

# Unit 24

## Hypothesis Tests about Means

Objectives:

- To recognize the difference between a paired  $t$  test and a two-sample  $t$  test
- To perform a paired  $t$  test
- To perform a two-sample  $t$  test

A measure of the amount of variation (dispersion) we expect to occur when using a statistic from a sample to estimate a parameter in a population is the *standard error* of the statistic. For instance, when using a sample mean  $\bar{x}$  to estimate a population mean  $\mu$ , the standard error of the sample mean  $\bar{x}$  can be estimated by  $s/\sqrt{n}$ . Recall that the  $t$  test statistic in a hypothesis test about a population mean is found simply by dividing the difference between the sample mean  $\bar{x}$  and the hypothesized mean by the estimate of the standard error of the mean  $s/\sqrt{n}$ , that is,

$$t_{n-1} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} .$$

Test statistics of similar form can be used in hypothesis tests to compare two means. We shall consider a  $t$  test statistic used with data treated as one sample of paired observations and a  $t$  test statistic used with data treated as two independent samples of observations.

Let us first consider a  $t$  test statistic used with paired data. When our interest is in the mean of the differences between paired observations, we can think of the observed differences as comprising one sample of observations. Consequently, a  $t$  test statistic which can be used with paired data is the  $t_{n-1}$  statistic with the sample mean and standard deviation of the differences between pairs. In practice, the hypothesized value of the mean difference is almost always zero, but it need not be in general; however, we shall exclusively focus on situations where zero is the hypothesized mean difference. If we denote the sample mean of the differences as  $\bar{d}$  and the sample standard deviation of the differences as  $s_d$ , we can write the test statistic for this hypothesis test as

$$t_{n-1} = \frac{\bar{d} - 0}{s_d/\sqrt{n}} = \frac{\bar{d}}{s_d/\sqrt{n}} .$$

We shall call this the *paired  $t$  test statistic*, and a hypothesis test which makes use of this test statistic can be called a *paired  $t$  test*.

This test will be appropriate when a simple random sample of paired observations is selected from a normally distributed population.

To illustrate the paired  $t$  test, let us consider using a 0.05 significance level to perform a hypothesis test to see if there is any evidence that the mean time to perform a specific task is reduced when background music is provided. Each of seven available subjects performs the task twice, once

without background music provided and once with background music, in random order, and the results are

Subject Number	Minutes without Background Music	Minutes without Background Music
1	13.8	13.4
2	14.3	15.2
3	12.7	12.6
4	10.9	9.3
5	11.0	8.7
6	9.6	10.2
7	10.3	9.0

recorded in Table 24-1. A paired  $t$  test is to be performed, since the data consist of one sample of paired observations.

In order to perform a paired  $t$  test, we must first obtain the differences between paired observations from the data. Since we are interested in whether or not mean time to perform a specific task is reduced when background music is provided, it is natural for us to choose to subtract the time with background music from the time without background music; however, the choice of order of subtraction in calculating the differences is really arbitrary. Enter these differences in the appropriate place in Table 24-2. (You should find that these differences are +0.4, -0.9, +0.1, +1.6, +2.3, -0.6, and +1.3.)

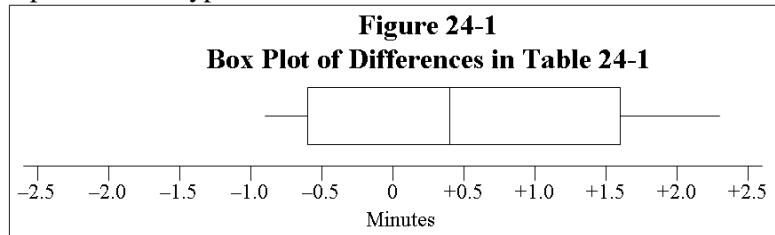
Notice that  $d$  is used in Table 24-2 to represent a difference which results from subtracting time with background music from the corresponding time without background music. Then,  $\sum d$  represents the sum of these differences and  $\bar{d}$  represents the mean of these differences. Enter the values of  $\sum d$  and  $\bar{d}$  in the appropriate places in Table 24-2. (You should find that  $\sum d = +4.2$  and  $\bar{d} = +0.6$ .) Also,  $\sum (d - \bar{d})^2$  represents the sum of the squared deviations of the differences from their mean,  $s_d^2$  represents the variance of the differences, and  $s_d$  represents the standard deviation of the differences. Enter the values of  $\sum (d - \bar{d})^2$ ,  $s_d^2$ , and  $s_d$  in the appropriate

<b>Table 24-2</b>			
<b>Mean and Standard Deviation of the Differences in Table 24-1</b>			
Differences			
$d = \text{No Music} - \text{Music}$			
13.8 - 13.4 =		$n =$	$\sum d =$
14.3 - 15.2 =			
12.7 - 12.6 =		$\bar{d} =$	
10.9 - 9.3 =			
11.0 - 8.7 =		$\sum (d - \bar{d})^2 =$	
9.6 - 10.2 =			
10.3 - 9.0 =		$s_d^2 =$	$s_d =$

places in Table 24-2. (You should find that  $\sum (d - \bar{d})^2 = 8.36$ ,  $s_d^2 = 1.393$ , and  $s_d = 1.180$ .)

