10 Inferences Involving Two Populations

10.1 Dependent and Independent Samples

Dependent and Independent Samples

Male students and female students are two populations. In this chapter we are going to study the procedures for making inferences about two populations.

When comparing two populations, we need two samples, one from each population.

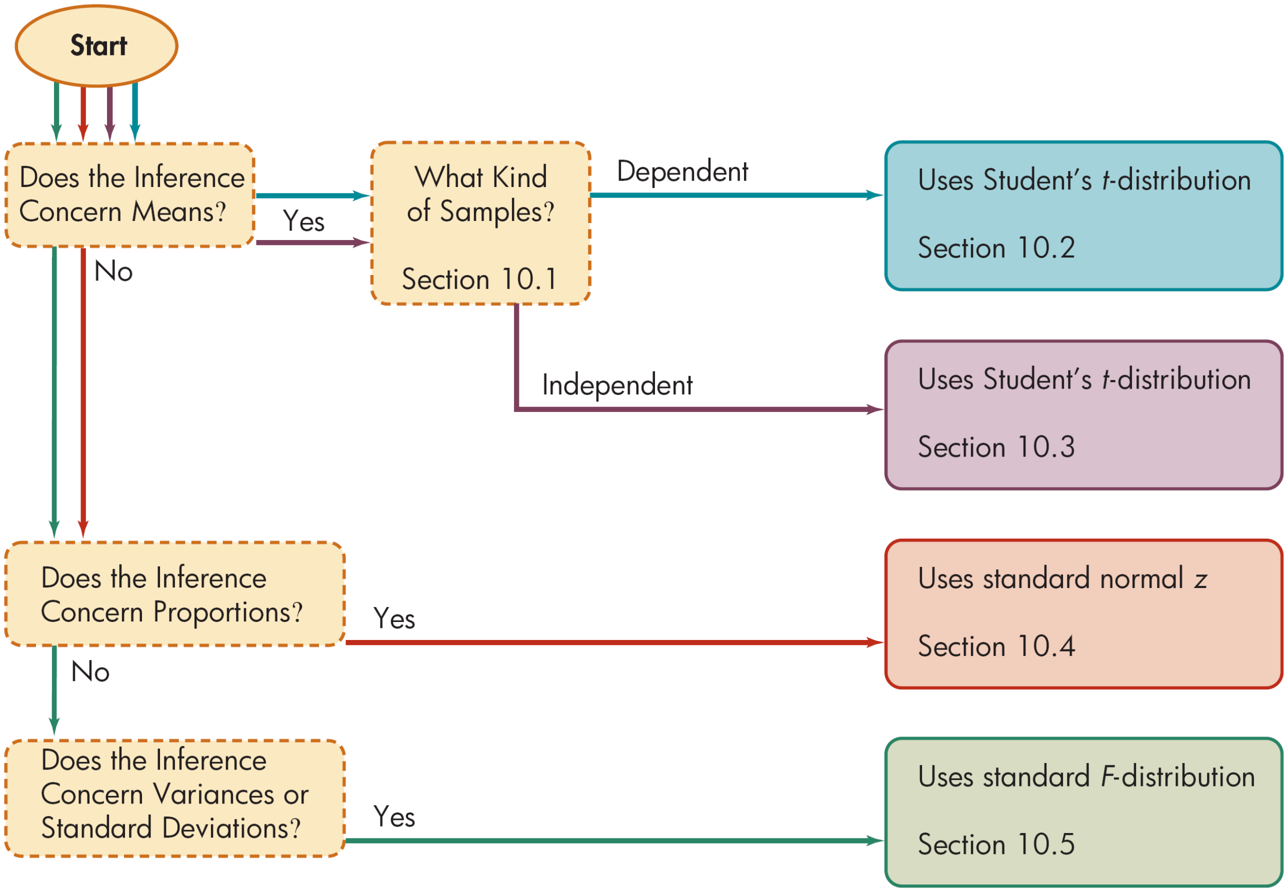
Two basic kinds of samples can be used: independent and dependent.

Dependent and Independent Samples

- The dependence or independence of two samples is determined by the sources of the data.
- A **source** can be a person, an object, or anything else that yields a data value. If the same set of sources or related sets are used to obtain the data representing both populations, we have **dependent samples**.
- If two unrelated sets of sources are used, one set from each population, we have **independent samples**.

Inference on Two populations

The following examples should clarify these ideas.



