
ONE-SAMPLE HYPOTHESES

This chapter will continue the discussion of Section 6.4 on how to draw inferences about population parameters by testing hypotheses about them using appropriate sample estimates. It will consider hypotheses about each of several population parameters: population mean, median, variance (or standard deviation), and coefficient of variation. The chapter will also introduce procedures for expressing the confidence one can have in estimating parameters from sample statistics.

7.1 TWO-TAILED HYPOTHESES CONCERNING THE MEAN

Section 6.4 introduced the concept of statistical testing using a pair of statistical hypotheses, the null and alternate hypotheses, as statements that a population mean (μ) is equal to some specified value (let's call it μ_0):

$$H_0: \mu = \mu_0;$$

$$H_A: \mu \neq \mu_0.$$

For example, let us consider the body temperatures of twenty-five intertidal crabs that we exposed to air at 24.3°C (Example 7.1). We may wish to ask whether the mean body temperature of members of this species of crab is the same as the ambient air temperature of 24.3°C. Therefore,

$$H_0: \mu = 24.3^\circ\text{C}, \text{ and}$$

$$H_A: \mu \neq 24.3^\circ\text{C},$$

where the null hypothesis states that the mean of the population of data from which this sample of twenty-five came is 24.3°C (i.e., μ is "no different from 24.3°C"), and the alternate hypothesis is that the population mean is not equal to (i.e., μ is different from) 24.3°C.

EXAMPLE 7.1 The two-tailed t test for difference between a population mean and a hypothesized population mean.

Body temperatures (measured in $^{\circ}\text{C}$) of twenty-five intertidal crabs placed in air at 24.3°C : 25.8, 24.6, 26.1, 22.9, 25.1, 27.3, 24.0, 24.5, 23.9, 26.2, 24.3, 24.6, 23.3, 25.5, 28.1, 24.8, 23.5, 26.3, 25.4, 25.5, 23.9, 27.0, 24.8, 22.9, 25.4.

$$H_0: \mu = 24.3^{\circ}\text{C}$$

$$H_A: \mu \neq 24.3^{\circ}\text{C}$$

$$\alpha = 0.05$$

$$n = 25$$

$$\bar{X} = 25.03^{\circ}\text{C}$$

$$s^2 = 1.80(^{\circ}\text{C})^2$$

$$s_{\bar{X}} = \sqrt{\frac{1.80(^{\circ}\text{C})^2}{25}} = 0.27^{\circ}\text{C}$$

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}} = \frac{25.03^{\circ}\text{C} - 24.3^{\circ}\text{C}}{0.27^{\circ}\text{C}} = 2.704$$

$$\nu = 24$$

$$t_{0.05(2), 24} = 2.064$$

As $|t| > t_{0.05(2), 24}$, reject H_0 and conclude that the sample of twenty-five body temperatures came from a population whose mean is not 24.3°C .

$$0.01 < P < 0.02 \quad [P = 0.012]$$

In Section 6.2 (Equation 6.16), $Z = (\bar{X} - \mu) / \sigma_{\bar{X}}$ was introduced as a normal deviate, and it was shown how one can determine the probability of obtaining a sample with mean \bar{X} from a population with a specified mean μ . And Section 6.4 discussed how the normal deviate can be used to test hypotheses about a population mean. Note, however, that the calculation of Z requires the knowledge of $\sigma_{\bar{X}}$, which we typically do not have. The best we can do is to calculate $s_{\bar{X}}$ as an estimate of $\sigma_{\bar{X}}$. If n is very, very large, then $s_{\bar{X}}$ is a good estimate of $\sigma_{\bar{X}}$, and we can calculate Z using this estimate. However, for most biological situations n is insufficiently large to consider $s_{\bar{X}}$ as an accurate estimate of $\sigma_{\bar{X}}$; but we can use, in place of the normal distribution (Z), a distribution known as t , the development of which was a major breakthrough in statistical methodology*:

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}} \quad (7.1)$$

*The t statistic is also referred to as "Student's t ." William Sealy Gosset (1876–1937), was an English statistician with the title "brewer" in the Guinness brewery of Dublin, who used the pseudonym "Student" to publish many noteworthy developments in statistical theory and practice, including ("Student," 1908) the presentation of the distribution that often bears his name. (See Irwin, 1978; Pearson, 1939; Pearson, Plackett, and Barnard, 1990.) Gosset referred to his distribution as z ; and between 1922 and 1925 R. A. Fisher (e.g., 1925a, 1925b: 106–113, 117–125; 1928) helped develop its potential in statistical testing while modifying it; Gosset called the modification " t " (Eisenhart, 1979). Although Gosset was a modest man, he is referred to as "one of the most original minds in contemporary science" by R. A. Fisher (1939a), himself one of the most insightful and influential statisticians of all time.

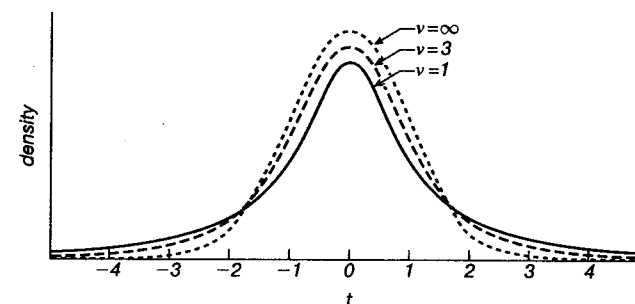


Figure 7.1 The t distribution for various degrees of freedom, ν . For $\nu = \infty$, the t distribution is identical to the normal distribution.

As do other distributions to be encountered among statistical methods, the t distribution has different shapes for different values of what is known as *degrees of freedom* (denoted by ν , the lowercase Greek nu).*

$$\nu = n - 1. \quad (7.2)$$

Recall that n is the size of the sample (i.e., the number of data from which \bar{X} has been calculated). The influence of ν on the shape of the t distribution is shown in Fig. 7.1. This distribution is leptokurtic (see Section 6.1), having a greater concentration of values around the mean and in the tails than does a normal distribution; but as n (and, therefore, ν) increases, the t distribution tends to resemble a normal distribution more closely, and for $\nu = \infty$ (i.e., for an infinitely large sample[†]), the t and normal distributions are identical; that is, $t_{\alpha, \infty} = Z_{\alpha}$.

The mean of the sample of twenty-five data shown in Example 7.1 is 25.03°C , and the sample variance is $1.80(^{\circ}\text{C})^2$. These statistics are estimates of the mean and variance of the population from which this sample came. However, this is only one of a very large number of samples of size twenty-five that could have been taken at random from the population. The distribution of the means of all possible samples with $n = 25$ is the t distribution for $\nu = 24$, which is represented by the curve of Fig. 7.2. In this figure, the mean of the t distribution (i.e., $t = 0$) represents the mean hypothesized in H_0 ; (i.e.,

*In early writings of the t distribution (during the 1920s and 1930s), the symbol n or f was used for degrees of freedom. This was often confusing because these letters had commonly been used to denote other quantities in statistics so Maurice G. Kendall (1943: 292) recommended ν .