Before beginning our paired  $t$  test, we examine a box plot of the differences, displayed as Figure 24-1, to see if there is any reason for us to think that a paired  $t$  test will not be appropriate. Since we see that there are no outliers, and that the box plot does not look extremely skewed, we have no reason to believe that a paired  $t$  test will not be appropriate. Of course the examination of a normal probability plot is preferable, and one would find in this case that the points appear to lie reasonably close to a straight line. This supports our belief that the paired  $t$  test is appropriate. We are now ready to perform the hypothesis test.

The first step is to state the null and alternative hypotheses, and choose a significance level. We shall state the hypotheses in terms of the mean difference. Although the order of subtraction to obtain the differences is arbitrary, once the choice is made our notation must reflect that choice.



Since we chose to subtract the time with background music from the time without background music, we denote the mean difference as  $\mu_{N-M}$  where the "N" in the subscript represents time with no background music, and the "M" in the subscript represents time with background music. This test is one-sided, since we are looking for a difference in only one direction. We complete the first step of the hypothesis test as follows:

$$H_0: \mu_{N-M} = 0 \text{ vs. } H_1: \mu_{N-M} > 0 \quad (\alpha = 0.05, \text{ one-sided test})$$

The second step is to collect data and calculate the value of the test statistic. In Table 24-2, you found that  $\bar{d} = +0.6$  minutes and  $s_d = 1.180$  minutes for the  $n = 7$  differences. We now calculate the value of our paired  $t$  test statistic as follows:

$$t_6 = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{+0.6}{1.180 / \sqrt{7}} = 1.345 .$$

The third step is to define the rejection region, decide whether or not to reject the null hypothesis, and obtain the  $p$ -value of the test. Figure 24-2 displays the rejection region graphically. The shaded area corresponds to values of the paired  $t$  test statistic where  $\bar{d}$  is larger than the hypothesized zero mean difference and which are unlikely to occur if  $H_0: \mu_{N-M} = 0$  is true. In order that this shaded area be equal to  $\alpha = 0.05$ , the rejection region is defined by the  $t$ -score with  $df = 6$  above which 0.05 of the area lies. From Table A.3, we find that  $t_{6,0.05} = 1.943$ , and we can then define our rejection region algebraically as

$$t_6 \geq 1.943 .$$

Since our test statistic, calculated in the second step, was found to be  $t_6 = 1.345$ , which is not in the rejection region, our decision is not to reject  $H_0: \mu_{N-M} = 0$ ; in other words, our data does not provide sufficient evidence to suggest that  $H_1: \mu_{N-M} > 0$  is true.

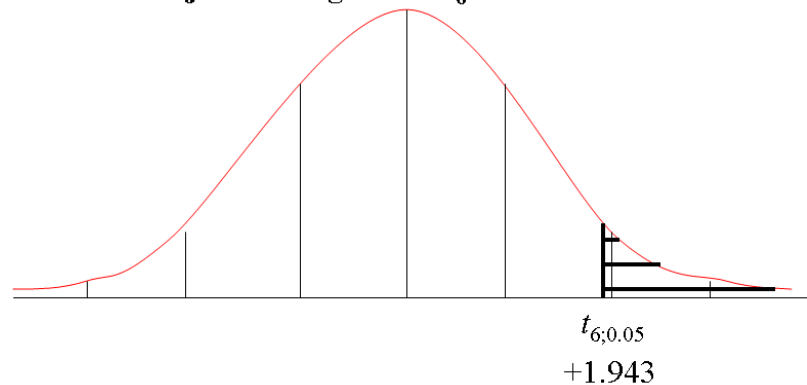
Figure 24-3 graphically illustrates the  $p$ -value. Our test statistic was found to be  $t_6 = 1.345$ . The  $p$ -value of this hypothesis test is the probability of

