9.12 Independent random samples selected from two normal populations produced the following sample means and standard deviations:

Sample 1	Sample 2
$n_1 = 17$	$n_2 = 12$
$\bar{x}_1 = 5.4$	$\bar{x}_2 = 7.9$
$s_1 = 3.4$	$s_2 = 4.8$

a. Assuming equal variances, conduct the test $H_0: (\mu_1 - \mu_2) = 0$ against $H_a: (\mu_1 - \mu_2) \neq 0$ using $\alpha = .05$. [redacted]

b. Find and interpret the 95% confidence interval for $(\mu_1 - \mu_2)$. [redacted]

Applying the Concepts—Basic

9.13 Shared leadership in airplane crews. *Human Factors* (Mar. 2014) published a study that examined the effect of shared leadership by the cockpit and cabin crews of a commercial airplane. Simulated flights were taken by 84 six-person crews, where each crew consisted of a two-person cockpit (captain and first officer) and a four-person cabin team (three flight attendants and a purser). During the simulation, smoke appeared in the cabin and the reactions of the crew were monitored for teamwork. Each crew was rated as working either successfully or unsuccessfully as a team. Also, each individual member was evaluated for leadership (measured as number of leadership functions exhibited per minute). The mean leadership values for successful and unsuccessful teams were compared. A summary of the test results for both captains and lead flight attendants is displayed in the table at the bottom of the page.

a. Consider the data for captains. Interpret the p-value for testing (at $\alpha = .05$) whether the mean leadership values for captains from successful and unsuccessful teams differ. reject H_0

b. Consider the data for flight attendants. Interpret the p-value for testing (at $\alpha = .05$) whether the mean leadership values for flight attendants from successful and unsuccessful teams differ. Fail to reject H_0

9.14 Last name and acquisition timing. The speed with which consumers decide to purchase a product was investigated in the *Journal of Consumer Research* (Aug. 2011). The researchers theorized that consumers with last names that begin with letters later in the alphabet will tend to acquire items faster than those whose last names are earlier in the alphabet—called the *last name effect*. MBA students were offered up to four free tickets to attend a top-ranked women's college basketball game for which there were a limited supply of tickets. The first letter of the last name of those who responded to an e-mail offer in time to receive the tickets was noted as well as the response time (measured in minutes). The researchers compared the response times for two groups of MBA students: (1) those with last names beginning with one of the first 9 letters of the alphabet and (2) those with last names beginning with one of the last 9 letters of the alphabet. Summary statistics for the two groups are provided in the table.

	First 9 Letters: A–I	Last 9 Letters: R–Z
Sample size	25	25
Mean response time (minutes)	25.08	19.38
Standard deviation (minutes)	10.41	7.12

Source: Carlson, K. A., & Conrad, J. M. "The last name effect: How last name influences acquisition timing." *Journal of Consumer Research*, Vol. 38, No. 2, Aug. 2011.

a. Construct a 95% confidence interval for the difference between the true mean response times for MBA students in the two groups. [redacted]

b. Based on the interval, part **a**, which group has the shortest mean response time? Does this result support the researchers' *last name effect* theory? Explain. [redacted]

9.15 Effectiveness of teaching software. The U.S. Department of Education (DOE) conducted a national study of the effectiveness of educational software. In one phase of the study, a sample of 1,516 first-grade students in classrooms that used educational software was compared to a sample of 1,103 first-grade students in classrooms that did not use the technology. In its *Report to Congress* (Mar. 2007), the DOE concluded that "[mean] test scores [of students on the SAT reading test] were not significantly higher in classrooms using reading ... software products" than in classrooms that did not use educational software.

a. Identify the parameter of interest to the DOE. $\mu_1 - \mu_2$

b. Specify the null and alternative hypotheses for the test conducted by the DOE. $H_0: \mu_1 - \mu_2 = 0$; $H_a: \mu_1 - \mu_2 > 0$

c. The p-value for the test was reported as .62. Based on this value, do you agree with the conclusion of the DOE? Explain. Yes; fail to reject H_0