"Road Map" to Two Population Inferences

Figure 10.1

Example 1 – *Dependent Versus Independent Samples*

A test will be conducted to see whether the participants in a physical-fitness class actually improve in their level of fitness. It is anticipated that approximately 500 people will sign up for this course.

The instructor decides that she will give 50 of the participants a set of tests before the course begins (a pre-test), and then she will give another set of tests to 50 participants at the end of the course (a post-test).

Example 1 – *Dependent Versus Independent Samples* cont'd

Two sampling procedures are proposed:

Plan A: Randomly select 50 participants from the list of those enrolled and give them the pre-test. At the end of the course, make a second random selection of size 50 and give them the post-test.

Plan B: Randomly select 50 participants and give them the pre-test; give the same set of 50 the post-test when they complete the course.

Example 1 – *Dependent Versus Independent Samples* cont'd

- Plan A illustrates independent sampling; the sources (the class participants) used for each sample (pre-test and post-test) were selected separately.

- Plan B illustrates dependent sampling; the sources used for both samples (pre-test and post-test) are the same.

- Typically, when both a pre-test and a post-test are used, the same subjects participate in the study.
- Thus, pre-test versus post-test (before versus after) studies usually use dependent samples.

Example 2 – *Dependent Versus Independent Samples*

- A test is being designed to compare the wearing quality of two brands of automobile tires.
- The automobiles will be selected and equipped with the new tires and then driven under “normal” conditions for 1 month. Then a measurement will be taken to determine how much wear took place.
- Two plans are proposed:

✓ Plan C: A sample of cars will be selected randomly, equipped with brand A tires, and driven for 1 month. Another sample of cars will be selected, equipped with brand B tires, and driven for 1 month.

Independent (unrelated sources)

✓ Plan D: A sample of cars will be selected randomly, equipped with one tire of brand A and one tire of brand B (the other two tires are not part of the test), and driven for 1 month.

dependent (common sources)

10.2 Inferences Concerning the Mean Difference Using Two Dependent Samples

Inferences Concerning the Mean Difference Using Two Dependent Samples

- When dependent samples are involved, the data are thought of as "paired data."
- The data may be paired as a result of being obtained from "before" and "after" studies or from matching two subjects with similar traits to form "matched pairs."
- The pairs of data values are compared directly to each other by using the difference in their numerical values.
- The resulting difference is called a **paired difference**

Paired Difference

$$d = x_1 - x_2 \tag{10.1}$$

Procedures and Assumptions for Inferences Involving Paired Data

Procedures and Assumptions for Inferences Involving Paired Data

- A test was conducted to compare the wearing quality of the tires produced by two tire companies.
- One tire of each brand was placed on each of six test cars. The position (left or right side, front or back) was determined with the aid of a random-number table.
- Next table lists the amounts of wear (in thousandths of an inch) that resulted from the test.

Car	1	2	3	4	5	6
Brand A	125	64	94	38	90	106
Brand B	133	65	103	37	102	115

- Since the various cars, drivers, and conditions were the same for each tire of a paired set of data, it makes sense to use the paired difference ***d***.

Procedures and Assumptions for Inferences Involving Paired Data

Car	1	2	3	4	5	6
Brand A	125	64	94	38	90	106
Brand B	133	65	103	37	102	115

Our two dependent samples of data may be combined into one set of *d* values, where *d* = B – A.

Car	1	2	3	4	5	6
$d = B - A$	8	1	9	−1	12	9

The difference between the two population means, when dependent samples are used (often called **dependent means**), is equivalent to the **mean of the paired differences**.

Procedures and Assumptions for Inferences Involving Paired Data

In order to make inferences about the mean of all possible paired differences μ_d, we need to know the *sampling distribution* of $\overline{d}$.

When paired observations are randomly selected from normal populations, the paired difference, $d = x_1 - x_2$ will be approximately normally distributed about a mean μ_d with a standard deviation of σ_d.

This is another situation in which the *t*-test for one mean is applied; namely, we wish to make inferences about an unknown mean (μ_d) where the random variable (d) involved has an approximately normal distribution with an unknown standard deviation (σ_d).

Procedures and Assumptions for Inferences Involving Paired Data

Inferences about the mean of all possible paired differences μ_d are based on samples of *n* dependent pairs of data and the ***t*-distribution** with $n - 1$ degrees of freedom (df), under the following assumption:

Assumption for inferences about the mean of paired differences μ_d The paired data are randomly selected from normally distributed populations.