†The modern symbol for infinity (∞) was introduced in 1655 by English mathematician John Wallis (1616–1703) (Cajori, 1929:44), who also introduced the notation of negative and fractional exponents, which his countryman, Sir Isaac Newton (1642–1727) wrote out in a form similar to modern usage (Cajori, 1928: 354–355):

$$X^{-a} = \frac{1}{X^a}; \quad X^{\frac{1}{2}} = \sqrt{X}.$$

The notion of fractional powers was conceived by French writer Nicole Oresme (ca. 1323–1382), a bishop in Normandy (Cajori, 1928: 91).

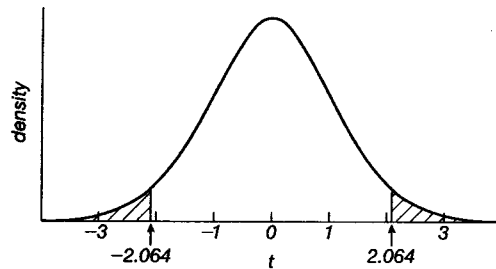


Figure 7.2 The t distribution for $\nu = 24$, showing the critical region (shaded area) for a two-tailed test using $\alpha = 0.05$. (The critical value of t is 2.064.)

$\mu = \mu_0 = 24.3^\circ\text{C}$), for, by Equation 7.1, $t = 0$ when $\bar{X} = \mu$. The shaded areas in this figure represent the extreme 5% of the total area under the curve (2.5% in each tail). Thus, an \bar{X} so far from μ that it lies in either of the shaded areas has a probability of less than 5% of occurring by chance alone, and we assume that it occurred because H_0 is, in fact, false. As explained in Section 6.4 regarding the Z distribution, because an extreme t value in either direction from μ will cause us to reject H_0 , we are said to be considering a “two-tailed” (or “two-sided”) test.

For $\nu = 24$, we can consult Appendix Table B.3 to find the following two-tailed probabilities (denoted as “ $\alpha(2)$ ”) of various values of t :

ν	$\alpha(2)$	0.50	0.20	0.10	0.05	0.02	0.01
24		0.685	1.318	1.711	2.064	2.492	2.797

Thus, for example, for a two-tailed α of 0.05, the shaded areas of the curve begin at 2.064 t units on either side of μ . Therefore, we can state:

$$P(|t| \geq 2.064) = 0.05.$$

That is, 2.064 and -2.064 are the *critical values* of t ; and if t (calculated from Equation 7.1) is equal to or greater than 2.064, or is equal to or less than -2.064 , that will be considered reasonable cause to reject H_0 and consider H_A to be a true statement. That portion of the t distribution beyond the critical values (i.e., the shaded areas in the figure) is called the *critical region*.^{*} For the sample of twenty-five body temperatures (see Example 7.1), $t = 2.704$. As 2.704 lies within the critical region (i.e., $2.704 > 2.064$), H_0 is rejected, and we conclude that the mean body temperature of crabs under the conditions of our experiment is not 24.3°C .

To summarize, the hypotheses for the two-tailed test are

$$H_0: \mu = \mu_0 \quad \text{and} \quad H_A: \mu \neq \mu_0,$$

where μ_0 denotes the hypothesized value to which we are comparing the population mean. (In the above example, $\mu_0 = 24.3^\circ\text{C}$.) The test statistic is calculated by Equation 7.1, and if its absolute value is larger than the two-tailed critical value of t from Appendix Table B.3, we reject H_0 and assume H_A to be true. The critical value of t can be

^{*}David (1995) traces the first use of this term to J. Neyman and E. S. Pearson in 1933.

abbreviated as $t_{\alpha(2), \nu}$, where $\alpha(2)$ refers to the two-tailed probability of α . Thus, for the preceding example, we could write $t_{0.02(2), 24} = 2.064$. In general, for a two-tailed t test,

$$\text{if } |t| \geq t_{\alpha(2), \nu}, \text{ then reject } H_0.$$

Example 7.1 presents the computations for the analysis of the crab data. A t of 2.704 is calculated, which for 24 degrees of freedom lies between the tabled critical values of $t_{0.02(2), 24} = 2.492$ and $t_{0.01(2), 24} = 2.797$. Therefore, if the null hypothesis, H_0 , is a true statement about the population we sampled, the probability of \bar{X} being at least this far from μ is between 0.01 and 0.02; that is $0.01 < P(|t| \geq 2.704) < 0.02$.^{*} As this probability is less than 0.05, we reject H_0 and declare it is not a true statement. For a consideration of the types of errors involved in rejecting or accepting the null hypothesis, refer to Section 6.4.

Frequently, the hypothesized value in the null and alternate hypotheses is zero. For example, the weights of twelve rats might be measured before and after the animals are placed on a regimen of forced exercise for one week. The change in weight of the animals (i.e., weight after minus weight before) could be recorded, and it might have been found that the mean weight change was -0.65 g (i.e., the mean weight change is a 0.65 g weight loss). If we wished to infer whether such exercise causes any significant change in rat weight, we could state $H_0: \mu = 0$ and $H_A: \mu \neq 0$; Example 7.2 summarizes the t test for this H_0 and H_A . This test is two tailed, for a large $\bar{X} - \mu$ difference in either direction will constitute grounds for rejecting the veracity of H_0 .[†]

The theoretical basis of the t testing utilized throughout this chapter assumes that sample data came from a normal population, assuring that the mean at hand came from a normal distribution of means. Fortunately, the t test is *robust*,[‡] meaning that its validity is not seriously affected by moderate deviations from this underlying assumption. The test also assumes—as other statistical tests typically do—that the data are a random sample (see Section 2.3).

A common situation in which one is dealing with a population known to be non-normal is the case where the data are percentages or proportions. Such data are known to be binomial, rather than normal, and the treatment of such data is discussed in Section 13.3.

The effect of nonnormality is greater for smaller α , and the effect decreases as n increases (Ractliffe, 1968). For symmetric distributions there is little effect of departure from normality (i.e., from mesokurtosis); for asymmetric distributions the test performs best with strong leptokurtosis present and poorly with platykurtosis and mesokurtosis; and the adverse effect of non-normality is much less for two-tailed testing than for one-tailed (Section 7.2) testing (Cicchitelli, 1989).

^{*}Some calculators and computer programs have the capability of determining the probability of a given t (e.g., sec. Boomsma and Molenaar, 1994; Guenther, 1977). For the present example, we would thereby find that $P(|t| \geq 2.704) = 0.012$.

[†]Data that result from the differences between pairs of data (such as measurements before and after an experimental treatment) are discussed further in Chapter 9.

[‡]The term “robustness” was introduced by Box in 1953 (David, 1995).

EXAMPLE 7.2 A two-tailed test for significant difference between a population mean and a hypothesized population mean of zero.

