CHAPTER 7

HYPOTHESIS TESTING

CHAPTER OVERVIEW

This chapter covers hypothesis testing, the second of two general areas of statistical inference. Hypothesis testing is a topic with which you as a student are likely to have some familiarity. Interval estimation, discussed in the preceding chapter, and hypothesis testing are based on similar concepts. In fact, confidence intervals may be used to arrive at the same conclusions that are reached through the use of hypothesis tests. This chapter provides a format, followed throughout the remainder of this book, for conducting a hypothesis test.

TOPICS

LEARNING OUTCOMES

After studying this chapter, the student will

1. understand how to correctly state a null and alternative hypothesis and carry out a structured hypothesis test.
2. understand the concepts of type I error, type II error, and the power of a test.
3. be able to calculate and interpret *z*, *t*, F, and chi-square test statistics for making statistical inferences.
4. understand how to calculate and interpret *p* values.

7.1 INTRODUCTION

One type of statistical inference, estimation, is discussed in the preceding chapter. The other type, hypothesis testing, is the subject of this chapter. As is true with estimation, the *purpose of hypothesis testing is to aid the clinician, researcher, or administrator in reaching a conclusion concerning a population by examining a sample from that population*. Estimation and hypothesis testing are not as different as they are made to appear by the fact that most textbooks devote a separate chapter to each. As we will explain later, one may use confidence intervals to arrive at the same conclusions that are reached by using the hypothesis testing procedures discussed in this chapter.

Basic Concepts In this section some of the basic concepts essential to an understanding of hypothesis testing are presented. The specific details of particular tests will be given in succeeding sections.

DEFINITION

A *hypothesis* may be defined simply as *a statement about one or more populations*.

The hypothesis is frequently concerned with the parameters of the populations about which the statement is made. A hospital administrator may hypothesize that the average length of stay of patients admitted to the hospital is 5 days; a public health nurse may hypothesize that a particular educational program will result in improved communication between nurse and patient; a physician may hypothesize that a certain drug will be effective in 90 percent of the cases for which it is used. By means of hypothesis testing one determines whether or not such statements are compatible with the available data.

Types of Hypotheses Researchers are concerned with two types of hypotheses—research hypotheses and statistical hypotheses.

DEFINITION

The *research hypothesis* is the conjecture or supposition that motivates the research.

It may be the result of years of observation on the part of the researcher. A public health nurse, for example, may have noted that certain clients responded more readily to a particular type of health education program. A physician may recall numerous instances in which certain combinations of therapeutic measures were more effective than any one of them alone. Research projects often result from the desire of such health practitioners to

determine whether or not their theories or suspicions can be supported when subjected to the rigors of scientific investigation.

Research hypotheses lead directly to statistical hypotheses.

DEFINITION

Statistical hypotheses **are hypotheses that are stated in such a way that they may be evaluated by appropriate statistical techniques.**

In this book the hypotheses that we will focus on are statistical hypotheses. We will assume that the research hypotheses for the examples and exercises have already been considered.

Hypothesis Testing Steps For convenience, hypothesis testing will be presented as a ten-step procedure. There is nothing magical or sacred about this particular format. It merely breaks the process down into a logical sequence of actions and decisions.

1. **Data.** The nature of the data that form the basis of the testing procedures must be understood, since this determines the particular test to be employed. Whether the data consist of counts or measurements, for example, must be determined.
2. **Assumptions.** As we learned in the chapter on estimation, different assumptions lead to modifications of confidence intervals. The same is true in hypothesis testing: A general procedure is modified depending on the assumptions. In fact, the same assumptions that are of importance in estimation are important in hypothesis testing. We have seen that these include assumptions about the normality of the population distribution, equality of variances, and independence of samples.
3. **Hypotheses.** There are two statistical hypotheses involved in hypothesis testing, and these should be stated explicitly. The *null hypothesis* is the *hypothesis to be tested*. It is designated by the symbol H_0. The null hypothesis is sometimes referred to as a *hypothesis of no difference,* since it is a statement of agreement with (or no difference from) conditions presumed to be true in the population of interest. In general, the null hypothesis is set up for the express purpose of being discredited. Consequently, the complement of the conclusion that the researcher is seeking to reach becomes the statement of the null hypothesis. In the testing process the null hypothesis either is rejected or is not rejected. If the null hypothesis is not rejected, we will say that the data on which the test is based do not provide sufficient evidence to cause rejection. If the testing procedure leads to rejection, we will say that the data at hand are not compatible with the null hypothesis, but are supportive of some other hypothesis. The *alternative hypothesis* is a statement of what we will believe is true if our sample data cause us to reject the null hypothesis. Usually the alternative hypothesis and the research hypothesis are the same, and in fact the two terms are used interchangeably. We shall designate the alternative hypothesis by the symbol H_A.

Rules for Stating Statistical Hypotheses When hypotheses are of the type considered in this chapter an indication of equality (either $=$, $\leq$, or $\geq$) must appear in the null hypothesis. Suppose, for example, that we want to answer the question: Can we conclude that a certain population mean is not 50? The null hypothesis is

$$H_0: \mu = 50$$

and the alternative is

$$H_A: \mu \neq 50$$

Suppose we want to know if we can conclude that the population mean is greater than 50. Our hypotheses are

$$H_0: \mu \leq 50 \qquad H_A: \mu > 50$$

If we want to know if we can conclude that the population mean is less than 50, the hypotheses are

$$H_0: \mu \geq 50 \qquad H_A: \mu < 50$$

In summary, we may state the following rules of thumb for deciding what statement goes in the null hypothesis and what statement goes in the alternative hypothesis:

(a) What you hope or expect to be able to conclude as a result of the test usually should be placed in the alternative hypothesis.

(b) The null hypothesis should contain a statement of equality, either $=$, $\leq$, or $\geq$.

(c) The null hypothesis is the hypothesis that is tested.

(d) The null and alternative hypotheses are complementary. That is, the two together exhaust all possibilities regarding the value that the hypothesized parameter can assume.

A Precaution It should be pointed out that neither hypothesis testing nor statistical inference, in general, leads to the proof of a hypothesis; it merely indicates whether the hypothesis is supported or is not supported by the available data. When we fail to reject a null hypothesis, therefore, we do not say that it is true, but that it may be true. When we speak of accepting a null hypothesis, we have this limitation in mind and do not wish to convey the idea that accepting implies proof.

4. **Test statistic.** The test statistic is some statistic that may be computed from the data of the sample. As a rule, there are many possible values that the test statistic may assume, the particular value observed depending on the particular sample drawn. As we will see, the test statistic serves as a decision maker, since the decision

to reject or not to reject the null hypothesis depends on the magnitude of the test statistic. An example of a test statistic is the quantity

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{7.1.1}$$

where μ_0 is a hypothesized value of a population mean. This test statistic is related to the statistic

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \tag{7.1.2}$$

with which we are already familiar.

General Formula for Test Statistic The following is a general formula for a test statistic that will be applicable in many of the hypothesis tests discussed in this book:

$$\text{test statistic} = \frac{\text{relevant statistic} - \text{hypothesized parameter}}{\text{standard error of the relevant statistic}}$$

In Equation 7.1.1, $\bar{x}$ is the relevant statistic, μ_0 is the hypothesized parameter, and $\sigma/\sqrt{n}$ is the standard error of $\bar{x}$, the relevant statistic.

5. **Distribution of test statistic.** It has been pointed out that the key to statistical inference is the sampling distribution. We are reminded of this again when it becomes necessary to specify the probability distribution of the test statistic. The distribution of the test statistic

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

for example, follows the standard normal distribution if the null hypothesis is true and the assumptions are met.

6. **Decision rule.** All possible values that the test statistic can assume are points on the horizontal axis of the graph of the distribution of the test statistic and are divided into two groups; one group constitutes what is known as the *rejection region* and the other group makes up the *nonrejection region*. The values of the test statistic forming the rejection region are those values that are less likely to occur if the null hypothesis is true, while the values making up the acceptance region are more likely to occur if the null hypothesis is true. *The decision rule tells us to reject the null hypothesis if the value of the test statistic that we compute from our sample is one of the values in the rejection region and to not reject the null hypothesis if the computed value of the test statistic is one of the values in the nonrejection region.*

Significance Level The decision as to which values go into the rejection region and which ones go into the nonrejection region is made on the basis of the desired *level of significance,* designated by α. The term *level of significance* reflects the fact that hypothesis tests are sometimes called significance tests, and a computed value of the test

statistic that falls in the rejection region is said to be *significant*. The level of significance, α, specifies the area under the curve of the distribution of the test statistic that is above the values on the horizontal axis constituting the rejection region.

DEFINITION

The *level of significance* α is a probability and, in fact, is the *probability of rejecting a true null hypothesis*.

Since to reject a true null hypothesis would constitute an error, it seems only reasonable that we should make the probability of rejecting a true null hypothesis small and, in fact, that is what is done. We select a small value of α in order to make the probability of rejecting a true null hypothesis small. The more frequently encountered values of α are .01, .05, and .10.

Types of Errors The error committed when a true null hypothesis is rejected is called the *type I error.* The *type II error* is the error committed when a false null hypothesis is not rejected. The probability of committing a type II error is designated by β.

Whenever we reject a null hypothesis there is always the concomitant risk of committing a type I error, rejecting a true null hypothesis. Whenever we fail to reject a null hypothesis the risk of failing to reject a false null hypothesis is always present. We make α small, but we generally exercise no control over β, although we know that in most practical situations it is larger than α.

We never know whether we have committed one of these errors when we reject or fail to reject a null hypothesis, since the true state of affairs is unknown. If the testing procedure leads to rejection of the null hypothesis, we can take comfort from the fact that we made α small and, therefore, the probability of committing a type I error was small. If we fail to reject the null hypothesis, we do not know the concurrent risk of committing a type II error, since β is usually unknown but, as has been pointed out, we do know that, in most practical situations, it is larger than α.

Figure 7.1.1 shows for various conditions of a hypothesis test the possible actions that an investigator may take and the conditions under which each of the two types of error will be made. The table shown in this figure is an example of what is generally referred to as a *confusion matrix.*

7. **Calculation of test statistic.** From the data contained in the sample we compute a value of the test statistic and compare it with the rejection and nonrejection regions that have already been specified.

		Condition of Null Hypothesis	
		True	False
Possible Action	Fail to reject H_0	Correct action	Type II error
	Reject H_0	Type I error	Correct action

FIGURE 7.1.1 Conditions under which type I and type II errors may be committed.

8. **Statistical decision.** The statistical decision consists of rejecting or of not rejecting the null hypothesis. It is rejected if the computed value of the test statistic falls in the rejection region, and it is not rejected if the computed value of the test statistic falls in the nonrejection region.
9. **Conclusion.** If H_0 is rejected, we conclude that H_A is true. If H_0 is not rejected, we conclude that H_0 may be true.
10. ***p* values.** The p value is a number that tells us how unusual our sample results are, given that the null hypothesis is true. A p value indicating that the sample results are not likely to have occurred, if the null hypothesis is true, provides justification for doubting the truth of the null hypothesis.

DEFINITION

A *p value* is the probability that the computed value of a test statistic is at least as extreme as a specified value of the test statistic when the null hypothesis is true. Thus, the *p value* is the smallest value of α for which we can reject a null hypothesis.

We emphasize that when the null hypothesis is not rejected one should not say that the null hypothesis is accepted. We should say that the null hypothesis is "not rejected." We avoid using the word "accept" in this case because we may have committed a type II error. Since, frequently, the probability of committing a type II error can be quite high, we do not wish to commit ourselves to accepting the null hypothesis.

Figure 7.1.2 is a flowchart of the steps that we follow when we perform a hypothesis test.

Purpose of Hypothesis Testing The purpose of hypothesis testing is to assist administrators and clinicians in making decisions. The administrative or clinical decision usually depends on the statistical decision. If the null hypothesis is rejected, the administrative or clinical decision usually reflects this, in that the decision is compatible with the alternative hypothesis. The reverse is usually true if the null hypothesis is not rejected. The administrative or clinical decision, however, may take other forms, such as a decision to gather more data.

We also emphasize that the hypothesis testing procedures highlighted in the remainder of this chapter generally examine the case of normally distributed data or cases where the procedures are appropriate because the central limit theorem applies. In practice, it is not uncommon for samples to be small relative to the size of the population, or to have samples that are highly skewed, and hence the assumption of normality is violated. Methods to handle this situation, that is *distribution-free* or *nonparametric methods,* are examined in detail in Chapter 13. Most computer packages include an analytical procedure (for example, the Shapiro-Wilk or Anderson-Darling test) for testing normality. It is important that such tests are carried out prior to analysis of data. Further, when testing two samples, there is an implicit assumption that the variances are equal. Tests for this assumption are provided in Section 7.8. Finally, it should be noted that hypothesis

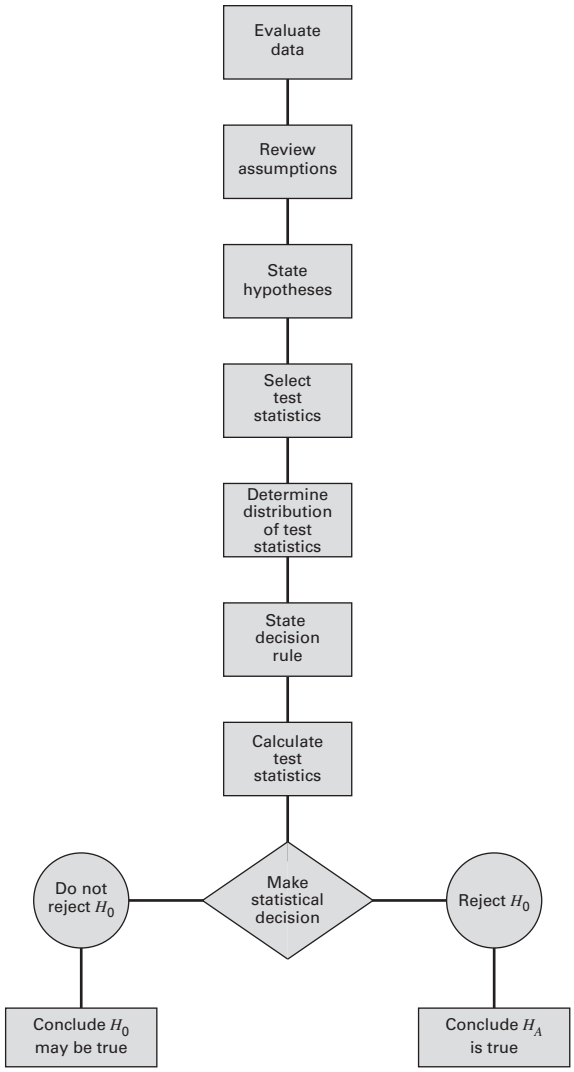

FIGURE 7.1.2 Steps in the hypothesis testing procedure.

tests, just like confidence intervals, are relatively sensitive to the size of the samples being tested, and caution should be taken when interpreting results involving very small sample sizes.

We must emphasize at this point, however, that the outcome of the statistical test is only one piece of evidence that influences the administrative or clinical decision. The statistical decision should not be interpreted as definitive but should be considered along with all the other relevant information available to the experimenter.

With these general comments as background, we now discuss specific hypothesis tests.

7.2 HYPOTHESIS TESTING: A SINGLE POPULATION MEAN

In this section we consider the testing of a hypothesis about a population mean under three different conditions: (1) when sampling is from a normally distributed population of values with known variance; (2) when sampling is from a normally distributed population with unknown variance, and (3) when sampling is from a population that is not normally distributed. Although the theory for conditions 1 and 2 depends on normally distributed populations, it is common practice to make use of the theory when relevant populations are only approximately normally distributed. This is satisfactory as long as the departure from normality is not drastic. When sampling is from a normally distributed population and the population variance is known, the test statistic for testing H_0: $\mu = \mu_0$ is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \tag{7.2.1}$$

which, when H_0 is true, is distributed as the standard normal. Examples 7.2.1 and 7.2.2 illustrate hypothesis testing under these conditions.

Sampling from Normally Distributed Populations: Population Variances Known As we did in Chapter 6, we again emphasize that situations in which the variable of interest is normally distributed with a known variance are rare. The following example, however, will serve to illustrate the procedure.

EXAMPLE 7.2.1

Researchers are interested in the mean age of a certain population. Let us say that they are asking the following question: Can we conclude that the mean age of this population is different from 30 years?

Solution: Based on our knowledge of hypothesis testing, we reply that they can conclude that the mean age is different from 30 if they can reject the null hypothesis that the mean is equal to 30. Let us use the ten-step hypothesis testing procedure given in the previous section to help the researchers reach a conclusion.

1. **Data.** The data available to the researchers are the ages of a simple random sample of 10 individuals drawn from the population of interest. From this sample a mean of $\bar{x} = 27$ has been computed.
2. **Assumptions.** It is assumed that the sample comes from a population whose ages are approximately normally distributed. Let us also assume that the population has a known variance of $\sigma^2 = 20$.
3. **Hypotheses.** The hypothesis to be tested, or null hypothesis, is that the mean age of the population is equal to 30. The alternative hypothesis is

that the mean age of the population is not equal to 30. Note that we are identifying with the alternative hypothesis the conclusion the researchers wish to reach, so that if the data permit rejection of the null hypothesis, the researchers' conclusion will carry more weight, since the accompanying probability of rejecting a true null hypothesis will be small. We will make sure of this by assigning a small value to α, the probability of committing a type I error. We may present the relevant hypotheses in compact form as follows:

$$H_0: \mu = 30$$
$$H_A: \mu \neq 30$$

4. **Test statistic.** Since we are testing a hypothesis about a population mean, since we assume that the population is normally distributed, and since the population variance is known, our test statistic is given by Equation 7.2.1.
5. **Distribution of test statistic.** Based on our knowledge of sampling distributions and the normal distribution, we know that the test statistic is normally distributed with a mean of 0 and a variance of 1, if H_0 is true. There are many possible values of the test statistic that the present situation can generate; one for every possible sample of size 10 that can be drawn from the population. Since we draw only one sample, we have only one of these possible values on which to base a decision.
6. **Decision rule.** The decision rule tells us to reject H_0 if the computed value of the test statistic falls in the rejection region and to fail to reject H_0 if it falls in the nonrejection region. We must now specify the rejection and nonrejection regions. We can begin by asking ourselves what magnitude of values of the test statistic will cause rejection of H_0. If the null hypothesis is false, it may be so either because the population mean is less than 30 or because the population mean is greater than 30. Therefore, either sufficiently small values or sufficiently large values of the test statistic will cause rejection of the null hypothesis. We want these extreme values to constitute the rejection region. How extreme must a possible value of the test statistic be to qualify for the rejection region? The answer depends on the significance level we choose, that is, the size of the probability of committing a type I error. Let us say that we want the probability of rejecting a true null hypothesis to be $\alpha = .05$. Since our rejection region is to consist of two parts, sufficiently small values and sufficiently large values of the test statistic, part of α will have to be associated with the large values and part with the small values. It seems reasonable that we should divide α equally and let $\alpha/2 = .025$ be associated with small values and $\alpha/2 = .025$ be associated with large values.

Critical Value of Test Statistic

What value of the test statistic is so large that, when the null hypothesis is true, the probability of obtaining a value this large or larger is .025? In other words, what is the value of z to the right of which lies .025 of the area under the standard normal distribution? The value of z to the right of which lies .025 of the area is the same value that has .975 of the area between it and $-\infty$. We look in the body of Appendix Table D until we find .975 or its closest value and read the corresponding marginal entries to obtain our z value. In the present example the value of z is 1.96. Similar reasoning will lead us to find -1.96 as the value of the test statistic so small that when the null hypothesis is true, the probability of obtaining a value this small or smaller is .025. Our rejection region, then, consists of all values of the test statistic equal to or greater than 1.96 and less than or equal to -1.96. The nonrejection region consists of all values in between. We may state the decision rule for this test as follows: *reject H_0 if the computed value of the test statistic is either ≥ 1.96 or ≤ -1.96.* Otherwise, do not reject H_0. The rejection and nonrejection regions are shown in Figure 7.2.1. The values of the test statistic that separate the rejection and nonrejection regions are called *critical values* of the test statistic, and the rejection region is sometimes referred to as the *critical region*.

The decision rule tells us to compute a value of the test statistic from the data of our sample and to reject H_0 if we get a value that is either equal to or greater than 1.96 or equal to or less than -1.96 and to fail to reject H_0 if we get any other value. The value of α and, hence, the decision rule should be decided on before gathering the data. This prevents our being accused of allowing the sample results to influence our choice of α. This condition of objectivity is highly desirable and should be preserved in all tests.

7. Calculation of test statistic. From our sample we compute

$$z = \frac{27 - 30}{\sqrt{20/10}} = \frac{-3}{1.4142} = -2.12$$

8. Statistical decision. Abiding by the decision rule, we are able to reject the null hypothesis since -2.12 is in the rejection region. We

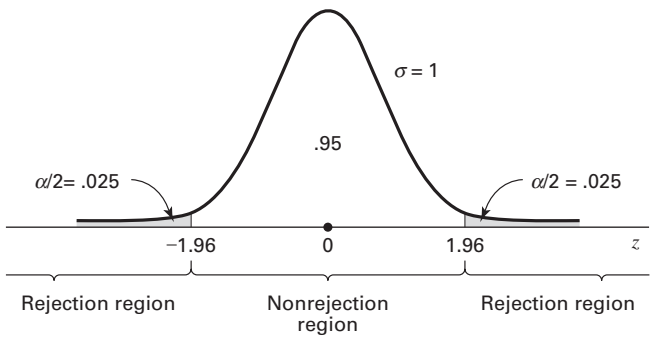

FIGURE 7.2.1 Rejection and nonrejection regions for Example 7.2.1.

can say that the computed value of the test statistic is significant at the .05 level.

9. **Conclusion.** We conclude that μ is not equal to 30 and let our administrative or clinical actions be in accordance with this conclusion.

10. ***p* values.** Instead of saying that an observed value of the test statistic is significant or is not significant, most writers in the research literature prefer to report the exact probability of getting a value as extreme as or more extreme than that observed if the null hypothesis is true. In the present instance these writers would give the computed value of the test statistic along with the statement $p = .0340$. The statement $p = .0340$ means that the probability of getting a value as extreme as 2.12 in either direction, when the null hypothesis is true, is .0340. The value .0340 is obtained from Appendix Table D and is the probability of observing a $z \geq 2.12$ or a $z \leq -2.12$ when the null hypothesis is true. That is, when H_0 is true, the probability of obtaining a value of z as large as or larger than 2.12 is .0170, and the probability of observing a value of z as small as or smaller than -2.12 is .0170. The probability of one or the other of these events occurring, when H_0 is true, is equal to the sum of the two individual probabilities, and hence, in the present example, we say that $p = .0170 + .0170 = .0340$.