obtaining a test statistic value  $t_6$  which represents a larger distance above the hypothesized zero mean difference than the value actually observed  $t_6 = 1.345$ . In Figure 24-3, the observed test statistic value  $t_6 = 1.345$  has been located on the horizontal axis. The shaded area corresponds to test statistic values which represent a larger distance above the hypothesized zero mean difference than the value actually observed  $t_6 = 1.345$ ; in other words, the shaded area in Figure 24-3 is the  $p$ -value of the hypothesis test. From Table A.3, we find that 1.345 is less than  $t_{6,0.10} = 1.440$ , which tells us that the area above the  $t$ -score 1.345 is greater than 0.10. Therefore, the shaded area in Figure 24-3, which is the  $p$ -value, must be greater than 0.10. We indicate this by writing  $0.10 < p$ -value. The fact that  $0.10 < p$ -value confirms to us that  $H_0$  is not rejected with  $\alpha = 0.05$ . However, it also tells us that  $H_0$  would not be rejected with  $\alpha = 0.10$  nor with  $\alpha = 0.01$ .

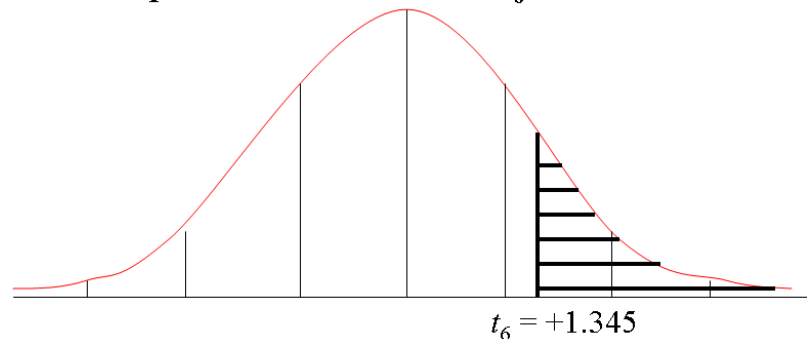
To complete the fourth step of the hypothesis test, we can summarize the results of the hypothesis test as follows:

Since  $t_6 = 1.345$  and  $t_{6,0.05} = 1.943$ , we do not have sufficient evidence to reject  $H_0$ . We conclude that the mean time to perform the task is not reduced when background music is provided ( $0.10 < p$ -value).

**Figure 24-2**  
**Rejection Region for  $t_6$  with  $\alpha = 0.05$**



**Figure 24-3**  
 **$p$ -value for Test Statistic  $t_6 = +1.345$**





represent the respective sample variances. A  $t$  test statistic to compare two means with independent simple random samples will be similar to the  $t$  test statistics we have already introduced; specifically, a  $t$  test statistic to compare two means with independent samples will be the difference between the sample means  $\bar{x}_1 - \bar{x}_2$  divided by an appropriate standard error.

One of two standard errors might be used, depending on whether or not the two samples are selected from populations having roughly equal standard deviations. Consequently, there are two  $t$  test statistics available, and statistical software on calculators and computers generally have the ability to calculate both of these test statistics, leaving it up to the user to decide which test statistic to make use of. These two  $t$  test statistics are

$$t_{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad \text{where} \quad s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

and

$$t_v = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \text{where} \quad v = \frac{(n_1-1)(n_2-1) \left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{(n_2-1) \left( \frac{s_1^2}{n_1} \right)^2 + (n_1-1) \left( \frac{s_2^2}{n_2} \right)^2}$$

Each of these test statistics can require a substantial amount of calculation. Since the calculations necessary for these statistics are readily available from statistical software and programmable calculators, we shall only briefly discuss the formulas.

The  $t_{n_1+n_2-2}$  statistic depends on  $s_p^2$  which is called the *pooled sample variance*; this test statistic is appropriate when we can reasonably assume that the populations being sampled have roughly equal standard deviations (i.e., the populations do not significantly differ in dispersion). The pooled sample variance is a weighted average of the two sample variances which provides a single estimate of the population variance common to both populations. We call  $t_{n_1+n_2-2}$  the *pooled two sample  $t$  test statistic*, and its degrees of freedom is  $n_1+n_2-2$ , which is the sum of the sample sizes minus the number of samples. If this test statistic is used when the two sampled populations have very different standard deviations, the results can be misleading especially if the sample sizes are not equal.

The  $t_v$  statistic depends on the calculation of the degrees of freedom  $v$  which requires a somewhat complex formula, and  $v$  generally turns out not to be an integer. This test statistic is designed to be appropriate no matter whether the populations being sampled have equal or very different standard deviations (i.e., regardless of whether the populations significantly differ in dispersion or not). We call  $t_v$  the *separate two sample  $t$  test statistic* (since the two sample standard deviations are each estimating a separate variance when the populations do have the same dispersion), and in practice we simply round  $v$  to the nearest integer value. In situations where the two sampled populations have equal, or close to equal, standard deviations, the statistics  $t_v$  and  $t_{n_1+n_2-2}$  will tend to be very close in value and in degrees of freedom.