Confidence Interval Procedure

Confidence Interval Procedure

- The $1 - \alpha$ **confidence interval for estimating the mean difference** μ_d is found using this formula:

Confidence Interval for Mean Difference (Dependent Samples)

$$\bar{d} - t(\text{df}, \alpha/2) \cdot \frac{s_d}{\sqrt{n}} \quad \text{to} \quad \bar{d} + t(\text{df}, \alpha/2) \cdot \frac{s_d}{\sqrt{n}}, \text{ where df} = n - 1 \qquad \textbf{(10.2)}$$

- where $\bar{d}$ is the mean of the sample differences:

$$\bar{d} = \frac{\Sigma d}{n} \qquad \textbf{(10.3)}$$

- and s_d is the standard deviation of the sample differences:

$$s_d = \sqrt{\frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n - 1}} \qquad \textbf{(10.4)}$$

Example 4 – *Constructing a Confidence Interval for μ_d*

- Construct the 95% confidence interval for the mean difference in the paired data on tire wear

Car	1	2	3	4	5	6
Brand A	125	64	94	38	90	106
Brand B	133	65	103	37	102	115

- Our two dependent samples of data may be combined into one set of *d* values, where d = B – A.

Car	1	2	3	4	5	6
$d = \text{B} - \text{A}$	8	1	9	−1	12	9

- The sample information is $n = 6$ pieces of paired data, $\bar{d} = 6.3$, and $s_d = 5.1$. Assume the amounts of wear are approximately normally distributed for both brands of tires.

Example 4 – *Solution*

Step 1 **Parameter of interest:** μ_d, the mean difference in the amounts of wear between the two brands of tires

Step 2 **a. Assumptions:** Both sampled populations are approximately normal.

b. Probability distribution: The *t* -distribution with df = 6 – 1 = 5 and formula (10.2) will be used.

c. Level of confidence: $1 - \alpha = 0.95$.

Example 4 – *Solution*

cont'd

Step 3 **Sample information:** n = 6, $\bar{d}$ = 6.3, and s_d = 5.1

The mean:

$$\bar{d} = \frac{\sum d}{n}: \qquad \bar{d} = \frac{38}{6} = 6.333 = \mathbf{6.3}$$

The standard deviation:

$$s_d = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}}: \qquad s_d = \sqrt{\frac{372 - \frac{(38)^2}{6}}{6-1}}$$

$$= \sqrt{26.27}$$

$$= 5.13$$

$$= \mathbf{5.1}$$

Example 4 – *Solution*

cont'd

Step 4 **a. Confidence coefficient:**

This is a two-tailed situation with $\alpha/2 = 0.025$ in one tail. From Table 6 in Appendix B,
$t(\text{df}, \alpha/2) = t(5, 0.025) = 2.57$.

Area in One Tail	0.25	0.10	0.05	0.025	0.01	0.005
Area in Two Tails						
df	**0.50**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**
3	0.765	1.64	2.35	3.18	4.54	5.84
4	0.741	1.53	2.13	2.78	3.75	4.60
5	0.727	1.48	2.02	2.57	3.36	4.03

b. Maximum error of estimate: Using the maximum error part of formula (10.2), we have

$$E = t(\text{df}, \alpha/2) \cdot \frac{s_d}{\sqrt{n}}: \qquad E = 2.57 \cdot \left(\frac{5.1}{\sqrt{6}}\right) = (2.57)(2.082)$$

$$= 5.351$$

$$= \mathbf{5.4}$$

Example 4 – *Solution*

cont'd

c. Lower/upper confidence limits:

$\bar{d} \pm E$:   6.3 $\pm$ 5.4

6.3 – 5.4 = **0.9** to 6.3 + 5.4 = **11.7**

Step 5 **a. Confidence interval:** 0.9 to 11.7 is the 95% confidence interval for μ_d.

b. That is, with 95% confidence we can say that the mean difference in the amounts of wear is between 0.9 and 11.7 thousandths of an inch. Or, in other words, the population mean tire wear for brand B is between 0.9 and 11.7 thousandths of an inch greater than the population mean tire wear for brand A.