Recall that the p value for a test may be defined also as the smallest value of α for which the null hypothesis can be rejected. Since, in Example 7.2.1, our p value is .0340, we know that we could have chosen an α value as small as .0340 and still have rejected the null hypothesis. If we had chosen an α smaller than .0340, we would not have been able to reject the null hypothesis. A general rule worth remembering, then, is this: *if the p value is less than or equal to α, we reject the null hypothesis; if the p value is greater than α, we do not reject the null hypothesis*.

The reporting of p values as part of the results of an investigation is more informative to the reader than such statements as "the null hypothesis is rejected at the .05 level of significance" or "the results were not significant at the .05 level." Reporting the p value associated with a test lets the reader know just how common or how rare is the computed value of the test statistic given that H_0 is true. ■

Testing H_0 by Means of a Confidence Interval Earlier, we stated that one can use confidence intervals to test hypotheses. In Example 7.2.1 we used a hypothesis testing procedure to test H_0: $\mu = 30$ against the alternative, H_A: $\mu \neq 30$. We were able to reject H_0 because the computed value of the test statistic fell in the rejection region.

Let us see how we might have arrived at this same conclusion by using a 100 $(1 - \alpha)$ percent confidence interval. The 95 percent confidence interval for μ is

$$27 \pm 1.96\sqrt{20/10}$$
$$27 \pm 1.96(1.414)$$
$$27 \pm 2.7714$$
$$24.2286,\ 29.7714$$

Since this interval does not include 30, we say 30 is not a candidate for the mean we are estimating and, therefore, μ is not equal to 30 and H_0 is rejected. This is the same conclusion reached by means of the hypothesis testing procedure.

If the hypothesized parameter, 30, had been within the 95 percent confidence interval, we would have said that H_0 is not rejected at the .05 level of significance. In general, *when testing a null hypothesis by means of a two-sided confidence interval, we reject H_0 at the α level of significance if the hypothesized parameter is not contained within the $100(1 - \alpha)$ percent confidence interval. If the hypothesized parameter is contained within the interval, H_0 cannot be rejected at the α level of significance.*

One-Sided Hypothesis Tests The hypothesis test illustrated by Example 7.2.1 is an example of a *two-sided test,* so called because the rejection region is split between the two sides or tails of the distribution of the test statistic. A hypothesis test may be *one-sided,* in which case all the rejection region is in one or the other tail of the distribution. Whether a one-sided or a two-sided test is used depends on the nature of the question being asked by the researcher.

If both large and small values will cause rejection of the null hypothesis, a two-sided test is indicated. When either sufficiently "small" values only or sufficiently "large" values only will cause rejection of the null hypothesis, a one-sided test is indicated.

EXAMPLE 7.2.2

Refer to Example 7.2.1. Suppose, instead of asking if they could conclude that $\mu \neq 30$, the researchers had asked: Can we conclude that $\mu < 30$? To this question we would reply that they can so conclude if they can reject the null hypothesis that $\mu \geq 30$.

Solution: Let us go through the ten-step procedure to reach a decision based on a one-sided test.

1. **Data.** See the previous example.
2. **Assumptions.** See the previous example.
3. **Hypotheses.**

$$H_0: \mu \geq 30$$
$$H_A: \mu < 30$$

The inequality in the null hypothesis implies that the null hypothesis consists of an infinite number of hypotheses. The test will be made only

at the point of equality, since it can be shown that if H_0 is rejected when the test is made at the point of equality it would be rejected if the test were done for any other value of μ indicated in the null hypothesis.

4. **Test statistic.**

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

5. **Distribution of test statistic.** See the previous example.

6. **Decision rule.** Let us again use $\alpha = .05$. To determine where to place the rejection region, let us ask ourselves what magnitude of values would cause rejection of the null hypothesis. If we look at the hypotheses, we see that sufficiently small values would cause rejection and that large values would tend to reinforce the null hypothesis. We will want our rejection region to be where the small values are—at the lower tail of the distribution. This time, since we have a one-sided test, all of α will go in the one tail of the distribution. By consulting Appendix Table D, we find that the value of z to the left of which lies .05 of the area under the standard normal curve is -1.645 after interpolating. Our rejection and nonrejection regions are now specified and are shown in Figure 7.2.2.

   Our decision rule tells us to reject H_0 if the computed value of the test statistic is less than or equal to -1.645.

7. **Calculation of test statistic.** From our data we compute

$$z = \frac{27 - 30}{\sqrt{20/10}} = -2.12$$

8. **Statistical decision.** We are able to reject the null hypothesis since $-2.12 < -1.645$.

9. **Conclusion.** We conclude that the population mean is smaller than 30 and act accordingly.

10. ***p* value.** The p value for this test is .0170, since $P(z \leq -2.12)$, when H_0 is true, is .0170 as given by Appendix Table D when we determine

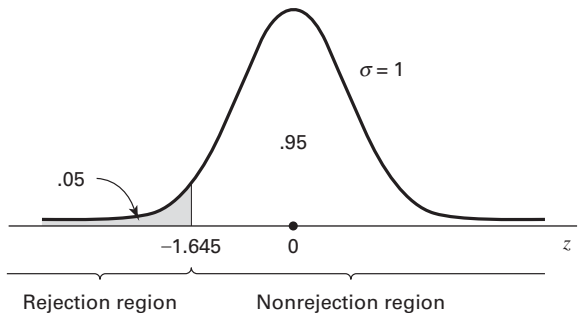

FIGURE 7.2.2 Rejection and nonrejection regions for Example 7.2.2.

the magnitude of the area to the left of -2.12 under the standard normal curve. One can test a one-sided null hypothesis by means of a one-sided confidence interval. However, we will not cover the construction and interpretation of this type of confidence interval in this book.

If the researcher's question had been, "Can we conclude that the mean is greater than 30?," following the above ten-step procedure would have led to a one-sided test with all the rejection region at the upper tail of the distribution of the test statistic and a critical value of $+1.645$. ■

Sampling from a Normally Distributed Population: Population Variance Unknown As we have already noted, the population variance is usually unknown in actual situations involving statistical inference about a population mean. When sampling is from an approximately normal population with an unknown variance, the test statistic for testing H_0: $\mu = \mu_0$ is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{7.2.2}$$

which, when H_0 is true, is distributed as Student's t with $n - 1$ degrees of freedom. The following example illustrates the hypothesis testing procedure when the population is assumed to be normally distributed and its variance is unknown. This is the usual situation encountered in practice.

EXAMPLE 7.2.3

Nakamura et al. (A-1) studied subjects with medial collateral ligament (MCL) and anterior cruciate ligament (ACL) tears. Between February 1995 and December 1997, 17 consecutive patients with combined acute ACL and grade III MCL injuries were treated by the same physician at the research center. One of the variables of interest was the length of time in days between the occurrence of the injury and the first magnetic resonance imaging (MRI). The data are shown in Table 7.2.1. We wish to know if we can conclude that the mean number of days between injury and initial MRI is not 15 days in a population presumed to be represented by these sample data.

TABLE 7.2.1 Number of Days Until MRI for Subjects with MCL and ACL Tears

Subject	Days	Subject	Days	Subject	Days	Subject	Days
1	14	6	0	11	28	16	14
2	9	7	10	12	24	17	9
3	18	8	4	13	24		
4	26	9	8	14	2		
5	12	10	21	15	3		

Source: Norimasa Nakamura, Shuji Horibe, Yukyoshi Toritsuka, Tomoki Mitsuoka, Hideki Yoshikawa, and Konsei Shino, "Acute Grade III Medial Collateral Ligament Injury of the Knee Associated with Anterior Cruciate Ligament Tear," *American Journal of Sports Medicine, 31* (2003), 261–267.

Solution: We will be able to conclude that the mean number of days for the population is not 15 if we can reject the null hypothesis that the population mean is equal to 15.

1. **Data.** The data consist of number of days until MRI on 17 subjects as previously described.
2. **Assumptions.** The 17 subjects constitute a simple random sample from a population of similar subjects. We assume that the number of days until MRI in this population is approximately normally distributed.
3. **Hypotheses.**

$$H_0: \mu = 15$$
$$H_A: \mu \neq 15$$

4. **Test statistic.** Since the population variance is unknown, our test statistic is given by Equation 7.2.2.
5. **Distribution of test statistic.** Our test statistic is distributed as Student's t with $n - 1 = 17 - 1 = 16$ degrees of freedom if H_0 is true.
6. **Decision rule.** Let $\alpha = .05$. Since we have a two-sided test, we put $\alpha/2 = .025$ in each tail of the distribution of our test statistic. The t values to the right and left of which .025 of the area lies are 2.1199 and -2.1199. These values are obtained from Appendix Table E. The rejection and nonrejection regions are shown in Figure 7.2.3.

   The decision rule tells us to compute a value of the test statistic and reject H_0 if the computed t is either greater than or equal to 2.1199 or less than or equal to -2.1199.
7. **Calculation of test statistic.** From our sample data we compute a sample mean of 13.2941 and a sample standard deviation of 8.88654. Substituting these statistics into Equation 7.2.2 gives

$$t = \frac{13.2941 - 15}{8.88654/\sqrt{17}} = \frac{-1.7059}{2.1553} = -.791$$

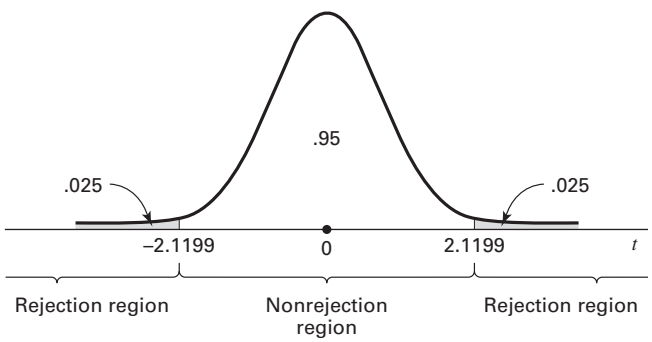

FIGURE 7.2.3 Rejection and nonrejection regions for Example 7.2.3.

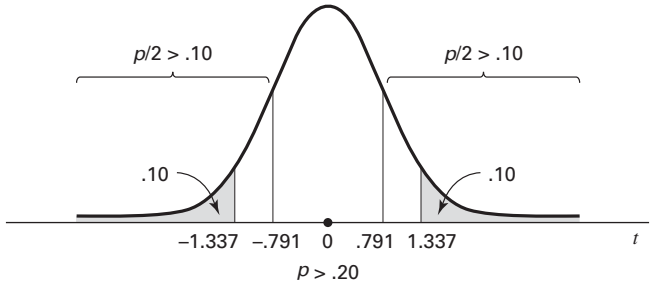

FIGURE 7.2.4 Determination of p value for Example 7.2.3.

8. **Statistical decision.** Do not reject H_0, since $-.791$ falls in the nonrejection region.
9. **Conclusion.** Our conclusion, based on these data, is that the mean of the population from which the sample came may be 15.
10. ***p* value.** The exact p value for this test cannot be obtained from Appendix Table E since it gives t values only for selected percentiles. The p value can be stated as an interval, however. We find that $-.791$ is less than -1.337, the value of t to the left of which lies .10 of the area under the t with 16 degrees of freedom. Consequently, when H_0 is true, the probability of obtaining a value of t as small as or smaller than $-.791$ is greater than .10. That is $P(t \leq -.791) > .10$. Since the test was two-sided, we must allow for the possibility of a computed value of the test statistic as large in the opposite direction as that observed. Appendix Table E reveals that $P(t \geq .791) > .10$ also. The p value, then, is $p > .20$. Figure 7.2.4 shows the p value for this example.

If in the previous example the hypotheses had been

$$H_0: \mu \geq 15$$
$$H_A: \mu < 15$$

the testing procedure would have led to a one-sided test with all the rejection region at the lower tail of the distribution, and if the hypotheses had been

$$H_0: \mu \leq 15$$
$$H_A: \mu > 15$$

we would have had a one-sided test with all the rejection region at the upper tail of the distribution. ■

Sampling from a Population That Is Not Normally Distributed

If, as is frequently the case, the sample on which we base our hypothesis test about a population mean comes from a population that is not normally distributed, we may, if our sample is large (greater than or equal to 30), take advantage of the central limit theorem and use $z = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$ as the test statistic. If the population standard deviation

is not known, the usual practice is to use the sample standard deviation as an estimate. The test statistic for testing H_0: $\mu = \mu_0$, then, is

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{7.2.3}$$

which, when H_0 is true, is distributed approximately as the standard normal distribution if n is large. The rationale for using s to replace σ is that the large sample, necessary for the central limit theorem to apply, will yield a sample standard deviation that closely approximates σ.

EXAMPLE 7.2.4

The goal of a study by Klingler et al. (A-2) was to determine how symptom recognition and perception influence clinical presentation as a function of race. They characterized symptoms and care-seeking behavior in African-American patients with chest pain seen in the emergency department. One of the presenting vital signs was systolic blood pressure. Among 157 African-American men, the mean systolic blood pressure was 146 mm Hg with a standard deviation of 27. We wish to know if, on the basis of these data, we may conclude that the mean systolic blood pressure for a population of African-American men is greater than 140.

Solution: We will say that the data do provide sufficient evidence to conclude that the population mean is greater than 140 if we can reject the null hypothesis that the mean is less than or equal to 140. The following test may be carried out:

1. **Data.** The data consist of systolic blood pressure scores for 157 African-American men with $\bar{x} = 146$ and $s = 27$.
2. **Assumptions.** The data constitute a simple random sample from a population of African-American men who report to an emergency department with symptoms similar to those in the sample. We are unwilling to assume that systolic blood pressure values are normally distributed in such a population.
3. **Hypotheses.**

$$H_0: \mu \leq 140$$
$$H_A: \mu > 140$$

4. **Test statistic.** The test statistic is given by Equation 7.2.3, since s is unknown.
5. **Distribution of test statistic.** Because of the central limit theorem, the test statistic is at worst approximately normally distributed with $\mu = 0$ if H_0 is true.
6. **Decision rule.** Let $\alpha = .05$. The critical value of the test statistic is 1.645. The rejection and nonrejection regions are shown in Figure 7.2.5. Reject H_0 if computed $z \geq 1.645$.

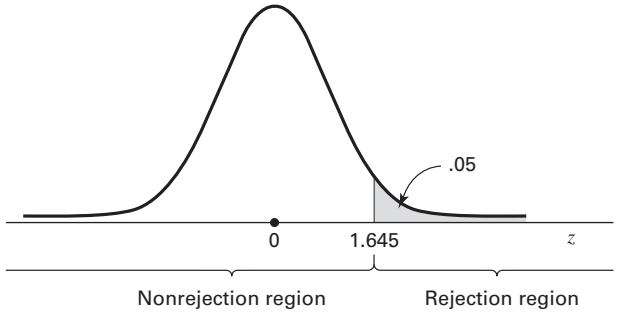

FIGURE 7.2.5 Rejection and nonrejection regions for Example 7.2.4.

7. **Calculation of test statistic.**

$$z = \frac{146 - 140}{27/\sqrt{157}} = \frac{6}{2.1548} = 2.78$$

8. **Statistical decision.** Reject H_0 since $2.78 > 1.645$.
9. **Conclusion.** Conclude that the mean systolic blood pressure for the sampled population is greater than 140.
10. ***p* value.** The p value for this test is $1 - .9973 = .0027$, since as shown in Appendix Table D, the area $(.0027)$ to the right of 2.78 is less than .05, the area to the right of 1.645. ■

Procedures for Other Conditions If the population variance had been known, the procedure would have been identical to the above except that the known value of σ, instead of the sample value s, would have been used in the denominator of the computed test statistic.

Depending on what the investigators wished to conclude, either a two-sided test or a one-sided test, with the rejection region at the lower tail of the distribution, could have been made using the above data.

When testing a hypothesis about a single population mean, we may use Figure 6.3.3 to decide quickly whether the test statistic is z or t.

Computer Analysis To illustrate the use of computers in testing hypotheses we consider the following example.

EXAMPLE 7.2.5

The following are the head circumferences (centimeters) at birth of 15 infants:

33.38	32.15	33.99	34.10	33.97
34.34	33.95	33.85	34.23	32.73
33.46	34.13	34.45	34.19	34.05

We wish to test $H_0: \mu = 34.5$ against $H_A: \mu \neq 34.5$.

Dialog box:

Stat ➤ Basic Statistics ➤ 1-Sample t

Type *C1* in **Samples in columns.** Type *34.5* in the **Test mean** box. Click **OK.**

Session command:

```
MTB > TTEST  34.5  C1
```

Output:

One-Sample T: C1

```
TEST OF MU = 34.5 VS NOT = 34.5
VARIABLE   N     MEAN   STDEV  SE MEAN        95% CI             T      P
C1        15  33.7980  0.6303   0.1627  (33.4490,  34.1470)  −4.31  0.001
```

FIGURE 7.2.6 MINITAB procedure and output for Example 7.2.5.

Solution: We assume that the assumptions for use of the t statistic are met. We enter the data into Column 1 and proceed as shown in Figure 7.2.6.

To indicate that a test is one-sided when in Windows, click on the **Options** button and then choose "less than" or "greater than" as appropriate in the **Alternative** box. If z is the appropriate test statistic, we choose 1-Sample z from the Basic Statistics menu. The remainder of the commands are the same as for the t test.

We learn from the printout that the computed value of the test statistic is -4.31 and the p value for the test is .0007. SAS® users may use the output from PROC MEANS or PROC UNIVARIATE to perform hypothesis tests.

When both the z statistic and the t statistic are inappropriate test statistics for use with the available data, one may wish to use a nonparametric technique to test a hypothesis about a single population measure of central tendency. One such procedure, the sign test, is discussed in Chapter 13. ■

EXERCISES

For each of the following exercises carry out the ten-step hypothesis testing procedure for the given significance level. For each exercise, as appropriate, explain why you chose a one-sided test or a two-sided test. Discuss how you think researchers and/or clinicians might use the results of your hypothesis test. What clinical and/or research decisions and/or actions do you think would be appropriate in light of the results of your test?

7.2.1 Escobar et al. (A-3) performed a study to validate a translated version of the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) questionnaire used with Spanish-speaking patients with hip or knee osteoarthritis. For the 76 women classified with severe hip pain, the

WOMAC mean function score (on a scale from 0 to 100 with a higher number indicating less function) was 70.7 with a standard deviation of 14.6. We wish to know if we may conclude that the mean function score for a population of similar women subjects with severe hip pain is less than 75. Let $\alpha = .01$.

7.2.2 A study by Thienprasiddhi et al. (A-4) examined a sample of 16 subjects with open-angle glaucoma and unilateral hemifield defects. The ages (years) of the subjects were:

62	62	68	48	51	60	51	57
57	41	62	50	53	34	62	61

Source: Phamornsak Thienprasiddhi, Vivienne C. Greenstein, Candice S. Chen, Jeffrey M. Liebmann, Robert Ritch, and Donald C. Hood, "Multifocal Visual Evoked Potential Responses in Glaucoma Patients with Unilateral Hemifield Defects," *American Journal of Ophthalmology, 136* (2003), 34–40.

Can we conclude that the mean age of the population from which the sample may be presumed to have been drawn is less than 60 years? Let $\alpha = .05$.

7.2.3 The purpose of a study by Lugliè et al. (A-5) was to investigate the oral status of a group of patients diagnosed with thalassemia major (TM). One of the outcome measures was the decayed, missing, and filled teeth index (DMFT). In a sample of 18 patients the mean DMFT index value was 10.3 with a standard deviation of 7.3. Is this sufficient evidence to allow us to conclude that the mean DMFT index is greater than 9.0 in a population of similar subjects? Let $\alpha = .10$.

7.2.4 A study was made of a sample of 25 records of patients seen at a chronic disease hospital on an outpatient basis. The mean number of outpatient visits per patient was 4.8, and the sample standard deviation was 2. Can it be concluded from these data that the population mean is greater than four visits per patient? Let the probability of committing a type I error be .05. What assumptions are necessary?

7.2.5 In a sample of 49 adolescents who served as the subjects in an immunologic study, one variable of interest was the diameter of skin test reaction to an antigen. The sample mean and standard deviation were 21 and 11 mm erythema, respectively. Can it be concluded from these data that the population mean is less than 30? Let $\alpha = .05$.

7.2.6 Nine laboratory animals were infected with a certain bacterium and then immunosuppressed. The mean number of organisms later recovered from tissue specimens was 6.5 (coded data) with a standard deviation of .6. Can one conclude from these data that the population mean is greater than 6? Let $\alpha = .05$. What assumptions are necessary?

7.2.7 A sample of 25 freshman nursing students made a mean score of 77 on a test designed to measure attitude toward the dying patient. The sample standard deviation was 10. Do these data provide sufficient evidence to indicate, at the .05 level of significance, that the population mean is less than 80? What assumptions are necessary?

7.2.8 We wish to know if we can conclude that the mean daily caloric intake in the adult rural population of a developing country is less than 2000. A sample of 500 had a mean of 1985 and a standard deviation of 210. Let $\alpha = .05$.

7.2.9 A survey of 100 similar-sized hospitals revealed a mean daily census in the pediatrics service of 27 with a standard deviation of 6.5. Do these data provide sufficient evidence to indicate that the population mean is greater than 25? Let $\alpha = .05$.

7.2.10 Following a week-long hospital supervisory training program, 16 assistant hospital administrators made a mean score of 74 on a test administered as part of the evaluation of the training program. The sample standard deviation was 12. Can it be concluded from these data that the population mean is greater than 70? Let $\alpha = .05$. What assumptions are necessary?

7.2.11 A random sample of 16 emergency reports was selected from the files of an ambulance service. The mean time (computed from the sample data) required for ambulances to reach their destinations was 13 minutes. Assume that the population of times is normally distributed with a variance of 9. Can we conclude at the .05 level of significance that the population mean is greater than 10 minutes?

7.2.12 The following data are the oxygen uptakes (milliliters) during incubation of a random sample of 15 cell suspensions:

14.0, 14.1, 14.5, 13.2, 11.2, 14.0, 14.1, 12.2,
11.1, 13.7, 13.2, 16.0, 12.8, 14.4, 12.9

Do these data provide sufficient evidence at the .05 level of significance that the population mean is not 12 ml? What assumptions are necessary?

7.2.13 Can we conclude that the mean maximum voluntary ventilation value for apparently healthy college seniors is not 110 liters per minute? A sample of 20 yielded the following values:

132, 33, 91, 108, 67, 169, 54, 203, 190, 133,
96, 30, 187, 21, 63, 166, 84, 110, 157, 138

Let $\alpha = .01$. What assumptions are necessary?

7.2.14 The following are the systolic blood pressures (mm Hg) of 12 patients undergoing drug therapy for hypertension:

183, 152, 178, 157, 194, 163, 144, 114, 178, 152, 118, 158

Can we conclude on the basis of these data that the population mean is less than 165? Let $\alpha = .05$. What assumptions are necessary?

