Understanding Independent Samples *t*-Test

EXERCISE 16

STATISTICAL TECHNIQUE IN REVIEW

The **independent samples *t*-test** is a parametric statistical technique used to determine significant differences between the scores obtained from two samples or groups. Since the *t*-test is considered fairly easy to calculate, researchers often use it in determining differences between two groups. The *t*-test examines the differences between the means of the two groups in a study and adjusts that difference for the variability (computed by the standard error) among the data. When interpreting the results of *t*-tests, the larger the calculated *t* ratio, in absolute value, the greater the difference between the two groups. The significance of a *t* ratio can be determined by comparison with the critical values in a statistical table for the *t* distribution using the degrees of freedom (*df*) for the study (see Appendix A Critical Values for Student's *t* Distribution at the back of this text). The formula for *df* for an independent *t*-test is as follows:

$$df = (\text{number of subjects in sample 1} + \text{number of subjects in sample 2}) - 2$$

$$\text{Example } df = (65 \text{ in sample 1} + 67 \text{ in sample 2}) - 2 = 132 - 2 = 130$$

The *t*-test should be conducted only once to examine differences between two groups in a study, because conducting multiple *t*-tests on study data can result in an inflated Type 1 error rate. A Type I error occurs when the researcher rejects the null hypothesis when it is in actuality true. Researchers need to consider other statistical analysis options for their study data rather than conducting multiple *t*-tests. However, if multiple *t*-tests are conducted, researchers can perform a Bonferróni procedure or more conservative post hoc tests like Tukey's honestly significant difference (HSD), Student-Newman-Keuls, or Scheffé test to reduce the risk of a Type I error. Only the Bonferroni procedure is covered in this text; details about the other, more stringent post hoc tests can be found in Plichta and Kelvin (2013) and Zar (2010).

The Bonferroni procedure is a simple calculation in which the alpha is divided by the number of *t*-tests conducted on different aspects of the study data. The resulting number is used as the alpha or level of significance for each of the *t*-tests conducted. The Bonferroni procedure formula is as follows: alpha (α) ÷ number of *t*-tests performed on study data = more stringent study α to determine the significance of study results. For example, if a study's α was set at 0.05 and the researcher planned on conducting five *t*-tests on the study data, the α would be divided by the five *t*-tests ($0.05 \div 5 = 0.01$), with a resulting α of 0.01 to be used to determine significant differences in the study.

The *t*-test for independent samples or groups includes the following assumptions:

1. The raw scores in the population are normally distributed.
2. The dependent variable(s) is(are) measured at the interval or ratio levels.

3. The two groups examined for differences have equal variance, which is best achieved by a random sample and random assignment to groups.
4. All scores or observations collected within each group are independent or not related to other study scores or observations.

The *t*-test is robust, meaning the results are reliable even if one of the assumptions has been violated. However, the *t*-test is not robust regarding between-samples or within-samples independence assumptions or with respect to extreme violation of the assumption of normality. Groups do not need to be of equal sizes but rather of equal variance. Groups are independent if the two sets of data were not taken from the same subjects and if the scores are not related (Grove, Burns, & Gray, 2013; Plichta & Kelvin, 2013). This exercise focuses on interpreting and critically appraising the *t*-tests results presented in research reports. Exercise 31 provides a step-by-step process for calculating the independent samples *t*-test.

RESEARCH ARTICLE

Source

Canbulat, N., Ayhan, F., & Inal, S. (2015). Effectiveness of external cold and vibration for procedural pain relief during peripheral intravenous cannulation in pediatric patients. *Pain Management Nursing, 16*(1), 33–39.

Introduction

Canbulat and colleagues (2015, p. 33) conducted an experimental study to determine the "effects of external cold and vibration stimulation via Buzzy on the pain and anxiety levels of children during peripheral intravenous (IV) cannulation." Buzzy is an 8 × 5 × 2.5 cm battery-operated device for delivering external cold and vibration, which resembles a bee in shape and coloring and has a smiling face. A total of 176 children between the ages of 7 and 12 years who had never had an IV insertion before were recruited and randomly assigned into the equally sized intervention and control groups. During IV insertion, "the control group received no treatment. The intervention group received external cold and vibration stimulation via Buzzy...Buzzy was administered about 5 cm above the application area just before the procedure, and the vibration continued until the end of the procedure" (Canbulat et al., 2015, p. 36). Canbulat et al. (2015, pp. 37–38) concluded that "the application of external cold and vibration stimulation were effective in relieving pain and anxiety in children during peripheral IV" insertion and were "quick-acting and effective nonpharmacological measures for pain reduction." The researchers concluded that the Buzzy intervention is inexpensive and can be easily implemented in clinical practice with a pediatric population.

Relevant Study Results

The level of significance for this study was set at $\alpha = 0.05$. "There were no differences between the two groups in terms of age, sex [gender], BMI, and preprocedural anxiety according to the self, the parents', and the observer's reports ($p > 0.05$) (Table 1). When the pain and anxiety levels were compared with an independent samples *t* test,...the children in the external cold and vibration stimulation [intervention] group had significantly lower pain levels than the control group according to their self-reports (both WBFC [Wong Baker Faces Scale] and VAS [visual analog scale] scores; $p < 0.001$) (Table 2). The external cold and vibration stimulation group had significantly lower fear and anxiety

levels than the control group, according to parents' and the observer's reports ($p < 0.001$) (Table 3)" (Canbulat et al., 2015, p. 36).

TABLE 1 COMPARISON OF GROUPS IN TERMS OF VARIABLES THAT MAY AFFECT PROCEDURAL PAIN AND ANXIETY LEVELS

Characteristic	Buzzy (n = 88)	Control (n = 88)	χ^2 p
Sex			
Female (%), n	11 (12.5)	13 (14.8)	.82
Male (%), n	77 (87.5)	75 (85.2)	.41
Characteristic	**Buzzy (n = 88)**	**Control (n = 88)**	t p
Age (mean ± *SD*)	8.25 ± 1.51	8.61 ± 1.69	−1.498 .136
BMI (mean ± *SD*)	25.41 ± 6.74	26.94 ± 8.68	−1.309 .192
Preprocedural anxiety			
Self-report (mean ± *SD*)	2.03 ± 1.29	2.11 ± 1.58	−0.364 .716
Parent report (mean ± *SD*)	2.11 ± 1.20	2.17 ± 1.42	−0.285 .776
Observer report (mean ± *SD*)	2.18 ± 1.17	2.24 ± 1.37	−0.295 .768

BMI, body mass index.

Canbulat, N., Ayban, F., & Inal, S. (2015). Effectiveness of external cold and vibration for procedural pain relief during peripheral intravenous cannulation in pediatric patients. *Pain Management Nursing, 16*(1), p. 36.

TABLE 2 COMPARISON OF GROUPS' PROCEDURAL PAIN LEVELS DURING PERIPHERAL IV CANNULATION

	Buzzy (n = 88)	Control (n = 88)	t p
Procedural self-reported pain with WBFS (mean ± *SD*)	2.75 ± 2.68	5.70 ± 3.31	−6.498 0.000
Procedural self-reported pain with VAS (mean ± *SD*)	1.66 ± 1.95	4.09 ± 3.21	−6.065 0.000

IV, intravenous; WBFS, Wong-Baker Faces Scale; *SD*, standard deviation; VAS, visual analog scale.

Canbulat, N., Ayban, F., & Inal, S. (2015). Effectiveness of external cold and vibration for procedural pain relief during peripheral intravenous cannulation in pediatric patients. *Pain Management Nursing, 16*(1), p. 37.