Hypothesis-Testing Procedure

Hypothesis-Testing Procedure

- When we test a null **hypothesis about the mean difference,** the test statistic used will be the difference between the sample mean $\bar{d}$ and the hypothesized value of μ_d, divided by the estimated **standard error**.
- The value of the **test statistic** $t\star$ is calculated as follows:

Test Statistic for Mean Difference (Dependent Samples)

$$t\star = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}}, \text{ where df} = n - 1 \qquad (10.5)$$

- **Note**: A hypothesized mean difference, μ_d, can be any specified value. The most common value specified is zero; however, the difference can be nonzero.

Example 6 – *Two-Tailed Hypothesis Test for* $\square_d$

Suppose the sample data in Table 10.1 were collected with the hope of showing that the two tire brands do not wear equally.

Car	1	2	3	4	5	6
Brand A	125	64	94	38	90	106
Brand B	133	65	103	37	102	115

Amount of Tire Wear **[TA10-01]**

Table 10.1

Do the data provide sufficient evidence for us to conclude that the two brands show unequal wear, at the 0.05 level of significance? Assume the amounts of wear are approximately normally distributed for both brands of tires.

Example 6 – *Solution*

Step 1 **a. Parameter of interest:** μ_d, the mean difference (reduction) in pulse rate from before to after using the calcium channel blocker for the time period of the test

b. Statement of hypotheses:

H_0: $\mu_d = 0$ (no difference) Remember: d = B – A

H_a: $\mu_d \neq 0$ (difference)

Step 2 **a. Assumptions:** The assumption of normality is included in the statement of this problem.

Example 6 – *Solution*

cont'd

b. Test statistic: The t-distribution with df = n – 1 = 6 – 1 = 5 and $t\bigstar = (\bar{d} - \mu_d)/(s_d/\sqrt{n})$

c. Level of confidence: α = 0.05.

Step 3 **a. Sample information:** n = 6, $\bar{d}$ = 6.3, and s_d = 5.1

b. Calculated test statistic:

$$t\bigstar = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}}: \qquad t\bigstar = \frac{6.3 - 0.0}{5.1/\sqrt{6}}$$

$$= \frac{6.3}{2.08}$$

$$= \mathbf{3.03}$$

Example 6 – *Solution*

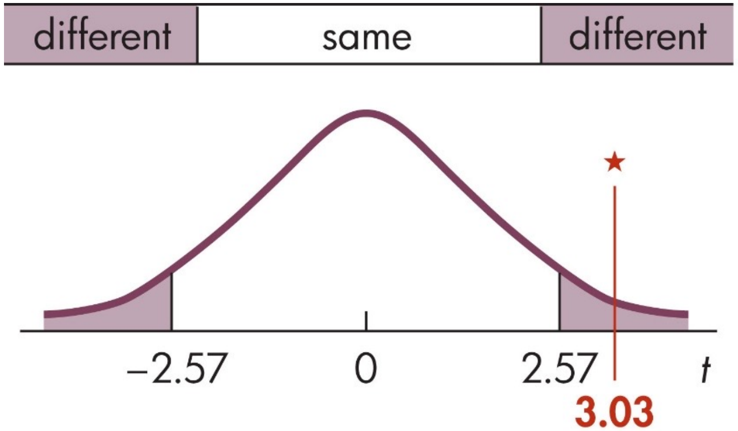
cont'd

Step 4 The Probability Distribution:

Classical:

a. The critical regions are both tails because H_a expresses concern for values related to "different form." The critical value is obtained from Table 6:
$t(5, 0.025) =$ **2.57**.

b. $t_\bigstar$ is in the critical region, as shown in **red** in the figure.



Example 6 – *Solution*

cont'd

Step 5 **a. Decision:** Reject H_o.

b. Conclusion: There is a significant mean difference in the amounts of wear at the 0.05 level of significance.

10.3 Inferences Concerning the Difference between Means Using Two Independent Samples

Inferences Concerning the Difference between Means Using Two Independent Samples

When comparing the means of two populations, we typically consider the difference between their means, $\mu_1 - \mu_2$ (often called "**independent means**").

The inferences about $\mu_1 - \mu_2$ will be based on the difference between the observed sample means, $\bar{x}_1 - \bar{x}_2$.

This observed difference, $\bar{x}_1 - \bar{x}_2$, belongs to a sampling distribution with the characteristics described in the following statement.