Weight change of twelve rats after being subjected to a regimen of forced exercise. Each weight change (in g) is the weight after exercise minus the weight before.

$$\begin{array}{ll}
 1.7 & H_0: \mu = 0 \\
 0.7 & H_A: \mu \neq 0 \\
 -0.4 & \\
 -1.8 & \alpha = 0.05 \\
 0.2 & n = 12 \\
 0.9 & \bar{X} = -0.65 \text{ g} \\
 -1.2 & s^2 = 1.5682 \text{ g}^2 \\
 -0.9 & \\
 -1.8 & s_{\bar{X}} = \sqrt{\frac{1.5682 \text{ g}^2}{12}} = 0.36 \text{ g} \\
 -1.4 & \\
 -1.8 & t = \frac{\bar{X} - \mu}{s_{\bar{X}}} = \frac{-0.65 \text{ g}}{0.36 \text{ g}} = -1.81 \\
 -2.0 & \\
 & \nu = n - 1 = 11 \\
 & t_{0.05(2), 11} = 2.201 \\
 & \text{Since } |t| < t_{0.05(2), 11}, \text{ do not reject } H_0. \\
 & 0.05 < P < 0.10 \quad [P = 0.098]
 \end{array}$$

It is important to appreciate that a sample used in statistical testing such as that discussed here must consist of truly replicated data. In Example 7.1, we desired to draw conclusions about a population of measurements representing a large number of animals (i.e., crabs). Therefore, the sample must consist of measurements (i.e., body temperatures) from n (i.e., twenty-five) animals; it would *not* be valid to obtain twenty-five body temperatures from a single animal. And, in Example 7.2, twelve individual rats must be used; it would *not* be valid to employ data obtained from subjecting the same animal to the experiment twelve times. Such invalid attempts at replication are discussed by Hurlbert (1984), who calls them *pseudoreplication*.

7.2 ONE-TAILED HYPOTHESES CONCERNING THE MEAN

In Section 7.1, we spoke of the hypotheses $H_0: \mu = \mu_0$ and $H_A: \mu \neq \mu_0$, because we were willing to consider a large deviation of \bar{X} in either direction from μ_0 as grounds for rejecting H_0 . However, in many instances, our interest lies only in whether \bar{X} is significantly larger (or significantly smaller) than μ_0 , and this is termed a "one-tailed" (or "one-sided") test situation. For example, one might be testing a drug hypothesized to cause weight reduction in humans. The investigator is interested only in whether a weight *loss* occurs after the drug is taken. (In Example 7.2, using a two-sided test, we

EXAMPLE 7.3 A one-tailed t test for the hypotheses $H_0: \mu \geq 0$ and $H_A: \mu < 0$.

The data are weight changes of humans, tabulated after administration of a drug proposed to result in weight loss. Each weight change (in kg) is the weight after minus the weight before drug administration.

$$\begin{array}{ll}
 0.2 & n = 12 \\
 -0.5 & \bar{X} = -0.61 \text{ kg} \\
 -1.3 & s^2 = 0.4008 \text{ kg}^2 \\
 -1.6 & \\
 -0.7 & s_{\bar{X}} = \sqrt{\frac{0.4008 \text{ kg}^2}{12}} = 0.18 \text{ kg} \\
 0.4 & \\
 -0.1 & t = \frac{\bar{X} - \mu}{s_{\bar{X}}} = \frac{-0.61 \text{ kg}}{0.18 \text{ kg}} = -3.389 \\
 0.0 & \\
 -0.6 & \nu = n - 1 = 11 \\
 -1.1 & t_{0.05(1), 11} = 1.796. \\
 -1.2 & \text{If } t \leq -t_{0.05(1), 11}, \text{ reject } H_0. \\
 -0.8 & \text{Conclusion: reject } H_0. \\
 & 0.0025 < P(t \leq -3.389) < 0.005 \quad [P = 0.0030]
 \end{array}$$

were interested in determining whether either weight loss or weight gain had occurred.) If there is either weight gain or no weight change, the drug will be considered a failure. Therefore, for this one-sided test, we should state $H_0: \mu \geq 0$ and $H_A: \mu < 0$. Here, the null hypothesis states that there is no mean weight loss (i.e., the mean weight change is greater than or equal to zero), and the alternate hypothesis states that there is a mean weight loss (i.e., the mean weight change is less than zero). By examining the alternate hypothesis, H_A , we see that H_0 will be rejected if t is in the left-hand critical region of the t distribution. In general,

$$\text{for } H_A: \mu < \mu_0,$$

$$\text{if } t \leq -t_{\alpha(1), \nu}, \text{ then reject } H_0.*$$

Example 7.3, summarizes such a set of twelve weight change data tested against this pair of hypotheses. From Appendix Table B.3 we find that $t_{0.05(1), 11} = 1.796$, and the critical region for this test is shown in Fig. 7.3. From this figure, and by examining Appendix Table B.3, we see that $t_{\alpha(1), \nu} = t_{2\alpha(2), \nu}$ or $t_{\alpha(2), \nu} = t_{\alpha/2(1), \nu}$; that is, for example, the critical value of t for one-sided test at $\alpha = 0.05$ is the same as the critical value of t for a two-sided test at $\alpha = 0.10$.

If we are interested in whether \bar{X} is significantly *greater* than some value, μ_0 , the hypotheses for the one-tailed test are $H_0: \mu \leq \mu_0$ and $H_A: \mu > \mu_0$. For example, a drug manufacturer might advertise that a product dissolves completely in gastric juice

*For one-tailed testing of this H_0 , probabilities of t up to 0.25 are indicated in Appendix Table B.3. If $t = 0$, then $P = 0.50$; so if $-t_{0.25(1), \nu} < t < 0$, then $0.25 < P < 0.50$; and if $t > 0$, then $P > 0.50$.

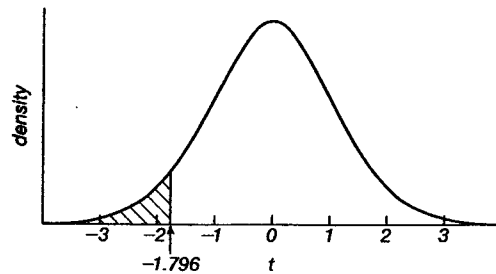


Figure 7.3 The distribution of t for $\nu = 11$, showing the critical region (shaded area) for a one-tailed test using $\alpha = 0.05$. (The critical value of t is -1.796 .)

within 45 sec. The hypotheses appropriate for testing this claim are $H_0: \mu \leq 45$ sec and $H_A: \mu > 45$ sec, because we are not particularly interested in the possibility that the product dissolves faster than is claimed, but we wish to determine whether its dissolving time is longer than advertised. Thus, the rejection region would be in the right-hand tail, rather than in the left-hand tail (the latter being the case in Example 7.3). The details of such a test are shown in Example 7.4. In general,

for $H_A: \mu > \mu_0$,

if $t \geq t_{\alpha(1), \nu}$, then reject H_0 .*

EXAMPLE 7.4 The one-tailed t test for the hypotheses $H_0: \mu \leq 45$ sec and $H_A: \mu > 45$ sec.