TABLE 3 COMPARISON OF GROUPS' PROCEDURAL ANXIETY LEVELS DURING PERIPHERAL IV CANNULATION

Procedural Child Anxiety	Buzzy (n = 88)	Control (n = 88)	t p
Parent reported (mean ± *SD*)	0.94 ± 1.06	2.09 ± 1.39	−6.135 0.000
Observer reported (mean ± *SD*)	0.92 ± 1.03	2.14 ± 1.34	−6.745 0.000

SD, standard deviation; IV, intravenous.

Canbulat, N., Ayban, F., & Inal, S. (2015). Effectiveness of external cold and vibration for procedural pain relief during peripheral intravenous cannulation in pediatric patients. *Pain Management Nursing, 16*(1), p. 37.

STUDY QUESTIONS

1. What type of statistical test was conducted by Canbulat et al. (2015) to examine group differences in the dependent variables of procedural pain and anxiety levels in this study? What two groups were analyzed for differences?

2. What did Canbulat et al. (2015) set the level of significance, or alpha (α), at for this study?

3. What are the *t* and *p* (probability) values for procedural self-reported pain measured with a visual analog scale (VAS)? What do these results mean?

4. What is the null hypothesis for observer-reported procedural anxiety for the two groups? Was this null hypothesis accepted or rejected in this study? Provide a rationale for your answer.

5. What is the *t*-test result for BMI? Is this result statistically significant? Provide a rationale for your answer. What does this result mean for the study?

6. What causes an increased risk for Type I errors when *t*-tests are conducted in a study? How might researchers reduce the increased risk for a Type I error in a study?

7. Assuming that the *t*-tests presented in Table 2 and Table 3 are all the *t*-tests performed by Canbulat et al. (2015) to analyze the dependent variables' data, calculate a Bonferroni procedure for this study.

8. Would the *t*-test for observer-reported procedural anxiety be significant based on the more stringent α calculated using the Bonferroni procedure in question 7? Provide a rationale for your answer.

9. The results in Table 1 indicate that the Buzzy intervention group and the control group were not significantly different for gender, age, body mass index (BMI), or preprocedural anxiety (as measured by self-report, parent report, or observer report). What do these results indicate about the equivalence of the intervention and control groups at the beginning of the study? Why are these results important?

10. Canbulat et al. (2015) conducted the χ^2 test to analyze the difference in sex or gender between the Buzzy intervention group and the control group. Would an independent samples *t*-test be appropriate to analyze the gender data in this study (review algorithm in Exercise 12)? Provide a rationale for your answer.

Answers to Study Questions

1. An independent samples *t*-test was conducted to examine group differences in the dependent variables in this study. The two groups analyzed for differences were the Buzzy experimental or intervention group and the control group.

2. The level of significance or alpha (α) was set at 0.05.

3. The result was $t = -6.065$, $p = 0.000$ for procedural self-reported pain with the VAS (see Table 2). The t value is statistically significant as indicated by the $p = 0.000$, which is less than $\alpha = 0.05$ set for this study. The t result means there is a significant difference between the Buzzy intervention group and the control group in terms of the procedural self-reported pain measured with the VAS. As a point of clarification, p values are never zero in a study. There is always some chance of error.

4. The null hypothesis is: There is no difference in observer-reported procedural anxiety levels between the Buzzy intervention and the control groups for school-age children. The $t = -6.745$ for observer-reported procedural anxiety levels, $p = 0.000$, which is less than $\alpha = 0.05$ set for this study. Since this study result was statistically significant, the null hypothesis was rejected.

5. The $t = -1.309$ for BMI. The nonsignificant $p = .192$ for BMI is greater than $\alpha = 0.05$ set for this study. The nonsignificant result means there is no statistically significant difference between the Buzzy intervention and control groups for BMI. The two groups need to be similar for demographic variables to decrease the potential for error and increase the likelihood that the results are an accurate reflection of reality.

6. The conduct of multiple *t*-tests causes an increased risk for Type I errors. If only one *t*-test is conducted on study data, the risk of Type I error does not increase. The Bonferroni procedure and the more stringent Tukey's honestly significant difference (HSD), Student Newman-Keuls, or Scheffé test can be calculated to reduce the risk of a Type I error (Plichta & Kelvin, 2013; Zar, 2010).

7. The Bonferroni procedure is calculated by alpha ÷ number of *t*-tests conducted on study variables' data. Note that researchers do not always report all *t*-tests conducted, especially if they were not statistically significant. The *t*-tests conducted on demographic data are not of concern. Canbulat et al. reported the results of four *t*-tests conducted to examine differences between the intervention and control groups for the dependent variables procedural self-reported pain with WBFS, procedural self-reported pain with VAS, parent-reported anxiety levels, and observer-reported anxiety levels. The Bonferroni calculation for this study: 0.05 (alpha) ÷ number of *t*-tests conducted = 0.05 ÷ 4 = 0.0125. The new α set for the study is 0.0125.

8. Based on the Bonferroni result = 0.0125 obtained in Question 7, the $t = -6.745$, $p = 0.000$, is still significant since it is less than 0.0125.

9. The intervention and control groups were examined for differences related to the demographic variables gender, age, and BMI and the dependent variable preprocedural anxiety that might have affected the procedural pain and anxiety posttest levels in the children 7 to 12 years old. These nonsignificant results indicate the intervention and control groups were similar or equivalent for these variables at the beginning of the study. Thus, Canbulat et al. (2015) can conclude the significant differences found between the two groups for procedural pain and anxiety levels were probably due to the effects of the intervention rather than sampling error or initial group differences.

10. No, the independent samples *t*-test would not have been appropriate to analyze the differences in gender between the Buzzy intervention and control groups. The demographic variable gender is measured at the nominal level or categories of females and males. Thus, the χ^2 test is the appropriate statistic for analyzing gender data (see Exercise 19). In contrast, the *t*-test is appropriate for analyzing data for the demographic variables age and BMI measured at the ratio level.

Questions to Be Graded

EXERCISE 16

Name: _____ Class: _____

Date: _____

Follow your instructor's directions to submit your answers to the following questions for grading. Your instructor may ask you to write your answers below and submit them as a hard copy for grading. Alternatively, your instructor may ask you to use the space below for notes and submit your answers online at http://evolve.elsevier.com/Grove/Statistics/ under "Questions to Be Graded."

1. What do degrees of freedom (*df*) mean? Canbulat et al. (2015) did not provide the *df*s in their study. Why is it important to know the *df* for a *t* ratio? Using the *df* formula, calculate the *df* for this study.

2. What are the means and standard deviations (*SD*s) for age for the Buzzy intervention and control groups? What statistical analysis is conducted to determine the difference in means for age for the two groups? Was this an appropriate analysis technique? Provide a rationale for your answer.

3. What are the *t* value and *p* value for age? What do these results mean?

4. What are the assumptions for conducting the independent samples *t*-test?

5. Are the groups in this study independent or dependent? Provide a rationale for your answer.

6. What is the null hypothesis for procedural self-reported pain measured with the Wong Baker Faces Scale (WBFS) for the two groups? Was this null hypothesis accepted or rejected in this study? Provide a rationale for your answer.

7. Should a Bonferroni procedure be conducted in this study? Provide a rationale for your answer.

8. What variable has a result of $t = -6.135$, $p = 0.000$? What does the result mean?

9. In your opinion, is it an expected or unexpected finding that both *t* values on Table 2 were found to be statistically significant. Provide a rationale for your answer.

10. Describe one potential clinical benefit for pediatric patients to receive the Buzzy intervention that combined cold and vibration during IV insertion.