7.2.15 Can we conclude that the mean age at death of patients with homozygous sickle-cell disease is less than 30 years? A sample of 50 patients yielded the following ages in years:

15.5	2.0	45.1	1.7	.8	1.1	18.2	9.7	28.1	18.2
27.6	45.0	1.0	66.4	2.0	67.4	2.5	61.7	16.2	31.7
6.9	13.5	1.9	31.2	9.0	2.6	29.7	13.5	2.6	14.4
20.7	30.9	36.6	1.1	23.6	.9	7.6	23.5	6.3	40.2
23.7	4.8	33.2	27.1	36.7	3.2	38.0	3.5	21.8	2.4

Let $\alpha = .05$. What assumptions are necessary?

7.2.16 The following are intraocular pressure (mm Hg) values recorded for a sample of 21 elderly subjects:

14.5	12.9	14.0	16.1	12.0	17.5	14.1	12.9	17.9	12.0
16.4	24.2	12.2	14.4	17.0	10.0	18.5	20.8	16.2	14.9
19.6									

Can we conclude from these data that the mean of the population from which the sample was drawn is greater than 14? Let $\alpha = .05$. What assumptions are necessary?

7.2.17 Suppose it is known that the IQ scores of a certain population of adults are approximately normally distributed with a standard deviation of 15. A simple random sample of 25 adults drawn from this population had a mean IQ score of 105. On the basis of these data can we conclude that the mean IQ score for the population is not 100? Let the probability of committing a type I error be .05.

7.2.18 A research team is willing to assume that systolic blood pressures in a certain population of males are approximately normally distributed with a standard deviation of 16. A simple random sample of 64 males from the population had a mean systolic blood pressure reading of 133. At the .05 level of significance, do these data provide sufficient evidence for us to conclude that the population mean is greater than 130?

7.2.19 A simple random sample of 16 adults drawn from a certain population of adults yielded a mean weight of 63 kg. Assume that weights in the population are approximately normally distributed with a variance of 49. Do the sample data provide sufficient evidence for us to conclude that the mean weight for the population is less than 70 kg? Let the probability of committing a type I error be .01.

7.3 HYPOTHESIS TESTING: THE DIFFERENCE BETWEEN TWO POPULATION MEANS

Hypothesis testing involving the difference between two population means is most frequently employed to determine whether or not it is reasonable to conclude that the two population means are unequal. In such cases, one or the other of the following hypotheses may be formulated:

1. $H_0: \mu_1 - \mu_2 = 0, \quad H_A: \mu_1 - \mu_2 \neq 0$
2. $H_0: \mu_1 - \mu_2 \geq 0, \quad H_A: \mu_1 - \mu_2 < 0$
3. $H_0: \mu_1 - \mu_2 \leq 0, \quad H_A: \mu_1 - \mu_2 > 0$

It is possible, however, to test the hypothesis that the difference is equal to, greater than or equal to, or less than or equal to some value other than zero.

As was done in the previous section, hypothesis testing involving the difference between two population means will be discussed in three different contexts: (1) when sampling is from normally distributed populations with known population variances, (2) when sampling is from normally distributed populations with unknown population variances, and (3) when sampling is from populations that are not normally distributed.

Sampling from Normally Distributed Populations: Population Variances Known When each of two independent simple random samples has been drawn from a normally distributed population with a known variance, the test statistic for testing the null hypothesis of equal population means is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{7.3.1}$$

where the subscript 0 indicates that the difference is a hypothesized parameter. When H_0 is true the test statistic of Equation 7.3.1 is distributed as the standard normal.

EXAMPLE 7.3.1

Researchers wish to know if the data they have collected provide sufficient evidence to indicate a difference in mean serum uric acid levels between normal individuals and individuals with Down's syndrome. The data consist of serum uric acid readings on 12 individuals with Down's syndrome and 15 normal individuals. The means are $\bar{x}_1 = 4.5$ mg/100 ml and $\bar{x}_2 = 3.4$ mg/100 ml.

Solution: We will say that the sample data do provide evidence that the population means are not equal if we can reject the null hypothesis that the population means are equal. Let us reach a conclusion by means of the ten-step hypothesis testing procedure.

1. **Data.** See problem statement.
2. **Assumptions.** The data constitute two independent simple random samples each drawn from a normally distributed population with a variance equal to 1 for the Down's syndrome population and 1.5 for the normal population.
3. **Hypotheses.**

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_A: \mu_1 - \mu_2 \neq 0$$

An alternative way of stating the hypotheses is as follows:

$$H_0: \mu_1 = \mu_2$$
$$H_A: \mu_1 \neq \mu_2$$

4. **Test statistic.** The test statistic is given by Equation 7.3.1.
5. **Distribution of test statistic.** When the null hypothesis is true, the test statistic follows the standard normal distribution.
6. **Decision rule.** Let $\alpha = .05$. The critical values of z are ± 1.96. Reject H_0 unless $-1.96 < z_{\text{computed}} < 1.96$. The rejection and nonrejection regions are shown in Figure 7.3.1.
7. **Calculation of test statistic.**

$$z = \frac{(4.5 - 3.4) - 0}{\sqrt{1/12 + 1.5/15}} = \frac{1.1}{.4282} = 2.57$$

8. **Statistical decision.** Reject H_0, since $2.57 > 1.96$.
9. **Conclusion.** Conclude that, on the basis of these data, there is an indication that the two population means are not equal.
10. ***p* value.** For this test, $p = .0102$.

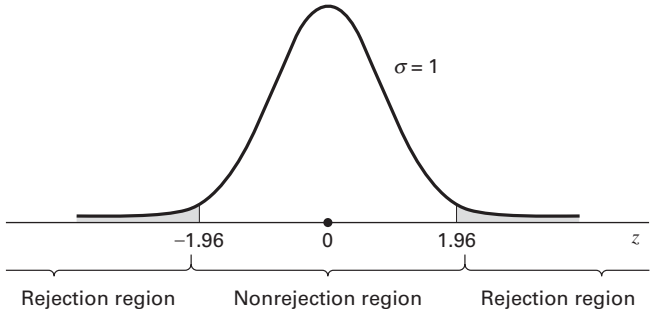

FIGURE 7.3.1 Rejection and nonrejection regions for Example 7.3.1. ■

A 95 Percent Confidence Interval for $\mu_1 - \mu_2$ In the previous chapter the 95 percent confidence interval for $\mu_1 - \mu_2$, computed from the same data, was found to be .26 to 1.94. Since this interval does not include 0, we say that 0 is not a candidate for the difference between population means, and we conclude that the difference is not zero. Thus we arrive at the same conclusion by means of a confidence interval.

Sampling from Normally Distributed Populations: Population Variances Unknown As we have learned, when the population variances are unknown, two possibilities exist. The two population variances may be equal or they may be unequal. We consider first the case where it is known, or it is reasonable to assume, that they are equal. A test of the hypothesis that two population variances are equal is described in Section 7.8.

Population Variances Equal When the population variances are unknown, but assumed to be equal, we recall from Chapter 6 that it is appropriate to pool the sample variances by means of the following formula:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

When each of two independent simple random samples has been drawn from a normally distributed population and the two populations have equal but unknown variances, the test statistic for testing H_0: $\mu_1 = \mu_2$ is given by

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \tag{7.3.2}$$

which, when H_0 is true, is distributed as Student's t with $n_1 + n_2 - 2$ degrees of freedom.

EXAMPLE 7.3.2

The purpose of a study by Tam et al. (A-6) was to investigate wheelchair maneuvering in individuals with lower-level spinal cord injury (SCI) and healthy controls (C). Subjects used a modified wheelchair to incorporate a rigid seat surface to facilitate the specified experimental measurements. Interface pressure measurement was recorded by using a high-resolution pressure-sensitive mat with a spatial resolution of four sensors per square centimeter taped on the rigid seat support. During static sitting conditions, average pressures were recorded under the ischial tuberosities (the bottom part of the pelvic bones). The data for measurements of the left ischial tuberosity (in mm Hg) for the SCI and control groups are shown in Table 7.3.1. We wish to know if we may conclude, on the basis of these data, that, in general, healthy subjects exhibit lower pressure than SCI subjects.

Solution:

1. **Data.** See statement of problem.
2. **Assumptions.** The data constitute two independent simple random samples of pressure measurements, one sample from a population of control subjects and the other sample from a population with lower-level spinal cord injury. We shall assume that the pressure measurements in both populations are approximately normally distributed. The population variances are unknown but are assumed to be equal.
3. **Hypotheses.** H_0: $\mu_C \geq \mu_{SCI}$, H_A: $\mu_C < \mu_{SCI}$.
4. **Test statistic.** The test statistic is given by Equation 7.3.2.
5. **Distribution of test statistic.** When the null hypothesis is true, the test statistic follows Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom.
6. **Decision rule.** Let $\alpha = .05$. The critical value of t is -1.7341. Reject H_0 unless $t_{\text{computed}} > -1.7341$.
7. **Calculation of test statistic.** From the sample data we compute

$$\bar{x}_C = 126.1, \qquad s_C = 21.8, \qquad \bar{x}_{SCI} = 133.1, \qquad s_{SCI} = 32.2$$

Next, we pool the sample variances to obtain

$$s_p^2 = \frac{9(21.8)^2 + 9(32.2)^2}{9 + 9} = 756.04$$

TABLE 7.3.1 Pressures (mm Hg) Under the Pelvis during Static Conditions for Example 7.3.2

Control	131	115	124	131	122	117	88	114	150	169
SCI	60	150	130	180	163	130	121	119	130	148

Source: Eric W. Tam, Arthur F. Mak, Wai Nga Lam, John H. Evans, and York Y. Chow, "Pelvic Movement and Interface Pressure Distribution During Manual Wheelchair Propulsion," *Archives of Physical Medicine and Rehabilitation, 84* (2003), 1466–1472.

We now compute

$$t = \frac{(126.1 - 133.1) - 0}{\sqrt{\frac{756.04}{10} + \frac{756.04}{10}}} = -.569$$

8. **Statistical decision.** We fail to reject H_0, since $-1.7341 < -.569$; that is, $-.569$ falls in the nonrejection region.
9. **Conclusion.** On the basis of these data, we cannot conclude that the population mean pressure is less for healthy subjects than for SCI subjects.
10. ***p* value.** For this test, $p > .10$ since $-1.330 < -.569$. ■

Population Variances Unequal When two independent simple random samples have been drawn from normally distributed populations with unknown and unequal variances, the test statistic for testing H_0: $\mu_1 = \mu_2$ is

$$t' = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{7.3.3}$$

The critical value of t' for an α level of significance and a two-sided test is approximately

$$t'_{1-(\alpha/2)} = \frac{w_1 t_1 + w_2 t_2}{w_1 + w_2} \tag{7.3.4}$$

where $w_1 = s_1^2/n_1$, $w_2 = s_2^2/n_2$, $t_1 = t_{1-(\alpha/2)}$ for $n_1 - 1$ degrees of freedom, and $t_2 = t_{1-(\alpha/2)}$ for $n_2 - 1$ degrees of freedom. The critical value of t' for a one-sided test is found by computing $t'_{1-\alpha}$ by Equation 7.3.4, using $t_1 = t_{1-\alpha}$ for $n_1 - 1$ degrees of freedom and $t_2 = t_{1-\alpha}$ for $n_2 - 1$ degrees of freedom.

For a two-sided test, reject H_0 if the computed value of t' is either greater than or equal to the critical value given by Equation 7.3.4 or less than or equal to the negative of that value.

For a one-sided test with the rejection region in the right tail of the sampling distribution, reject H_0 if the computed t' is equal to or greater than the critical t'. For a one-sided test with a left-tail rejection region, reject H_0 if the computed value of t' is equal to or smaller than the negative of the critical t' computed by the indicated adaptation of Equation 7.3.4.

EXAMPLE 7.3.3

Dernellis and Panaretou (A-7) examined subjects with hypertension and healthy control subjects. One of the variables of interest was the aortic stiffness index. Measures of this variable were calculated from the aortic diameter evaluated by M-mode echocardiography and blood pressure measured by a sphygmomanometer. Generally, physicians wish

to reduce aortic stiffness. In the 15 patients with hypertension (group 1), the mean aortic stiffness index was 19.16 with a standard deviation of 5.29. In the 30 control subjects (group 2), the mean aortic stiffness index was 9.53 with a standard deviation of 2.69. We wish to determine if the two populations represented by these samples differ with respect to mean aortic stiffness index.

Solution:

1. **Data.** The sample sizes, means, and sample standard deviations are:

$$n_1 = 15, \quad \bar{x}_1 = 19.16, \quad s_1 = 5.29$$
$$n_2 = 30, \quad \bar{x}_2 = 9.53, \quad s_2 = 2.69$$

2. **Assumptions.** The data constitute two independent random samples, one from a population of subjects with hypertension and the other from a control population. We assume that aortic stiffness values are approximately normally distributed in both populations. The population variances are unknown and unequal.
3. **Hypotheses.**

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_A: \mu_1 - \mu_2 \neq 0$$

4. **Test statistic.** The test statistic is given by Equation 7.3.3.
5. **Distribution of test statistic.** The statistic given by Equation 7.3.3 does not follow Student's t distribution. We, therefore, obtain its critical values by Equation 7.3.4.
6. **Decision rule.** Let $\alpha = .05$. Before computing t' we calculate $w_1 = (5.29)^2/15 = 1.8656$ and $w_2 = (2.69)^2/30 = .2412$. In Appendix Table E we find that $t_1 = 2.1448$ and $t_2 = 2.0452$. By Equation 7.3.4 we compute

$$t' = \frac{1.8656(2.1448) + .2412(2.0452)}{1.8656 + .2412} = 2.133$$

Our decision rule, then, is reject H_0 if the computed t is either ≥ 2.133 or ≤ -2.133.

7. **Calculation of test statistic.** By Equation 7.3.3 we compute

$$t' = \frac{(19.16 - 9.53) - 0}{\sqrt{\frac{(5.29)^2}{15} + \frac{(2.69)^2}{30}}} = \frac{9.63}{1.4515} = 6.63$$

8. **Statistical decision.** Since $6.63 > 2.133$, we reject H_0.
9. **Conclusion.** On the basis of these results we conclude that the two population means are different.
10. ***p* value.** For this test $p < .05$. ■

Sampling from Populations That Are Not Normally Distributed

When sampling is from populations that are not normally distributed, the results of the central limit theorem may be employed if sample sizes are large (say, ≥ 30). This will allow the use of normal theory since the distribution of the difference between sample means will be approximately normal. When each of two large independent simple random samples has been drawn from a population that is not normally distributed, the test statistic for testing H_0: $\mu_1 = \mu_2$ is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{7.3.5}$$

which, when H_0 is true, follows the standard normal distribution. If the population variances are known, they are used; but if they are unknown, as is the usual case, the sample variances, which are necessarily based on large samples, are used as estimates. Sample variances are not pooled, since equality of population variances is not a necessary assumption when the z statistic is used.

EXAMPLE 7.3.4

The objective of a study by Sairam et al. (A-8) was to identify the role of various disease states and additional risk factors in the development of thrombosis. One focus of the study was to determine if there were differing levels of the anticardiolipin antibody IgG in subjects with and without thrombosis. Table 7.3.2 summarizes the researchers' findings:

TABLE 7.3.2 IgG Levels for Subjects with and Without Thrombosis for Example 7.3.4

Group	Mean IgG Level (ml/unit)	Sample Size	Standard Deviation
Thrombosis	59.01	53	44.89
No thrombosis	46.61	54	34.85

Source: S. Sairam, B. A. Baethge and T. McNearney, "Analysis of Risk Factors and Comorbid Diseases in the Development of Thrombosis in Patients with Anticardiolipin Antibodies," *Clinical Rheumatology, 22* (2003), 24–29.

We wish to know if we may conclude, on the basis of these results, that, in general, persons with thrombosis have, on the average, higher IgG levels than persons without thrombosis.

Solution:

1. **Data.** See statement of example.
2. **Assumptions.** The statistics were computed from two independent samples that behave as simple random samples from a population of persons with thrombosis and a population of persons who do not have thrombosis. Since the population variances are unknown, we will use the sample variances in the calculation of the test statistic.
3. **Hypotheses.**

$$H_0: \mu_T - \mu_{NT} \leq 0$$
$$H_A: \mu_T - \mu_{NT} > 0$$

or, alternatively,

$$H_0: \mu_T \leq \mu_{NT}$$
$$H_A: \mu_T > \mu_{NT}$$

4. **Test statistic.** Since we have large samples, the central limit theorem allows us to use Equation 7.3.5 as the test statistic.
5. **Distribution of test statistic.** When the null hypothesis is true, the test statistic is distributed approximately as the standard normal.
6. **Decision rule.** Let $\alpha = .01$. This is a one-sided test with a critical value of z equal to 2.33. Reject H_0 if $z_{\text{computed}} \geq 2.33$.
7. **Calculation of test statistic.**

$$z = \frac{59.01 - 46.61}{\sqrt{\frac{44.89^2}{53} + \frac{34.85^2}{54}}} = 1.59$$

8. **Statistical decision.** Fail to reject H_0, since $z = 1.59$ is in the nonrejection region.
9. **Conclusion.** These data indicate that on the average, persons with thrombosis and persons without thrombosis may not have differing IgG levels.
10. ***p* value.** For this test, $p = .0559$. When testing a hypothesis about the difference between two populations means, we may use Figure 6.4.1 to decide quickly whether the test statistic should be z or t. ■

We may use MINITAB to perform two-sample t tests. To illustrate, let us refer to the data in Table 7.3.1. We put the data for control subjects and spinal cord

Dialog box:

Stat ➤ Basic Statistics ➤ 2-Sample t

Choose **Samples in different columns.** Type *C1* in **First** and *C2* in **Second.** Click the **Options** box and select "less than" in the **Alternatives** box. Check **Assume equal variances.** Click **OK.**

Session command:

```
MTB > TwoSample 95.0 C1 C2
SUBC> Alternative -1,
SUBC> Pooled.
```

Output:

```
Two-Sample T-Test and CI: C, SCI

Two-sample T for C vs SCI
        N     Mean    StDev    SE Mean
C      10    126.1     21.8        6.9
SCI    10    133.1     32.2         10

Difference = mu C - mu SCI
Estimate for difference: -7.0
95% upper bound for difference: 14.3
T-Test of difference = 0 (vs <): T-Value = -0.57 P-Value = 0.288
DF = 18
Both use Pooled StDev = 27.5
```

FIGURE 7.3.2 MINITAB procedure and output for two-sample *t* test, Example 7.3.2 (data in Table 7.3.1).

injury subjects in Column 1 and Column 2, respectively, and proceed as shown in Figure 7.3.2.

The SAS® statistical package performs the *t* test for equality of population means under both assumptions regarding population variances: that they are equal and that they are not equal. Note that SAS® designates the *p* value as $\Pr > |t|$. The default output is a *p* value for a two-sided test. The researcher using SAS® must divide this quantity in half when the hypothesis test is one-sided. The SAS® package also tests for equality of population variances as described in Section 7.8. Figure 7.3.3 shows the SAS® output for Example 7.3.2.

Alternatives to *z* and *t* Sometimes neither the z statistic nor the t statistic is an appropriate test statistic for use with the available data. When such is the case, one may wish to use a nonparametric technique for testing a hypothesis about the difference between two population measures of central tendency. The Mann-Whitney test statistic and the median test, discussed in Chapter 13, are frequently used alternatives to the z and t statistics.

```
The SAS System
The TTEST Procedure

                              Statistics          Lower CL      Upper CL
                        Lower CL    Upper CL       Std     Std     Std     Std
Variable  group        N   Mean   Mean    Mean     Dev     Dev     Dev     Err
------------------------------------------------------------------------------
pressure  C           10  110.49  126.1  141.71  15.008   21.82  39.834     6.9
pressure  SCI         10  110.08  133.1  156.12  22.133  32.178  58.745  10.176
pressure  Diff (1-2)      -32.83     -7   18.83  20.773  27.491  40.655  12.294

                                    T-Tests
------------------------------------------------------------------------------
Variable     Method          Variances     DF       t Value     Pr > |t|

pressure     Pooled          Equal         18       -0.57       0.5761
pressure     Satterthwaite   Unequal     15.8       -0.57       0.5771

                              Equality of Variances
------------------------------------------------------------------------------
Variable     Method          Num DF        Den DF   F Value     Pr > F
pressure     Folded F             9             9      2.17      0.2626
```

FIGURE 7.3.3 SAS® output for Example 7.3.2 (data in Table 7.3.1).

EXERCISES

In each of the following exercises, complete the ten-step hypothesis testing procedure. State the assumptions that are necessary for your procedure to be valid. For each exercise, as appropriate, explain why you chose a one-sided test or a two-sided test. Discuss how you think researchers or clinicians might use the results of your hypothesis test. What clinical or research decisions or actions do you think would be appropriate in light of the results of your test?

7.3.1 Subjects in a study by Dabonneville et al. (A-9) included a sample of 40 men who claimed to engage in a variety of sports activities (multisport). The mean body mass index (BMI) for these men was 22.41 with a standard deviation of 1.27. A sample of 24 male rugby players had a mean BMI of 27.75 with a standard deviation of 2.64. Is there sufficient evidence for one to claim that, in general, rugby players have a higher BMI than the multisport men? Let $\alpha = .01$.

7.3.2 The purpose of a study by Ingle and Eastell (A-10) was to examine the bone mineral density (BMD) and ultrasound properties of women with ankle fractures. The investigators recruited 31 postmenopausal women with ankle fractures and 31 healthy postmenopausal women to serve as controls. One of the baseline measurements was the stiffness index of the lunar Achilles. The mean stiffness index for the ankle fracture group was 76.9 with a standard deviation of 12.6. In the control group, the mean was 90.9 with a standard deviation of 12.5. Do these data provide sufficient evidence to allow you to conclude that, in general, the mean stiffness index is higher

in healthy postmenopausal women than in postmenopausal women with ankle fractures? Let $\alpha = .05$.

7.3.3 Hoekema et al. (A-11) studied the craniofacial morphology of 26 male patients with obstructive sleep apnea syndrome (OSAS) and 37 healthy male subjects (non–OSAS). One of the variables of interest was the length from the most superoanterior point of the body of the hyoid bone to the Frankfort horizontal (measured in millimeters).

Length (mm) Non–OSAS				Length (mm) OSAS		
96.80	97.00	101.00	88.95	105.95	114.90	113.70
100.70	97.70	88.25	101.05	114.90	114.35	116.30
94.55	97.00	92.60	92.60	110.35	112.25	108.75
99.65	94.55	98.25	97.00	123.10	106.15	113.30
109.15	106.45	90.85	91.95	119.30	102.60	106.00
102.75	94.55	95.25	88.95	110.00	102.40	101.75
97.70	94.05	88.80	95.75	98.95	105.05	
92.10	89.45	101.40		114.20	112.65	
91.90	89.85	90.55		108.95	128.95	
89.50	98.20	109.80		105.05	117.70	

Source: A. Hoekema, D.D.S. Used with permission.

Do these data provide sufficient evidence to allow us to conclude that the two sampled populations differ with respect to length from the hyoid bone to the Frankfort horizontal? Let $\alpha = .01$.