Inferences Concerning the Difference between Means Using Two Independent Samples

If independent samples of sizes n_1 and n_2 are drawn randomly from large populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, then the sampling distribution of $\bar{x}_1 - \bar{x}_2$, the difference between the sample means, has

1. mean $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$ and

2. standard error $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\left(\frac{\sigma_1^2}{n_1}\right) + \left(\frac{\sigma_2^2}{n_2}\right)}.$ **(10.6)**

If both populations have normal distributions, then the sampling distribution of $\bar{x}_1 - \bar{x}_2$ will also be normally distributed.

The preceding statement is true for all sample sizes as long as the populations involved are normal and the population variances σ_1^2 and σ_2^2 are known quantities. However, as with inferences about one mean, the variance of a population is generally an unknown quantity.

Therefore, it will be necessary to estimate the standard error by replacing the variances, σ_1^2 and σ_2^2, in formula (10.6) with the best estimates available—namely, the sample variances s_1^2, and s_2^2. The *estimated standard error* will be found using the following formula:

$$\text{estimated standard error} = \sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)} \qquad \textbf{(10.7)}$$

Inferences Concerning the Difference between Means Using Two Independent Samples

Inferences about the difference between two population means, $\mu_1 - \mu_2$, will be based on the following assumptions.

Assumptions for inferences about the difference between two means, $\mu_1 - \mu_2$ The samples are randomly selected from normally distributed populations, and the samples are selected in an independent manner.

No assumptions are made about the population variances.

Inferences Concerning the Difference between Means Using Two Independent Samples

- Since the samples provide the information for determining the standard error, the ***t*-distribution** will be used as the test statistic. The inferences are divided into two cases.

- Case 1: The *t*-distribution will be used, and the number of degrees of freedom will be calculated using computer.
    - ➢ a number between the smaller of $n_1 - 1$ or $n_2 - 1$, and $n_1 + n_2 - 2$

- Case 2: The *t*-distribution will be used, and the number of degrees of freedom will be approximated.
    - ➢ The value of df will be the smaller of $n_1 - 1$ or $n_2 - 1$

Note

$A > B$ ("A is greater than B") is equivalent to $B < A$("B is less than A"). When the difference between A and B is being discussed, it is customary to express the difference as "larger – smaller" so that the resulting difference is positive: $A - B > 0$.

Expressing the difference as "smaller – larger" results in $B - A < 0$ (the difference is negative) and is usually unnecessarily confusing. Therefore, it is recommended that the difference be expressed as "larger – smaller."

Confidence Interval Procedure

Confidence Interval Procedure

We will use the following formula to calculate the end points of the **1 – α confidence interval**.

Confidence Interval for the Difference between Two Means (Independent Samples)

$$(\bar{x}_1 - \bar{x}_2) - t(\text{df}, \alpha/2) \cdot \sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)} \quad \text{to} \quad (\bar{x}_1 - \bar{x}_2) + t(\text{df}, \alpha/2) \cdot \sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)} \qquad \textbf{(10.8)}$$

where df is either calculated or is the smaller of $n_1 - 1$ or $n_2 - 1$.

Example 8 – *Constructing a Confidence Interval for the Difference between Two Means*

The heights (in inches) of 20 randomly selected women and 30 randomly selected men were independently obtained from the student body of a certain college in order to estimate the difference in their mean heights. The sample information is given in Table 10.2.

Sample	Number	Mean	Standard Deviation
Female (*f*)	20	63.8	2.18
Male (*m*)	30	69.8	1.92

Sample Information on Student Heights

Table 10.2

Example 8 – *Constructing a Confidence Interval for the Difference between Two Means*

cont'd

Assume that heights are approximately normally distributed for both populations.

Find the 95% confidence interval for the difference between the mean heights, $\mu_m - \mu_f$.

Solution:

Step 1 **Parameter of interest:** $\mu_m - \mu_f$, the difference between the mean height of male students and the mean height of female students

Example 8 – *Solution*

Step 2 **a. Assumptions:** Both populations are approximately normal, and the samples were randomly and independently selected.

b. Probability distribution: The *t* -distribution with df = 19, the smaller of $n_m - 1 = 30 - 1 = 29$ or $n_f - 1 = 20 - 1 = 19$ and formula (10.8)

c. Level of confidence: $1 - \alpha = 0.95$

Step 3 **Sample information:** See Table 10.2.