Dissolving times (in sec) of a drug in gastric juice: 42.7, 43.4, 44.6, 45.1, 45.6, 45.9, 46.8, 47.6.

$$H_0: \mu \leq 45 \text{ sec}$$

$$H_A: \mu > 45 \text{ sec}$$

$$\alpha = 0.05$$

$$n = 8$$

$$\bar{X} = 45.21 \text{ sec}$$

$$SS = 18.8288 \text{ sec}^2$$

$$s^2 = 2.6898 \text{ sec}^2$$

$$s_{\bar{X}} = 0.58 \text{ sec}$$

$$t = \frac{45.21 \text{ sec} - 45 \text{ sec}}{0.58 \text{ sec}} = 0.36$$

$$\nu = 7$$

$$t_{0.05(1), 7} = 1.895$$

If $t \geq t_{0.05(1), 7}$, reject H_0 .

Conclusion: do not reject H_0 .

$$P(t \geq 0.36) > 0.25 \quad [P = 0.36]$$

7.3 CONFIDENCE LIMITS FOR THE POPULATION MEAN

We learned in Section 7.1 that 5% of all possible sample means from a population with mean μ will yield t values—where $t = (\bar{X} - \mu)/s_{\bar{X}}$ —that are either larger than $t_{0.05(2), \nu}$, or smaller than $-t_{0.05(2), \nu}$ (i.e., $|t| > t_{0.05(2), \nu}$). This means that 95% of all t values

*For this H_0 , if $t = 0$, then $P = 0.50$; therefore, if $0 < t < t_{0.25(1), \nu}$, then $0.25 < P < 0.50$, and if $t < 0$, then $P > 0.50$.

TWO-SAMPLE HYPOTHESES

Among the most commonly employed biostatistical procedures is the comparison of two samples to infer whether differences exist between the two populations sampled. This chapter will consider hypotheses comparing two population means, medians, variances (or standard deviations), coefficients of variation, and indices of diversity. In doing so, we introduce another very important sampling distribution, the F distribution—named for its discoverer, R. A. Fisher (by Snedecor, 1934: 15)—and will demonstrate further use of Student's t distribution.

The objective of many two-sample hypotheses is to make inferences about population parameters by examining sample statistics. Other hypothesis-testing procedures, however, draw inferences about populations without referring to parameters. Such procedures are called *nonparametric* methods, and several will be discussed in this and following chapters.

8.1 TESTING FOR DIFFERENCE BETWEEN TWO MEANS

Example 8.1 presents the results of an experiment in which thirteen persons were divided at random into two groups, one group of six and one group of seven.*

The members of the first group were given one kind of drug (called "B"), and the members of the second group were given another kind of drug (called "G"). Blood is to be taken from each person and the time it takes the blood to clot is to be recorded. The

*Sir Ronald Aylmer Fisher (1890–1962) introduced the important concept of assigning subjects *at random* to groups for different experimental treatments (Bartlett, 1965).

EXAMPLE 8.1 A two-sample t test for the two-tailed hypotheses, $H_0: \mu_1 = \mu_2$ and $H_A: \mu_1 \neq \mu_2$ (which could also be stated as $H_0: \mu_1 - \mu_2 = 0$ and $H_A: \mu_1 - \mu_2 \neq 0$). The data are human blood-clotting times (in minutes) of individuals given one of two different drugs.

$$H_0: \mu_1 = \mu_2$$

$$H_A: \mu_1 \neq \mu_2$$

Given drug B	Given drug G
8.8	9.9
8.4	9.0
7.9	11.1
8.7	9.6
9.1	8.7
9.6	10.4
	9.5
$n_1 = 6$	$n_2 = 7$
$v_1 = 5$	$v_2 = 6$
$\bar{X}_1 = 8.75$ min	$\bar{X}_2 = 9.74$ min
$SS_1 = 1.6950$ min ²	$SS_2 = 4.0171$ min ²

$$s_p^2 = \frac{SS_1 + SS_2}{v_1 + v_2} = \frac{1.6950 + 4.0171}{5 + 6} = \frac{5.7121}{11} = 0.5193 \text{ min}^2$$

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \sqrt{\frac{0.5193}{6} + \frac{0.5193}{7}} = \sqrt{0.0866 + 0.0742} \\ = \sqrt{0.1608} = 0.40 \text{ min}$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}} = \frac{8.75 - 9.74}{0.40} = \frac{-0.99}{0.40} = -2.475$$

$$t_{0.05(2), v} = t_{0.05(2), 11} = 2.201$$

Therefore, reject H_0 .

$$0.02 < P(|t| \geq 2.475) < 0.05 \quad [P = 0.030]$$

two-tailed hypotheses, $H_0: \mu_1 - \mu_2 = 0$ and $H_A: \mu_1 - \mu_2 \neq 0$, can be proposed to ask whether, in the population sampled, blood of persons treated with drug B has the same mean clotting time as does blood from persons treated with drug G. These hypotheses are commonly expressed in their equivalent forms: $H_0: \mu_1 = \mu_2$ and $H_A: \mu_1 \neq \mu_2$. The data from this experiment are presented in Example 8.1.

If the two samples came from normal populations, and if the two populations have equal variances, then a t value may be calculated in a manner analogous to the t test introduced in Section 7.1. The t value for testing the preceding hypotheses concerning the difference between two population means is

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}} \quad (8.1)$$

The quantity $\bar{X}_1 - \bar{X}_2$ is simply the difference between the two means, and $s_{\bar{X}_1 - \bar{X}_2}$ is the standard error of the difference between the sample means.

The quantity $s_{\bar{X}_1 - \bar{X}_2}$, along with $s_{\bar{X}_1 - \bar{X}_2}^2$, the variance of the difference between the means, is new to us, and we need to consider it further. Both $s_{\bar{X}_1 - \bar{X}_2}^2$ and $s_{\bar{X}_1 - \bar{X}_2}$ are statistics that can be calculated from the sample data and are estimates of the population parameters, $\sigma_{\bar{X}_1 - \bar{X}_2}^2$ and $\sigma_{\bar{X}_1 - \bar{X}_2}$, respectively. It can be shown mathematically that the variance of the difference between two independent variables is equal to the sum of the variances of the two variables, so that $\sigma_{\bar{X}_1 - \bar{X}_2}^2 = \sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2$. Independence means that there is no correlation between the two variables.* As $\sigma_{\bar{X}}^2 = \sigma^2/n$, we can write

$$\sigma_{\bar{X}_1 - \bar{X}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \quad (8.2)$$

As the two-sample t test requires that we assume $\sigma_1^2 = \sigma_2^2$, we can write

$$\sigma_{\bar{X}_1 - \bar{X}_2}^2 = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}. \quad (8.3)$$

Thus, to calculate the estimate of $\sigma_{\bar{X}_1 - \bar{X}_2}^2$, we must have an estimate of σ^2 . Since both s_1^2 and s_2^2 are assumed to estimate σ^2 , we compute the *pooled variance*, s_p^2 , which is then used as the best estimate of σ^2 :

$$s_p^2 = \frac{SS_1 + SS_2}{v_1 + v_2}, \quad (8.4)$$

and

$$s_{\bar{X}_1 - \bar{X}_2}^2 = \frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}. \quad (8.5)$$

Thus,

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}, \quad (8.6)$$

and Equation 8.1 becomes

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}, \quad (8.7a)$$

which for equal sample sizes (i.e., $n_1 = n_2$, so each sample size may be referred to as n),

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{2s_p^2}{n}}}. \quad (8.7b)$$

*If there is a correlation between them, then see Chapter 9.