7.3.4 Can we conclude that patients with primary hypertension (PH), on the average, have higher total cholesterol levels than normotensive (NT) patients? This was one of the inquiries of interest for Rossi et al. (A-12). In the following table are total cholesterol measurements (mg/dl) for 133 PH patients and 41 NT patients. Can we conclude that PH patients have, on average, higher total cholesterol levels than NT patients? Let $\alpha = .05$.

Total Cholesterol (mg/dl)						
Primary Hypertensive Patients					Normotensive Patients	
207	221	212	220	190	286	189
172	223	260	214	245	226	196
191	181	210	215	171	187	142
221	217	265	206	261	204	179
203	208	206	247	182	203	212
241	202	198	221	162	206	163
208	218	210	199	182	196	196
199	216	211	196	225	168	189
185	168	274	239	203	229	142
235	168	223	199	195	184	168
214	214	175	244	178	186	121
134	203	203	214	240	281	
226	280	168	236	222	203	

(*Continued*)

Total Cholesterol (mg/dl)						
Primary Hypertensive Patients					Normotensive Patients	
222	203	178	249	117	177	135
213	225	217	212	252	179	161
272	227	200	259	203	194	
185	239	226	189	245	206	
181	265	207	235	218	219	
238	228	232	239	152	173	
141	226	182	239	231	189	
203	236	215	210	237	194	
222	195	239	203		196	
221	284	210	188		212	
180	183	207	237		168	
276	266	224	231		188	
226	258	251	222		232	
224	214	212	174		242	
206	260	201	219		200	

Source: Gian Paolo Rossi, M.D., F.A.C.C., F.A.H.A. Used with permission.

7.3.5 Garção and Cabrita (A-13) wanted to evaluate the community pharmacist's capacity to positively influence the results of antihypertensive drug therapy through a pharmaceutical care program in Portugal. Eighty-two subjects with essential hypertension were randomly assigned to an intervention or a control group. The intervention group received monthly monitoring by a research pharmacist to monitor blood pressure, assess adherence to treatment, prevent, detect, and resolve drug-related problems, and encourage nonpharmacologic measures for blood pressure control. The changes after 6 months in diastolic blood pressure (pre − post, mm Hg) are given in the following table for patients in each of the two groups.

Intervention Group				Control Group			
20	4	12	16	0	4	12	0
2	24	6	10	12	2	2	8
36	6	24	16	18	2	0	10
26	−2	42	10	0	8	0	14
2	8	20	6	8	10	−4	8
20	8	14	6	10	0	12	0
2	16	−2	2	8	6	4	2
14	14	10	8	14	10	28	−8
30	8	2	16	4	−2	−18	16
18	20	18	−12	−2	2	12	12
6				−6			

Source: José Garção, M.S., Pharm.D. Used with permission.

On the basis of these data, what should the researcher conclude? Let $\alpha = .05$.

7.3.6 A test designed to measure mothers' attitudes toward their labor and delivery experiences was given to two groups of new mothers. Sample 1 (attenders) had attended prenatal classes held at

the local health department. Sample 2 (nonattenders) did not attend the classes. The sample sizes and means and standard deviations of the test scores were as follows:

Sample	n	$\bar{x}$	s
1	15	4.75	1.0
2	22	3.00	1.5

Do these data provide sufficient evidence to indicate that attenders, on the average, score higher than nonattenders? Let $\alpha = .05$.

7.3.7 Cortisol level determinations were made on two samples of women at childbirth. Group 1 subjects underwent emergency cesarean section following induced labor. Group 2 subjects delivered by either cesarean section or the vaginal route following spontaneous labor. The sample sizes, mean cortisol levels, and standard deviations were as follows:

Sample	n	$\bar{x}$	s
1	10	435	65
2	12	645	80

Do these data provide sufficient evidence to indicate a difference in the mean cortisol levels in the populations represented? Let $\alpha = .05$.

7.3.8 Protoporphyrin levels were measured in two samples of subjects. Sample 1 consisted of 50 adult male alcoholics with ring sideroblasts in the bone marrow. Sample 2 consisted of 40 apparently healthy adult nonalcoholic males. The mean protoporphyrin levels and standard deviations for the two samples were as follows:

Sample	$\bar{x}$	s
1	340	250
2	45	25

Can one conclude on the basis of these data that protoporphyrin levels are higher in the represented alcoholic population than in the nonalcoholic population? Let $\alpha = .01$.

7.3.9 A researcher was interested in knowing if preterm infants with late metabolic acidosis and preterm infants without the condition differ with respect to urine levels of a certain chemical. The mean levels, standard deviations, and sample sizes for the two samples studied were as follows:

Sample	n	$\bar{x}$	s
With condition	35	8.5	5.5
Without condition	40	4.8	3.6

What should the researcher conclude on the basis of these results? Let $\alpha = .05$.

7.3.10 Researchers wished to know if they could conclude that two populations of infants differ with respect to mean age at which they walked alone. The following data (ages in months) were collected:

Sample from population A: 9.5, 10.5, 9.0, 9.75, 10.0, 13.0, 10.0, 13.5, 10.0, 9.5, 10.0, 9.75

Sample from population B: 12.5, 9.5, 13.5, 13.75, 12.0, 13.75, 12.5, 9.5, 12.0, 13.5, 12.0, 12.0

What should the researchers conclude? Let $\alpha = .05$.

7.3.11 Does sensory deprivation have an effect on a person's alpha-wave frequency? Twenty volunteer subjects were randomly divided into two groups. Subjects in group A were subjected to a 10-day period of sensory deprivation, while subjects in group B served as controls. At the end of the experimental period, the alpha-wave frequency component of subjects' electroencephalograms was measured. The results were as follows:

Group A: 10.2, 9.5, 10.1, 10.0, 9.8, 10.9, 11.4, 10.8, 9.7, 10.4

Group B: 11.0, 11.2, 10.1, 11.4, 11.7, 11.2, 10.8, 11.6, 10.9, 10.9

Let $\alpha = .05$.

7.3.12 Can we conclude that, on the average, lymphocytes and tumor cells differ in size? The following are the cell diameters (μm) of 40 lymphocytes and 50 tumor cells obtained from biopsies of tissue from patients with melanoma:

Lymphocytes							
9.0	9.4	4.7	4.8	8.9	4.9	8.4	5.9
6.3	5.7	5.0	3.5	7.8	10.4	8.0	8.0
8.6	7.0	6.8	7.1	5.7	7.6	6.2	7.1
7.4	8.7	4.9	7.4	6.4	7.1	6.3	8.8
8.8	5.2	7.1	5.3	4.7	8.4	6.4	8.3

Tumor Cells									
12.6	14.6	16.2	23.9	23.3	17.1	20.0	21.0	19.1	19.4
16.7	15.9	15.8	16.0	17.9	13.4	19.1	16.6	18.9	18.7
20.0	17.8	13.9	22.1	13.9	18.3	22.8	13.0	17.9	15.2
17.7	15.1	16.9	16.4	22.8	19.4	19.6	18.4	18.2	20.7
16.3	17.7	18.1	24.3	11.2	19.5	18.6	16.4	16.1	21.5

Let $\alpha = .05$.

7.4 PAIRED COMPARISONS

In our previous discussion involving the difference between two population means, it was assumed that the samples were independent. A method frequently employed for assessing the effectiveness of a treatment or experimental procedure is one that makes

use of related observations resulting from nonindependent samples. A hypothesis test based on this type of data is known as a *paired comparisons* test.

Reasons for Pairing It frequently happens that true differences do not exist between two populations with respect to the variable of interest, but the presence of extraneous sources of variation may cause rejection of the null hypothesis of no difference. On the other hand, true differences also may be masked by the presence of extraneous factors.

Suppose, for example, that we wish to compare two sunscreens. There are at least two ways in which the experiment may be carried out. One method would be to select a simple random sample of subjects to receive sunscreen A and an independent simple random sample of subjects to receive sunscreen B. We send the subjects out into the sunshine for a specified length of time, after which we will measure the amount of damage from the rays of the sun. Suppose we employ this method, but inadvertently, most of the subjects receiving sunscreen A have darker complexions that are naturally less sensitive to sunlight. Let us say that after the experiment has been completed we find that subjects receiving sunscreen A had less sun damage. We would not know if they had less sun damage because sunscreen A was more protective than sunscreen B or because the subjects were naturally less sensitive to the sun.

A better way to design the experiment would be to select just one simple random sample of subjects and let each member of the sample receive both sunscreens. We could, for example, randomly assign the sunscreens to the left or the right side of each subject's back with each subject receiving both sunscreens. After a specified length of exposure to the sun, we would measure the amount of sun damage to each half of the back. If the half of the back receiving sunscreen A tended to be less damaged, we could more confidently attribute the result to the sunscreen, since in each instance both sunscreens were applied to equally pigmented skin.

The objective in paired comparisons tests is to eliminate a maximum number of sources of extraneous variation by making the pairs similar with respect to as many variables as possible.

Related or paired observations may be obtained in a number of ways. The same subjects may be measured before and after receiving some treatment. Litter mates of the same sex may be assigned randomly to receive either a treatment or a placebo. Pairs of twins or siblings may be assigned randomly to two treatments in such a way that members of a single pair receive different treatments. In comparing two methods of analysis, the material to be analyzed may be divided equally so that one-half is analyzed by one method and one-half is analyzed by the other. Or pairs may be formed by matching individuals on some characteristic, for example, digital dexterity, which is closely related to the measurement of interest, say, posttreatment scores on some test requiring digital manipulation.

Instead of performing the analysis with individual observations, we use d_i, the difference between pairs of observations, as the variable of interest.

When the n sample differences computed from the n pairs of measurements constitute a simple random sample from a normally distributed population of differences, the test statistic for testing hypotheses about the population mean difference μ_d is

$$t = \frac{\bar{d} - \mu_{d_0}}{s_{\bar{d}}} \tag{7.4.1}$$

where $\bar{d}$ is the sample mean difference, μ_{d_0} is the hypothesized population mean difference, $s_{\bar{d}} = s_d/\sqrt{n}$, n is the number of sample differences, and s_d is the standard deviation of the sample differences. When H_0 is true, the test statistic is distributed as Student's t with $n - 1$ degrees of freedom.

Although to begin with we have two samples—say, before levels and after levels—we do not have to worry about equality of variances, as with independent samples, since our variable is the difference between readings in the same individual, or matched individuals, and, hence, only one variable is involved. The arithmetic involved in performing a paired comparisons test, therefore, is the same as for performing a test involving a single sample as described in Section 7.2.

The following example illustrates the procedures involved in a paired comparisons test.

EXAMPLE 7.4.1

John M. Morton et al. (A-14) examined gallbladder function before and after fundoplication—a surgery used to stop stomach contents from flowing back into the esophagus (reflux)—in patients with gastroesophageal reflux disease. The authors measured gallbladder functionality by calculating the gallbladder ejection fraction (GBEF) before and after fundoplication. The goal of fundoplication is to increase GBEF, which is measured as a percent. The data are shown in Table 7.4.1. We wish to know if these data provide sufficient evidence to allow us to conclude that fundoplication increases GBEF functioning.

Solution: We will say that sufficient evidence is provided for us to conclude that the fundoplication is effective if we can reject the null hypothesis that the population mean change μ_d is different from zero in the appropriate direction. We may reach a conclusion by means of the ten-step hypothesis testing procedure.

1. **Data.** The data consist of the GBEF for 12 individuals, before and after fundoplication. We shall perform the statistical analysis on the differences in preop and postop GBEF. We may obtain the differences in one of two ways: by subtracting the preop percents from the postop percents or by subtracting the postop percents from the preop percents. Let us obtain the differences by subtracting the preop

TABLE 7.4.1 Gallbladder Function in Patients with Presentations of Gastroesophageal Reflux Disease Before and After Treatment

Preop (%)	22	63.3	96	9.2	3.1	50	33	69	64	18.8	0	34
Postop (%)	63.5	91.5	59	37.8	10.1	19.6	41	87.8	86	55	88	40

Source: John M. Morton, Steven P. Bowers, Tananchai A. Lucktong, Samer Mattar, W. Alan Bradshaw, Kevin E. Behrns, Mark J. Koruda, Charles A. Herbst, William McCartney, Raghuveer K. Halkar, C. Daniel Smith, and Timothy M. Farrell, "Gallbladder Function Before and After Fundoplication," *Journal of Gastrointestinal Surgery, 6* (2002), 806–811.

percents from the postop percents. The d_i = postop − preop differences are:

41.5, 28.2, −37.0, 28.6, 7.0, −30.4, 8.0, 18.8, 22.0, 36.2, 88.0, 6.0

2. **Assumptions.** The observed differences constitute a simple random sample from a normally distributed population of differences that could be generated under the same circumstances.
3. **Hypotheses.** The way we state our null and alternative hypotheses must be consistent with the way in which we subtract measurements to obtain the differences. In the present example, we want to know if we can conclude that the fundoplication is useful in increasing GBEF percentage. If it is effective in improving GBEF, we would expect the postop percents to tend to be higher than the preop percents. If, therefore, we subtract the preop percents from the postop percents (postop − preop), we would expect the differences to tend to be positive. Furthermore, we would expect the mean of a population of such differences to be positive. So, under these conditions, asking if we can conclude that the fundoplication is effective is the same as asking if we can conclude that the population mean difference is positive (greater than zero).

   The null and alternative hypotheses are as follows:

$$H_0: \mu_d \leq 0$$
$$H_A: \mu_d > 0$$

   If we had obtained the differences by subtracting the postop percents from the preop weights (preop − postop), our hypotheses would have been

$$H_0: \mu_d \geq 0$$
$$H_A: \mu_d < 0$$

   If the question had been such that a two-sided test was indicated, the hypotheses would have been

$$H_0: \mu_d = 0$$
$$H_A: \mu_d \neq 0$$

   regardless of the way we subtracted to obtain the differences.
4. **Test statistic.** The appropriate test statistic is given by Equation 7.4.1.
5. **Distribution of test statistic.** If the null hypothesis is true, the test statistic is distributed as Student's t with $n - 1$ degrees of freedom.
6. **Decision rule.** Let $\alpha = .05$. The critical value of t is 1.7959. Reject H_0 if computed t is greater than or equal to the critical value. The rejection and nonrejection regions are shown in Figure 7.4.1.

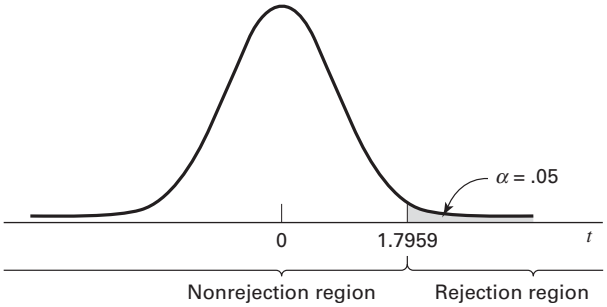

FIGURE 7.4.1 Rejection and nonrejection regions for Example 7.4.1.

7. **Calculation of test statistic.** From the $n = 12$ differences d_i, we compute the following descriptive measures:

$$\bar{d} = \frac{\sum d_i}{n} = \frac{(41.5) + (28.2) + (-37.0) + \cdots + (6.0)}{12} = \frac{216.9}{12} = 18.075$$

$$s_d^2 = \frac{\sum (d_i - \bar{d})^2}{n-1} = \frac{n\sum d_i^2 - (\sum d_i)^2}{n(n-1)} = \frac{12(15669.49) - (216.9)^2}{(12)(11)} = 1068.0930$$

$$t = \frac{18.075 - 0}{\sqrt{1068.0930/12}} = \frac{18.075}{9.4344} = 1.9159$$

8. **Statistical decision.** Reject H_0, since 1.9159 is in the rejection region.
9. **Conclusion.** We may conclude that the fundoplication procedure increases GBEF functioning.
10. ***p* value.** For this test, $.025 < p < .05$, since $1.7959 < 1.9159 < 2.2010$. ■

A Confidence Interval for μ_d A 95 percent confidence interval for μ_d may be obtained as follows:

$$\bar{d} \pm t_{1-(\alpha/2)} s_{\bar{d}}$$
$$18.075 \pm 2.2010\sqrt{1068.0930/12}$$
$$18.075 \pm 20.765$$
$$-2.690, 38.840$$

The Use of *z* If, in the analysis of paired data, the population variance of the differences is known, the appropriate test statistic is

$$z = \frac{\bar{d} - \mu_d}{\sigma_d/\sqrt{n}} \tag{7.4.2}$$

It is unlikely that σ_d will be known in practice.

Paired T-Test and CI: C2, C1

```
Paired T for C2 - C1

                N     Mean     StDev    SE Mean
C2             12  56.6083  27.8001     8.0252
C1             12  38.5333  30.0587     8.6772
Difference     12  18.0750  32.6817     9.4344



95% lower bound for mean difference: 1.1319
T-Test of mean difference = 0 (vs > 0): T-Value = 1.92 P-Value =
0.041
```

FIGURE 7.4.2 MINITAB procedure and output for paired comparisons test, Example 7.4.1 (data in Table 7.4.1).

If the assumption of normally distributed d_i's cannot be made, the central limit theorem may be employed if n is large. In such cases, the test statistic is Equation 7.4.2, with s_d used to estimate σ_d when, as is generally the case, the latter is unknown.

We may use MINITAB to perform a paired t-test. The output from this procedure is given in Figure 7.4.2.

Disadvantages The use of the paired comparisons test is not without its problems. If different subjects are used and randomly assigned to two treatments, considerable time and expense may be involved in our trying to match individuals on one or more relevant variables. A further price we pay for using paired comparisons is a loss of degrees of freedom. If we do not use paired observations, we have $2n - 2$ degrees of freedom available as compared to $n - 1$ when we use the paired comparisons procedure.

In general, in deciding whether or not to use the paired comparisons procedure, one should be guided by the economics involved as well as by a consideration of the gains to be realized in terms of controlling extraneous variation.

Alternatives If neither z nor t is an appropriate test statistic for use with available data, one may wish to consider using some nonparametric technique to test a hypothesis about a median difference. The sign test, discussed in Chapter 13, is a candidate for use in such cases.

EXERCISES

In the following exercises, carry out the ten-step hypothesis testing procedure at the specified significance level. For each exercise, as appropriate, explain why you chose a one-sided test or a two-sided test. Discuss how you think researchers or clinicians might use the results of your hypothesis test.

What clinical or research decisions or actions do you think would be appropriate in light of the results of your test?

7.4.1 Ellen Davis Jones (A-15) studied the effects of reminiscence therapy for older women with depression. She studied 15 women 60 years or older residing for 3 months or longer in an assisted living long-term care facility. For this study, depression was measured by the Geriatric Depression Scale (GDS). Higher scores indicate more severe depression symptoms. The participants received reminiscence therapy for long-term care, which uses family photographs, scrapbooks, and personal memorabilia to stimulate memory and conversation among group members. Pre-treatment and post-treatment depression scores are given in the following table. Can we conclude, based on these data, that subjects who participate in reminiscence therapy experience, on average, a decline in GDS depression scores? Let $\alpha = .01$.

Pre–GDS:	12	10	16	2	12	18	11	16	16	10	14	21	9	19	20
Post–GDS:	11	10	11	3	9	13	8	14	16	10	12	22	9	16	18

Source: Ellen Davis Jones, N.D., R.N., FNP-C. Used with permission.

7.4.2 Beney et al. (A-16) evaluated the effect of telephone follow-up on the physical well-being dimension of health-related quality of life in patients with cancer. One of the main outcome variables was measured by the physical well-being subscale of the Functional Assessment of Cancer Therapy Scale–General (FACT-G). A higher score indicates higher physical well-being. The following table shows the baseline FACT-G score and the follow-up score to evaluate the physical well-being during the 7 days after discharge from hospital to home for 66 patients who received a phone call 48–72 hours after discharge that gave patients the opportunity to discuss medications, problems, and advice. Is there sufficient evidence to indicate that quality of physical well-being significantly decreases in the first week of discharge among patients who receive a phone call? Let $\alpha = .05$.

Subject	Baseline FACT-G	Follow-up FACT-G	Subject	Baseline FACT-G	Follow-up FACT-G
1	16	19	34	25	14
2	26	19	35	21	17
3	13	9	36	14	22
4	20	23	37	23	22
5	22	25	38	19	16
6	21	20	39	19	15
7	20	10	40	18	23
8	15	20	41	20	21
9	25	22	42	18	11
10	20	18	43	22	22
11	11	6	44	7	17
12	22	21	45	23	9
13	18	17	46	19	16
14	21	13	47	17	16
15	25	25	48	22	20
16	17	21	49	19	23
17	26	22	50	5	17
18	18	22	51	22	17
19	7	9	52	12	6

(*Continued*)

Subject	Baseline FACT-G	Follow-up FACT-G	Subject	Baseline FACT-G	Follow-up FACT-G
20	25	24	53	19	19
21	22	15	54	17	20
22	15	9	55	7	6
23	19	7	56	27	10
24	23	20	57	22	16
25	19	19	58	16	14
26	21	24	59	26	24
27	24	23	60	17	19
28	21	15	61	23	22
29	28	27	62	23	23
30	18	26	63	13	3
31	25	26	64	24	22
32	25	26	65	17	21
33	28	28	66	22	21

Source: Johnny Beney, Ph.D. and E. Beth Devine, Pharm.D., M.B.A. et al. Used with permission.

7.4.3 The purpose of an investigation by Morley et al. (A-17) was to evaluate the analgesic effectiveness of a daily dose of oral methadone in patients with chronic neuropathic pain syndromes. The researchers used a visual analogue scale (0–100 mm, higher number indicates higher pain) ratings for maximum pain intensity over the course of the day. Each subject took either 20 mg of methadone or a placebo each day for 5 days. Subjects did not know which treatment they were taking. The following table gives the mean maximum pain intensity scores for the 5 days on methadone and the 5 days on placebo. Do these data provide sufficient evidence, at the .05 level of significance, to indicate that in general the maximum pain intensity is lower on days when methadone is taken?

Subject	Methadone	Placebo
1	29.8	57.2
2	73.0	69.8
3	98.6	98.2
4	58.8	62.4
5	60.6	67.2
6	57.2	70.6
7	57.2	67.8
8	89.2	95.6
9	97.0	98.4
10	49.8	63.2
11	37.0	63.6

Source: John S. Morley, John Bridson, Tim P. Nash, John B. Miles, Sarah White, and Matthew K. Makin, "Low-Dose Methadone Has an Analgesic Effect in Neuropathic Pain: A Double-Blind Randomized Controlled Crossover Trial," *Palliative Medicine, 17* (2003), 576–587.

7.4.4 Woo and McKenna (A-18) investigated the effect of broadband ultraviolet B (UVB) therapy and topical calcipotriol cream used together on areas of psoriasis. One of the outcome variables is the Psoriasis Area and Severity Index (PASI). The following table gives the PASI scores for 20 subjects measured at baseline and after eight treatments. Do these data provide sufficient evidence, at the .01 level of significance, to indicate that the combination therapy reduces PASI scores?