Understanding Paired or Dependent Samples *t*-Test

EXERCISE 17

STATISTICAL TECHNIQUE IN REVIEW

The **paired or dependent samples *t*-test** is a parametric statistical procedure calculated to determine differences between two sets of repeated measures data from one group of people. The scores used in the analysis might be obtained from the same subjects under different conditions, such as the one group pretest–posttest design. With this type of design, a single group of subjects experiences the pretest, treatment, and posttest. Subjects are referred to as serving as their own control during the pretest, which is then compared with the posttest scores following the treatment. Paired scores also result from a one-group repeated measures design, where one group of participants is exposed to different levels of an intervention. For example, one group of participants might be exposed to two different doses of a medication and the outcomes for each participant for each dose of medication are measured, resulting in paired scores. The one group design is considered a weak quasi-experimental design because it is difficult to determine the effects of a treatment without a comparison to a separate control group (Shadish, Cook, & Campbell, 2002).

A less common type of paired groups is when the groups are matched as part of the design to ensure similarities between the two groups and thus reduce the effect of extraneous variables (Grove, Burns, & Gray, 2013; Shadish et al., 2002). For example, two groups might be matched on demographic variables such as gender, age, and severity of illness to reduce the extraneous effects of these variables on the study results. The assumptions for the paired samples *t*-test are as follows:

1. The distribution of scores is normal or approximately normal.
2. The dependent variable(s) is(are) measured at interval or ratio levels.
3. Repeated measures data are collected from one group of subjects, resulting in paired scores.
4. The differences between the paired scores are independent.

RESEARCH ARTICLE

Source

Lindseth, G. N., Coolahan, S. E., Petros, T. V., & Lindseth, P. D. (2014). Neurobehavioral effects of aspartame consumption. *Research in Nursing & Health, 37*(3), 185–193.

Introduction

Despite the widespread use of the artificial sweetener aspartame in drinks and food, there are concern and controversy about the mixed research evidence on its neurobehavioral

effects. Thus Lindseth and colleagues (2014) conducted a one-group repeated measures design to determine the neurobehavioral effects of consuming both low- and high-aspartame diets in a sample of 28 college students. "The participants served as their own controls....A random assignment of the diets was used to avoid an error of variance for possible systematic effects of order" (Lindseth et al., 2014, p. 187). "Healthy adults who consumed a study-prepared high-aspartame diet (25 mg/kg body weight/day) for 8 days and a low-aspartame diet (10 mg/kg body weight/day) for 8 days, with a 2-week washout between the diets, were examined for within-subject differences in cognition, depression, mood, and headache. Measures included weight of foods consumed containing aspartame, mood and depression scales, and cognitive tests for working memory and spatial orientation. When consuming high-aspartame diets, participants had more irritable mood, exhibited more depression, and performed worse on spatial orientation tests. Aspartame consumption did not influence working memory. Given that the higher intake level tested here was well below the maximum acceptable daily intake level of 40–50 mg/kg body weight/day, careful consideration is warranted when consuming food products that may affect neurobehavioral health" (Lindseth et al., 2014, p. 185).

Relevant Study Results

"The mean age of the study participants was 20.8 years (SD = 2.5). The average number of years of education was 13.4 (SD = 1.0), and the mean body mass index was 24.1 (SD = 3.5)...Based on Vandenberg MRT scores, spatial orientation scores were significantly better for participants after their low-aspartame intake period than after their high intake period (Table 2). Two participants had clinically significant cognitive impairment after consuming high-aspartame diets....Participants were significantly more depressed after they consumed the high-aspartame diet compared to when they consumed the low-aspartame diet (Table 2)....Only one participant reported a headache; no difference in headache incidence between high- and low-aspartame intake periods could be established" (Lindseth et al., 2014, p. 190).

TABLE 2 WITHIN-SUBJECT DIFFERENCES IN NEUROBEHAVIOR SCORES AFTER HIGH AND LOW ASPARTAME INTAKE (N = 28)

Variable	*M*	*SD*	Paired *t*-Test	*p*
Spatial orientation				
High-aspartame	14.1	4.2	2.4	.03*
Low-aspartame	16.6	4.3		
Working memory				
High-aspartame	730.0	152.7	1.5	N.S.
Low-aspartame	761.1	201.6		
Mood (irritability)				
High-aspartame	33.4	9.0	3.4	.002**
Low-aspartame	30.5	7.3		
Depression				
High-aspartame	36.8	7.0	3.8	.001**
Low-aspartame	34.4	6.2		

*$p < .05$.
**$p < .01$.
M = Mean; SD = Standard deviation; N.S. = Nonsignificant.

Lindseth, G. N., Coolahan, S. E., Petros, T. V., & Lindseth, P. D. (2014). Neurobehavioral effects of aspartame consumption. *Research in Nursing & Health, 37*(3), p. 190.

STUDY QUESTIONS

1. Are independent or dependent (paired) scores examined in this study? Provide a rationale for your answer.

2. What independent (intervention) and dependent (outcome) variables were included in this study?

3. What inferential statistical technique was calculated to examine differences in the participants when they received the high-aspartame diet intervention versus the low-aspartame diet? Is this technique appropriate? Provide a rationale for your answer.

4. What statistical techniques were calculated to describe spatial orientation for the participants consuming low- and high-aspartame diets? Were these techniques appropriate? Provide a rationale for your answer.

5. What was the dispersion of the scores for spatial orientation for the high- and low-aspartame diets? Is the dispersion of these scores similar or different? Provide a rationale for your answer.

6. What is the paired *t*-test value for spatial orientation between the participants' consumption of high- and low-aspartame diets? Are these results significant? Provide a rationale for your answer.

7. State the null hypothesis for spatial orientation for this study. Was this hypothesis accepted or rejected? Provide a rationale for your answer.

8. Discuss the meaning of the results regarding spatial orientation for this study. What is the clinical importance of this result? Document your answer.

9. Was there a significant difference in the participants' reported headaches between the high- and low-aspartame intake periods? What does the result indicate?

10. What additional research is needed to determine the neurobehavioral effects of aspartame consumption?

Answers to Study Questions

1. This study was conducted using one group of 28 college students who consumed both high- and low- aspartame diets and differences in their responses to these two diets (interventions) were examined. Lindseth et al. (2014, p. 187) stated that "the participants served as their own controls" in this study, indicating the scores from the one group are paired. In Table 2, the *t*-tests are identified as paired *t*-tests, which are conducted on dependent or paired samples.

2. The interventions were high-aspartame diet (25 mg/kg body weight/day) and low-aspartame diet (10 mg/kg body weight/day). The dependent or outcome variables were spatial orientation, working memory, mood (irritability), depression, and headaches (see Table 2 and narrative of results).

3. Differences were examined with the paired *t*-test (see Table 2). This statistical technique is appropriate since the study included one group and the participants served as their own control (Plichta & Kelvin, 2013). The dependent variables were measured at least at the interval level for each subject following their consumption of high- and low-aspartame diets and were then examined for differences to determine the effects of the two aspartame diets.

4. Means and standard deviations (*SD*s) were used to describe spatial orientation for high- and low-aspartame diets. The data in the study were considered at least interval level, so means and *SD*s are the appropriate analysis techniques for describing the study dependent variables (Grove et al., 2013).

5. Standard deviation (*SD*) is a measure of dispersion that was reported in this study. Spatial orientation following a high-aspartame diet had an $SD = 4.2$ and an $SD = 4.3$ for a low-aspartame diet. These *SD*s are very similar, indicating similar dispersions of spatial orientation scores following the two aspartame diets.

6. Paired *t*-test = 2.4 for spatial orientation, which is a statistically significant result since $p = .03$*. The single asterisk (*) directs the reader to the footnote at the bottom of the table, which identifies * $p < .05$. Since the study result of $p = .03$ is less than $\alpha = .05$ set for this study, then the result is statistically significant.

7. There is no significant difference in spatial orientation scores for participants following consumption of a low-aspartame diet versus a high-aspartame diet. The null hypothesis was rejected because of the significant difference found for spatial orientation (see the answer to Question 6). Significant results cause the rejection of the null hypothesis and lend support to the research hypothesis that the levels of aspartame do effect spatial orientation.