Having now introduced the pooled two sample  $t$  test statistic and the separate two sample  $t$  test statistic, the immediate question which comes to mind concerns when to use which test. However, rather than having to decide which of the two sample  $t$  test statistics to use in a given situation, one approach would be just to always use the separate two sample  $t$  test statistic without concerning oneself with whether or not the standard deviations are substantially different. This is a very reasonable approach, since both test statistics will generally be almost identical when the population standard deviations are equal, or close to equal. We shall take this approach and call the corresponding hypothesis test the *separate two sample  $t$  test*.

To illustrate separate two sample  $t$  test, let us consider using a 0.10 significance level to perform a hypothesis test to see if there is any evidence of a difference in mean lifetime between the two brands of light bulbs: Sunn and Brighto. Four bulbs from the brand named Sunn comprise one sample and three bulbs from the brand named Brighto comprise a second sample. The hours of lifetime for each bulb are recorded in Table 24-3.

We might consider displaying this data with two contiguous box plots, one for each brand, but we decided instead to display the data with the back-to-back stem-and-leaf displays in Figure 24-4. The middle column in Figure 24-4 consists of the stems, which represent the first two digits of the observations; on each side of this middle column are the leaves which represent the third digit of the observations.

We need a two sample test here, since the data is treated as two independent simple random samples, each selected from a different brand of light bulb. We shall assume that the light bulb lifetimes for each brand are roughly normality distributed, and we shall perform the separate two sample  $t$  test.

The first step is to state the null and alternative hypotheses, and choose a significance level. We shall state the hypotheses in terms of the difference between means; the order of subtraction is an arbitrary choice, and we shall choose to subtract the mean for Brighto from the mean for Sunn. This test is two-sided, since we are looking for a difference in either direction. We complete the first step of the hypothesis test as follows:

$$H_0: \mu_S - \mu_B = 0 \text{ vs. } H_1: \mu_S - \mu_B \neq 0$$

( $\alpha = 0.10$ , two-sided test)

The second step is to collect data and calculate the value of the test statistic. Our data is displayed in Table 24-3. We shall not discuss in detail the calculation of the separate two sample  $t$  statistic, because, as stated previously, the ability to do these calculations easily is readily available with readily available from statistical software and programmable calculators. Readers interested in the details of performing these calculations may refer to Section B.2 of Appendix B, where the data of Table 24-3 is used as an example of how to do the calculations. Using either some appropriate statistical software or programmable calculator, or using the results in Section B.2 of Appendix B, we find

that the separate two sample  $t$  statistic has degrees of freedom  $\nu = 3.9 \approx 4$ , and that the test statistic is  $t_4 = 2.104$ . This of course is based on subtracting the mean for Brighto from the mean for Sunn.

The third step is to define the rejection region, decide whether or not to reject the null hypothesis, and obtain the  $p$ -value of the test. Figure 24-5 displays the rejection region graphically. The shaded area corresponds to values of the separate two-sample  $t$  test statistic where  $\bar{x}_S - \bar{x}_B$  is away from the hypothesized zero difference between means and which are unlikely to occur if  $H_0: \mu_S - \mu_B = 0$  is true. In order that this shaded area be equal to  $\alpha = 0.10$ , the rejection region is defined by the  $t$ -score with  $df = 4$  above which 0.05 of the area lies; below the negative of this  $t$ -score lies 0.05 of the area, making a total area of 0.10. From Table A.3, we find that  $t_{4;0.05} = 2.132$ . We can then define our rejection region algebraically as

$$t_4 \geq 2.132 \text{ or } t_4 \leq -2.132 .$$

**Table 24-3**  
**Hours of Lifetime of Light Bulbs for Two Brands**

<u>Sunn</u>			
575	611	572	602
<u>Brighto</u>			
568	569	528	

**Figure 24-4**  
**Hours of Lifetime of Light Bulbs for Two Brands**

<u>Sunn</u>		<u>Brighto</u>
	51	
	52	8
	53	
	54	
	55	
	56	8 9
5 2	57	
	58	
	59	
2	60	
1	61	
	62	

Since our test statistic, calculated in the second step, was found to be  $t_4 = 2.104$ , which is not in the rejection region, our decision is not to reject  $H_0: \mu_S - \mu_B = 0$  in other words, our data does not provide sufficient evidence to suggest that  $H_1: \mu_S - \mu_B \neq 0$  is true.

