in a certain city are asked if they favor an additional 4% gasoline sales tax to provide badly needed revenues for street repairs. If more than 220 but fewer than 260 favor the sales tax, we shall conclude that 60% of the voters are for it.

(a) Find the probability of committing a type I error if 60% of the voters favor the increased tax.

(b) What is the probability of committing a type II error using this test procedure if actually only 48% of the voters are in favor of the additional gasoline tax?

10.13 Suppose, in Exercise 10.12, we conclude that 60% of the voters favor the gasoline sales tax if more than 214 but fewer than 266 voters in our sample favor it. Show that this new critical region results in a smaller value for α at the expense of increasing β.

10.14 A manufacturer has developed a new fishing line, which the company claims has a mean breaking strength of 15 kilograms with a standard deviation of 0.5 kilogram. To test the hypothesis that $\mu = 15$ kilograms against the alternative that $\mu < 15$ kilograms, a random sample of 50 lines will be tested. The critical region is defined to be $\bar{x} < 14.9$.

(a) Find the probability of committing a type I error when H_0 is true.

(b) Evaluate β for the alternatives $\mu = 14.8$ and $\mu = 14.9$ kilograms.

10.15 A soft-drink machine at a steak house is regulated so that the amount of drink dispensed is approximately normally distributed with a mean of 200 milliliters and a standard deviation of 15 milliliters. The machine is checked periodically by taking a sample of 9 drinks and computing the average content. If $\bar{x}$ falls in the interval $191 < \bar{x} < 209$, the machine is thought to be operating satisfactorily; otherwise, we conclude that $\mu \neq 200$ milliliters.

(a) Find the probability of committing a type I error when $\mu = 200$ milliliters.

(b) Find the probability of committing a type II error when $\mu = 215$ milliliters.

10.16 Repeat Exercise 10.15 for samples of size $n = 25$. Use the same critical region.

10.17 A new curing process developed for a certain type of cement results in a mean compressive strength of 5000 kilograms per square centimeter with a standard deviation of 120 kilograms. To test the hypothesis that $\mu = 5000$ against the alternative that $\mu < 5000$, a random sample of 50 pieces of cement is tested. The critical region is defined to be $\bar{x} < 4970$.

(a) Find the probability of committing a type I error when H_0 is true.

(b) Evaluate β for the alternatives $\mu = 4970$ and $\mu = 4960$.

10.18 If we plot the probabilities of failing to reject H_0 corresponding to various alternatives for μ (including the value specified by H_0) and connect all the points by a smooth curve, we obtain the **operating characteristic curve** of the test criterion, or simply the OC curve. Note that the probability of failing to reject H_0 when it is true is simply $1 - \alpha$. Operating characteristic curves are widely used in industrial applications to provide a visual display of the merits of the test criterion. With reference to Exercise 10.15, find the probabilities of failing to reject H_0 for the following 9 values of μ and plot the OC curve: 184, 188, 192, 196, 200, 204, 208, 212, and 216.