8. The researchers reported, "Based on Vandenberg MRT scores, spatial orientation scores were significantly better for participants after their low-aspartame intake period than after their high intake period (Table 2)" (Lindseth et al., 2014, p. 190). This result is clinically important since the high-aspartame diet significantly reduced the participants' spatial orientation.

Healthcare providers need to be aware of this finding, since it is consistent with previous research, and encourage people to consume fewer diet drinks and foods with aspartame. The American Heart Association and the American Diabetic Association have provided a statement about the effects of aspartame that can be found on the National Guideline Clearinghouse website at http://www.guideline.gov/content.aspx?id=38431&search=effects+aspartame.

9. There was no significant difference in reported headaches based on the level (high or low) of aspartame diet consumed. Additional research is needed to determine if this result is an accurate reflection of reality or is due to design weaknesses, sampling or data collection errors, or chance (Grove et al., 2013).

10. Additional studies are needed with larger samples to determine the effects of aspartame in the diet. Lindseth et al. (2014) conducted a power analysis that indicated the sample size should have been at least 30 participants. Thus, the sample size was small at $N = 28$, which increased the potential for a Type II error. Diets higher in aspartame (40–50 mg/kg body weight/day) should be examined for neurobehavioral effects. Longitudinal studies to examine the effects of aspartame over more than 8 days are needed. Future research needs to examine the length of washout period needed between the different levels of aspartame diets. Researchers also need to examine the measurement methods to ensure they have strong validity and reliability. Could a stronger test of working memory be used in future research?

Questions to Be Graded

EXERCISE 17

Name: _____ Class: _____

Date: _____

Follow your instructor's directions to submit your answers to the following questions for grading. Your instructor may ask you to write your answers below and submit them as a hard copy for grading. Alternatively, your instructor may ask you to use the space below for notes and submit your answers online at http://evolve.elsevier.com/Grove/Statistics/ under "Questions to Be Graded."

1. What are the assumptions for conducting a paired or dependent samples *t*-test in a study? Which of these assumptions do you think were met by the Lindseth et al. (2014) study?

2. In the introduction, Lindseth et al. (2014) described a "2-week washout between diets." What does this mean? Why is this important?

3. What is the paired *t*-test value for mood (irritability) between the participants' consumption of high- versus low-aspartame diets? Is this result statistically significant? Provide a rationale for your answer.

4. State the null hypothesis for mood (irritability) that was tested in this study. Was this hypothesis accepted or rejected? Provide a rationale for your answer.

5. Which *t* value in Table 2 represents the greatest relative or standardized difference between the high- and low-aspartame diets? Is this *t* value statistically significant? Provide a rationale for your answer.

6. Discuss why the larger *t* values are more likely to be statistically significant.

7. Discuss the meaning of the results regarding depression for this study. What is the clinical importance of this result?

8. What is the smallest, paired *t*-test value in Table 2? Why do you think the smaller *t* values are not statistically significant?

9. Discuss the clinical importance of these study results about the consumption of aspartame. Document your answer with a relevant source.

10. Are these study findings related to the consumption of high- and low-aspartame diets ready for implementation in practice? Provide a rationale for your answer.

PART 2

Conducting and Interpreting Statistical Analyses

Selecting Appropriate Analysis Techniques for Studies

EXERCISE 23

Multiple factors are involved in determining the suitability of a statistical procedure for a particular study. Some of these factors are related to the nature of the study, some to the nature of the researcher, and others to the nature of statistical theory. Specific factors include the following: (1) purpose of the study; (2) study hypotheses, questions, or objectives; (3) study design; (4) level of measurement of variables in a study; (5) previous experience in statistical analysis; (6) statistical knowledge level; (7) availability of statistical consultation; (8) financial resources; and (9) access and knowledge of statistical software. Use items 1 to 4 to identify statistical procedures that meet the requirements of a particular study, and then further narrow your options through the process of elimination based on items 5 through 9.

The most important factor to examine when choosing a statistical procedure is the study hypothesis or primary research question. The hypothesis that is clearly stated indicates the statistic(s) needed to test it. An example of a clearly developed hypothesis is, "There is a difference in employment rates between veterans who receive vocational rehabilitation and veterans who are on a wait-list control." This statement tells the researcher that a statistic to determine differences between two groups is appropriate to address this hypothesis. This statement also informs the researcher that the dependent variable is a rate, which is dichotomous: employed or unemployed.

One approach to selecting an appropriate statistical procedure or judging the appropriateness of an analysis technique is to use a decision tree. A decision tree directs your choices by gradually narrowing your options through the decisions you make. A decision tree developed by Cipher (2013) that can be helpful in selecting statistical procedures is presented in Figure 23-1. This is the same decision tree presented and reviewed in Exercise 12 for determining the appropriateness of the statistical tests and results presented in research reports.

One disadvantage of decision trees is that if you make an incorrect or uninformed decision (guess), you can be led down a path in which you might select an inappropriate statistical procedure for your study. Decision trees are often constrained by space and therefore do not include all the information needed to make an appropriate selection of a statistical procedure for a study. The following examples of questions are designed to guide the selection or evaluation of statistical procedures that are reflected in Figure 23-1. Each question confronts you with a decision, and the decision you make narrows the field of available statistical procedures.

1. Is the research question/hypothesis descriptive, associational (correlational), or difference-oriented?
2. How many variables are involved?
3. What is the measurement scale of the independent and dependent variable(s)?