Sample	Number	Mean	Standard Deviation
Female (f)	20	63.8	2.18
Male (m)	30	69.8	1.92

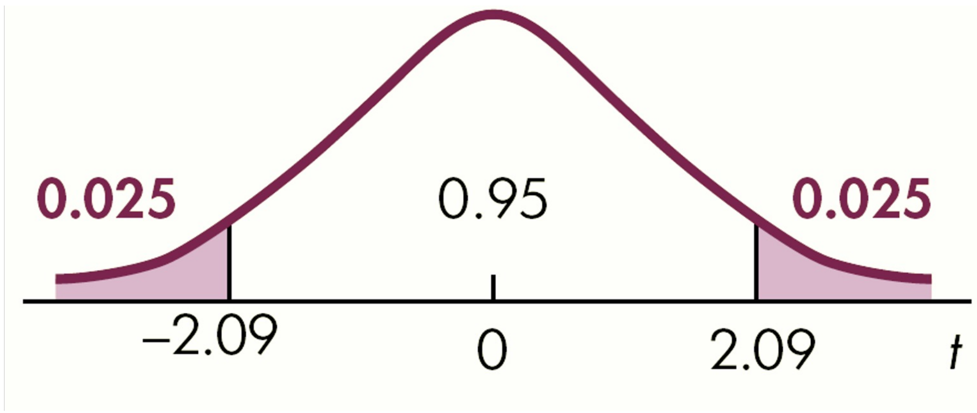
Sample Information on Student Heights

Table 10.2

Example 8 – *Solution*

cont'd

Step 4 **a. Confidence coefficient:** We have a two-tailed situation with $\alpha/2 = 0.025$ in one tail and df = 19.



From Table 6 in Appendix B,
$t(\text{df}, \alpha/2) = t(19, 0.025) = 2.09$. See the figure.

Area in One Tail

	0.25	0.10	0.05	0.025	0.01	0.005
Area in Two Tails						
df	**0.50**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**
16	0.690	1.34	1.75	2.12	2.58	2.92
17	0.689	1.33	1.74	2.11	2.57	2.90
18	0.688	1.33	1.73	2.10	2.55	2.88
19	0.688	1.33	1.73	2.09	2.54	2.86
20	0.687	1.33	1.72	2.09	2.53	2.85

Example 8 – *Solution*

cont'd

b. Maximum error of estimate: Using the maximum error part of formula (10.8), we have

$$E = t(\text{df}, \alpha/2) \cdot \sqrt{\left(\frac{s_m^2}{n_m}\right) + \left(\frac{s_f^2}{n_f}\right)}: \quad E = 2.09 \cdot \sqrt{\left(\frac{1.92^2}{30}\right) + \left(\frac{2.18^2}{20}\right)}$$

$$= (2.09)(0.60) = \mathbf{1.25}$$

c. Lower and upper confidence limits:

$$\left(\bar{x}_m - \bar{x}_f\right) \pm E$$

(69.8 - 63.8) □ 1.25

6.00 – 1.25 = **4.75** to 6.00 + 1.25 = **7.25**

Example 8 – *Solution*

cont'd

Step 5 **a. Confidence interval.**

4.75 to 7.25 is the 95% confidence interval for $\mu_m - \mu_f$.

b. That is, with 95% confidence, we can say that the difference between the mean heights of the male and female students is between 4.75 and 7.25 inches; that is, the mean height of male students is between 4.75 and 7.25 inches greater than the mean height of female students.

Hypothesis-Testing Procedure

Hypothesis-Testing Procedure

When we test a null **hypothesis about the difference between two population means**, the test statistic used will be the difference between the observed difference of the sample means and the hypothesized difference of the population means, divided by the estimated standard error.

The test statistic is assumed to have approximately a *t*-distribution when the null hypothesis is true and the normality assumption has been satisfied.

Hypothesis-Testing Procedure

The calculated value of the **test statistic** is found using this formula:

Test Statistic for the Difference between Two Means (Independent Samples)

$$t\star = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}} \qquad \textbf{(10.9)}$$

where df is either calculated or is the smaller of $n_1 - 1$ or $n_2 - 1$.

Note: A hypothesized difference between the two population means, $\mu_1 - \mu_2$, can be any specified value. The most common value specified is zero; however, the difference can be nonzero.

Suppose that we are interested in comparing the academic success of college students who belong to fraternal organizations with the academic success of those who do not belong to fraternal organizations.

The reason for the comparison is the recent concern that fraternity members, on average, are achieving at a lower academic level than nonfraternal students achieve. (Cumulative GPA is used to measure academic success.) Random samples of size 40 are taken from each population.

The sample results are listed in Table 10.3.