Example 8.1 summarizes the procedure for testing the hypotheses under consideration. The critical value to be obtained from Appendix Table B.3 is $t_{\alpha(2), (v_1 + v_2)}$, the two-tailed t value for the α significance level, with $v_1 + v_2$ degrees of freedom. We shall also write this as $t_{\alpha(2), \nu}$, defining the pooled degrees of freedom to be

$$\nu = v_1 + v_2 \quad \text{or, equivalently,} \quad \nu = n_1 + n_2 - 2. \quad (8.8)$$

In the two-tailed test, H_0 will be rejected if either $t \geq t_{\alpha(2), \nu}$ or $t \leq -t_{\alpha(2), \nu}$. Another way of stating this is that H_0 will be rejected if $|t| \geq t_{\alpha(2), \nu}$.

One-tailed hypotheses can be tested in situations where the investigator is interested in detecting a difference in only one direction. For example, a gardener may use a particular fertilizer for a particular kind of plant, and a new fertilizer is advertised as being an improvement. Let us say that plant height at maturity is an important characteristic of this kind of plant, with taller plants being preferable. An experiment was run, raising ten plants on the present fertilizer and eight on the new one, with the resultant eighteen plant heights shown in Example 8.2. If the new fertilizer produces plants that are shorter than, or the same height as, plants grown with the present fertilizer, then we shall decide that the advertising claims are unfounded; therefore, the statements of $\mu_1 > \mu_2$ and $\mu_1 = \mu_2$ belong in the same hypothesis, namely the null hypothesis, H_0 . If, however, mean plant height is indeed greater with the newer fertilizer, then it shall be declared to be distinctly better, with the alternate hypothesis ($H_A: \mu_1 < \mu_2$) concluded to be the true statement. The t statistic is calculated by Equation 8.1, just as for the two-tailed test. But this calculated t is then compared with the critical value $t_{\alpha(1), \nu}$, rather than with $t_{\alpha(2), \nu}$.

In other cases, the one-tailed hypotheses, $H_0: \mu_1 \leq \mu_2$ and $H_A: \mu_1 > \mu_2$, may be appropriate. Just as introduced in the one-sample testing of Sections 7.1, and 7.2, the following summary of procedures applies to two-sample t testing:

For $H_A: \mu_1 \neq \mu_2$, if $|t| \geq t_{\alpha(2), \nu}$, then reject H_0 .

For $H_A: \mu_1 < \mu_2$, if $t \leq -t_{\alpha(1), \nu}$, then reject H_0 .*

For $H_A: \mu_1 > \mu_2$, if $t \geq t_{\alpha(1), \nu}$, then reject H_0 .†

As indicated in Section 6.4, the null and alternate hypotheses are to be decided upon *before* the data are collected.

Note that $H_0: \mu_1 = \mu_2$ may be written $H_0: \mu_1 - \mu_2 = 0$ and $H_A: \mu_1 \neq \mu_2$ as $H_A: \mu_1 - \mu_2 \neq 0$; the generalized two-tailed hypotheses are $H_0: \mu_1 - \mu_2 = \mu_0$ and $H_A: \mu_1 - \mu_2 \neq \mu_0$, tested as

$$t = \frac{|\bar{X}_1 - \bar{X}_2| - \mu_0}{s_{\bar{X}_1 - \bar{X}_2}}, \quad (8.9)$$

where μ_0 may be any hypothesized difference between population means.

*For this one-tailed hypothesis test, probabilities of t up to 0.25 are indicated in Appendix Table B.3. If $t = 0$, then $P = 0.50$; so if $-t_{0.25(1), \nu} < t < 0$, then $0.25 < P < 0.50$; and if $t > 0$ then $P > 0.50$.

†For this one-tailed hypothesis test, $t = 0$ indicates $P = 0.50$; therefore, if $0 < t < t_{0.25(1), \nu}$, then $0.25 < P < 0.50$; and if $t < 0$, then $P > 0.50$.

EXAMPLE 8.2 A two-sample t test for the one-tailed hypotheses, $H_0: \mu_1 \geq \mu_2$ and $H_A: \mu_1 < \mu_2$ (which could also be stated as $H_0: \mu_1 - \mu_2 \geq 0$ and $H_A: \mu_1 - \mu_2 < 0$). The data are heights of plants, each grown with one of two different fertilizers.

$$H_0: \mu_1 \geq \mu_2$$

$$H_A: \mu_1 < \mu_2$$

Present fertilizer	Newer fertilizer
48.2 cm	52.3 cm
54.6	57.4
58.3	55.6
47.8	53.2
51.4	61.3
52.0	58.0
55.2	59.8
49.1	54.8
49.9	
52.6	
$n_1 = 10$	$n_2 = 8$
$v_1 = 9$	$v_2 = 7$
$\bar{X}_1 = 51.91$ cm	$\bar{X}_2 = 56.55$ cm
$SS_1 = 102.23$ cm ²	$SS_2 = 69.20$ cm ²

$$s_p^2 = \frac{102.23 + 69.20}{9 + 7} = \frac{171.43}{16} = 10.71 \text{ cm}^2$$

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{10.71}{10} + \frac{10.71}{8}} = \sqrt{2.41} = 1.55 \text{ cm}$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}} = \frac{51.91 - 56.55}{1.55} = \frac{-4.64}{1.55} = -2.99$$

$$t_{0.05(1), 16} = 1.746$$

As t of -2.99 is less than -1.746 , H_0 is rejected.

$$0.0025 < P < 0.005 \quad [P = 0.0043]$$

Also, $H_0: \mu_1 \leq \mu_2$ and $H_A: \mu_1 > \mu_2$ may be written as $H_0: \mu_1 - \mu_2 \leq 0$ and $H_A: \mu_1 - \mu_2 > 0$, respectively. The generalized hypotheses for this type of one-tailed test are $H_0: \mu_1 - \mu_2 \leq \mu_0$ and $H_A: \mu_1 - \mu_2 > \mu_0$, for which the t is

$$t = \frac{\bar{X}_1 - \bar{X}_2 - \mu_0}{s_{\bar{X}_1 - \bar{X}_2}}, \quad (8.10)$$

and μ_0 may be any specified value of $\mu_1 - \mu_2$.