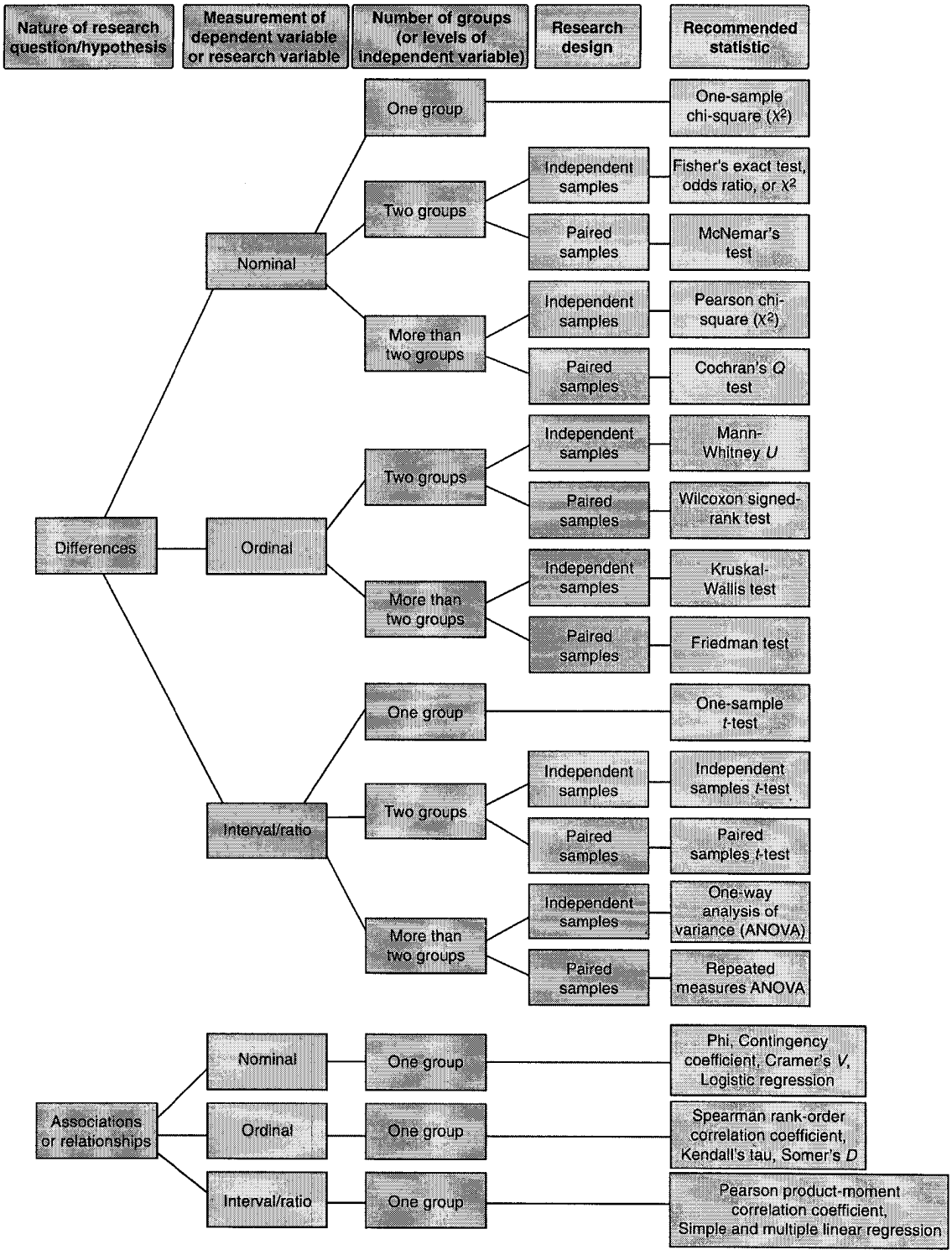


FIGURE 23-1 ▪ STATISTICAL SELECTION DECISION TREE.

4. What is the distribution of the dependent variables (normal, non-normal)?
5. Do the data meet the assumptions for a parametric statistic?

As you can see, selecting and evaluating statistical procedures requires that you make a number of judgments regarding the nature of the data and what you want to know. Knowledge of the statistical procedures and their assumptions is necessary for selecting appropriate procedures. You must weigh the advantages and disadvantages of various statistical options. Access to a statistician can be invaluable in selecting the appropriate procedures.

STATISTICS TO ADDRESS BASIC DIFFERENCE RESEARCH QUESTIONS AND HYPOTHESES

The following statistics address research questions or hypotheses that involve differences between groups or assessments. This list is by no means exhaustive, but it represents the most common parametric and nonparametric inferential statistics involving differences. Exercise 12 discusses the concepts of parametric versus nonparametric statistics and inferential versus descriptive statistics.

t-Test for Independent Samples

One of the most common statistical tests chosen to investigate differences between two independent samples is the independent samples *t*-test, which is a parametric inferential statistic. The independent samples *t*-test only compares two groups at a time, and the dependent variable must be continuous and normally distributed. The independent samples *t*-test is reviewed in Exercise 16, and the process for conducting the independent samples *t*-test is in Exercise 31.

t-Test for Paired Samples

A paired samples *t*-test (also referred to as a dependent samples *t*-test) is a statistical procedure that compares two sets of data from one group of people (or naturally occurring pairs, such as siblings or spouses). This *t*-test is a parametric inferential statistical test. The dependent variable in a paired samples *t*-test must be continuous and normally distributed. The term *paired samples* refers to a research design that repeatedly assesses the same group of people, an approach commonly referred to as *repeated measures*. The *t*-test for paired samples is reviewed in Exercise 17, and the process for conducting the paired samples *t*-test is in Exercise 32.

One-Way Analysis of Variance (ANOVA)

The one-way analysis of variance (ANOVA) is a parametric inferential statistical procedure that compares data between two or more groups or conditions to investigate the presence of differences between those groups on some continuous, normally distributed dependent variable. There are many types of ANOVAs. The one-way ANOVA involvês testing one independent variable and one dependent variable (as opposed to other types of ANOVAs, such as factorial ANOVAs, that incorporate multiple independent variables). ANOVA is reviewed in Exercise 18, and the process for conducting ANOVA is presented in Exercise 33.

Repeated-Measures ANOVA

A repeated-measures ANOVA is a statistical procedure that compares multiple setsof data from one group of people. The dependent variable in a repeated-measures ANOVA must

be continuous and normally distributed. The term *repeated measures* refers to a research design that repeatedly assesses the same group of people over time. Repeated measures can also refer to naturally occurring pairs, such as siblings or spouses.

Mann-Whitney *U*

The Mann-Whitney *U* test is the nonparametric alternative to an independent samples *t*-test. Like the independent samples *t*-test, the Mann-Whitney *U* test is a statistical procedure that compares differences between two independent samples. However, the Mann-Whitney *U* is the preferred test over the independent samples *t*-test when the distribution of the dependent variable significantly deviates from normality, or the dependent variable is ordinal and cannot be treated as an interval/ratio scaled variable. The Mann-Whitney *U* is reviewed in Exercise 21.

Kruskal-Wallis Test

The Kruskal-Wallis test is the nonparametric alternative to the one-way ANOVA. Like the one-way ANOVA, the Kruskal-Wallis test is a statistical procedure that compares differences between two or more groups. However, the Kruskal-Wallis test is the preferred test over the ANOVA when the distribution of the dependent variable significantly deviates from normality, or the dependent variable is ordinal and cannot be treated as an interval/ratio scaled variable.

Friedman Test

The Friedman test is the nonparametric alternative to a repeated-measures ANOVA. Like the repeated-measures ANOVA, the Friedman test is a statistical procedure that compares multiple sets of data from one group of people. However, the Friedman test is the preferred test over the repeated-measures ANOVA when the dependent variable data significantly deviate from a normal distribution, or the dependent variable is ordinal and cannot be treated as an interval/ratio scaled variable.

Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test is the nonparametric alternative to the paired samples *t*-test. Like the paired samples *t*-test, the Wilcoxon signed-rank test is a statistical procedure that compares two sets of data from one group of people. However, the Wilcoxon signed-rank test is the preferred test over the paired samples *t*-test when the dependent variable data significantly deviate from a normal distribution, or the dependent variable is ordinal and cannot be treated as an interval/ratio scaled variable. The Wilcoxon signed-rank test is reviewed in Exercise 22.

Pearson Chi-Square Test

The Pearson chi-square test (χ^2) is a nonparametric inferential statistical test that compares differences between groups on variables measured at the nominal level. The χ^2 compares the frequencies that are observed with the frequencies that are expected. When a study requires that researchers compare proportions (percentages) in one category versus another category, the χ^2 is a statistic that will reveal if the difference in proportion is statistically improbable. A one-way χ^2 is a statistic that compares different levels of one variable only. A two-way χ^2 is a statistic that tests whether proportions in levels of one nominal variable are significantly different from proportions of the second nominal variable. The Pearson χ^2 test is reviewed in Exercise 19, and Exercise 35 provides the steps for conducting this test.

STATISTICS TO ADDRESS BASIC ASSOCIATIONAL/CORRELATIONAL RESEARCH QUESTIONS AND HYPOTHESES

The following statistics address research questions or hypotheses that involve associations or correlations between variables. This list is by no means exhaustive, but it represents the most common parametric and nonparametric inferential statistics involving associations, predictive and otherwise.

Pearson Product-Moment Correlation Coefficient (*r*)

The Pearson product-moment correlation coefficient is a parametric inferential statistic computed between two continuous, normally distributed variables. The Pearson correlation is represented by the statistic *r*, and the value of the *r* is always between −1.00 and +1.00. A value of zero indicates absolutely no relationship between the two variables; a positive correlation indicates that higher values of *x* are associated with higher values of *y*; and a negative, or inverse, correlation indicates that higher values of *x* are associated with lower values of *y*. The Pearson *r* is reviewed in Exercise 13, and Exercise 28 provides steps for conducting this analysis.