Figure 24-6 graphically illustrates the  $p$ -value. Our test statistic was found to be  $t_4 = 2.104$ . The  $p$ -value of this hypothesis test is the probability of obtaining a test statistic value  $t_4$  which represents a larger distance away from the hypothesized zero difference between means than the value actually observed  $t_4 = 2.104$ . In Figure 24-6, the observed test statistic value  $t_4 = 2.104$  and the value  $-2.104$  have both been located on the horizontal axis. The shaded area corresponds to test statistic values  $t_4$  which represent a larger distance away from the hypothesized zero difference between means than value actually observed  $t_4 = 2.104$ ; in other words, the shaded area in Figure 24-6 is the  $p$ -value of the hypothesis test. From Table A.3, we find that 2.104 is between  $t_{4;0.10} = 1.533$  and  $t_{4;0.05} = 2.132$ , which tells us that the area above the  $t$ -score 2.171 is between 0.05 and 0.10. Therefore, the shaded area in Figure 24-6, which is the  $p$ -value, must be between 0.10 and 0.20. We indicate this by writing  $0.10 < p\text{-value} < 0.20$ . The fact that  $0.10 < p\text{-value} < 0.20$  confirms to us that  $H_0$  is not rejected with  $\alpha = 0.10$ . However, it also tells us that  $H_0$  would not be rejected with  $\alpha = 0.05$  nor with  $\alpha = 0.01$ . Although we obtained our  $p$ -value from Table A.3, you will most likely find that any statistical software or programmable calculator which gives you the separate two sample  $t$  statistic also gives you an exact  $p$ -value.

To complete the fourth step of the hypothesis test, we can summarize the results of the hypothesis test as follows:

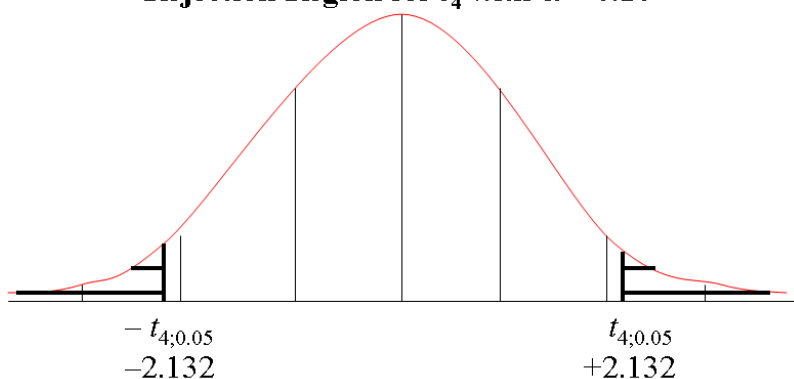
Since  $t_4 = 2.104$  and  $t_{4;0.05} = 2.132$ , we do not have sufficient evidence to reject  $H_0$ . We conclude that there is no difference in mean lifetime between the two brands of light bulbs Sunn and Brighto ( $0.10 < p\text{-value} < 0.20$ ).

(Just for the sake of comparison, we note that the pooled two sample  $t$  statistic has degrees of freedom  $n_S + n_B - 2 = 4 + 3 - 2 = 5$ , that this test statistic is  $t_5 = 2.171$ , and that for this test statistic  $0.05 < p\text{-value} < 0.10$ . The fact that the two tests do not give the same result may be a result of the standard deviations being substantially different for the two brands of bulbs. Once again, we point out our suggested strategy of simply always choosing to use the separate two sample  $t$  test statistic without concerning oneself with whether or not the standard deviations are substantially different.)

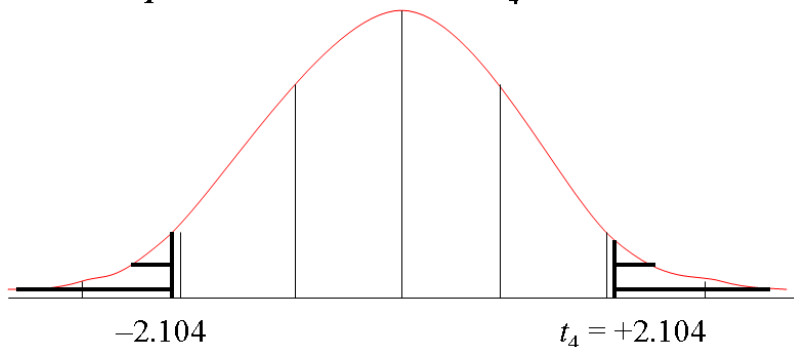
Before we conclude we conclude our discussion of the paired one sample  $t$  test and the separate two sample  $t$  test, we want to briefly remind you that it is always important to keep in mind the difference between statistical significance and clinical significance. To illustrate, suppose a bus is scheduled to arrive at the Funville terminal at 11:00pm on Fridays and again at 11:00pm on Saturdays, and our interest is in whether or not the mean number of minutes the bus is late is different for Fridays and Saturdays. From a simple random sample of observations for Fridays we find the sample mean to be 15.2 minutes, and from a simple random sample of observations for Saturdays we find the sample mean to be 17.2 minutes.