Subject	Baseline	After 8 Treatments
1	5.9	5.2
2	7.6	12.2
3	12.8	4.6
4	16.5	4.0
5	6.1	0.4
6	14.4	3.8
7	6.6	1.2
8	5.4	3.1
9	9.6	3.5
10	11.6	4.9
11	11.1	11.1
12	15.6	8.4
13	6.9	5.8
14	15.2	5.0
15	21.0	6.4
16	5.9	0.0
17	10.0	2.7
18	12.2	5.1
19	20.2	4.8
20	6.2	4.2

Source: W. K. Woo, M.D. Used with permission.

7.4.5 One of the purposes of an investigation by Porcellini et al. (A-19) was to investigate the effect on CD4 T cell count of administration of intermittent interleukin (IL-2) in addition to highly active antiretroviral therapy (HAART). The following table shows the CD4 T cell count at baseline and then again after 12 months of HAART therapy with IL-2. Do the data show, at the .05 level, a significant change in CD4 T cell count?

Subject	1	2	3	4	5	6	7
CD4 T cell count at entry ($\times 10^6$/L)	173	58	103	181	105	301	169
CD4 T cell count at end of follow-up ($\times 10^6$/L)	257	108	315	362	141	549	369

Source: Simona Procellini, Giuliana Vallanti, Silvia Nozza, Guido Poli, Adraino Lazzarin, Guiseppe Tabussi, and Antonio Grassia, "Improved Thymopoietic Potential in Aviremic HIV-Infected Individuals with HAART by Intermittent IL-2 Administration," *AIDS, 17* (2003), 1621–1630.

7.5 HYPOTHESIS TESTING: A SINGLE POPULATION PROPORTION

Testing hypotheses about population proportions is carried out in much the same way as for means when the conditions necessary for using the normal curve are met. One-sided or two-sided tests may be made, depending on the question being asked. When a

sample sufficiently large for application of the central limit theorem as discussed in Section 5.5 is available for analysis, the test statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \tag{7.5.1}$$

which, when H_0 is true, is distributed approximately as the standard normal.

EXAMPLE 7.5.1

Wagenknecht et al. (A-20) collected data on a sample of 301 Hispanic women living in San Antonio, Texas. One variable of interest was the percentage of subjects with impaired fasting glucose (IFG). IFG refers to a metabolic stage intermediate between normal glucose homeostasis and diabetes. In the study, 24 women were classified in the IFG stage. The article cites population estimates for IFG among Hispanic women in Texas as 6.3 percent. Is there sufficient evidence to indicate that the population of Hispanic women in San Antonio has a prevalence of IFG higher than 6.3 percent?

Solution:

1. **Data.** The data are obtained from the responses of 301 individuals of which 24 possessed the characteristic of interest; that is, $\hat{p} = 24/301 = .080$.
2. **Assumptions.** The study subjects may be treated as a simple random sample from a population of similar subjects, and the sampling distribution of $\hat{p}$ is approximately normally distributed in accordance with the central limit theorem.
3. **Hypotheses.**

   $$H_0: p \leq .063$$
   $$H_A: p > .063$$

   We conduct the test at the point of equality. The conclusion we reach will be the same as we would reach if we conducted the test using any other hypothesized value of p greater than .063. If H_0 is true, $p = .063$ and the standard error $\sigma_{\hat{p}} = \sqrt{(.063)(.937)/301}$. Note that we use the hypothesized value of p in computing $\sigma_{\hat{p}}$. We do this because the entire test is based on the assumption that the null hypothesis is true. To use the sample proportion, $\hat{p}$, in computing $\sigma_{\hat{p}}$ would not be consistent with this concept.
4. **Test statistic.** The test statistic is given by Equation 7.5.1.
5. **Distribution of test statistic.** If the null hypothesis is true, the test statistic is approximately normally distributed with a mean of zero.
6. **Decision rule.** Let $\alpha = .05$. The critical value of z is 1.645. Reject H_0 if the computed z is ≥ 1.645.

7. **Calculation of test statistic.**

$$z = \frac{.080 - .063}{\sqrt{\frac{(.063)(.937)}{301}}} = 1.21$$

8. **Statistical decision.** Do not reject H_0 since $1.21 < 1.645$.
9. **Conclusion.** We cannot conclude that in the sampled population the proportion who are IFG is higher than 6.3 percent.
10. ***p* value.** $p = .1131$. ■

Tests involving a single proportion can be carried out using a variety of computer programs. Outputs from MINITAB and NCSS, using the data from Example 7.5.1, are shown in Figure 7.5.1. It should be noted that the results will vary slightly, because of rounding errors, if calculations are done by hand. It should also be noted that some programs, such as NCSS, use a continuity correction in calculating the z-value, and therefore the test statistic values and corresponding p values differ slightly from the MINITAB output.

MINITAB Output

Test and CI for One Proportion

```
Test of p = 0.063 vs p > 0.063

                              95% Lower
Sample       X     N    Sample p     Bound     Z-Value     P-Value
1           24   301    0.079734   0.054053      1.19       0.116

Using the normal approximation.
```

NCSS Output

Normal Approximation using (P0)

Alternative Hypothesis	Z-Value	Prob Level	Decision (5%)
P<>P0	1.0763	0.281780	Accept H0
P<P0	1.0763	0.859110	Accept H0
P>P0	1.0763	0.140890	Accept H0

FIGURE 7.5.1 MINITAB and partial NCSS output for the data in Example 7.5.1

EXERCISES

For each of the following exercises, carry out the ten-step hypothesis testing procedure at the designated level of significance. For each exercise, as appropriate, explain why you chose a one-sided test or a two-sided test. Discuss how you think researchers or clinicians might use the results of your hypothesis test. What clinical or research decisions or actions do you think would be appropriate in light of the results of your test?

7.5.1 Jacquemyn et al. (A-21) conducted a survey among gynecologists-obstetricians in the Flanders region and obtained 295 responses. Of those responding, 90 indicated that they had performed at least one cesarean section on demand every year. Does this study provide sufficient evidence for us to conclude that less than 35 percent of the gynecologists-obstetricians in the Flanders region perform at least one cesarean section on demand each year? Let $\alpha = .05$.

7.5.2 In an article in the journal *Health and Place,* Hui and Bell (A-22) found that among 2428 boys ages 7 to 12 years, 461 were overweight or obese. On the basis of this study, can we conclude that more than 15 percent of the boys ages 7 to 12 in the sampled population are obese or overweight? Let $\alpha = .05$.

7.5.3 Becker et al. (A-23) conducted a study using a sample of 50 ethnic Fijian women. The women completed a self-report questionnaire on dieting and attitudes toward body shape and change. The researchers found that five of the respondents reported at least weekly episodes of binge eating during the previous 6 months. Is this sufficient evidence to conclude that less than 20 percent of the population of Fijian women engage in at least weekly episodes of binge eating? Let $\alpha = .05$.

7.5.4 The following questionnaire was completed by a simple random sample of 250 gynecologists. The number checking each response is shown in the appropriate box.

1. When you have a choice, which procedure do you prefer for obtaining samples of endometrium?
- **(a)** Dilation and curettage **175**
- **(b)** Vobra aspiration **75**

2. Have you seen one or more pregnant women during the past year whom you knew to have elevated blood lead levels?
- **(a)** Yes **25**
- **(b)** No **225**

3. Do you routinely acquaint your pregnant patients who smoke with the suspected hazards of smoking to the fetus?
- **(a)** Yes **238**
- **(b)** No **12**

Can we conclude from these data that in the sampled population more than 60 percent prefer dilation and curettage for obtaining samples of endometrium? Let $\alpha = .01$.

7.5.5 Refer to Exercise 7.5.4. Can we conclude from these data that in the sampled population fewer than 15 percent have seen (during the past year) one or more pregnant women with elevated blood lead levels? Let $\alpha = .05$.

7.5.6 Refer to Exercise 7.5.4. Can we conclude from these data that more than 90 percent acquaint their pregnant patients who smoke with the suspected hazards of smoking to the fetus? Let $\alpha = .05$.

7.6 HYPOTHESIS TESTING: THE DIFFERENCE BETWEEN TWO POPULATION PROPORTIONS

The most frequent test employed relative to the difference between two population proportions is that their difference is zero. It is possible, however, to test that the difference is equal to some other value. Both one-sided and two-sided tests may be made.

When the null hypothesis to be tested is $p_1 - p_2 = 0$, we are hypothesizing that the two population proportions are equal. We use this as justification for combining the results of the two samples to come up with a pooled estimate of the hypothesized common proportion. If this procedure is adopted, one computes

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}, \text{ and } \bar{q} = 1 - \bar{p}$$

where x_1 and x_2 are the numbers in the first and second samples, respectively, possessing the characteristic of interest. This pooled estimate of $p = p_1 = p_2$ is used in computing $\hat{\sigma}_{\hat{p}_1 - \hat{p}_2}$, the estimated standard error of the estimator, as follows:

$$\hat{\sigma}_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\bar{p}(1 - \bar{p})}{n_1} + \frac{\bar{p}(1 - \bar{p})}{n_2}} \tag{7.6.1}$$

The test statistic becomes

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)_0}{\hat{\sigma}_{\hat{p}_1 - \hat{p}_2}} \tag{7.6.2}$$

which is distributed approximately as the standard normal if the null hypothesis is true.

EXAMPLE 7.6.1

Noonan syndrome is a genetic condition that can affect the heart, growth, blood clotting, and mental and physical development. Noonan et al. (A-24) examined the stature of men and women with Noonan syndrome. The study contained 29 male and 44 female adults. One of the cut-off values used to assess stature was the third percentile of adult height. Eleven of the males fell below the third percentile of adult male height, while 24 of the females fell below the third percentile of female adult height. Does this study provide sufficient evidence for us to conclude that among subjects with Noonan syndrome, females are more likely than males to fall below the respective third percentile of adult height? Let $\alpha = .05$.

Solution:

1. **Data.** The data consist of information regarding the height status of Noonan syndrome males and females as described in the statement of the example.
2. **Assumptions.** We assume that the patients in the study constitute independent simple random samples from populations of males and females with Noonan syndrome.
3. **Hypotheses.**

$$H_0: p_F \leq p_M \quad \text{or} \quad p_F - p_M \leq 0$$
$$H_A: p_F > p_M \quad \text{or} \quad p_F - p_M > 0$$

   where p_F is the proportion of females below the third percentile of female adult height and p_M is the proportion of males below the third percentile of male adult height.
4. **Test statistic.** The test statistic is given by Equation 7.6.2.
5. **Distribution of test statistic.** If the null hypothesis is true, the test statistic is distributed approximately as the standard normal.
6. **Decision rule.** Let $\alpha = .05$. The critical value of z is 1.645. Reject H_0 if computed z is greater than 1.645.
7. **Calculation of test statistic.** From the sample data we compute $\hat{p}_F = 24/44 = .545$, $\hat{p}_M = 11/29 = .379$, and $\bar{p} = (24 + 11)/(44 + 29) = .479$. The computed value of the test statistic, then, is

$$z = \frac{(.545 - .379)}{\sqrt{\dfrac{(.479)(.521)}{44} + \dfrac{(.479)(.521)}{29}}} = 1.39$$

8. **Statistical decision.** Fail to reject H_0 since $1.39 < 1.645$.
9. **Conclusion.** In the general population of adults with Noonan syndrome there may be no difference in the proportion of males and females who have heights below the third percentile of adult height.
10. ***p* value.** For this test $p = .0823$. ■

Tests involving two proportions, using the data from Example 7.6.1, can be carried out with a variety of computer programs. Outputs from MINITAB and NCSS are shown in Figure 7.6.1. Again, it should be noted that, because of rounding errors, the results will vary slightly if calculations are done by hand.

MINITAB Output

Test and CI for Two Proportions

```
Sample    X    N   Sample p
1        24   44   0.545455
2        11   29   0.379310

Difference = p (1) − p (2)
Estimate for difference: 0.166144
95% lower bound for difference: −0.0267550
Test for difference = 0 (vs > 0):  Z = 1.39  P-Value = 0.082
```

NCSS Output

Test Name	Test Statistic's Distribution	Test Statistic Value	Prob Level	Conclude H1 at 5% Significance?
Z-Test	Normal	1.390	0.0822	No

FIGURE 7.6.1 MINITAB and partial NCSS output for the data in Example 7.6.1

EXERCISES

In each of the following exercises use the ten-step hypothesis testing procedure. For each exercise, as appropriate, explain why you chose a one-sided test or a two-sided test. Discuss how you think researchers or clinicians might use the results of your hypothesis test. What clinical or research decisions or actions do you think would be appropriate in light of the results of your test?

7.6.1 Ho et al. (A-25) used telephone interviews of randomly selected respondents in Hong Kong to obtain information regarding individuals' perceptions of health and smoking history. Among 1222 current male smokers, 72 reported that they had "poor" or "very poor" health, while 30 among 282 former male smokers reported that they had "poor" or "very poor" health. Is this sufficient evidence to allow one to conclude that among Hong Kong men there is a difference between current and former smokers with respect to the proportion who perceive themselves as having "poor" and "very poor" health? Let $\alpha = .01$.

7.6.2 Landolt et al. (A-26) examined rates of posttraumatic stress disorder (PTSD) in mothers and fathers. Parents were interviewed 5 to 6 weeks after an accident or a new diagnosis of cancer or diabetes mellitus type I for their child. Twenty-eight of the 175 fathers interviewed and 43 of the

180 mothers interviewed met the criteria for current PTSD. Is there sufficient evidence for us to conclude that fathers are less likely to develop PTSD than mothers when a child is traumatized by an accident, cancer diagnosis, or diabetes diagnosis? Let $\alpha = .05$.

7.6.3 In a *Kidney International* article, Avram et al. (A-27) reported on a study involving 529 hemodialysis patients and 326 peritoneal dialysis patients. They found that at baseline 249 subjects in the hemodialysis treatment group were diabetic, while at baseline 134 of the subjects in the peritoneal dialysis group were diabetic. Is there a significant difference in diabetes prevalence at baseline between the two groups of this study? Let $\alpha = .05$. What does your finding regarding sample significance imply about the populations of subjects?

7.6.4 In a study of obesity the following results were obtained from samples of males and females between the ages of 20 and 75:

	n	Number Overweight
Males	150	21
Females	200	48

Can we conclude from these data that in the sampled populations there is a difference in the proportions who are overweight? Let $\alpha = .05$.

7.7 HYPOTHESIS TESTING: A SINGLE POPULATION VARIANCE

In Section 6.9 we examined how it is possible to construct a confidence interval for the variance of a normally distributed population. The general principles presented in that section may be employed to test a hypothesis about a population variance. When the data available for analysis consist of a simple random sample drawn from a normally distributed population, the test statistic for testing hypotheses about a population variance is

$$\chi^2 = (n - 1)s^2/\sigma^2 \tag{7.7.1}$$

which, when H_0 is true, is distributed as χ^2 with $n - 1$ degrees of freedom.

EXAMPLE 7.7.1

The purpose of a study by Wilkins et al. (A-28) was to measure the effectiveness of recombinant human growth hormone (rhGH) on children with total body surface area burns > 40 percent. In this study, 16 subjects received daily injections at home of rhGH. At baseline, the researchers wanted to know the current levels of insulin-like growth factor (IGF-I) prior to administration of rhGH. The sample variance of IGF-I levels (in ng/ml) was 670.81. We wish to know if we may conclude from these data that the population variance is not 600.

Solution:

1. **Data.** See statement in the example.
2. **Assumptions.** The study sample constitutes a simple random sample from a population of similar children. The IGF-I levels are normally distributed.
3. **Hypotheses.**

$$H_0: \sigma^2 = 600$$
$$H_A: \sigma^2 \neq 600$$

4. **Test statistic.** The test statistic is given by Equation 7.7.1.
5. **Distribution of test statistic.** When the null hypothesis is true, the test statistic is distributed as χ^2 with $n - 1$ degrees of freedom.
6. **Decision rule.** Let $\alpha = .05$. Critical values of χ^2 are 6.262 and 27.488. Reject H_0 unless the computed value of the test statistic is between 6.262 and 27.488. The rejection and nonrejection regions are shown in Figure 7.7.1.
7. **Calculation of test statistic.**

$$\chi^2 = \frac{15(670.81)}{600} = 16.77$$

8. **Statistical decision.** Do not reject H_0 since $6.262 < 16.77 < 27.488$.
9. **Conclusion.** Based on these data we are unable to conclude that the population variance is not 600.
10. ***p* value.** The determination of the p value for this test is complicated by the fact that we have a two-sided test and an asymmetric sampling distribution. When we have a two-sided test and a symmetric sampling distribution such as the standard normal or t, we may, as we have seen, double the one-sided p value. Problems arise when we attempt to do this with an asymmetric sampling distribution such as the

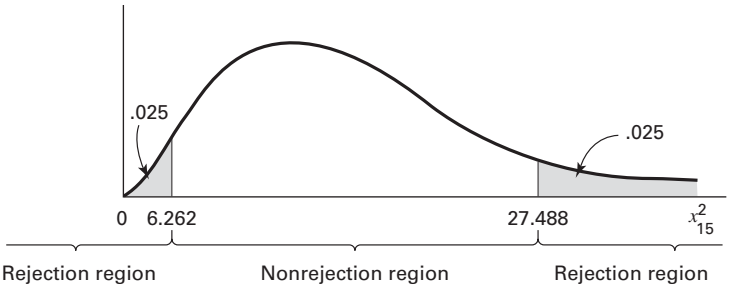

FIGURE 7.7.1 Rejection and nonrejection regions for Example 7.7.1.

chi-square distribution. In this situation the one-sided p value is reported along with the direction of the observed departure from the null hypothesis. In fact, this procedure may be followed in the case of symmetric sampling distributions. Precedent, however, seems to favor doubling the one-sided p value when the test is two-sided and involves a symmetric sampling distribution.

For the present example, then, we may report the p value as follows: $p > .05$ (two-sided test). A population variance greater than 600 is suggested by the sample data, but this hypothesis is not strongly supported by the test.

If the problem is stated in terms of the population standard deviation, one may square the sample standard deviation and perform the test as indicated above. ■

One-Sided Tests Although this was an example of a two-sided test, one-sided tests may also be made by logical modification of the procedure given here.

$$\text{For } H_A: \sigma^2 > \sigma_0^2, \text{ reject } H_0 \text{ if computed } \chi^2 \geq \chi^2_{1-\alpha}$$
$$\text{For } H_A: \sigma^2 < \sigma_0^2, \text{ reject } H_0 \text{ if computed } \chi^2 \leq \chi^2_{\alpha}$$

Tests involving a single population variance can be carried out using MINITAB software. Most other statistical computer programs lack procedures for carrying out these tests directly. The output from MINITAB, using the data from Example 7.7.1, is shown in Figure 7.7.2.

Test and CI for One Variance

```
Statistics

 N  StDev  Variance
16   25.9       671

95% Confidence Intervals

              CI for      CI for
Method        StDev       Variance
Standard  (19.1, 40.1)  (366, 1607)

Tests

Method    Chi-Square    DF    P-Value
Standard       16.77    15      0.666
```

FIGURE 7.7.2 MINITAB output for the data in Example 7.7.1.

EXERCISES

In each of the following exercises, carry out the ten-step testing procedure. For each exercise, as appropriate, explain why you chose a one-sided test or a two-sided test. Discuss how you think researchers or clinicians might use the results of your hypothesis test. What clinical or research decisions or actions do you think would be appropriate in light of the results of your test?

7.7.1 Recall Example 7.2.3, where Nakamura et al. (A-1) studied subjects with acute medial collateral ligament injury (MCL) with anterior cruciate ligament tear (ACL). The ages of the 17 subjects were:

31, 26, 21, 15, 26, 16, 19, 21, 28, 27, 22, 20, 25, 31, 20, 25, 15

Use these data to determine if there is sufficient evidence for us to conclude that in a population of similar subjects, the variance of the ages of the subjects is not 20 years. Let $\alpha = .01$.

7.7.2 Robinson et al. (A-29) studied nine subjects who underwent baffle procedure for transposition of the great arteries (TGA). At baseline, the systemic vascular resistance (SVR) (measured in $WU \times m^2$) values at rest yielded a standard deviation of 28. Can we conclude from these data that the SVR variance of a population of similar subjects with TGA is not 700? Let $\alpha = .10$.

7.7.3 Vital capacity values were recorded for a sample of 10 patients with severe chronic airway obstruction. The variance of the 10 observations was .75. Test the null hypothesis that the population variance is 1.00. Let $\alpha = .05$.

7.7.4 Hemoglobin (g percent) values were recorded for a sample of 20 children who were part of a study of acute leukemia. The variance of the observations was 5. Do these data provide sufficient evidence to indicate that the population variance is greater than 4? Let $\alpha = .05$.

7.7.5 A sample of 25 administrators of large hospitals participated in a study to investigate the nature and extent of frustration and emotional tension associated with the job. Each participant was given a test designed to measure the extent of emotional tension he or she experienced as a result of the duties and responsibilities associated with the job. The variance of the scores was 30. Can it be concluded from these data that the population variance is greater than 25? Let $\alpha = .05$.

7.7.6 In a study in which the subjects were 15 patients suffering from pulmonary sarcoid disease, blood gas determinations were made. The variance of the Pao_2 (mm Hg) values was 450. Test the null hypothesis that the population variance is greater than 250. Let $\alpha = .05$.

7.7.7 Analysis of the amniotic fluid from a simple random sample of 15 pregnant women yielded the following measurements on total protein (grams per 100 ml) present:

.69, 1.04, .39, .37, .64, .73, .69, 1.04,
.83, 1.00, .19, .61, .42, .20, .79

Do these data provide sufficient evidence to indicate that the population variance is greater than .05? Let $\alpha = .05$. What assumptions are necessary?

7.8 HYPOTHESIS TESTING: THE RATIO OF TWO POPULATION VARIANCES

As we have seen, the use of the *t* distribution in constructing confidence intervals and in testing hypotheses for the difference between two population means assumes that the population variances are equal. As a rule, the only hints available about the magnitudes

of the respective variances are the variances computed from samples taken from the populations. We would like to know if the difference that, undoubtedly, will exist between the sample variances is indicative of a real difference in population variances, or if the difference is of such magnitude that it could have come about as a result of chance alone when the population variances are equal.

Two methods of chemical analysis may give the same results on the average. It may be, however, that the results produced by one method are more variable than the results of the other. We would like some method of determining whether this is likely to be true.

Variance Ratio Test Decisions regarding the comparability of two population variances are usually based *on the variance ratio test,* which is a test of the null hypothesis that two population variances are equal. When we test the hypothesis that two population variances are equal, we are, in effect, testing the hypothesis that their ratio is equal to 1.