Sample	Number	Mean	Standard Deviation
Fraternity members (*f*)	40	2.03	0.68
Nonmembers (*n*)	40	2.21	0.59

Sample Information on Academic Success

Table 10.3

Complete a hypothesis test using $\alpha = 0.05$. Assume that the GPAs for both groups are approximately normally distributed.

Example 9 – *Solution*

Step 1 **a. Parameter of interest:** $\mu_n - \mu_f$, the difference between the mean GPAs for the nonfraternity members and the fraternity members

b. Statement of hypotheses:

H_0: $\mu_n - \mu_f = 0$ ($\leq$)(fraternity averages are no lower)

H_a: $\mu_n - \mu_f > 0$ (fraternity averages are lower)

Step 2 **a. Assumptions:** Both populations are approximately normal, and random samples were selected. Since the two populations are separate, the samples are independent.

Example 9 – *Solution*

cont'd

b. Test statistic: The t-distribution with df = the smaller of $n_n - 1$ or $n_f - 1$; since both n's are 40, df = 40 – 1 = **39**; and $t\star$ is calculated using formula (10.9)

c. Level of significance: $\alpha = 0.05$

Step 3 **a. Sample information:** See Table 10.3.

Sample	Number	Mean	Standard Deviation
Fraternity members (f)	40	2.03	0.68
Nonmembers (n)	40	2.21	0.59

Sample Information on Academic Success

Table 10.3

Example 9 – *Solution*

cont'd

b. Calculated test statistic:

Sample	Number	Mean	Standard Deviation
Fraternity members (*f*)	40	2.03	0.68
Nonmembers (*n*)	40	2.21	0.59

$$t^* = \frac{(\bar{x}_n - \bar{x}_f) - (\mu_n - \mu_f)}{\sqrt{\left(\frac{s_n^2}{n_n}\right) + \left(\frac{s_f^2}{n_f}\right)}}:$$

$$t\bigstar = \frac{(2.21 - 2.03) - (0.00)}{\sqrt{\left(\frac{0.59^2}{40}\right) + \left(\frac{0.68^2}{40}\right)}}$$

$$= \frac{0.18}{\sqrt{0.00870 + 0.01156}}$$

$$= \frac{0.18}{0.1423}$$

$$= \mathbf{1.26}$$

Example 9 – *Solution*

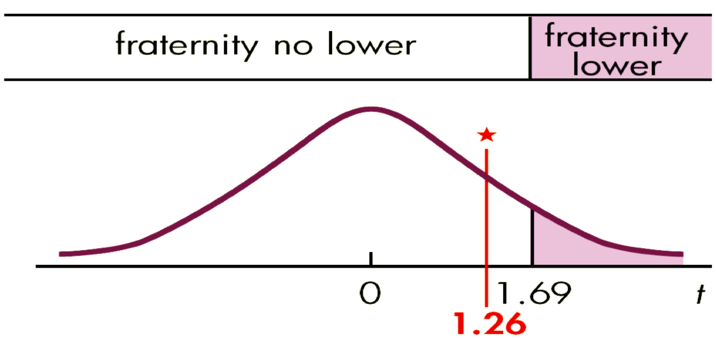
cont'd

Step 4 Probability Distribution:

Classical:

a. The critical region is the right-hand tail because H_a expresses concern for values related to "greater than." The critical value is obtained from Table 6:

$t(39, 0.05) = \mathbf{1.69}$.



Area in One Tail	0.25	0.10	0.05	0.025	0.01	0.005
Area in Two Tails						
df	**0.50**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**
28	0.683	1.31	1.70	2.05	2.47	2.76
29	0.683	1.31	1.70	2.05	2.46	2.76
30	0.683	1.31	1.70	2.04	2.46	2.75
35	0.682	1.31	1.69	2.03	2.44	2.72
40	0.681	1.30	1.68	2.02	2.42	2.70
50	0.679	1.30	1.68	2.01	2.40	2.68
70	0.678	1.29	1.67	1.99	2.38	2.65
100	0.677	1.29	1.66	1.98	2.36	2.63
df > 100	0.675	1.28	1.65	1.96	2.33	2.58

b. $t\bigstar$ is not in the critical region, as shown in **red** on the figure.

Example 9 – *Solution*

cont'd

Step 5 **a. Decision:** Fail to reject H_o.

b. Conclusion: At the 0.05 level of significance, the claim that the fraternity members achieve at a lower level than nonmembers is not supported by the sample data.