Whether or not this difference of 2.0 minutes represents a clinically significant difference is a question of whether or not the size of this difference has any practical impact, and this involves at least some subjective judgment. However, we can not possibly know whether or not this represents a statistically significant

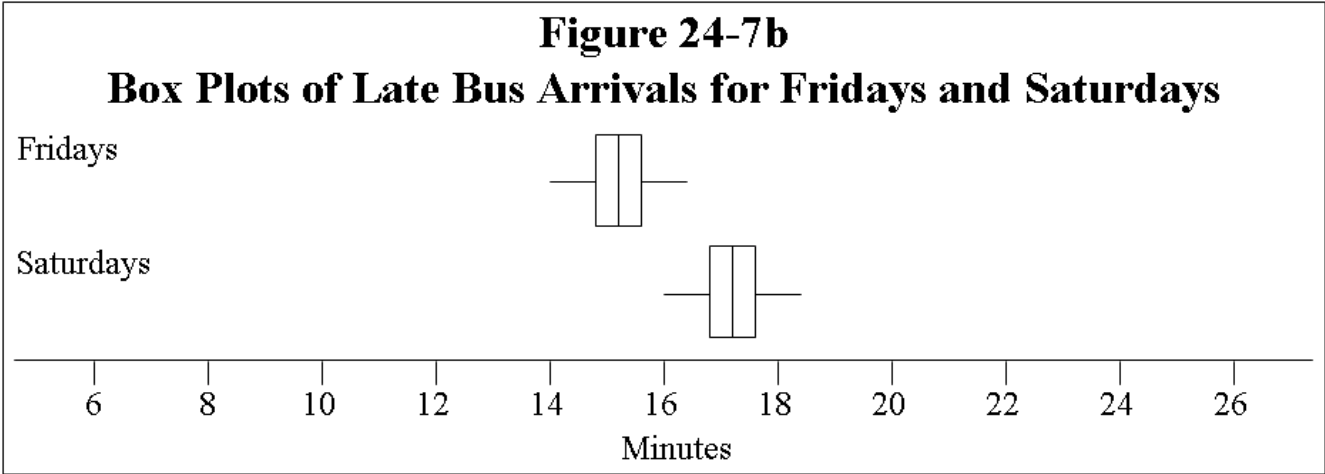
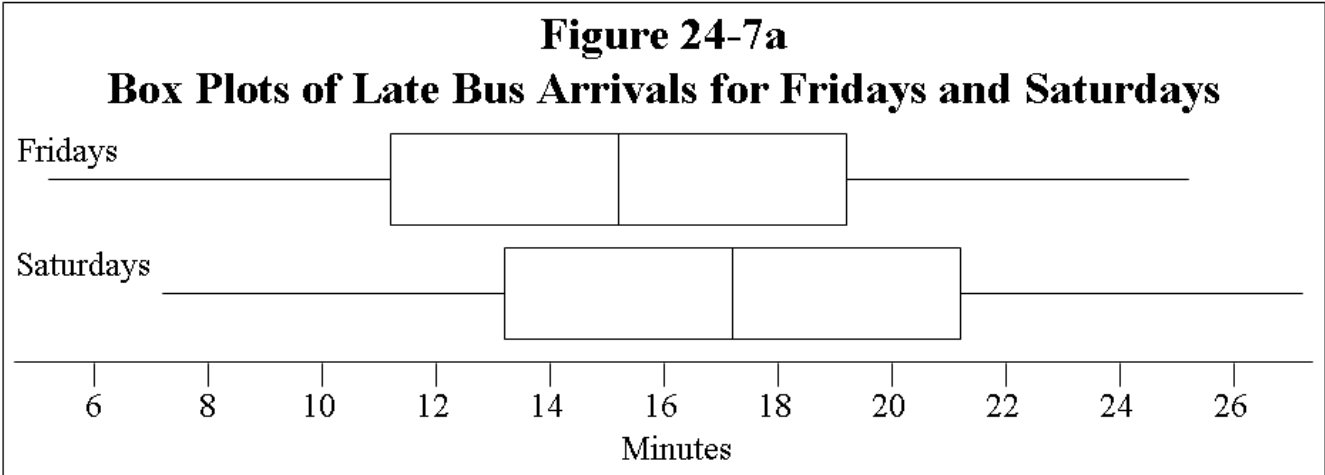
**Figure 24-5**  
**Rejection Region for  $t_4$  with  $\alpha = 0.10$**