9.16 Cognitive impairment of schizophrenics. A study of the differences in cognitive function between normal individuals and patients diagnosed with schizophrenia was published in the *American Journal of Psychiatry* (Apr. 2010). The total time (in minutes) a subject spent on the Trail Making Test (a standard psychological test) was used as a measure of cognitive function. The researchers theorize that the mean time on the Trail Making Test for schizophrenics will be larger than the corresponding mean for normal subjects. The data for independent random

Table for Exercise 9.13

	Successful Teams ($n = 60$)		Unsuccessful Teams ($n = 24$)			
	Mean	Std. Dev.	Mean	Std. Dev.	t-value	p-value
Captain	.66	.10	.50	.20	3.72	.000
Flight Att.	.40	.24	.39	.13	0.12	.907

Source: Bienefeld, N., & Grote, G. "Shared leadership in multiteam systems: How cockpit and cabin crews lead each other to safety." *Human Factors*, Vol. 65, No. 2, Mar. 2014 (Table 2).

samples of 41 schizophrenics and 49 normal individuals yielded the following results:

	Schizophrenia	Normal
Sample size	41	49
Mean time	104.23	62.24
Standard deviation	45.45	16.34

Based on Perez-Iglesias, R., et al. "White matter integrity and cognitive impairment in first-episode psychosis." *American Journal of Psychiatry*, Vol. 167, No. 4, Apr. 2010 (Table 1).

a. Define the parameter of interest to the researchers.
b. Set up the null and alternative hypothesis for testing the researchers' theory.
c. The researchers conducted the test, part b, and reported a p-value of .001. What conclusions can you draw from this result? (Use $\alpha = .01$.)
d. Find a 99% confidence interval for the target parameter. Interpret the result. Does your conclusion agree with that of part c?

9.17 Children's recall of TV ads. Children's recall and recognition of television advertisements was studied in the *Journal of Advertising* (Spring 2006). Two groups of children were shown a 60-second commercial for Sunkist FunFruit Rock-n-Roll Shapes. One group (the A/V group) was shown the ad with both audio and video; the second group (the video-only group) was shown only the video portion of the commercial. Following the viewing, the children were asked to recall 10 specific items from the ad. The number of items recalled correctly by each child is summarized in the accompanying table. The researchers theorized that "children who receive an audiovisual presentation will have the same level of mean recall of ad information as those who receive only the visual aspects of the ad."

Video-Only Group	A/V Group
$n_1 = 20$	$n_2 = 20$
$\bar{x}_1 = 3.70$	$\bar{x}_2 = 3.30$
$s_1 = 1.98$	$s_2 = 2.13$

Based on Maher, J. K., Hu, M. Y., and Kolbe, R. H. "Children's recall of television ad elements." *Journal of Advertising*, Vol. 35, No. 1, Spring 2006 (Table 1).

a. Set up the appropriate null and alternative hypotheses to test the researchers' theory.
b. Find the value of the test statistic.
c. Give the rejection region for $\alpha = .10$.
d. Make the appropriate inference. What can you say about the researchers' theory?
e. The researchers reported the p-value of the test as p-value $= .62$. Interpret this result.
f. What conditions are required for the inference to be valid?

9.18 Comparing taste test rating protocols. Taste testers of new food products are presented with several competing food samples and asked to rate the taste of each on a 9-point scale (where 1 = "dislike extremely" and 9 = "like extremely"). In the *Journal of Sensory Studies* (June 2014), food scientists compared two different taste testing protocols. The sequential monadic (SM) method presented the samples one at a time to the taster in a random order, while the rank rating (RR) method presented the samples to the taster all at once side by side. In one experiment, 108 consumers of peach jam were asked to taste-test five different varieties. Half the testers used the SM protocol, and half used the RR protocol during testing. In a second experiment, 108 consumers of cheese were asked to taste-test four different varieties. Again, half the testers used the SM protocol, and half used the RR protocol during testing. For each product (peach jam and cheese), the mean taste scores of the two protocols (SM and RR) were compared. The results are shown in the accompanying tables.