Spearman Rank-Order Correlation Coefficient

The Spearman rank-order correlation coefficient is the nonparametric alternative to the Pearson *r*. Like the Pearson correlation, the Spearman rank-order correlation is a statistical procedure that examines the association between two continuous variables. However, the Spearman rank-order correlation is the preferred test over the Pearson *r* when one or both of the variables of data significantly deviate from a normal distribution, or the variables are ordinal and cannot be treated as an interval/ratio scaled variable. The Spearman rank-order correlation is reviewed in Exercise 20.

Phi/Cramer's *V*

The phi (Φ) coefficient is the nonparametric alternative to the Pearson *r* when the two variables being correlated are both dichotomous. Like the Pearson *r* and Spearman rank-order correlation coefficient, the phi yields a value between −1.0 and 1.0, where a zero represents no association between the variables. The Cramer's *V* is the nonparametric alternative to the Pearson *r* when the two variables being correlated are both nominal. The Cramer's *V* yields a value between 0 and 1, where a zero represents no association between the variables and a 1 represents a perfect association between the variables.

Odds Ratio

When both the predictor and the dependent variables are dichotomous (variables that have only two values), the odds ratio is a commonly used statistic to obtain an indication of association. The odds ratio (*OR*) is defined as the ratio of the odds of an event occurring in one group to the odds of it occurring in another group. The *OR* can also be computed when the dependent variable is dichotomous and the predictor is continuous and would be computed by performing logistic regression analysis. Logistic regression analysis tests a predictor (or set of predictors) with a dichotomous dependent variable. Whether the predictor is dichotomous or continuous, an *OR* of 1.0 indicates that the predictor does not affect the odds of the outcome. An *OR* of >1.0 indicates that the predictor is associated with a higher odds of the outcome, and an *OR* of <1.0 indicates that predictor is associated with a lower odds of the outcome. The steps for conducting the odds ratio are presented in Exercise 36.

Simple and Multiple Linear Regression

Linear regression is a procedure that provides an estimate of the value of a dependent variable based on the value of an independent variable or set of independent variables, also referred to as predictors. Knowing that estimate with some degree of accuracy, we can use regression analysis to predict the value of one variable if we know the value of the other variable. The regression equation is a mathematical expression of the influence that a predictor (or set of predictors) has on a dependent variable, based on a theoretical framework. A regression equation can be generated with a data set containing subjects' x and y values. The score on variable y (dependent variable, or outcome) is predicted from the same subject's known score on variable x (independent variable, or predictor). Simple linear regression is reviewed in Exercise 14, and Exercise 29 provides the steps for conducting this analysis. Multiple linear regression is reviewed in Exercise 15, and Exercise 30 provides steps for conducting this analysis.

APPLYING INFERENTIAL STATISTICS TO POPULATION DATA

Secondary data analyses of large state and national data sets can be an economically feasible and time-efficient way for nurse researchers to address important health, epidemiological, and clinical research questions. However, data that were collected on a national level using complex sampling procedures, such as those survey data made available to researchers by the Centers for Disease Control and Prevention or the Centers for Medicare and Medicaid Services, require specific weighting procedures prior to computing inferential statistics. Aponte (2010) conducted a thorough review of publicly available population data sets with nurse researchers in mind and elaborated on the advantages and disadvantages of using population survey data.

When the researcher is analyzing secondary population-based data sets, much attention must be given to the quality of the data, the extent of missing data, and the manner in which the data were sampled. Analyses of secondary population data involve much data cleaning, which can include recoding, missing data imputation, and weighting for complex sampling approaches. "Point and click" software programs such as SPSS Statistics Base (*IBM SPSS Statistics for Windows, Version 22.0.* Armonk, NY: IBM Corp.) can be used by the researcher for data cleaning; however, they cannot be used to adjust the data for complex sampling. There are only a handful of statistical software programs that can address the adjustments for sampling required by population survey data (Aday & Cornelius, 2006; Aponte, 2010). When such data are analyzed without adjusting for sampling, the results are at risk for Type I error and are generally considered invalid (Aday & Cornelius, 2006). In this text, the exercises focused on calculating statistical analyses (Exercises 28–36) involve data from samples and not data from populations.

STUDY QUESTIONS

1. What statistic would be appropriate for an associational research question or hypothesis involving the correlation between two normally distributed continuous variables?

2. What statistic would be appropriate for a difference research question involving the comparison of two repeated assessments from one group of participants, where the dependent variable is measured at the interval/ratio level or is continuous and the data are normally distributed?

3. A nurse educator is interested in the difference between traditional clinical instruction in pediatrics and simulated pediatrics instruction in a BSN (Bachelor's of Science in Nursing) program. She randomizes students to receiving 50 hours of either traditional clinical rotations in a pediatrics department or 50 hours of simulated instruction in pediatrics. At the end of the 50 hours, the students are assessed for clinical competency in pediatrics using a standardized instrument that yields a continuous score, where higher values represent higher levels of competency. Her research question is: Is there a difference between the traditional clinical group and the simulation group on clinical competency? What is the appropriate statistic to address the research question, if the scores are normally distributed?

4. What is the appropriate statistic to address the research question in Question 3 above if the scores are *NOT* normally distributed?

5. What statistic would be appropriate for a difference research question involving the comparison of two groups on a dichotomous dependent variable?

6. A diabetes educator wants to track diabetic patients' progress throughout an educational program for glycemic control. She collects each of her patients' HbA1c levels at baseline, again halfway through the educational program, and once again at the end of the program (a total of three HbA1c values). Her research question is: Did the patients' glycemic control (as measured by HbA1c values) change over the course of the program, from baseline to end of program? What is the appropriate statistic to address the research question, if the HbA1c values are normally distributed?

7. What is the appropriate statistic to address the research question in Question 6 above if the scores are *NOT* normally distributed?

8. In the scenario presented in Question 6 above, the researcher also collected data on the participants, such as gender, duration of disease, age, and depression levels. Her research question is: Do gender, duration of disease, age, and depression levels predict baseline HbA1c levels? What is the appropriate statistic to address the research question, if the baseline HbA1c levels are normally distributed?

9. What statistic would be appropriate for a difference research question involving the comparison of three independent groups on a normally distributed continuous dependent variable?

10. What statistic would be appropriate for a difference research question involving the comparison of three independent groups on a non-normally distributed, skewed, continuous dependent variable?

Answers to Study Questions

1. A Pearson *r* is an appropriate statistic to test an association between two normally distributed continuous variables.

2. A paired samples *t*-test is an appropriate statistic to compare two repeated assessments from one group of participants, where the dependent variable is continuous or measured at the interval/ratio level and the data are normally distributed.

3. An independent samples *t*-test is an appropriate statistic to compare two groups on a normally distributed continuous dependent variable (clinical competency scores).

4. Mann-Whitney *U* is an appropriate statistic to compare two independent groups on a non-normally distributed continuous dependent variable.

5. A Pearson chi-square test is an appropriate statistic to compare two groups on a dichotomous dependent variable.

6. A repeated-measures ANOVA is an appropriate statistic to compare three or more repeated assessments of the HbA1C from one group of diabetic patients, where the dependent variable is normally distributed.

7. A Friedman test is an appropriate statistic to compare three or more repeated assessments from one group of participants, where the dependent variable is not normally distributed.

8. Multiple linear regression is the appropriate statistical procedure that tests the extent to which a set of variables predicts a normally distributed dependent variable.

9. A one-way ANOVA is an appropriate statistic to compare three independent groups on a normally distributed continuous dependent variable.

10. A Kruskal-Wallis test is an appropriate statistic to compare three independent groups on a non-normally, skewed continuous dependent variable.