Lastly, $H_0: \mu_1 \geq \mu_2$ and $H_A: \mu_1 < \mu_2$ may be written as $H_0: \mu_1 - \mu_2 \geq 0$ and $H_A: \mu_1 - \mu_2 < 0$, and the generalized one-tailed hypotheses of this type are $H_0: \mu_1 - \mu_2 \geq$

μ_0 and $H_A: \mu_1 - \mu_2 < \mu_0$, with the appropriate t statistic being that of Equation 8.10. For example, the gardener collecting the data of Example 8.2 may have decided, because the newer fertilizer is more expensive than the other, that it should be used only if the plants grown with it averaged at least 5.0 cm taller than plants grown with the present fertilizer. Then, $\mu_0 = \mu_1 - \mu_2 = -5.0$ cm and, by Equation 8.10, we would calculate $t = (51.91 - 56.55 + 5.0)/1.55 = 0.36/1.55 = 0.232$, which is not \geq the critical value shown in Example 8.2; so $H_0: \mu_1 - \mu_2 \geq -5.0$ cm is not rejected. The following summary of procedures applies to these general hypotheses:

For $H_A: \mu_1 - \mu_2 \neq \mu_0$, if $|t| \geq t_{\alpha(2), v}$, then reject H_0 .

For $H_A: \mu_1 - \mu_2 < \mu_0$, if $t \leq -t_{\alpha(1), v}$, then reject H_0 .

For $H_A: \mu_1 - \mu_2 > \mu_0$, if $t \geq t_{\alpha(1), v}$, then reject H_0 .

If two sample means are graphed along with the sample standard deviations, sample standard errors, or confidence intervals for the means (as in Figs. 7.4 and 7.5), it is tempting for some to conclude whether the means are significantly different based upon whether the measures of dispersion overlap. The efficacy of such a procedure is discussed by Andrews, Snee, and Sarner (1980), Browne (1979), and Simpson, Roe, and Lewontin (1960: 350–354).

By the procedure of Section 8.9, one can test whether the measurements in one population are a specified amount as large as those in a second population.

Martín Andrés et al. (1995) show how correlation analysis (Section 19.1) may be employed to test the difference between two means.

Violations of the Two-Sample t Test Assumptions. The two-sample t test assumes, by dint of its underlying theory, that both samples come at random from normal populations with equal variances. The biological researcher cannot, however, always be assured that these assumptions are correct. Fortunately, numerous studies have shown that the t test is robust enough to stand considerable departures from its theoretical assumptions, especially if the sample sizes are equal or nearly equal, and especially when two-tailed hypotheses are considered (e.g., Boneau, 1960; Box, 1953; Cochran, 1947; Posten, Yeh, and Owen, 1982; Srivastava, 1958).

The larger the samples, the more robust the test. If the underlying populations are markedly skewed, then one should be wary of one-tailed testing, and if there is considerable non-normality in the populations, then very small significance levels (say, $\alpha < 0.01$) should not be depended upon.

The power of the two-tailed t is affected very little by skewness in the sampled populations, but there can be a serious effect on one-tailed tests. The actual power of the test is less than that discussed in Section 8.4 when the sampled populations are platykurtic and greater when the populations are leptokurtic, especially for small sample sizes (Glass, Peckham, and Sanders, 1972).

If the population variances are unequal, then the probability of a Type I error will tend to be greater than the stated α ; but if the sample sizes are equal, then the t test is quite robust for moderate to large sample sizes, as shown in Table 8.1. If $n_1 \neq n_2$, then

TABLE 8.1 Maximum Probabilities of Type I Error When Applying the t Test to Two Samples of Various Sizes, $n_1 = n_2 = n$, from Normal Populations Having Various Variance Ratios, σ_1^2/σ_2^2

σ_1^2/σ_2^2	n :	3	5	10	15	16	20	30
For $\alpha = 0.05$:								
3.33 or 0.300		0.059	0.056	0.054	0.052		0.052	0.051
5.00 or 0.200		0.064	0.061	0.056	0.054		0.053	0.052
10.00 or 0.100			0.068	0.059	0.056		0.055	0.053
∞ or 0				0.065	0.060		0.057	0.055
For $\alpha = 0.01$:								
3.33 or 0.300		0.013	0.013	0.012		0.011	0.011	0.011
5.00 or 0.200		0.015	0.015	0.013		0.012	0.011	0.011
10.00 or 0.100		0.020	0.019	0.015		0.013	0.012	0.012
∞ or 0				0.018		0.015	0.014	0.013

These probabilities are gleaned from the extensive analysis of Ramsey (1980).

the probability of a Type I error will be less than α if the larger σ^2 is associated with the larger sample, and this probability will be greater than α if the smaller sample came from the population with the larger variance. The greater the difference between variances the greater will be the departure from α , with there being only a slight effect if the sample sizes are no more than 10% or so from equality. And robustness is compromised more if the smaller variance is associated with the larger sample than if the reverse is true (Kohr and Games, 1974; Posten, 1992; Posten, Yeh, and Owen, 1982; Ramsey, 1980). Ramsey (1980) also referred to an observation by Hsu (1938) of remarkable robustness in the presence of unequal variances if $n_1 = n_2 + 1$ and $\sigma_1^2 > \sigma_2^2$; so—if we have good estimates of the population variances—it is wise to plan experiments that have samples that are unequal in size by 1, where the larger sample comes from the population with the larger variance.

The comparison of two means from normal population without assuming equal variances is known as the “Behrens-Fisher problem,” referring to the solution by Behrens (1929) and Fisher (1939b); and numerous other studies of it have ensued (e.g., Cochran, 1964; Cochran and Cox, 1957: 100–102; Dixon and Massey, 1969: 119; Fisher and Yates, 1963: 3–4, 60–61; Gill, 1971; Lee and Fineberg, 1991; Lee and Gurland, 1975; Satterthwaite, 1946). One of the easiest, yet reliable, of such procedures is that attributed to Smith (1936) and also known as “Welch’s approximate t ”;[†] (Davenport and Webster, 1975; Mehta and Srinivasan, 1970; Scheffé, 1970; Wang, 1971; Welch, 1936, 1938). The test statistic is

$$t' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \quad (8.11)$$

*In Fisher and Yates (1963), s refers to the standard error, not the standard deviation.

[†]B[ernard] L[ewis] Welch (1911–1989), English statistician. (See Mardia, 1990.)

and the critical value is Student’s t with degrees of freedom of

$$v' = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}. \quad (8.12)$$

The degrees of freedom thus computed are usually not integer, in which case the next smaller integer should be used. If $n_1 \neq n_2$, and the population variances are very different, then t' will provide a better test than t . If the population variances are very similar, then t is the better (i.e., more powerful) test. If $n_1 = n_2$, or $s_1^2 = s_2^2$, then t' (Equation 8.11) is identical to t (Equation 8.17). If $n_1 = n_2$ and $s_1^2 = s_2^2$, then $t' = t$ and $v' = v$. Welch (1938) suggested that an improved test is obtained by employing $n_1 - 3$ and $n_2 - 3$ in place of $n_1 - 1$ and $n_2 - 1$, respectively, and this has been confirmed by Fenstad (1983); the two procedures have identical results if $n_1 = n_2$.