***Peach jam* (taste means)**

Variety	RR	SM	p-value
A	5.8	5.5	0.503
B	5.9	5.7	0.665
C	4.4	4.4	0.962
D	5.9	5.6	0.414
E	5.4	5.1	0.535

***Cheese* (taste means)**

Variety	RR	SM	p-value
A	6.3	5.5	0.017
B	6.8	6.2	0.076
C	6.6	5.7	0.001
D	6.5	5.7	0.034

Source: Gutierrez-Salomon, A. L., Gambaro, A., and Angulo, O. "Influence of sample presentation protocol on the results of consumer tests." *Journal of Sensory Studies*, Vol. 29, No. 3, June 2014 (Tables 1 and 4).

a. Consider the five varieties of peach jam. Identify the varieties for which you can conclude that "the mean taste scores of the two protocols (SM and RR) differ significantly at $\alpha = .05$."
b. Consider the four varieties of cheese. Identify the varieties for which you can conclude that "the mean taste scores of the two protocols (SM and RR) differ significantly at $\alpha = .05$."
c. Explain why the taste test scores do not need to be normally distributed in order for the inferences, parts a and b, to be valid.

9.19 Bulimia study. The "fear of negative evaluation" (FNE) scores for 11 female students known to suffer from the eating disorder bulimia and 14 female students with normal eating habits, first presented in Exercise 2.44 (p. 80), are reproduced in the next table. (Recall that the higher the score, the greater is the fear of a negative evaluation.)

D BULIMIA

Bulimic students:	21 13 10 20 25 19 16 21 24 13 14
Normal students:	13 6 16 13 8 19 23 18 11 19 7 10 15 20

Based on Randles, R. H. "On neutral responses (zeros) in the sign test and ties in the Wilcoxon-Mann-Whitney test." *The American Statistician*, Vol. 55, No. 2, May 2001 (Figure 3).

a. Locate a 95% confidence interval for the difference between the population means of the FNE scores for bulimic and normal female students on the MINITAB printout shown on page 478. Interpret the result.

SPSS performed these calculations and obtained the interval ($−10,537.50, $11,337.50), highlighted in Figure 9.14.

Notice that the independent samples interval includes 0. Consequently, if we were to use this interval to make an inference about $(\mu_1 - \mu_2)$, we would incorrectly conclude that the mean starting salaries of males and females do not differ! You can see that the confidence interval for the independent sampling experiment is about 35 times wider than for the corresponding paired difference confidence interval. Blocking out the variability due to differences in majors and grade point averages significantly increases the information about the difference in males' and females' mean starting salaries by providing a much more accurate (a smaller confidence interval for the same confidence coefficient) estimate of $(\mu_1 - \mu_2)$.

You may wonder whether a paired difference experiment is always superior to an independent samples experiment. The answer is, most of the time, but not always. We sacrifice half the degrees of freedom in the t-statistic when a paired difference design is used instead of an independent samples design. This is a loss of information, and unless that loss is more than compensated for by the reduction in variability obtained by blocking (pairing), the paired difference experiment will result in a net loss of information about $(\mu_1 - \mu_2)$. Thus, we should be convinced that the pairing will significantly reduce variability before performing a paired difference experiment. Most of the time, this will happen.

One final note: The pairing of the observations is determined *before* the experiment is performed (i.e., by the *design* of the experiment). A paired difference experiment is *never* obtained by pairing the sample observations *after* the measurements have been acquired.

Teaching Tip

Point out that the paired difference experiment is a useful tool in statistics and should be attempted whenever appropriate. The disadvantage discussed here is small relative to the large benefits the experiment can provide.