Questions to Be Graded

EXERCISE 23

Name: _____ Class: _____

Date: _____

Follow your instructor's directions to submit your answers to the following questions for grading. Your instructor may ask you to write your answers below and submit them as a hard copy for grading. Alternatively, your instructor may ask you to use the space below for notes and submit your answers online at http://evolve.elsevier.com/Grove/Statistics/ under "Questions to Be Graded."

1. A researcher surveyed two groups of professionals, nurse practitioners and physicians, and asked them whether or not they supported expanding the role of nurse practitioners' (NPs) prescribing privileges, answered as either "yes" or "no." Her research question is: Is there a difference between NPs and physicians on proportions of support for expanded prescription privileges? What is the appropriate statistic to address the research question?

2. What statistic would be appropriate for an associational research question involving the correlation between two non-normally distributed, skewed continuous variables?

3. A researcher is interested in the extent to which years of practice among NPs predicts level of support for expanded prescription privileges, measured on a 10-point Likert scale. She finds that both variables, years of practice and level of support, are normally distributed. Her research question is: Does years of practice among NPs predict level of support for expanded prescription privileges? What is the appropriate statistic to address the research question?

4. What statistic would be appropriate for a difference research question involving the comparison of two independent groups on a normally distributed continuous dependent variable?

5. A statistics professor tests her students' level of knowledge at the beginning of the semester and administers the same test at the end of the semester. She compares the two sets of continuous scores. Her research question is: Is there a difference in statistics knowledge from the beginning to the end of the semester? What is the appropriate statistic to address the research question, if the scores are normally distributed?

6. What is the appropriate statistic to address the research question in Question 3 if the scores are *NOT* normally distributed?

7. What is the appropriate statistic to identify the association between two dichotomous variables, where the researcher is interested in identifying the odds of an outcome occurring?

8. What statistic would be appropriate for a difference research question involving the comparison of two independent groups on a non-normally distributed, skewed continuous dependent variable?

9. A nurse educator is interested in the difference between traditional clinical instruction in pediatrics and simulated pediatrics instruction in a BSN (Bachelor's of Science in Nursing) program. She randomizes students to receiving 50 hours of either traditional clinical rotations in a pediatrics department or 50 hours of simulated instruction in pediatrics. At the end of the 50 hours, the students are assessed for clinical competency in pediatrics using a standardized instrument that yields a pass/fail result. Her research question is: Is there a difference in between the traditional clinical group and the simulation group on rates of passing the competency assessment? What is the appropriate statistic to address the research question?

10. What statistic would be appropriate for an associational research question involving the extent to which a set of variables predict a continuous, normally distributed dependent variable?

Calculating *t*-tests for Independent Samples

EXERCISE 31

One of the most common statistical tests chosen to investigate significant differences between two independent samples is the independent samples *t*-test. The samples are independent if the study participants in one group are unrelated or different participants than those in the second group. The dependent variable in an independent samples *t*-test must be scaled as interval or ratio. If the dependent variable is measured with a Likert scale, and the frequency distribution is approximately normally distributed, these data are usually considered interval level measurement and are appropriate for an independent samples *t*-test (de Winter & Dodou, 2010; Rasmussen, 1989).

RESEARCH DESIGNS APPROPRIATE FOR THE INDEPENDENT SAMPLES *t*-TEST

Research designs that may utilize the independent samples *t*-test include the randomized experimental, quasi-experimental, and comparative designs (Gliner, Morgan, & Leech, 2009). The independent variable (the "grouping" variable for the *t*-test) may be active or attributional. An **active independent variable** refers to an intervention, treatment, or program. An **attributional independent variable** refers to a characteristic of the participant, such as gender, diagnosis, or ethnicity. Regardless of the nature of the independent variable, the independent samples *t*-test only compares two groups at a time.

Example 1: Researchers conduct a randomized experimental study where the participants are randomized to either a novel weight loss intervention or a placebo. The number of pounds lost from baseline to post-treatment for both groups is measured. The research question is: "Is there a difference between the two groups in weight lost?" The active independent variable is the weight loss intervention, and the dependent variable is number of pounds lost over the treatment span.

Null hypothesis: There is no difference between the intervention and the control groups in weight lost.

Example 2: Researchers conduct a retrospective comparative descriptive study where a chart review of patients is done to identify patients who recently underwent a colonoscopy. The patients were divided into two groups: those who used statins continuously in the past year, and those who did not. The dependent variable is the number of polyps found during the colonoscopy, and the independent variable is statin use. Her research question is: "Is there a significant difference between the statin users and nonusers in number of colon polyps found?"

Null hypothesis: There is no difference between the two groups in number of colon polyps found.

STATISTICAL FORMULA AND ASSUMPTIONS

Use of the independent samples *t*-test involves the following assumptions (Zar, 2010):

1. Sample means from the population are normally distributed.
2. The dependent variable is measured at the interval/ratio level.
3. The two samples have equal variance.
4. All observations within each sample are independent.

The formula and calculation of the independent samples *t*-test are presented in this exercise.

The formula for the independent samples *t*-test is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{X1-X2}}$$

where

$\bar{X}_1$ = mean of group 1
$\bar{X}_2$ = mean of group 2
"s_{X1-X2}" = to the standard error of the difference between the two groups

To compute the *t*-test, one must compute the denominator in the formula, which is the standard error of the difference between the means. If the two groups have different sample sizes, then one must use this formula:

$$s_{\bar{X}_1-\bar{X}_2}\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

where

n_1 = group 1 sample size
n_2 = group 2 sample size
s_1 = group 1 variance
s_2 = group 2 variance

If the two groups have the same number of subjects in each group, then one can use this simplified formula:

$$s_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{s_1^2+s_2^2}{n}}$$

where

n = the sample size in each group and not the total sample of both groups

HAND CALCULATIONS

A randomized experimental study examined the impact of a special type of vocational rehabilitation on employment variables among spinal cord injured veterans, in which

post-treatment wages and hours worked were examined (Ottomanelli et al., 2012). Participants were randomized to receive supported employment or treatment as usual (subsequently referred to as "Control"). Supported employment refers to a type of specialized interdisciplinary vocational rehabilitation designed to help people with disabilities obtain and maintain community-based competitive employment in their chosen occupation (Bond, 2004).

The data from this study are presented in Table 31-1. A simulated subset was selected for this example so that the computations would be small and manageable. In actuality, studies involving independent samples *t*-tests need to be adequately powered (Aberson, 2010; Cohen, 1988). See Exercises 24 and 25 for more information regarding statistical power.

TABLE 31-1 POST-TREATMENT WEEKLY HOURS WORKED BY TREATMENT AND CONTROL GROUPS

Participant # Treatment Group	Weekly Hours Worked Treatment Group	Participant # Control Group	Weekly Hours Worked Control Group
1	25	11	8
2	27	12	9
3	19	13	15
4	20	14	17
5	40	15	24
6	25	16	15
7	40	17	18
8	35	18	9
9	30	19	18
10	15	20	16

The independent variable in this example is the treatment (or intervention) Supported Employment Vocational Rehabilitation, and the treatment group is compared with the control group that received usual or standard care. The dependent variable was number of weekly hours worked post-treatment. The null hypothesis is: "There is no difference between the treatment and control groups in post-treatment weekly hours worked among veterans with spinal cord injuries."

The computations for the independent samples *t*-test are as follows:

Step 1: Compute means for both groups, which involves the sum of scores for each group divided by the number in the group.

The mean for Group 1, Treatment Group: $\bar{X}_1 = 27.6$

The mean for Group 2, Control Group: $\bar{X}_2 = 14.9$

Step 2: Compute the numerator of the *t*-test:

$$27.6 - 14.9 = \mathbf{12.7}$$

It does not matter which group is designated as "Group 1" or "Group 2."