We learned in the preceding chapter that, when certain assumptions are met, the quantity $(s_1^2/\sigma_1^2)/(s_2^2/\sigma_2^2)$ is distributed as F with $n_1 - 1$ numerator degrees of freedom and $n_2 - 1$ denominator degrees of freedom. If we are hypothesizing that $\sigma_1^2 = \sigma_2^2$, we assume that the hypothesis is true, and the two variances cancel out in the above expression leaving s_1^2/s_2^2, which follows the same F distribution. The ratio s_1^2/s_2^2 will be designated V.R. for variance ratio.

For a two-sided test, we follow the convention of placing the larger sample variance in the numerator and obtaining the critical value of F for $\alpha/2$ and the appropriate degrees of freedom. However, for a one-sided test, which of the two sample variances is to be placed in the numerator is predetermined by the statement of the null hypothesis. For example, for the null hypothesis that σ_1^2/σ_2^2, the appropriate test statistic is V.R. $= s_1^2/s_2^2$. The critical value of F is obtained for α (not $\alpha/2$) and the appropriate degrees of freedom. In like manner, if the null hypothesis is that $\sigma_1^2 \geq \sigma_2^2$, the appropriate test statistic is V.R. $= s_2^2/s_1^2$. In all cases, the decision rule is to reject the null hypothesis if the computed V.R. is equal to or greater than the critical value of F.

EXAMPLE 7.8.1

Borden et al. (A-30) compared meniscal repair techniques using cadaveric knee specimens. One of the variables of interest was the load at failure (in newtons) for knees fixed with the FasT-FIX technique (group 1) and the vertical suture method (group 2). Each technique was applied to six specimens. The standard deviation for the FasT-FIX method was 30.62, and the standard deviation for the vertical suture method was 11.37. Can we conclude that, in general, the variance of load at failure is higher for the FasT-FIX technique than the vertical suture method?

Solution:

1. **Data.** See the statement of the example.
2. **Assumptions.** Each sample constitutes a simple random sample of a population of similar subjects. The samples are independent. We assume the loads at failure in both populations are approximately normally distributed.

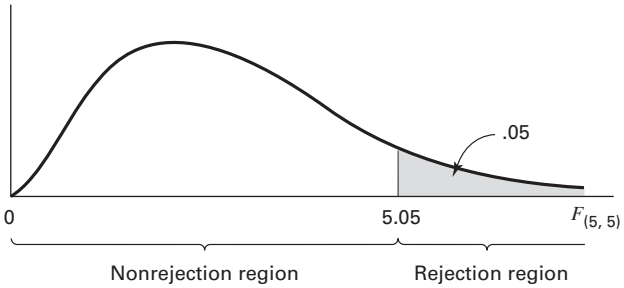

FIGURE 7.8.1 Rejection and nonrejection regions, Example 7.8.1.

3. **Hypotheses.**

$$H_0: \sigma_1^2 \leq \sigma_2^2$$
$$H_A: \sigma_1^2 > \sigma_2^2$$

4. **Test statistic.**

$$\text{V.R.} = \frac{s_1^2}{s_2^2} \tag{7.8.1}$$

5. **Distribution of test statistic.** When the null hypothesis is true, the test statistic is distributed as F with $n_1 - 1$ numerator and $n_2 - 1$ denominator degrees of freedom.
6. **Decision rule.** Let $\alpha = .05$. The critical value of F, from Appendix Table G, is 5.05. Note that if Table G does not contain an entry for the given numerator degrees of freedom, we use the column closest in value to the given numerator degrees of freedom. Reject H_0 if V.R. ≥ 5.05. The rejection and nonrejection regions are shown in Figure 7.8.1.
7. **Calculation of test statistic.**

$$\text{V.R.} = \frac{(30.62)^2}{(11.37)^2} = 7.25$$

8. **Statistical decision.** We reject H_0, since $7.25 > 5.05$; that is, the computed ratio falls in the rejection region.
9. **Conclusion.** The failure load variability is higher when using the FasT-FIX method than the vertical suture method.
10. ***p* value.** Because the computed V.R. of 7.25 is greater than 5.05, the p value for this test is less than 0.05. ■

Several computer programs can be used to test the equality of two variances. Outputs from these programs will differ depending on the test that is used. We saw in Figure 7.3.3,

for example, that the SAS system uses a folded F-test procedure. MINITAB uses two different tests. The first is an F-test under the assumption of normality, and the other is a modified Levene's test (1) that is used when normality cannot be assumed. SPSS uses an unmodified Levene's test (2). Regardless of the options, these tests are generally considered superior to the variance ratio test that is presented in Example 7.8.1. Discussion of the mathematics behind these tests is beyond the scope of this book, but an example is given to illustrate these procedures, since results from these tests are often provided automatically as outputs when a computer program is used to carry out a *t*-test.

EXAMPLE 7.8.2

Using the data from Example 7.3.2, we are interested in testing whether the assumption of the equality of variances can be assumed prior to performing a *t*-test. For ease of discussion, the data are reproduced below (Table 7.8.1):

TABLE 7.8.1 Pressures (mm Hg) Under the Pelvis During Static Conditions for Example 7.3.2

Control	131	115	124	131	122	117	88	114	150	169
SCI	60	150	130	180	163	130	121	119	130	148

Partial outputs for MINITAB, SAS, and SPSS are shown in Figure 7.8.2. Regardless of the test or program that is used, we fail to reject the null hypothesis of equal variances ($H_0: \sigma_1^2 = \sigma_2^2$) because all p values > 0.05. We may now proceed with a *t*-test under the assumption of equal variances. ■

MINITAB Output

F-Test	
Test Statistic	0.46
P-Value	0.263
Levene's Test	
Test Statistic	0.49
P-Value	0.495

SPSS Output

Levene's Test for Equality of Variances	
F	Sig.
.664	.482

SAS Output

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
pressure	Folded F	9	9	2.17	0.2626

FIGURE 7.8.2 Partial MINITAB, SPSS, and SAS outputs for testing the equality of two population variances.

EXERCISES

In the following exercises perform the ten-step test. For each exercise, as appropriate, explain why you chose a one-sided test or a two-sided test. Discuss how you think researchers or clinicians might use the results of your hypothesis test. What clinical or research decisions or actions do you think would be appropriate in light of the results of your test?

7.8.1 Dora et al. (A-31) investigated spinal canal dimensions in 30 subjects symptomatic with disc herniation selected for a discectomy and 45 asymptomatic individuals. The researchers wanted to know if spinal canal dimensions are a significant risk factor for the development of sciatica. Toward that end, they measured the spinal canal dimension between vertebrae L3 and L4 and obtained a mean of 17.8 mm in the discectomy group with a standard deviation of 3.1. In the control group, the mean was 18.5 mm with a standard deviation of 2.8 mm. Is there sufficient evidence to indicate that in relevant populations the variance for subjects symptomatic with disc herniation is larger than the variance for control subjects? Let $\alpha = .05$.

7.8.2 Nagy et al. (A-32) studied 50 stable patients who were admitted for a gunshot wound that traversed the mediastinum. Of these, eight were deemed to have a mediastinal injury and 42 did not. The standard deviation for the ages of the eight subjects with mediastinal injury was 4.7 years, and the standard deviation of ages for the 42 without injury was 11.6 years. Can we conclude from these data that the variance of age is larger for a population of similar subjects without injury compared to a population with mediastinal injury? Let $\alpha = .05$.

7.8.3 A test designed to measure level of anxiety was administered to a sample of male and a sample of female patients just prior to undergoing the same surgical procedure. The sample sizes and the variances computed from the scores were as follows:

$$\text{Males:} \quad n = 16, s^2 = 150$$
$$\text{Females:} \quad n = 21, s^2 = 275$$

Do these data provide sufficient evidence to indicate that in the represented populations the scores made by females are more variable than those made by males? Let $\alpha = .05$.

7.8.4 In an experiment to assess the effects on rats of exposure to cigarette smoke, 11 animals were exposed and 11 control animals were not exposed to smoke from unfiltered cigarettes. At the end of the experiment, measurements were made of the frequency of the ciliary beat (beats/min at 20°C) in each animal. The variance for the exposed group was 3400 and 1200 for the unexposed group. Do these data indicate that in the populations represented the variances are different? Let $\alpha = .05$.

7.8.5 Two pain-relieving drugs were compared for effectiveness on the basis of length of time elapsing between administration of the drug and cessation of pain. Thirteen patients received drug 1, and 13 received drug 2. The sample variances were $s_1^2 = 64$ and $s_2^2 = 16$. Test the null hypothesis that the two populations variances are equal. Let $\alpha = .05$.

7.8.6 Packed cell volume determinations were made on two groups of children with cyanotic congenital heart disease. The sample sizes and variances were as follows:

Group	n	s^2
1	10	40
2	16	84

Do these data provide sufficient evidence to indicate that the variance of population 2 is larger than the variance of population 1? Let $\alpha = .05$.

7.8.7 Independent simple random samples from two strains of mice used in an experiment yielded the following measurements on plasma glucose levels following a traumatic experience:

Strain A:	54, 99, 105, 46, 70, 87, 55, 58, 139, 91
Strain B:	93, 91, 93, 150, 80, 104, 128, 83, 88, 95, 94, 97

Do these data provide sufficient evidence to indicate that the variance is larger in the population of strain A mice than in the population of strain B mice? Let $\alpha = .05$. What assumptions are necessary?

7.9 THE TYPE II ERROR AND THE POWER OF A TEST

In our discussion of hypothesis testing our focus has been on α, the probability of committing a type I error (rejecting a true null hypothesis). We have paid scant attention to β, the probability of committing a type II error (failing to reject a false null hypothesis). There is a reason for this difference in emphasis. For a given test, α is a single number assigned by the investigator in advance of performing the test. It is a measure of the acceptable risk of rejecting a true null hypothesis. On the other hand, β may assume one of many values. Suppose we wish to test the null hypothesis that some population parameter is equal to some specified value. If H_0 is false and we fail to reject it, we commit a type II error. If the hypothesized value of the parameter is not the true value, the value of β (the probability of committing a type II error) depends on several factors: (1) the true value of the parameter of interest, (2) the hypothesized value of the parameter, (3) the value of α, and (4) the sample size, n. For fixed α and n, then, we may, before performing a hypothesis test, compute many values of β by postulating many values for the parameter of interest given that the hypothesized value is false.

For a given hypothesis test it is of interest to know how well the test controls type II errors. If H_0 is in fact false, we would like to know the probability that we will reject it. The *power* of a test, designated $1 - \beta$, provides this desired information. The quantity $1 - \beta$ is the probability that we will reject a false null hypothesis; it may be computed for any alternative value of the parameter about which we are testing a hypothesis. Therefore, $1 - \beta$ is the probability that we will take the correct action when H_0 is false because the true parameter value is equal to the one for which we computed $1 - \beta$. For a given test we may specify any number of possible values of the parameter of interest and for each compute the value of $1 - \beta$. The result is called a *power function.* The graph of a power function, called a *power curve,* is a helpful device for quickly assessing the nature of the power of a given test. The following example illustrates the procedures we use to analyze the power of a test.

EXAMPLE 7.9.1

Suppose we have a variable whose values yield a population standard deviation of 3.6. From the population we select a simple random sample of size $n = 100$. We select a value of $\alpha = .05$ for the following hypotheses:

$$H_0: \mu = 17.5, \quad H_A: \mu \neq 17.5$$

Solution: When we study the power of a test, we locate the rejection and nonrejection regions on the $\bar{x}$ scale rather than the z scale. We find the critical values of $\bar{x}$ for a two-sided test using the following formulas:

$$\bar{x}_{\mathrm{U}} = \mu_0 + z\frac{\sigma}{\sqrt{n}} \tag{7.9.1}$$

and

$$\bar{x}_{\mathrm{L}} = \mu_0 - z\frac{\sigma}{\sqrt{n}} \tag{7.9.2}$$

where $\bar{x}_{\mathrm{U}}$ and $\bar{x}_{\mathrm{L}}$ are the upper and lower critical values, respectively, of $\bar{x}$; $+z$ and $-z$ are the critical values of z; and μ_0 is the hypothesized value of μ. For our example, we have

$$\begin{aligned}\bar{x}_{\mathrm{U}} &= 17.50 + 1.96\frac{(3.6)}{(10)} = 17.50 + 1.96(.36)\\ &= 17.50 + .7056 = 18.21\end{aligned}$$

and

$$\bar{x}_{\mathrm{L}} = 17.50 - 1.96(.36) = 17.50 - .7056 = 16.79$$

Suppose that H_0 is false, that is, that μ is not equal to 17.5. In that case, μ is equal to some value other than 17.5. We do not know the actual value of μ. But if H_0 is false, μ is one of the many values that are greater than or smaller than 17.5. Suppose that the true population mean is $\mu_1 = 16.5$. Then the sampling distribution of $\bar{x}_1$ is also approximately normal, with $\mu_{\bar{x}} = \mu = 16.5$. We call this sampling distribution $f(\bar{x}_1)$, and we call the sampling distribution under the null hypothesis $f(\bar{x}_0)$.

β, the probability of the type II error of failing to reject a false null hypothesis, is the area under the curve of $f(\bar{x}_1)$ that overlaps the nonrejection region specified under H_0. To determine the value of β, we find the area under $f(\bar{x}_1)$, above the $\bar{x}$ axis, and between $\bar{x} = 16.79$ and $\bar{x} = 18.21$. The value of β is equal to $P(16.79 \le \bar{x} \le 18.21)$ when $\mu = 16.5$. This is the same as

$$\begin{aligned}P\left(\frac{16.79 - 16.5}{.36} \le z \le \frac{18.21 - 16.5}{.36}\right) &= P\left(\frac{.29}{.36} \le z \le \frac{1.71}{.36}\right)\\ &= P(.81 \le z \le 4.75)\\ &\approx 1 - .7910 = .2090\end{aligned}$$

Thus, the probability of taking an appropriate action (that is, rejecting H_0) when the null hypothesis states that $\mu = 17.5$, but in fact $\mu = 16.5$, is

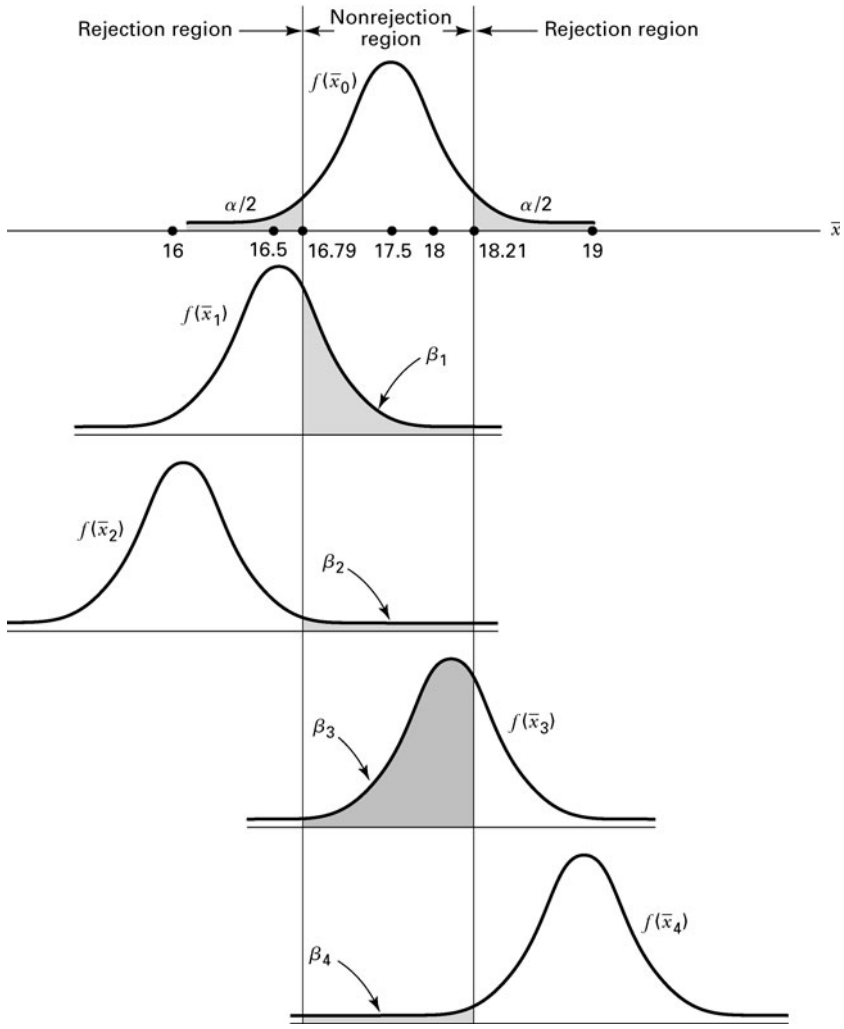

FIGURE 7.9.1 Size of β for selected values for H_1 for Example 7.9.1.

$1 - .2090 = .7910$. As we noted, μ may be one of a large number of possible values when H_0 is false. Figure 7.9.1 shows a graph of several such possibilities. Table 7.9.1 shows the corresponding values of β and $1 - \beta$ (which are approximate), along with the values of β for some additional alternatives.

Note that in Figure 7.9.1 and Table 7.9.1 those values of μ under the alternative hypothesis that are closer to the value of μ specified by H_0 have larger associated β values. For example, when $\mu = 18$ under the alternative hypothesis, $\beta = .7190$; and when $\mu = 19.0$ under H_A, $\beta = .0143$. The power of the test for these two alternatives, then, is $1 - .7190 = .2810$ and $1 - .0143 = .9857$, respectively. We show the power of the test graphically

TABLE 7.9.1 Values of β and $1 - \beta$ for Selected Alternative Values of μ_1, Example 7.9.1

Possible Values of μ Under H_A When H_0 is False	β	$1 - \beta$
16.0	0.0143	0.9857
16.5	0.2090	0.7910
17.0	0.7190	0.2810
18.0	0.7190	0.2810
18.5	0.2090	0.7910
19.0	0.0143	0.9857

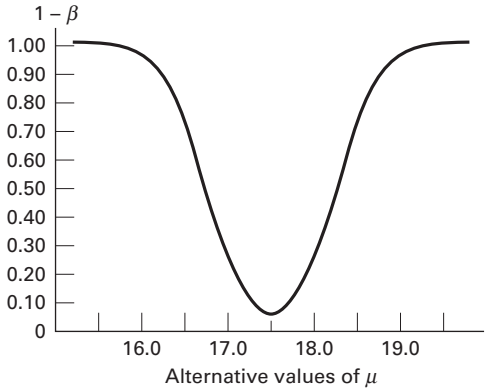

FIGURE 7.9.2 Power curve for Example 7.9.1.

in a power curve, as in Figure 7.9.2. Note that the higher the curve, the greater the power. ■

Although only one value of α is associated with a given hypothesis test, there are many values of β, one for each possible value of μ if μ_0 is not the true value of μ as hypothesized. Unless alternative values of μ are much larger or smaller than μ_0, β is relatively large compared with α. Typically, we use hypothesis-testing procedures more often in those cases in which, when H_0 is false, the true value of the parameter is fairly close to the hypothesized value. In most cases, β, the computed probability of failing to reject a false null hypothesis, is larger than α, the probability of rejecting a true null hypothesis. These facts are compatible with our statement that a decision based on a rejected null hypothesis is more conclusive than a decision based on a null hypothesis that is not rejected. The probability of being wrong in the latter case is generally larger than the probability of being wrong in the former case.

Figure 7.9.2 shows the V-shaped appearance of a power curve for a two-sided test. In general, a two-sided test that discriminates well between the value of the parameter in H_0 and values in H_1 results in a narrow V-shaped power curve. A wide V-shaped curve

indicates that the test discriminates poorly over a relatively wide interval of alternative values of the parameter.

Power Curves for One-Sided Tests The shape of a power curve for a one-sided test with the rejection region in the upper tail is an elongated S. If the rejection region of a one-sided test is located in the lower tail of the distribution, the power curve takes the form of a reverse elongated S. The following example shows the nature of the power curve for a one-sided test.

EXAMPLE 7.9.2

The mean time laboratory employees now take to do a certain task on a machine is 65 seconds, with a standard deviation of 15 seconds. The times are approximately normally distributed. The manufacturers of a new machine claim that their machine will reduce the mean time required to perform the task. The quality-control supervisor designs a test to determine whether or not she should believe the claim of the makers of the new machine. She chooses a significance level of $\alpha = 0.01$ and randomly selects 20 employees to perform the task on the new machine. The hypotheses are

$$H_0: \mu \geq 65, \quad H_A: \mu < 65$$

The quality-control supervisor also wishes to construct a power curve for the test.

Solution: The quality-control supervisor computes, for example, the following value of $1 - \beta$ for the alternative $\mu = 55$. The critical value of $1 - \beta$ for the test is

$$65 - 2.33\left(\frac{15}{\sqrt{20}}\right) = 57$$

We find β as follows:

$$\beta = P(\bar{x} > 57 \mid \mu = 55) = P\left(z > \frac{57 - 55}{15/\sqrt{20}}\right) = P(z > .60)$$

$$= 1 - .7257 = .2743$$

Consequently, $1 - \beta = 1 - .2743 = .7257$. Figure 7.9.3 shows the calculation of β. Similar calculations for other alternative values of μ also yield

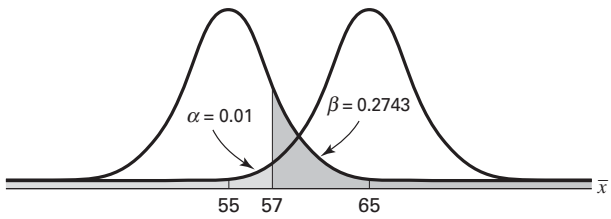

FIGURE 7.9.3 β calculated for $\mu = 55$.

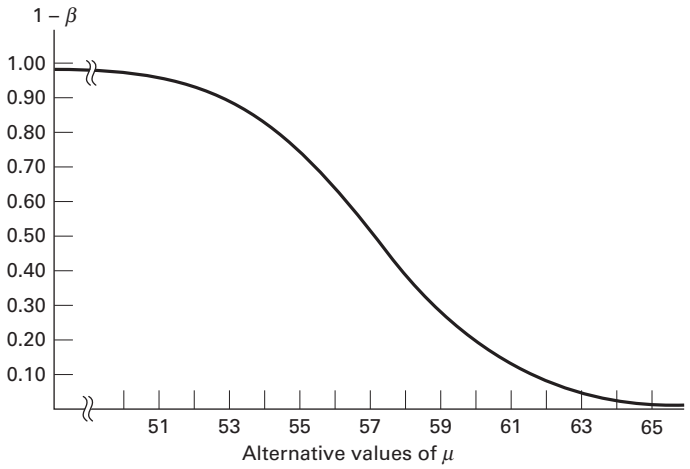

FIGURE 7.9.4 Power curve for Example 7.9.2.

values of $1 - \beta$. When plotted against the values of μ, these give the power curve shown in Figure 7.9.4. ■

Operating Characteristic Curves Another way of evaluating a test is to look at its *operating characteristic* (OC) *curve*. To construct an OC curve, we plot values of β, rather than $1 - \beta$, along the vertical axis. Thus, an OC curve is the complement of the corresponding power curve.