Some authors have recommended that the two variances should be compared and concluded to be similar (Section 8.5), prior to employing the t test. However, considering that the t test is so robust, and that the variance-comparison test performs so poorly when the distributions are non-normal (see Section 8.5), the routine test of variances is not recommended. Markowski and Markowski (1990) and Gans (1991) enlarge upon this conclusion, and Moser and Stevens (1992) explain that there is no circumstance when this two-step procedure performs better than using either t or t' (whichever is appropriate).

If there are severe deviations from the normality and/or equality-of-variance assumptions, the nonparametric test of Section 8.10 could be employed, as it is not adversely affected by violations of these assumptions, and some researchers would prefer that procedure to the modified t test above.

Replication of Data. It is important to employ data that are true replicates of the variable to be tested. In Example 8.1 the purpose of the experiment was to ask whether there is a difference in blood-clotting times between persons administered two different drugs. This necessitates having a blood measurement on each of n_1 individuals in the first sample and n_2 individuals in the second sample. It would *not* be valid to use n_1 measurements from a single person and n_2 measurements from another person, and to do so would be engaging in what Hurlbert (1984) discusses as *pseudoreplication*.

8.2 CONFIDENCE LIMITS FOR POPULATION MEANS

In Section 7.3, we defined the confidence interval for a population mean as $\bar{X} \pm t_{\alpha(2), v} s_{\bar{X}}$, where $s_{\bar{X}}$ is the best estimate of $\sigma_{\bar{X}}$ and is calculated as $\sqrt{s^2/n}$. For the two-sample situation, where we assume that $\sigma_1^2 = \sigma_2^2$, the confidence interval for either μ_1 or μ_2 is calculated using s_p^2 (rather than either s_1^2 or s_2^2) as the best estimate of σ^2 , and we use

EXAMPLE 8.19 (continued)

$$H'_1 = \frac{n \log n - \sum f_i \log f_i}{n} = \frac{197.5679 - 144.0751}{99} \\ = 0.5403$$

$$s^2_{H'_1} = \frac{\sum f_i \log^2 f_i - (\sum f_i \log f_i)^2/n}{n^2} = 0.00137602$$

Diet item	Louisiana Blue Jays		
	f_i	$f_i \log f_i$	$f_i \log^2 f_i$
Oak	48	80.6996	135.6755
Pine	23	31.3197	42.6489
Grape	11	11.4553	11.9294
Corn	13	14.4813	16.1313
Blueberry	8	7.2247	6.5246
Other	2	0.6021	0.1812
$s_2 = 6$	$n_2 = \sum f_i = 105$	$\sum f_i \log f_i = 145.7827$	$\sum f_i \log^2 f_i = 213.0909$

$$H'_2 = \frac{n \log n - \sum f_i \log f_i}{n} = \frac{212.2249 - 145.7827}{105} = 0.6328$$

$$s^2_{H'_2} = \frac{\sum f_i \log^2 f_i - (\sum f_i \log f_i)^2/n}{n^2} = 0.00096918$$

$$s_{H'_1 - H'_2} = \sqrt{s^2_{H'_1} + s^2_{H'_2}} = \sqrt{0.00137602 + 0.00096918} = 0.0484$$

$$t = \frac{H'_1 - H'_2}{s_{H'_1 - H'_2}} = \frac{-0.0925}{0.0484} = -1.911$$

$$v = \frac{\left(\frac{s^2_{H'_1}}{n_1} + \frac{s^2_{H'_2}}{n_2} \right)^2}{\frac{\left(\frac{s^2_{H'_1}}{n_1} \right)^2}{n_1} + \frac{\left(\frac{s^2_{H'_2}}{n_2} \right)^2}{n_2}} = \frac{(0.00137602 + 0.00096918)^2}{\frac{(0.00137602)^2}{99} + \frac{(0.00096918)^2}{105}} \\ = \frac{0.000005499963}{0.00000028071} = 196$$

$$t_{0.05(2), 196} = 1.972$$

Therefore, do not reject H_0 .

$$0.05 < P < 0.10 \quad [P = 0.057]$$

EXERCISES

8.1 Using the following data, test the null hypothesis that male and female turtles have the same mean serum cholesterol concentrations.

Serum Cholesterol (mg/100 ml)	
Male	Female
220.1	223.4
218.6	221.5
229.6	230.2
228.8	224.3
222.0	223.8
224.1	230.8
226.5	

8.2 It is proposed that animals with a northerly distribution have shorter appendages than animals from a southerly distribution. Test an appropriate hypothesis (by computing t), using the following wing length data for birds (data are in millimeters).

Northern	Southern
120	116
113	117
125	121
118	114
116	116
114	118
119	123
	120

- 8.3 If $\bar{X}_1 = 4.6$ kg, $s_1^2 = 3.88$ kg², $n_1 = 18$, $\bar{X}_2 = 6.0$ kg, $s_2^2 = 4.35$ kg², and $n_2 = 26$, test the hypotheses $H_0: \mu_1 \geq \mu_2$ and $H_A: \mu_1 < \mu_2$.
- 8.4 If $\bar{X}_1 = 334.6$ g, $\bar{X}_2 = 349.8$ g, $SS_1 = 364.34$ g², $SS_2 = 286.78$ g², $n_1 = 19$, and $n_2 = 24$, test the hypothesis that the mean weight of population 2 is more than 10 g greater than the mean weight of population 1.
- 8.5 If the null hypothesis in Exercise 8.1 is rejected, compute the 95% confidence limits for μ_1 , μ_2 , and $\mu_1 - \mu_2$. If H_0 is not rejected, calculate the 95% confidence limits for the common population mean.
- 8.6 A sample is to be taken from each of two populations from which previous samples of size fourteen have had $SS_1 = 244.66$ (km/hr)² and $SS_2 = 289.18$ (km/hr)². What size sample should be taken from each population in order to estimate $\mu_1 - \mu_2$ to within 2.0 km/hr, with 95% confidence?
- 8.7 Consider the populations described in Exercise 8.6.
- (a) How large a sample should we take from each population if we wish to detect a difference between μ_1 and μ_2 of at least 5.0 km/hr, using a 5% significance level and a t test with 90% power?
- (b) If we take a sample of twenty from one population and twenty-two from the other, what is the smallest difference between μ_1 and μ_2 that we have a 90% probability of detecting with a t test using $\alpha = 0.05$?

9

PAIRED-SAMPLE HYPOTHESES

The two-sample testing procedures discussed in Chapter 8 apply when the two samples are independent, independence implying that each datum in one sample is in no way associated with any specific datum in the other sample. However, there are instances when each observation in Sample 1 is in some way correlated with an observation in Sample 2, so that the data may be said to occur in pairs.