Ethics IN Statistics

In a two-group analysis, intentionally pairing observations after the data have been collected in order to produce a desired result is considered *unethical statistical practice*.

Teaching Tip

In general the nonparametric analyses of Chapter 14 are useful whenever the assumptions are not satisfied.

What Do You Do When the Assumption of a Normal Distribution for the Population of Differences Is Not Satisfied?

Answer: Use the Wilcoxon signed rank test for the paired difference design (Chapter 14).

Exercises 9.31–9.54

Understanding the Principles

9.31 In a paired difference experiment, when should the observations be paired, before or after the data are collected?

9.32 What are the advantages of using a paired difference experiment over an independent samples design?

9.33 *True or False.* In a paired difference experiment, $\bar{x}_d = \bar{x}_1 - \bar{x}_2$. True

9.34 What conditions are required for valid large-sample inferences about μ_d? small-sample inferences?

Learning the Mechanics

9.35 NW A paired difference experiment yielded n_d pairs of observations. In each case, what is the rejection region for testing H_0: $\mu_d = 2$ against H_a: $\mu_d > 2$?

a. $n_d = 10, \alpha = .05$ $t > 1.833$
b. $n_d = 20, \alpha = .10$ $t > 1.328$
c. $n_d = 5, \alpha = .025$ $t > 2.776$
d. $n_d = 9, \alpha = .01$ $t > 2.896$

9.36 A paired difference experiment produced the following data:

$$n_d = 16 \quad \bar{x}_1 = 143 \quad \bar{x}_2 = 150 \quad \bar{x}_d = -7 \quad s_d^2 = 64$$

a. Determine the values of t for which the null hypothesis $\mu_1 - \mu_2 = 0$ would be rejected in favor of the alternative hypothesis $\mu_1 - \mu_2 < 0$. Use $\alpha = .10$. [redacted]

b. Conduct the paired difference test described in part **a.** Draw the appropriate conclusions. [redacted]

c. What assumptions are necessary so that the paired difference test will be valid?

d. Find a 90% confidence interval for the mean difference μ_d. [redacted]

e. Which of the two inferential procedures, the confidence interval of part **d** or the test of hypothesis of part **b**, provides more information about the difference between the population means?

9.37 NW D L09037 The data for a random sample of six paired observations are shown in the following table.

Pair	Sample from Population 1	Sample from Population 2
1	7	4
2	3	1
3	9	7
4	6	2
5	4	4
6	8	7

a. Calculate the difference between each pair of observations by subtracting observation 2 from observation 1. Use the differences to calculate $\bar{x}_d$ and s_d^2.

b. If μ_1 and μ_2 are the means of populations 1 and 2, respectively, express μ_d in terms of μ_1 and μ_2. $\mu_d = \mu_1 - \mu_2$

c. Form a 95% confidence interval for μ_d. 2 − 1.484

d. Test the null hypothesis $H_0: \mu_d = 0$ against the alternative hypothesis $H_a: \mu_d \neq 0$. Use $\alpha = .05$.

9.38 The data for a random sample of 10 paired observations are shown in the following table.
L09038

Pair	Population 1	Population 2
1	19	24
2	25	27
3	31	36
4	52	53
5	49	55
6	34	34
7	59	66
8	47	51
9	17	20
10	51	55

a. If you wish to test whether these data are sufficient to indicate that the mean for population 2 is larger than that for population 1, what are the appropriate null and alternative hypotheses? Define any symbols you use.

b. Conduct the test from part **a**, using $\alpha = .10$. What is your decision?

c. Find a 90% confidence interval for μ_d. Interpret this interval.

d. What assumptions are necessary to ensure the validity of the preceding analysis?

9.39 A paired difference experiment yielded the following results:

$$n_d = 49, \ \bar{x}_d = 13, \ s_d = 6.$$

a. Test $H_0: \mu_d = 10$ against $H_a: \mu_d \neq 10$, where $\mu_d = (\mu_1 - \mu_2)$. Use $\alpha = .05$.

b. Report the p-value for the test you conducted in part **a**. Interpret the p-value.