EXERCISES

Construct and graph the power function for each of the following situations.

7.9.1 $H_0: \mu \leq 516, \quad H_A: \mu > 516, \quad n = 16, \sigma = 32, \alpha = 0.05.$

7.9.2 $H_0: \mu = 3, \quad H_A: \mu \neq 3, \quad n = 100, \sigma = 1, \alpha = 0.05.$

7.9.3 $H_0: \mu \leq 4.25, \quad H_A: \mu > 4.25, \quad n = 81, \sigma = 1.8, \alpha = 0.01.$

7.10 DETERMINING SAMPLE SIZE TO CONTROL TYPE II ERRORS

You learned in Chapter 6 how to find the sample sizes needed to construct confidence intervals for population means and proportions for specified levels of confidence. You learned in Chapter 7 that confidence intervals may be used to test hypotheses. The method of determining sample size presented in Chapter 6 takes into account the probability of a type I error, but not a type II error since the level of confidence is determined by the confidence coefficient, $1 - \alpha$.

In many statistical inference procedures, the investigator wishes to consider the type II error as well as the type I error when determining the sample size. To illustrate the procedure, we refer again to Example 7.9.2.

EXAMPLE 7.10.1

In Example 7.9.2, the hypotheses are

$$H_0: \mu \geq 65, \quad H_A: \mu < 65$$

The population standard deviation is 15, and the probability of a type I error is set at .01. Suppose that we want the probability of failing to reject $H_0(\beta)$ to be .05 if H_0 is false because the true mean is 55 rather than the hypothesized 65. How large a sample do we need in order to realize, simultaneously, the desired levels of α and β?

Solution: For $\alpha = .01$ and $n = 20$, β is equal to .2743. The critical value is 57. Under the new conditions, the critical value is unknown. Let us call this new critical value C. Let μ_0 be the hypothesized mean and μ_1 the mean under the alternative hypothesis. We can transform each of the relevant sampling distributions of $\bar{x}$, the one with a mean of μ_0 and the one with a mean of μ_1 to a z distribution. Therefore, we can convert C to a z value on the horizontal scale of each of the two standard normal distributions. When we transform the sampling distribution of $\bar{x}$ that has a mean of μ_0 to the standard normal distribution, we call the z that results z_0. When we transform the sampling distribution $\bar{x}$ that has a mean of μ_1 to the standard normal distribution, we call the z that results z_1. Figure 7.10.1 represents the situation described so far.

We can express the critical value C as a function of z_0 and μ_0 and also as a function of z_1 and μ_1. This gives the following equations:

$$C = \mu_0 - z_0 \frac{\sigma}{\sqrt{n}} \tag{7.10.1}$$

$$C = \mu_1 + z_1 \frac{\sigma}{\sqrt{n}} \tag{7.10.2}$$

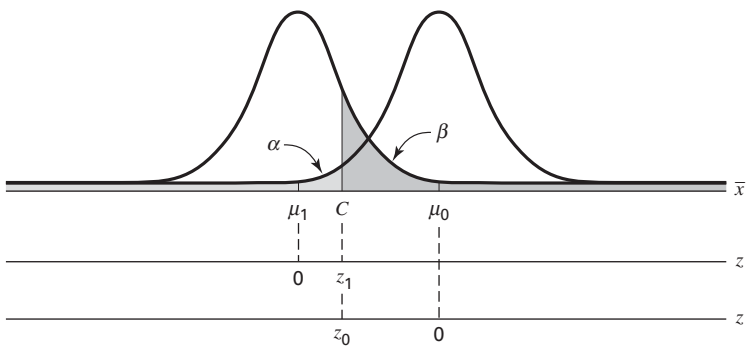

FIGURE 7.10.1 Graphic representation of relationships in determination of sample size to control both type I and type II errors.

We set the right-hand sides of these equations equal to each other and solve for n, to obtain

$$n = \left[\frac{(z_0 + z_1)\sigma}{(\mu_0 - \mu_1)}\right]^2 \tag{7.10.3}$$

To find n for our illustrative example, we substitute appropriate quantities into Equation 7.10.3. We have $\mu_0 = 65$, $\mu_1 = 55$, and $\sigma = 15$. From Appendix Table D, the value of z that has .01 of the area to its left is -2.33. The value of z that has .05 of the area to its right is 1.645. Both z_0 and z_1 are taken as positive. We determine whether C lies above or below either μ_0 or μ_1 when we substitute into Equations 7.10.1 and 7.10.2. Thus, we compute

$$n = \left[\frac{(2.33 + 1.645)(15)}{(65 - 55)}\right]^2 = 35.55$$

We would need a sample of size 36 to achieve the desired levels of α and β when we choose $\mu_1 = 55$ as the alternative value of μ.

We now compute C, the critical value for the test, and state an appropriate decision rule. To find C, we may substitute known numerical values into either Equation 7.10.1 or Equation 7.10.2. For illustrative purposes, we solve both equations for C. First we have

$$C = 65 - 2.33\left(\frac{15}{\sqrt{36}}\right) = 59.175$$

From Equation 7.10.2, we have

$$C = 55 - 1.645\left(\frac{15}{\sqrt{36}}\right) = 59.1125$$

The difference between the two results is due to rounding error.

The decision rule, when we use the first value of C, is as follows:

Select a sample of size 36 and compute $\bar{x}$, if $\bar{x} \leq 59.175$, reject H_0. If $\bar{x} > 59.175$, do not reject H_0.

We have limited our discussion of the type II error and the power of a test to the case involving a population mean. The concepts extend to cases involving other parameters. ■

EXERCISES

7.10.1 Given H_0: $\mu = 516$, H_A: $\mu > 516$, $n = 16$, $\sigma = 32$, $\alpha = .05$. Let $\beta = .10$ and $\mu_1 = 520$, and find n and C. State the appropriate decision rule.

7.10.2 Given H_0: $\mu \leq 4.500$, H_A: $\mu > 4.500$, $n = 16$, $\sigma = .020$, $\alpha = .01$. Let $\beta = .05$ and $\mu_1 = 4.52$, and find n and C. State the appropriate decision rule.

7.10.3 Given H_0: $\mu \leq 4.25$, H_A: $\mu > 4.25$, $n = 81$, $\sigma = 1.8$, $\alpha = .01$. Let $\beta = .03$ and $\mu_1 = 5.00$, and find n and C. State the appropriate decision rule.

7.11 SUMMARY

In this chapter the general concepts of hypothesis testing are discussed. A general procedure for carrying out a hypothesis test consisting of the following ten steps is suggested.

1. Description of data.
2. Statement of necessary assumptions.
3. Statement of null and alternative hypotheses.
4. Specification of the test statistic.
5. Specification of the distribution of the test statistic.
6. Statement of the decision rule.
7. Calculation of test statistic from sample data.
8. The statistical decision based on sample results.
9. Conclusion.
10. Determination of p value.

A number of specific hypothesis tests are described in detail and illustrated with appropriate examples. These include tests concerning population means, the difference between two population means, paired comparisons, population proportions, the difference between two population proportions, a population variance, and the ratio of two population variances. In addition we discuss the power of a test and the determination of sample size for controlling both type I and type II errors.

SUMMARY OF FORMULAS FOR CHAPTER 7

Formula Number	Name	Formula
7.1.1, 7.1.2, 7.2.1	z-transformation (using either μ or μ_0)	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
7.2.2	t-transformation	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
7.2.3	Test statistic when sampling from a population that is not normally distributed	$z = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
7.3.1	Test statistic when sampling from normally distributed populations: population variances known	$z = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$

(*Continued*)

7.3.2	Test statistic when sampling from normally distributed populations: population variances unknown and equal	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\dfrac{s_p^2}{n_1} + \dfrac{s_p^2}{n_2}}}$, where $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$
7.3.3, 7.3.4	Test statistic when sampling from normally distributed populations: population variances unknown and unequal	$t' = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$, where $t'_{(1-\alpha/2)} = \dfrac{w_1 t_1 + w_2 t_2}{w_1 + w_2}$
7.3.5	Sampling from populations that are not normally distributed	$z = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$
7.4.1	Test statistic for paired differences when the population variance is unknown	$t = \dfrac{\bar{d} - \mu_{d_0}}{s_{\bar{d}}}$
7.4.2	Test statistic for paired differences when the population variance is known	$z = \dfrac{\bar{d} - \mu_{d_0}}{\sigma_d/\sqrt{n}}$
7.5.1	Test statistic for a single population proportion	$z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0 q_0}{n}}}$
7.6.1, 7.6.2	Test statistic for the difference between two population proportions	$z = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)_0}{\hat{\sigma}_{\hat{p}_1 - \hat{p}_2}}$, where $\bar{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$, and $\hat{\sigma}_{\hat{p}_1 - \hat{p}_2} = \sqrt{\dfrac{\bar{p}(1 - \bar{p})}{n_1} + \dfrac{\bar{p}(1 - \bar{p})}{n_2}}$
7.7.1	Test statistic for a single population variance	$\chi^2 = \dfrac{(n - 1)s^2}{\sigma^2}$
7.8.1	Variance ratio	$V.R. = \dfrac{s_1^2}{s_2^2}$

7.9.1, 7.9.2	Upper and lower critical values for $\bar{x}$	$\bar{X}_U = \mu_0 + z\frac{\sigma}{\sqrt{n}}$ $\bar{X}_L = \mu_0 - z\frac{\sigma}{\sqrt{n}}$
7.10.1, 7.10.2	Critical value for determining sample size to control type II errors	$C = \mu_0 - z_0\frac{\sigma}{\sqrt{n}} = \mu_1 + z_1\frac{\sigma}{\sqrt{n}}$
7.10.3	Sample size to control type II errors	$n = \left[\frac{(z_0 + z_1)\sigma}{(\mu_0 - \mu_1)}\right]^2$
Symbol Key	• α = type 1 error rate • C = critical value • χ^2 = chi-square distribution • $\bar{d}$ = average difference • μ = mean of population • μ_0 = hypothesized mean • n = sample size • p = proportion for population • $\bar{p}$ = average proportion • q = $(1-p)$ • $\hat{p}$ = estimated proportion for sample • σ^2 = population variance • σ = population standard deviation • $\sigma_{\bar{d}}$ = standard error of difference • $\sigma_{\bar{x}}$ = standard error • s = standard deviation of sample • $s_{\bar{d}}$ = standard deviation of the difference • s_p = pooled standard deviation • t = Student's t-transformation • t' = Cochran's correction to t • $\bar{x}$ = mean of sample • $\bar{x}_L$ = lower limit of critical value for $\bar{x}$ • $\bar{x}_U$ = upper limit of critical value for $\bar{x}$ • z = standard normal transformation	

REVIEW QUESTIONS AND EXERCISES

1. What is the purpose of hypothesis testing?

2. What is a hypothesis?

3. List and explain each step in the ten-step hypothesis testing procedure.

4. Define:
- **(a)** Type I error
- **(b)** Type II error
- **(c)** The power of a test
- **(d)** Power function
- **(e)** Power curve
- **(f)** Operating characteristic curve

5. Explain the difference between the power curves for one-sided tests and two-sided tests.

6. Explain how one decides what statement goes into the null hypothesis and what statement goes into the alternative hypothesis.

7. What are the assumptions underlying the use of the t statistic in testing hypotheses about a single mean? The difference between two means?

8. When may the z statistic be used in testing hypotheses about
- **(a)** a single population mean?
- **(b)** the difference between two population means?
- **(c)** a single population proportion?
- **(d)** the difference between two population proportions?

9. In testing a hypothesis about the difference between two population means, what is the rationale behind pooling the sample variances?

10. Explain the rationale behind the use of the paired comparisons test.

11. Give an example from your field of interest where a paired comparisons test would be appropriate. Use real or realistic data and perform an appropriate hypothesis test.

12. Give an example from your field of interest where it would be appropriate to test a hypothesis about the difference between two population means. Use real or realistic data and carry out the ten-step hypothesis testing procedure.

13. Do Exercise 12 for a single population mean.

14. Do Exercise 12 for a single population proportion.

15. Do Exercise 12 for the difference between two population proportions.

16. Do Exercise 12 for a population variance.

17. Do Exercise 12 for the ratio of two population variances.

18. Ochsenkühn et al. (A-33) studied birth as a result of in vitro fertilization (IVF) and birth from spontaneous conception. In the sample, there were 163 singleton births resulting from IVF with a mean birth weight of 3071 g and sample standard deviation of 761 g. Among the 321 singleton births resulting from spontaneous conception, the mean birth weight was 3172 g with a standard deviation of 702 g. Determine if these data provide sufficient evidence for us to conclude that the mean birth weight in grams of singleton births resulting from IVF is lower, in general, than the mean birth weight of singleton births resulting from spontaneous conception. Let $\alpha = .10$.

19. William Tindall (A-34) performed a retrospective study of the records of patients receiving care for hypercholesterolemia. The following table gives measurements of total cholesterol for patients before and 6 weeks after taking a statin drug. Is there sufficient evidence at the $\alpha = .01$ level of significance for us to conclude that the drug would result in reduction in total cholesterol in a population of similar hypercholesterolemia patients?

Id. No.	Before	After	Id. No.	Before	After	Id. No.	Before	After
1	195	125	37	221	191	73	205	151
2	208	164	38	245	164	74	298	163
3	254	152	39	250	162	75	305	171
4	226	144	40	266	180	76	262	129
5	290	212	41	240	161	77	320	191
6	239	171	42	218	168	78	271	167
7	216	164	43	278	200	79	195	158
8	286	200	44	185	139	80	345	192
9	243	190	45	280	207	81	223	117
10	217	130	46	278	200	82	220	114
11	245	170	47	223	134	83	279	181
12	257	182	48	205	133	84	252	167
13	199	153	49	285	161	85	246	158
14	277	204	50	314	203	86	304	190
15	249	174	51	235	152	87	292	177
16	197	160	52	248	198	88	276	148
17	279	205	53	291	193	89	250	169
18	226	159	54	231	158	90	236	185
19	262	170	55	208	148	91	256	172
20	231	180	56	263	203	92	269	188
21	234	161	57	205	156	93	235	172
22	170	139	58	230	161	94	184	151
23	242	159	59	250	150	95	253	156
24	186	114	60	209	181	96	352	219
25	223	134	61	269	186	97	266	186
26	220	166	62	261	164	98	321	206
27	277	170	63	255	164	99	233	173
28	235	136	64	275	195	100	224	109
29	216	134	65	239	169	101	274	109
30	197	138	66	298	177	102	222	136
31	253	181	67	265	217	103	194	131
32	209	147	68	220	191	104	293	228
33	245	164	69	196	129	105	262	211
34	217	159	70	177	142	106	306	192
35	187	139	71	211	138	107	239	174
36	265	171	72	244	166			

Source: William Tindall, Ph.D. and the Wright State University Consulting Center. Used with permission.

20. The objective of a study by van Vollenhoven et al. (A-35) was to examine the effectiveness of Etanercept alone and Etanercept in combination with methotrexate in the treatment of rheumatoid arthritis. They performed a retrospective study using data from the STURE database, which collects efficacy and safety data for all patients starting biological treatments at the major hospitals in Stockholm, Sweden. The researchers identified 40 subjects who were prescribed Etanercept only and 57 who were given Etanercept with methotrexate. One of the outcome measures was the number of swollen joints. The following table gives the mean number of swollen joints in the two groups as well as the standard error of the mean. Is there sufficient evidence at the $\alpha = .05$ level of significance for us to conclude that there is a difference in mean swollen joint counts in the relevant populations?

Treatment	Mean	Standard Error of Mean
Etanercept	5.56	0.84
Etanercept plus methotrexate	4.40	0.57

21. Miyazaki et al. (A-36) examined the recurrence-free rates of stripping with varicectomy and stripping with sclerotherapy for the treatment of primary varicose veins. The varicectomy group consisted of 122 limbs for which the procedure was done, and the sclerotherapy group consisted of 98 limbs for which that procedure was done. After 3 years, 115 limbs of the varicectomy group and 87 limbs of the sclerotherapy group were recurrence-free. Is this sufficient evidence for us to conclude there is no difference, in general, in the recurrence-free rate between the two procedures for treating varicose veins? Let $\alpha = .05$.

22. Recall the study, reported in Exercise 7.8.1, in which Dora et al. (A-37) investigated spinal canal dimensions in 30 subjects symptomatic with disc herniation selected for a discectomy and 45 asymptomatic individuals (control group). One of the areas of interest was determining if there is a difference between the two groups in the spinal canal cross-sectional area (cm^2) between vertebrae L5/S1. The data in the following table are simulated to be consistent with the results reported in the paper. Do these simulated data provide evidence for us to conclude that a difference in the spinal canal cross-sectional area exists between a population of subjects with disc herniations and a population of those who do not have disc herniations? Let $\alpha = .05$.

Herniated Disc Group					Control Group					
2.62	2.57	1.98	3.21	3.59	3.72	4.30	2.87	3.87	2.73	5.28
1.60	1.80	3.91	2.56	1.53	1.33	2.36	3.67	1.64	3.54	3.63
2.39	2.67	3.53	2.26	2.82	4.26	3.08	3.32	4.00	2.76	3.58
2.05	1.19	3.01	2.39	3.61	3.11	3.94	4.39	3.73	2.22	2.73
2.09	3.79	2.45	2.55	2.10	5.02	3.62	3.02	3.15	3.57	2.37
2.28	2.33	2.81	3.70	2.61	5.42	3.35	2.62	3.72	4.37	5.28
					4.97	2.58	2.25	3.12	3.43	
					3.95	2.98	4.11	3.08	2.22	

Source: Simulated data.

23. Iannelo et al. (A-38) investigated differences between triglyceride levels in healthy obese (control) subjects and obese subjects with chronic active B or C hepatitis. Triglyceride levels of 208 obese controls had a mean value of 1.81 with a standard error of the mean of .07 mmol/L. The 19 obese hepatitis subjects had a mean of .71 with a standard error of the mean of .05. Is this sufficient evidence for us to conclude that, in general, a difference exists in average triglyceride levels between obese healthy subjects and obese subjects with hepatitis B or C? Let $\alpha = .01$.

24. Kindergarten students were the participants in a study conducted by Susan Bazyk et al. (A-39). The researchers studied the fine motor skills of 37 children receiving occupational therapy. They used an index of fine motor skills that measured hand use, eye–hand coordination, and manual

dexterity before and after 7 months of occupational therapy. Higher values indicate stronger fine motor skills. The scores appear in the following table.

Subject	Pre	Post	Subject	Pre	Post
1	91	94	20	76	112
2	61	94	21	79	91
3	85	103	22	97	100
4	88	112	23	109	112
5	94	91	24	70	70
6	112	112	25	58	76
7	109	112	26	97	97
8	79	97	27	112	112
9	109	100	28	97	112
10	115	106	29	112	106
11	46	46	30	85	112
12	45	41	31	112	112
13	106	112	32	103	106
14	112	112	33	100	100
15	91	94	34	88	88
16	115	112	35	109	112
17	59	94	36	85	112
18	85	109	37	88	97
19	112	112			

Source: Susan Bazyk, M.H.S. Used with permission.

Can one conclude on the basis of these data that after 7 months, the fine motor skills in a population of similar subjects would be stronger? Let $\alpha = .05$. Determine the p value.

25. A survey of 90 recently delivered women on the rolls of a county welfare department revealed that 27 had a history of intrapartum or postpartum infection. Test the null hypothesis that the population proportion with a history of intrapartum or postpartum infection is less than or equal to .25. Let $\alpha = .05$. Determine the p value.

26. In a sample of 150 hospital emergency admissions with a certain diagnosis, 128 listed vomiting as a presenting symptom. Do these data provide sufficient evidence to indicate, at the .01 level of significance, that the population proportion is less than .92? Determine the p value.

27. A research team measured tidal volume in 15 experimental animals. The mean and standard deviation were 45 and 5 cc, respectively. Do these data provide sufficient evidence to indicate that the population mean is greater than 40 cc? Let $\alpha = .05$.

28. A sample of eight patients admitted to a hospital with a diagnosis of biliary cirrhosis had a mean IgM level of 160.55 units per milliliter. The sample standard deviation was 50. Do these data provide sufficient evidence to indicate that the population mean is greater than 150? Let $\alpha = .05$. Determine the p value.

29. Some researchers have observed a greater airway resistance in smokers than in nonsmokers. Suppose a study, conducted to compare the percent of tracheobronchial retention of particles in smoking-discordant monozygotic twins, yielded the following results:

Percent Retention		Percent Retention	
Smoking Twin	Nonsmoking Twin	Smoking Twin	Nonsmoking Twin
60.6	47.5	57.2	54.3
12.0	13.3	62.7	13.9
56.0	33.0	28.7	8.9
75.2	55.2	66.0	46.1
12.5	21.9	25.2	29.8
29.7	27.9	40.1	36.2

Do these data support the hypothesis that tracheobronchial clearance is slower in smokers? Let $\alpha = .05$. Determine the p value for this test.

30. Circulating levels of estrone were measured in a sample of 25 postmenopausal women following estrogen treatment. The sample mean and standard deviation were 73 and 16, respectively. At the .05 significance level can one conclude on the basis of these data that the population mean is higher than 70?

31. Systemic vascular resistance determinations were made on a sample of 16 patients with chronic, congestive heart failure while receiving a particular treatment. The sample mean and standard deviation were 1600 and 700, respectively. At the .05 level of significance do these data provide sufficient evidence to indicate that the population mean is less than 2000?

32. The mean length at birth of 14 male infants was 53 cm with a standard deviation of 9 cm. Can one conclude on the basis of these data that the population mean is not 50 cm? Let the probability of committing a type I error be .10.

For each of the studies described in Exercises 33 through 38, answer as many of the following questions as possible: (a) What is the variable of interest? (b) Is the parameter of interest a mean, the difference between two means (independent samples), a mean difference (paired data), a proportion, or the difference between two proportions (independent samples)? (c) What is the sampled population? (d) What is the target population? (e) What are the null and alternative hypotheses? (f) Is the alternative one-sided (left tail), one-sided (right tail), or two-sided? (g) What type I and type II errors are possible? (h) Do you think the null hypothesis was rejected? Explain why or why not.

33. During a one-year period, Hong et al. (A-40) studied all patients who presented to the surgical service with possible appendicitis. One hundred eighty-two patients with possible appendicitis were randomized to either clinical assessment (CA) alone or clinical evaluation and abdominal/pelvic CT. A true-positive case resulted in a laparotomy that revealed a lesion requiring operation. A true-negative case did not require an operation at one-week follow-up evaluation. At the close of the study, they found no significant difference in the hospital length of stay for the two treatment groups.

34. Recall the study reported in Exercise 7.8.2 in which Nagy et al. (A-32) studied 50 stable patients admitted for a gunshot wound that traversed the mediastinum. They found that eight of the subjects had a mediastinal injury, while 42 did not have such an injury. They performed a student's t test to determine if there was a difference in mean age (years) between the two groups. The reported p value was .59.