**Figure 24-6**  
 **$p$ -value for Test Statistic  $t_4 = +2.104$**



difference without further information. A difference will be statistically significant when it is sufficiently larger than the standard error in the denominator of the appropriate test statistic. Without knowing the standard error it is impossible to tell if the difference is going to be significant or not. Figures 24-7a and 24-7b demonstrate two extreme situations. In both figures, it appears that each distribution is at least close to symmetric, implying that the mean and median will be close. Both figures appear suggest that the sample mean for Fridays was just about 15.2 minutes and that the sample mean for Saturdays was just about 17.2 minutes. Even though the difference between means is the same in both figures, Figure 24-7a gives the impression that this difference will not be statistically significant, while Figure 24-7b gives the impression that this difference will be statistically significant. The reason for this is that there is much more dispersion present in Figure 24-7a than in Figure 24-7b, which implies that the standard error of the difference between means will be considerably larger in Figure 24-7a than in Figure 24-7b; consequently, the difference between means in Figure 24-7a does not appear to be sufficiently larger than the standard error to be statistically significant, but the difference between means in Figure 24-7b does appear to be sufficiently larger than the standard error to be statistically significant.







### Answers to Self-Test Problems

- 24-1** (a) Two measurements of hours, one for TV and one for radio, are recorded on each individual, making this one sample of paired data.
- (b) Step 1:  $H_0: \mu_{T-R} = 0$ ,  $H_1: \mu_{T-R} \neq 0$ , ( $\alpha = 0.10$ , two-sided)  
Step 2:  $n = 30$ ,  $\bar{d} = -4.9$  hours,  $s_d = 10.1755$  hours, and  $t_{29} = -2.638$   
Step 3: The rejection is  $t_{29} \geq +1.699$  or  $t_{29} \leq -1.699$  (which can be written as  $|t_{29}| \geq 1.699$ ).  $H_0$  is rejected;  $0.01 < p\text{-value} < 0.02$ .  
Step 4: Since  $t_{29} = -2.638$  and  $t_{29;0.05} = 1.699$ , we have sufficient evidence to reject  $H_0$ . We conclude that there is a difference between the mean time spent watching TV weekly and the mean time spent listening to the radio weekly among voters in the state ( $0.01 < p\text{-value} < 0.02$ ). The data suggest that the mean weekly radio hours is higher than the mean weekly TV hours.
- (c) The five-number summary is  $-21, -13, -3.5, 0, +23$ ;  $+23$  is a potential outlier, which might make us question whether the  $t$  statistic is appropriate.
- (d) Since  $H_0$  is rejected, a Type I error is possible, which is concluding that  $\mu_{T-R} \neq 0$  when really  $\mu_{T-R} = 0$ .
- (e)  $H_0$  would not have been rejected with  $\alpha = 0.01$  but would have been rejected with  $\alpha = 0.05$ .
- 24-2** (a) Measurements of millions of customers are recorded once on each restaurant in one of two separate groups, the southern chain and the northern chain, making this two independent samples of data.
- (b) Step 1:  $H_0: \mu_S - \mu_N = 0$ ,  $H_1: \mu_S - \mu_N \neq 0$ , ( $\alpha = 0.05$ , two-sided)  
Step 2:  $n_S = 8$ ,  $\bar{x}_S = 2.975$  million customers,  $s_S = 1.4859$  million customers,  $n_N = 15$ ,  $\bar{x}_N = 3.9$  million customers,  $s_N = 0.4629$  million customers,  $t_8 = -1.717$ .  
Step 3: The rejection is  $t_8 \geq +2.306$  or  $t_8 \leq -2.306$  (which can be written as  $|t_8| \geq 2.306$ ).  $H_0$  is not rejected;  $0.10 < p\text{-value} < 0.20$ .  
Step 4: Since  $t_8 = -1.717$  and  $t_{8;0.025} = 2.306$ , we do not have sufficient evidence to reject  $H_0$ . We conclude that there is no difference in the mean number of customers between the southern and northern parts of the chain ( $0.10 < p\text{-value} < 0.20$ ).
- (c) The five-number summary for the southern chain is  $0.5, 1.95, 3.2, 3.95, 5.1$ ; there are no outliers. The five-number summary for the northern chain is  $3.1, 3.5, 3.8, 4.4, 4.7$ ; there are no outliers. Since there are no outliers, and neither of the distributions looks extremely skewed, there is no reason to believe that the two sample  $t$  test statistic is not appropriate.
- (d) Since  $H_0$  is not rejected, a Type II error is possible, which is concluding that  $\mu_S - \mu_N = 0$  when really  $\mu_S - \mu_N \neq 0$ .
- (e)  $H_0$  would not have been rejected with  $\alpha = 0.01$  nor with  $\alpha = 0.10$ .
- (f) **Optional:** The pooled two sample  $t$  statistic is  $t_{21} = -2.254$  which is very different from the separate two sample  $t$  test statistic, because the box plots and sample standard deviations appear to suggest that dispersion is much greater in the southern chain than in the northern chain.