Applying the Concepts—Basic

9.40 Summer weight-loss camp. Camp Jump Start is an 8-week summer camp for overweight and obese adolescents. Counselors develop a weight-management program for each camper that centers on nutrition education and physical activity. In a study published in *Pediatrics* (Apr. 2010), the body mass index (BMI) was measured for each of 76 campers both at the start and end of camp. Summary statistics on BMI measurements are shown in the table.

	Mean	Standard Deviation
Starting BMI	34.9	6.9
Ending BMI	31.6	6.2
Paired Differences	3.3	1.5

Based on Huelsing, J., Kanafani, N., Mao, J., and White, N. H. "Camp Jump Start: Effects of a residential summer weight-loss camp for older children and adolescents." *Pediatrics*, Vol. 125, No. 4, Apr. 2010 (Table 3).

a. Give the null and alternative hypothesis for determining whether the mean BMI at the end of camp is less than the mean BMI at the start of camp.

b. How should the data be analyzed, as an independent-samples t-test or as a paired-difference t-test? Explain.

c. Calculate the test statistic using the formula for an independent-samples t-test. (*Note:* This is *not* how the test should be conducted.)

d. Calculate the test statistic using the formula for a paired-difference t-test.

e. Compare the test statistics, parts **c** and **d**. Which test statistic provides more evidence in support of the alternative hypothesis?

f. The p-value of the test, part **d**, was reported as $p < .0001$. Interpret this result assuming $\alpha = .01$.

g. Do the differences in BMI values need to be normally distributed in order for the inference, part **f**, to be valid? Explain.

h. Find a 99% confidence interval for the true mean change in BMI for Camp Jump Start campers. Interpret the result.

9.41 Packaging of a children's health food. Refer to the *Journal of Consumer Behaviour* (Vol. 10, 2011) study of packaging of a children's health food product, Exercise 8.42 (p. 419). Recall that a fictitious brand of a healthy food product—sliced apples—was packaged to appeal to children (a smiling cartoon apple on the front of the package). The researchers compared the appeal of this fictitious brand to a commercially available brand of sliced apples that was not packaged for children. Each of 408 schoolchildren rated both brands on a 5-point "willingness to eat" scale, with 1 = "not willing at all" and 5 = "very willing." The fictitious brand had a sample mean score of 3.69, while the commercially available brand had a sample mean score of 3.00. The researchers wanted to compare the population mean score for the fictitious brand, μ_F, to the population mean score for the commercially available brand, μ_C. They theorized that μ_F will be greater than μ_C.

a. Specify the null and alternative hypothesis for the test.

b. Explain how the researchers should analyze the data and why.

c. The researchers reported a test statistic value of 5.71. Interpret this result. Use $\alpha = .05$ to draw your conclusion.

d. Find the approximate p-value of the test.

e. Could the researchers have tested at $\alpha = .01$ and arrived at the same conclusion?

9.42 Twinned drill holes. A traditional method of verifying mineralization grades in mining is to drill twinned holes, i.e., the drilling of a new hole, or "twin," next to an earlier drillhole. The use of twinned drill holes was investigated in *Exploration and Mining Geology* (Vol. 18, 2009). Geologists use data collected at both holes to estimate the total amount of heavy minerals (THM) present at the drilling site. The data in the next table (based on information provided in the journal article) represent THM percentages for a sample of 15 twinned holes drilled at a diamond mine in Africa. The geologists want to know if there is any evidence of a difference in the true THM means of all original holes and their twin holes drilled at the mine.
DRILL2

a. Explain why the data should be analyzed as paired differences.

b. Compute the difference between the "1st hole" and "2nd hole" measurements for each drilling location.

c. Find the mean and standard deviation of the differences, part **b**.