35. Dykstra et al. (A-41) studied 15 female patients with urinary frequency with or without incontinence. The women were treated with botulinum toxin type B (BTX-B). A t test of the

pre/post-difference in frequency indicated that these 15 patients experienced an average of 5.27 fewer frequency episodes per day after treatment with BTX-B. The p value for the test was less than 0.001.

36. Recall the study reported in Exercise 6.10.2 in which Horesh et al. (A-42) investigated suicidal behavior among adolescents. In addition to impulsivity, the researchers studied hopelessness among the 33 subjects in the suicidal group and the 32 subjects in the nonsuicidal group. The means for the two groups on the Beck Hopelessness Scale were 11.6 and 5.2, respectively, and the t value for the test was 5.13.

37. Mauksch et Al. (A-43) surveyed 500 consecutive patients (ages 18 to 64 years) in a primary care clinic serving only uninsured, low-income patients. They used self-report questions about why patients were coming to the clinic, and other tools to classify subjects as either having or not having major mental illness. Compared with patients without current major mental illness, patients with a current major mental illness reported significantly ($p < .001$) more concerns, chronic illnesses, stressors, forms of maltreatment, and physical symptoms.

38. A study by Hosking et al. (A-44) was designed to compare the effects of alendronate and risedronate on bone mineral density (BMD). One of the outcome measures was the percent increase in BMD at 12 months. Alendronate produced a significantly higher percent change (4.8 percent) in BMD than risedronate (2.8 percent) with a p value $< .001$.

39. For each of the following situations, identify the type I and type II errors and the correct actions.

(a) H_0: A new treatment is not more effective than the traditional one.
- **(1)** Adopt the new treatment when the new one is more effective.
- **(2)** Continue with the traditional treatment when the new one is more effective.
- **(3)** Continue with the traditional treatment when the new one is not more effective.
- **(4)** Adopt the new treatment when the new one is not more effective.

(b) H_0: A new physical therapy procedure is satisfactory.
- **(1)** Employ a new procedure when it is unsatisfactory.
- **(2)** Do not employ a new procedure when it is unsatisfactory.
- **(3)** Do not employ a new procedure when it is satisfactory.
- **(4)** Employ a new procedure when it is satisfactory.

(c) H_0: A production run of a drug is of satisfactory quality.
- **(1)** Reject a run of satisfactory quality.
- **(2)** Accept a run of satisfactory quality.
- **(3)** Reject a run of unsatisfactory quality.
- **(4)** Accept a run of unsatisfactory quality.

For each of the studies described in Exercises 40 through 55, do the following:

(a) Perform a statistical analysis of the data (including hypothesis testing and confidence interval construction) that you think would yield useful information for the researchers.

(b) State all assumptions that are necessary to validate your analysis.

(c) Find p values for all computed test statistics.

(d) Describe the population(s) about which you think inferences based on your analysis would be applicable.

40. A study by Bell (A-45) investigated the hypothesis that alteration of the vitamin D–endocrine system in blacks results from reduction in serum 25-hydroxyvitamin D and that the alteration is reversed by oral treatment with 25-hydroxyvitamin D_3. The eight subjects (three men and five women) were studied while on no treatment (control) and after having been given 25-hydroxyvitamin D_3 for 7 days

($25\text{-}OHD_3$). The following are the urinary calcium (mg/d) determinations for the eight subjects under the two conditions.

Subject	Control	$25\text{-}OHD_3$
A	66	98
B	115	142
C	54	78
D	88	101
E	82	134
F	115	158
G	176	219
H	46	60

Source: Dr. Norman H. Bell. Used with permission.

41. Montner et al. (A-46) conducted studies to test the effects of glycerol-enhanced hyperhydration (GEH) on endurance in cycling performance. The 11 subjects, ages 22–40 years, regularly cycled at least 75 miles per week. The following are the pre-exercise urine output volumes (ml) following ingestion of glycerol and water:

Subject #	Experimental, ml (Glycerol)	Control, ml (Placebo)
1	1410	2375
2	610	1610
3	1170	1608
4	1140	1490
5	515	1475
6	580	1445
7	430	885
8	1140	1187
9	720	1445
10	275	890
11	875	1785

Source: Dr. Paul Montner. Used with permission.

42. D'Alessandro et al. (A-47) wished to know if preexisting airway hyperresponsiveness (HR) predisposes subjects to a more severe outcome following exposure to chlorine. Subjects were healthy volunteers between the ages of 18 and 50 years who were classified as with and without HR. The following are the FEV_1 and specific airway resistance (Sraw) measurements taken on the subjects before and after exposure to appropriately diluted chlorine gas:

	Hyperreactive Subjects			
	Pre-Exposure		Post-Exposure	
Subject	FEV_1	Sraw	FEV_1	Sraw
1	3.0	5.80	1.8	21.4
2	4.1	9.56	3.7	12.5
3	3.4	7.84	3.0	14.3
4	3.3	6.41	3.0	10.9
5	3.3	9.12	3.0	17.1

Normal Subjects				
Subject	Pre-Exposure FEV_1	Sraw	Post-Exposure FEV_1	Sraw
1	4.3	5.52	4.2	8.70
2	3.9	6.43	3.7	6.94
3	3.6	5.67	3.3	10.00
4	3.6	3.77	3.5	4.54
5	5.1	5.53	4.9	7.37

Source: Dr. Paul Blanc. Used with permission.

43. Noting the paucity of information on the effect of estrogen on platelet membrane fatty acid composition, Ranganath et al. (A-48) conducted a study to examine the possibility that changes may be present in postmenopausal women and that these may be reversible with estrogen treatment. The 31 women recruited for the study had not menstruated for at least 3 months or had symptoms of the menopause. No woman was on any form of hormone replacement therapy (HRT) at the time she was recruited. The following are the platelet membrane linoleic acid values before and after a period of HRT:

Subject	Before	After	Subject	Before	After	Subject	Before	After
1	6.06	5.34	12	7.65	5.55	23	5.04	4.74
2	6.68	6.11	13	4.57	4.25	24	7.89	7.48
3	5.22	5.79	14	5.97	5.66	25	7.98	6.24
4	5.79	5.97	15	6.07	5.66	26	6.35	5.66
5	6.26	5.93	16	6.32	5.97	27	4.85	4.26
6	6.41	6.73	17	6.12	6.52	28	6.94	5.15
7	4.23	4.39	18	6.05	5.70	29	6.54	5.30
8	4.61	4.20	19	6.31	3.58	30	4.83	5.58
9	6.79	5.97	20	4.44	4.52	31	4.71	4.10
10	6.16	6.00	21	5.51	4.93			
11	6.41	5.35	22	8.48	8.80			

Source: Dr. L. Ranganath. Used with permission.

44. The purpose of a study by Goran et al. (A-49) was to examine the accuracy of some widely used body-composition techniques for children through the use of the dual-energy X-ray absorptiometry (DXA) technique. Subjects were children between the ages of 4 and 10 years. The following are fat mass measurements taken on the children by three techniques—DXA, skinfold thickness (ST), and bioelectrical resistance (BR):

DXA	ST	BR	Sex (1 = Male, 0 = Female)
3.6483	4.5525	4.2636	1
2.9174	2.8234	6.0888	0
7.5302	3.8888	5.1175	0
6.2417	5.4915	8.0412	0
10.5891	10.4554	14.1576	0
9.5756	11.1779	12.4004	0

(*Continued*)

DXA	ST	BR	Sex (1 = Male, 0 = Female)
2.4424	3.5168	3.7389	1
3.5639	5.8266	4.3359	1
1.2270	2.2467	2.7144	1
2.2632	2.4499	2.4912	1
2.4607	3.1578	1.2400	1
4.0867	5.5272	6.8943	0
4.1850	4.0018	3.0936	1
2.7739	5.1745	*	1
4.4748	3.6897	4.2761	0
4.2329	4.6807	5.2242	0
2.9496	4.4187	4.9795	0
2.9027	3.8341	4.9630	0
5.4831	4.8781	5.4468	0
3.6152	4.1334	4.1018	1
5.3343	3.6211	4.3097	0
3.2341	2.0924	2.5711	1
5.4779	5.3890	5.8418	0
4.6087	4.1792	3.9818	0
2.8191	2.1216	1.5406	1
4.1659	4.5373	5.1724	1
3.7384	2.5182	4.6520	1
4.8984	4.8076	6.5432	1
3.9136	3.0082	3.2363	1
12.1196	13.9266	16.3243	1
15.4519	15.9078	18.0300	0
20.0434	19.5560	21.7365	0
9.5300	8.5864	4.7322	1
2.7244	2.8653	2.7251	1
3.8981	5.1352	5.2420	0
4.9271	8.0535	6.0338	0
3.5753	4.6209	5.6038	1
6.7783	6.5755	6.6942	1
3.2663	4.0034	3.2876	0
1.5457	2.4742	3.6931	0
2.1423	2.1845	2.4433	1
4.1894	3.0594	3.0203	1
1.9863	2.5045	3.2229	1
3.3916	3.1226	3.3839	1
2.3143	2.7677	3.7693	1
1.9062	3.1355	12.4938	1
3.7744	4.0693	5.9229	1
2.3502	2.7872	4.3192	0
4.6797	4.4804	6.2469	0
4.7260	5.4851	7.2809	0
4.2749	4.4954	6.6952	0
2.6462	3.2102	3.8791	0

(*Continued*)

DXA	ST	BR	Sex (1 = Male, 0 = Female)
2.7043	3.0178	5.6841	0
4.6148	4.0118	5.1399	0
3.0896	3.2852	4.4280	0
5.0533	5.6011	4.3556	0
6.8461	7.4328	8.6565	1
11.0554	13.0693	11.7701	1
4.4630	4.0056	7.0398	0
2.4846	3.5805	3.6149	0
7.4703	5.5016	9.5402	0
8.5020	6.3584	9.6492	0
6.6542	6.8948	9.3396	1
4.3528	4.1296	6.9323	0
3.6312	3.8990	4.2405	1
4.5863	5.1113	4.0359	1
2.2948	2.6349	3.8080	1
3.6204	3.7307	4.1255	1
2.3042	3.5027	3.4347	1
4.3425	3.7523	4.3001	1
4.0726	3.0877	5.2256	0
1.7928	2.8417	3.8734	1
4.1428	3.6814	2.9502	1
5.5146	5.2222	6.0072	0
3.2124	2.7632	3.4809	1
5.1687	5.0174	3.7219	1
3.9615	4.5117	2.7698	1
3.6698	4.9751	1.8274	1
4.3493	7.3525	4.8862	0
2.9417	3.6390	3.4951	1
5.0380	4.9351	5.6038	0
7.9095	9.5907	8.5024	0
1.7822	3.0487	3.0028	1
3.4623	3.3281	2.8628	1
11.4204	14.9164	10.7378	1
1.2216	2.2942	2.6263	1
2.9375	3.3124	3.3728	1
4.6931	5.4706	5.1432	0
8.1227	7.7552	7.7401	0
10.0142	8.9838	11.2360	0
2.5598	2.8520	4.5943	0
3.7669	3.7342	4.7384	0
4.2059	2.6356	4.0405	0
6.7340	6.6878	8.1053	0
3.5071	3.4947	4.4126	1
2.2483	2.8100	3.6705	0
7.1891	5.4414	6.6332	0
6.4390	3.9532	5.1693	0

* Missing data.

Source: Dr. Michael I. Goran. Used with permission.

45. Hartard et al. (A-50) conducted a study to determine whether a certain training regimen can counteract bone density loss in women with postmenopausal osteopenia. The following are strength measurements for five muscle groups taken on 15 subjects before (B) and after (A) 6 months of training:

	Leg Press		Hip Flexor		Hip Extensor	
Subject	**(B)**	**(A)**	**(B)**	**(A)**	**(B)**	**(A)**
1	100	180	8	15	10	20
2	155	195	10	20	12	25
3	115	150	8	13	12	19
4	130	170	10	14	12	20
5	120	150	7	12	12	15
6	60	140	5	12	8	16
7	60	100	4	6	6	9
8	140	215	12	18	14	24
9	110	150	10	13	12	19
10	95	120	6	8	8	14
11	110	130	10	12	10	14
12	150	220	10	13	15	29
13	120	140	9	20	14	25
14	100	150	9	10	15	29
15	110	130	6	9	8	12

	Arm Abductor		Arm Adductor	
Subject	**(B)**	**(A)**	**(B)**	**(A)**
1	10	12	12	19
2	7	20	10	20
3	8	14	8	14
4	8	15	6	16
5	8	13	9	13
6	5	13	6	13
7	4	8	4	8
8	12	15	14	19
9	10	14	8	14
10	6	9	6	10
11	8	11	8	12
12	8	14	13	15
13	8	19	11	18
14	4	7	10	22
15	4	8	8	12

Source: Dr. Manfred Hartard. Used with permission.

46. Vitacca et al. (A-51) conducted a study to determine whether the supine position or sitting position worsens static, forced expiratory flows and measurements of lung mechanics. Subjects were aged

persons living in a nursing home who were clinically stable and without clinical evidence of cardiorespiratory diseases. Among the data collected were the following FEV_1 percent values for subjects in sitting and supine postures:

Sitting	Supine	Sitting	Supine
64	56	103	94
44	37	109	92
44	39	−99	−99
40	43	169	165
32	32	73	66
70	61	95	94
82	58	−99	−99
74	48	73	58
91	63		

Source: Dr. M. Vitacca. Used with permission.

47. The purpose of an investigation by Young et al. (A-52) was to examine the efficacy and safety of a particular suburethral sling. Subjects were women experiencing stress incontinence who also met other criteria. Among the data collected were the following pre- and postoperative cystometric capacity (ml) values:

Pre	Post	Pre	Post	Pre	Post	Pre	Post
350	321	340	320	595	557	475	344
700	483	310	336	315	221	427	277
356	336	361	333	363	291	405	514
362	447	339	280	305	310	312	402
361	214	527	492	200	220	385	282
304	285	245	330	270	315	274	317
675	480	313	310	300	230	340	323
367	330	241	230	792	575	524	383
387	325	313	298	275	140	301	279
535	325	323	349	307	192	411	383
328	250	438	345	312	217	250	285
557	410	497	300	375	462	600	618
569	603	302	335	440	414	393	355
260	178	471	630	300	250	232	252
320	362	540	400	379	335	332	331
405	235	275	278	682	339	451	400
351	310	557	381				

Source: Dr. Stephen B. Young. Used with permission.

48. Diamond et al. (A-53) wished to know if cognitive screening should be used to help select appropriate candidates for comprehensive inpatient rehabilitation. They studied a sample of geriatric rehabilitation patients using standardized measurement strategies. Among the data collected were the following admission and discharge scores made by the subjects on the Mini Mental State Examination (MMSE):

Admission	Discharge	Admission	Discharge
9	10	24	26
11	11	24	30
14	19	24	28
15	15	25	26
16	17	25	22
16	15	26	26
16	17	26	28
16	17	26	26
17	14	27	28
17	18	27	28
17	21	27	27
18	21	27	27
18	21	27	27
19	21	28	28
19	25	28	29
19	21	28	29
19	22	28	29
19	19	29	28
20	22	29	28
21	23	29	30
22	22	29	30
22	19	29	30
22	26	29	30
23	21	29	30
24	21	30	30
24	20		

Source: Dr. Stephen N. Macciocchi. Used with permission.

49. In a study to explore the possibility of hormonal alteration in asthma, Weinstein et al. (A-54) collected data on 22 postmenopausal women with asthma and 22 age-matched, postmenopausal, women without asthma. The following are the dehydroepiandrosterone sulfate (DHEAS) values collected by the investigators:

Without Asthma	With Asthma	Without Asthma	With Asthma
20.59	87.50	15.90	166.02
37.81	111.52	49.77	129.01
76.95	143.75	25.86	31.02
77.54	25.16	55.27	47.66
19.30	68.16	33.83	171.88
35.00	136.13	56.45	241.88
146.09	89.26	19.91	235.16
166.02	96.88	24.92	25.16
96.58	144.34	76.37	78.71
24.57	97.46	6.64	111.52
53.52	82.81	115.04	54.69

Source: Dr. Robert E. Weinstein. Used with permission.

50. The motivation for a study by Gruber et al. (A-55) was a desire to find a potentially useful serum marker in rheumatoid arthritis (RA) that reflects underlying pathogenic mechanisms. They measured, among other variables, the circulating levels of gelatinase B in the serum and synovial fluid (SF) of patients with RA and of control subjects. The results were as follows:

Serum		Synovial Fluid		Serum		Synovial Fluid	
RA	**Control**	**RA**	**Control**	**RA**	**Control**	**RA**	**Control**
26.8	23.4	71.8	3.0	36.7			
19.1	30.5	29.4	4.0	57.2			
249.6	10.3	185.0	3.9	71.3			
53.6	8.0	114.0	6.9	25.2			
66.1	7.3	69.6	9.6	46.7			
52.6	10.1	52.3	22.1	30.9			
14.5	17.3	113.1	13.4	27.5			
22.7	24.4	104.7	13.3	17.2			
43.5	19.7	60.7		10.3			
25.4	8.4	116.8		7.5			
29.8	20.4	84.9		31.6			
27.6	16.3	215.4		30.0			
106.1	16.5	33.6		42.0			
76.5	22.2	158.3		20.3			

Source: Dr. Darius Sorbi. Used with permission.

51. Benini et al. (A-56) conducted a study to evaluate the severity of esophageal acidification in achalasia following successful dilatation of the cardias and to determine which factors are associated with pathological esophageal acidification in such patients. Twenty-two subjects, of whom seven were males; ranged in ages from 28 to 78 years. On the basis of established criteria they were classified as refluxers or nonrefluxers. The following are the acid clearance values (min/reflux) for the 22 subjects:

Refluxers	Nonrefluxers
8.9	2.3
30.0	0.2
23.0	0.9
6.2	8.3
11.5	0.0
	0.9
	0.4
	2.0
	0.7
	3.6
	0.5
	1.4
	0.2
	0.7
	17.9
	2.1
	0.0

Source: Dr. Luigi Benini. Used with permission.

52. The objective of a study by Baker et al. (A-57) was to determine whether medical deformation alters in vitro effects of plasma from patients with preeclampsia on endothelial cell function to produce a paradigm similar to the in vivo disease state. Subjects were 24 nulliparous pregnant women before delivery, of whom 12 had preeclampsia and 12 were normal pregnant patients. Among the data collected were the following gestational ages (weeks) at delivery:

Preeclampsia	Normal Pregnant
38	40
32	41
42	38
30	40
38	40
35	39
32	39
38	41
39	41
29	40
29	40
32	40

Source: Dr. James M. Roberts. Used with permission.

53. Zisselman et al. (A-58) conducted a study to assess benzodiazepine use and the treatment of depression before admission to an inpatient geriatric psychiatry unit in a sample of elderly patients. Among the data collected were the following behavior disorder scores on 27 patients treated with benzodiazepines (W) and 28 who were not (WO).

W		WO	
.00	1.00	.00	.00
.00	1.00	.00	10.00
.00	.00	.00	.00
.00	.00	.00	18.00
.00	10.00	.00	.00
.00	2.00	.00	2.00
.00		.00	5.00
.00		.00	
.00		4.00	
.00		1.00	
4.00		2.00	
3.00		.00	
2.00		6.00	
.00		.00	
10.00		.00	
2.00		1.00	
.00		2.00	
9.00		1.00	
.00		22.00	
1.00		.00	
16.00		.00	

Source: Dr. Yochi Shmuely. Used with permission.

54. The objective of a study by Reinecke et al. (A-59) was to investigate the functional activity and expression of the sarcolemmal Na^+/Ca^{2+} exchange in the failing human heart. The researchers obtained left ventricular samples from failing human hearts of 11 male patients (mean age 51 years) undergoing cardiac transplantation. Nonfailing control hearts were obtained from organ donors (four females, two males, mean age 41 years) whose hearts could not be transplanted for noncardiac reasons. The following are the Na^+/Ca^{2+} exchanger activity measurements for the patients with end-stage heart failure (CHF) and nonfailing controls (NF).

NF	CHF
0.075	0.221
0.073	0.231
0.167	0.145
0.085	0.112
0.110	0.170
0.083	0.207
	0.112
	0.291
	0.164
	0.195
	0.185

Source: Dr. Hans Reinecke. Used with permission.

55. Reichman et al. (A-60) conducted a study with the purpose of demonstrating that negative symptoms are prominent in patients with Alzheimer's disease and are distinct from depression. The following are scores made on the Scale for the Assessment of Negative Symptoms in Alzheimer's Disease by patients with Alzheimer's disease (PT) and normal elderly, cognitively intact, comparison subjects (C).

PT	C
19	6
5	5
36	10
22	1
1	1
18	0
24	5
17	5
7	4
19	6
5	6
2	7
14	5
9	3
34	5
13	12

(*Continued*)

PT	C
0	0
21	5
30	1
43	2
19	3
31	19
21	3
41	5
	24
	3

Source: Dr. Andrew C. Coyne. Used with permission.

Exercises for Use with Large Data Sets Available on the Following Website: www.wiley.com/college/daniel

1. Refer to the creatine phosphokinase data on 1005 subjects (PCKDATA). Researchers would like to know if psychologically stressful situations cause an increase in serum creatine phosphokinase (CPK) levels among apparently healthy individuals. To help the researchers reach a decision, select a simple random sample from this population, perform an appropriate analysis of the sample data, and give a narrative report of your findings and conclusions. Compare your results with those of your classmates.

2. Refer to the prothrombin time data on 1000 infants (PROTHROM). Select a simple random sample of size 16 from each of these populations and conduct an appropriate hypothesis test to determine whether one should conclude that the two populations differ with respect to mean prothrombin time. Let $\alpha = .05$. Compare your results with those of your classmates. What assumptions are necessary for the validity of the test?

3. Refer to the head circumference data of 1000 matched subjects (HEADCIRC). Select a simple random sample of size 20 from the population and perform an appropriate hypothesis test to determine if one can conclude that subjects with the sex chromosome abnormality tend to have smaller heads than normal subjects. Let $\alpha = .05$. Construct a 95 percent confidence interval for the population mean difference. What assumptions are necessary? Compare your results with those of your classmates.

4. Refer to the hemoglobin data on 500 children with iron deficiency anemia and 500 apparently healthy children (HEMOGLOB). Select a simple random sample of size 16 from population A and an independent simple random sample of size 16 from population B. Does your sample data provide sufficient evidence to indicate that the two populations differ with respect to mean Hb value? Let $\alpha = .05$. What assumptions are necessary for your procedure to be valid? Compare your results with those of your classmates.

5. Refer to the manual dexterity scores of 500 children with learning disabilities and 500 children with no known learning disabilities (MANDEXT). Select a simple random sample of size 10 from population A and an independent simple random sample of size 15 from population B. Do your samples provide sufficient evidence for you to conclude that learning-disabled children, on the average, have lower manual dexterity scores than children without a learning disability? Let $\alpha = .05$. What assumptions are necessary in order for your procedure to be valid? Compare your results with those of your classmates.