(continued next page)

### Answers to Self-Test Problems

(continued from previous page)

- 24-3** (a) Measurements of blood pressure are recorded once on each employee in one of two separate groups, the Nuketown factory and the Highville factory, making this two independent samples of data.
- (b) Step 1:  $H_0: \mu_N - \mu_H = 0$ ,  $H_1: \mu_N - \mu_H > 0$ , ( $\alpha = 0.01$ , one-sided)
- Step 2:  $n_N = 35$ ,  $\bar{x}_N = 126.2$ ,  $s_N = 8.9$ ,  $n_H = 40$ ,  $\bar{x}_H = 117.9$ ,  $s_H = 10.4$ ,  $t_{73} = 3.724$ .
- Step 3: The rejection is  $t_{73} \geq +2.390$ .  $H_0$  is rejected;  $p$ -value  $< 0.0005$ .
- Step 4: Since  $t_{73} = 3.724$  and  $t_{73;0.01} = 2.306$ , we have sufficient evidence to reject  $H_0$ . We conclude that mean blood pressure is higher for employees at a Nuketown factory than for employees at a Highville factory ( $p$ -value  $< 0.0005$ ).
- (c) The presence of one or more outliers in the data would suggest that the  $t$  statistic may not be appropriate.
- (d) Since  $H_0$  is rejected, a Type I error is possible, which is concluding that  $\mu_N - \mu_H > 0$  when really  $\mu_N - \mu_H = 0$ .
- (e)  $H_0$  would have been rejected with both  $\alpha = 0.05$  and  $\alpha = 0.10$ .
- (f) **Optional:** The pooled two sample  $t$  statistic is  $t_{73} = +3.685$  which is very close to the separate two sample  $t$  test statistic, because the sample standard deviations appear to suggest that dispersion is roughly the same in the Nuketown and Highville factories.

### Summary

A measure of the amount of variation (dispersion) we expect to occur when using a statistic from a sample to estimate a parameter in a population is the *standard error* of the statistic. The  $t$  test statistic is found simply by dividing the difference between the sample mean  $\bar{x}$  and the hypothesized mean by the estimate of the standard error of the mean  $s/\sqrt{n}$ . In a hypothesis test concerning the mean of the differences between paired observations, we can think of the observed differences as comprising one sample of observations. In practice, the hypothesized value of the mean difference in such a hypothesis test is almost always zero. A *paired  $t$  test* is a one sample hypothesis test where the  $t_{n-1}$  statistic is used with the sample mean and standard deviation of the differences between pairs. Denoting the sample mean of the differences as  $\bar{d}$  and the sample standard deviation of the differences as  $s_d$ , we can write the test statistic for this hypothesis test as

$$t_{n-1} = \frac{\bar{d} - 0}{s_d / \sqrt{n}} = \frac{\bar{d}}{s_d / \sqrt{n}} .$$

which we call the *paired  $t$  test statistic*. This test will be appropriate when a simple random sample of differences of paired observations is either selected from a normally distributed population or is of a sufficiently large size.

With data treated as two independent samples of observations, a hypothesis test concerning the difference between two means is based on comparing the mean of one sample with the mean of the other sample. In practice, the hypothesized value of the difference between means is almost always zero. We let  $n_1$  and  $n_2$  represent the two sample sizes, let  $\bar{x}_1$  and  $\bar{x}_2$  represent the respective sample means, and let  $s_1^2$  and  $s_2^2$  represent the respective sample variances. A  $t$  test statistic to compare two means with independent simple random samples will be the difference between the sample means  $\bar{x}_1 - \bar{x}_2$  divided by an appropriate standard error. One of two standard errors might be used, depending on whether or not the two samples are selected from populations having roughly equal standard deviations. The two possible test statistics are the *pooled two sample  $t$  test statistic* and the *separate two sample  $t$  test statistic*, each of which can require a substantial amount of calculation. Section B.2 of Appendix B provides a detailed description of how to do the calculations, but in

practice these statistics are readily available from statistical software and programmable calculators. Rather than having to decide which of the two sample  $t$  tests to use, one approach would be just to always use the separate two sample  $t$  test statistic without concerning oneself with whether or not the standard deviations are substantially different. This is a very reasonable approach, since both test statistics will generally be almost identical when the population standard deviations are equal, or close to equal.