For example, we might wish to test the null hypothesis that the left foreleg and left hindleg lengths of deer are equal. We could make these two measurements on a number of deer, but we would have to remember that the variation among the data might be owing to two possible factors. First, the null hypothesis might be false, there being, in fact, a difference between foreleg and hindleg length. Second, deer are of different sizes, and for each deer the hindleg length is correlated with the foreleg length (i.e., a deer with a large front leg is likely to have a large hind leg). Thus, as Example 9.1 shows, the data can be tabulated in pairs, one pair (i.e., one hindleg measurement and one foreleg measurement) per animal.

9.1 TESTING MEAN DIFFERENCE

The two-tailed hypotheses implied by Example 9.1 are $H_0: \mu_1 - \mu_2 = 0$ and $H_A: \mu_1 - \mu_2 \neq 0$ (which, as pointed out in Section 8.1, could also be stated $H_0: \mu_1 = \mu_2$ and $H_A: \mu_1 \neq \mu_2$). However, we can define a mean population difference, μ_d , as $\mu_1 - \mu_2$, and write the hypotheses as $H_0: \mu_d = 0$ and $H_A: \mu_d \neq 0$. Although the use of either μ_d or $\mu_1 - \mu_2$ is correct, the former will be used hereafter when it implies the paired-sample situation.

EXAMPLE 9.1 The two-tailed paired-sample t test.

$$H_0: \mu_d = 0.$$

$$H_A: \mu_d \neq 0.$$

$$\alpha = 0.05$$

Deer (j)	Hindleg length (cm) (X_{1j})	Foreleg length (cm) (X_{2j})	Difference (cm) ($d_j = X_{1j} - X_{2j}$)
1	142	138	4
2	140	136	4
3	144	147	-3
4	144	139	5
5	142	143	-1
6	146	141	5
7	149	143	6
8	150	145	5
9	142	136	6
10	148	146	2

$$n = 10 \quad \bar{d} = 3.3 \text{ cm}$$

$$v = n - 1 = 9 \quad t = \frac{\bar{d}}{s_{\bar{d}}} = \frac{3.3}{0.97} = 3.402$$

$$s_d^2 = 9.3444 \text{ cm}^2$$

$$s_{\bar{d}} = 0.97 \text{ cm} \quad t_{0.05(2),9} = 2.262$$

Therefore, reject H_0 .

$$0.005 < P(|t| \geq 3.402) < 0.01 \quad [P = 0.008]$$

The test statistic for the null hypothesis is

$$t = \frac{\bar{d}}{s_{\bar{d}}} \quad (9.1)$$

Therefore, we do not use the original measurements for the two samples, but only the difference within each pair of measurements. One deals, then, with a sample of d_j values, whose mean is \bar{d} and whose variance, standard deviation, and standard error are denoted as s_d^2 , s_d , and $s_{\bar{d}}$ respectively. Thus, the *paired-sample t test*, as this procedure may be called, is essentially a one-sample t test, analogous to that described in Sections 7.1 and 7.2. In the paired-sample t test, n is the number of differences (i.e., the number of pairs of data), and $v = n - 1$. Note that the hypotheses used in Example 9.1 are special cases of the general hypotheses $H_0: \mu_d = \mu_0$ and $H_A: \mu_d \neq \mu_0$, where μ_0 is usually, but not always, zero.

For one-tailed hypotheses with paired samples, one can test either $H_0: \mu_d \geq \mu_0$ and $H_A: \mu_d < \mu_0$, or $H_0: \mu_d \leq \mu_0$ and $H_A: \mu_d > \mu_0$, depending on the question to be asked. Example 9.2 presents data from an experiment designed to test whether a new fertilizer results in an increase of more than 250 kg/ha in crop yield over the old fertilizer. For

testing this hypothesis, eighteen test plots of the crop were set up. It is probably unlikely to find eighteen field plots having exactly the same conditions of soil, moisture, wind, etc., but it should be possible to set up two plots with similar environmental conditions. If so, then the experimenter would be wise to set up nine pairs of plots, applying the new fertilizer to one plot of each pair and the old fertilizer to the other plot of that pair. As Example 9.2 shows, the statistical hypotheses to be tested are $H_0: \mu_d \leq 250$ kg/ha and $H_A: \mu_d > 250$ kg/ha.

EXAMPLE 9.2 A one-tailed paired-sample t test.

$$H_0: \mu_d \leq 250 \text{ kg/ha}$$

$$H_A: \mu_d > 250 \text{ kg/ha}$$

$$\alpha = 0.05$$

Plot (j)	Crop Yield (kg/ha)		d_j
	With new fertilizer (X_{1j})	With old fertilizer (X_{2j})	
1	2250	1920	330
2	2410	2020	390
3	2260	2060	200
4	2200	1960	240
5	2360	1960	400
6	2320	2140	180
7	2240	1980	260
8	2300	1940	360
9	2090	1790	300

$$n = 9 \quad \bar{d} = 295.6 \text{ kg/ha}$$

$$v = n - 1 = 8 \quad t = \frac{\bar{d} - 250}{s_{\bar{d}}} = 1.695$$

$$s_d = 80.6 \text{ kg/ha}$$

$$s_{\bar{d}} = 26.9 \text{ kg/ha} \quad t_{0.05(1),8} = 1.860$$

Therefore, do not reject H_0

$$0.05 < P < 0.10 \quad [P = 0.064]$$

Paired-sample t -testing requires that each datum in one sample is correlated with one, *but only one*, datum in the other sample. So, in the last example, each yield using new fertilizer is paired with only one yield using old fertilizer; and it would have been inappropriate to have some tracts of land large enough to collect two or more crop yields using each of the fertilizers.

The paired-sample t test does not have the normality and equality of variances assumptions of the two-sample t test, but assumes instead that the differences, d_j , come from a normally distributed population of differences. If there is, in fact, pairwise

correlation of data from the two samples, then the paired-sample t test will be more powerful than the two-sample t test. If no such correlation exists, then the two-sample t test will be the more powerful procedure. Hines (1996) showed that, unless n is very tiny, only a small correlation is needed to make the paired-sample test advantageous. If the data from Example 9.1 were subjected (inappropriately) to the two-sample t test, rather than to the paired-sample t test, a difference would not have been concluded, and a Type II error would have been committed.

EXERCISES

9.1 Concentrations of nitrogen oxides and of hydrocarbons were determined in a certain urban area (recorded in $\mu\text{g}/\text{m}^3$).

- (a) Test the hypothesis that both classes of air pollutants were present in the same concentration.

Day	Nitrogen oxides	Hydrocarbons
1	104	108
2	116	118
3	84	89
4	77	71
5	61	66
6	84	83
7	81	88
8	72	76
9	61	68
10	97	96
11	84	81

- (b) Calculate the 95% confidence interval for μ_d .