Another possible correct method for Step 2 is to subtract Group 1's mean from Group 2's mean, such as: $\bar{X}_2 - \bar{X}_1$: $14.9 - 27.6 = \mathbf{-12.7}$

This will result in the exact same *t*-test results and interpretation, only the *t*-test value will be negative instead of positive. The sign of the *t*-test does not matter in the interpretation of the results—only the *magnitude* of the *t*-test.

Step 3: Compute the standard error of the difference.

a. Compute the variances for each group:

$$s^2 \text{ for Group } 1 = 74.71$$

$$s^2 \text{ for Group } 2 = 24.99$$

b. Plug into the standard error of the difference formula:

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2 + s_2^2}{n}}$$

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{74.71 + 24.99}{10}}$$

$$s_{\bar{X}_1 - \bar{X}_2} = \mathbf{3.16}$$

Step 4: Compute *t* value:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

$$t = \frac{12.7}{3.16}$$

$$\boldsymbol{t = 4.02}$$

Step 5: Compute the degrees of freedom:

$$df = n_1 + n_2 - 2$$

$$df = 10 + 10 - 2$$

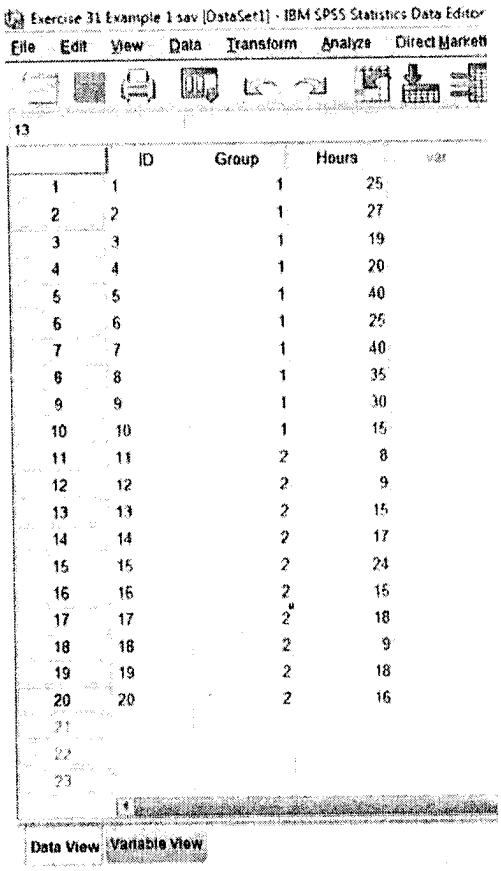
$$df = 18$$

Step 6: Locate the critical *t* value in the *t* distribution table (Appendix A) and compare it to the obtained *t* value.

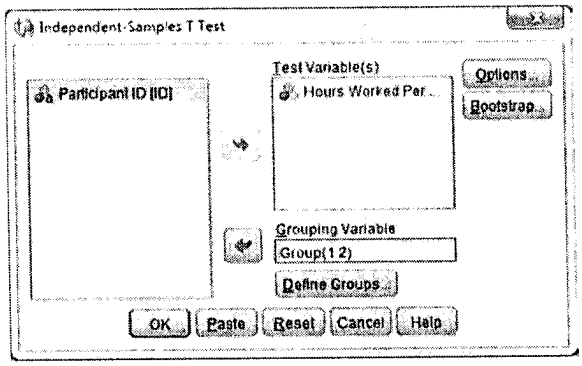
The critical *t* value for a two-tailed test with 18 degrees of freedom at alpha (α) = 0.05 is 2.101. This means that if we viewed the *t* distribution for $df = 18$, the middle 95% of the distribution would be marked by −2.10 and 2.10. The obtained *t* is 4.02, exceeding the critical value, which means our *t*-test is statistically significant and represents a real difference between the two groups. Therefore, we can reject our null hypothesis. It should be noted that if the obtained *t* was −4.02, it would also be considered statistically significant, because the absolute value of the obtained *t* is compared to the critical *t* tabled value.

SPSS COMPUTATIONS

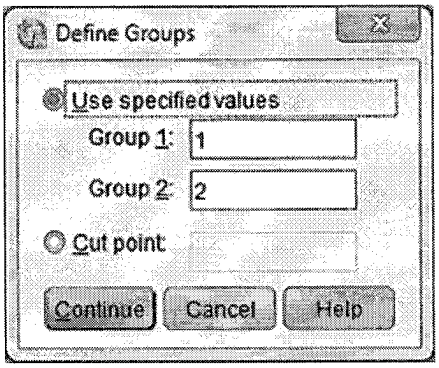
This is how our data set looks in SPSS.



Step 1: From the "Analyze" menu, choose "Compare Means" and "Independent Samples T Test". Move the dependent variable, Hours Worked, over to the right, like the window below.



Step 2: Move the independent variable, Group, into the space titled "Grouping Variable." Click "Define Groups" and enter 1 and 0 to represent the coding chosen to differentiate between the two groups. Click Continue and OK.



INTERPRETATION OF SPSS OUTPUT

The following tables are generated from SPSS. The first table contains descriptive statistics for hours worked, separated by the two groups. The second table contains the *t*-test results.

T-Test

Group Statistics

	Treatment Group	N	Mean	Std. Deviation	Std. Error Mean
Hours Worked Per Week	Control	10	27.60	8.644	2.733
	Supported Employment	10	14.90	4.999	1.581

Observe the means for the two groups

The first table displays descriptive statistics that allow us to observe the means for both groups. This table is important because it indicates that the participants in the treatment group (Supported Employment) worked 27.60 weekly hours on average, and the participants in the Control group worked 14.90 weekly hours on average.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Hours Worked Per Week	Equal variances assumed	3.267	.087	4.022	18	.001	12.700	3.158	6.066	19.334
	Equal variances not assumed			4.022	14.415	.001	12.700	3.158	5.946	19.454

Observe that the *t* value is 4.022 **The exact *p* value is .001**

The last table contains the actual *t*-test value, the *p* value, along with the values that compose the *t*-test formula. The first value in the table is the Levene's test for equality of variances. The Levene's test is a statistical test of the equal variances assumption. The *p* value is 0.087, indicating the there was no significant difference between the two groups' variances. If there had been a significant difference, the second row of the table, titled "Equal variances not assumed," would be reported in the results.

Following the Levene's test results are the *t*-test value of 4.022 and the *p* value of .001, otherwise known as the probability of obtaining a statistical value at least as extreme or as close to the one that was actually observed, assuming that the null hypothesis is true. Following the *t*-test value, the next value in the table is 12.70, which is the mean difference that we computed in Step 2 of our hand calculations. The next value in the table is 3.16, the value we computed in Steps 3a and 3b of our hand calculations.

FINAL INTERPRETATION IN AMERICAN PSYCHOLOGICAL ASSOCIATION (APA) FORMAT

The following interpretation is written as it might appear in a research article, formatted according to APA guidelines (APA, 2010). An independent samples *t*-test computed on hours worked revealed that veterans who received supported employment worked significantly more hours post-treatment than veterans in the control group, $t(18) = 4.02$, $p = 0.001$; $\bar{X} = 27.6$ versus 14.9, respectively. Thus, the particular type of vocational rehabilitation approach implemented to increase the work activity of spinal cord injured veterans appeared to have been more effective than standard practice.

STUDY QUESTIONS

1. Is the dependent variable in the Ottomanelli et al. (2012) example normally distributed? Provide a rationale for your answer.

2. If you have access to SPSS, compute the Shapiro-Wilk test of normality for the dependent variable, hours worked (as demonstrated in Exercise 26). What do the results indicate?

3. Do the data meet criteria for "independent samples"? Provide a rationale for your answer.

4. What is the null hypothesis in the example?

5. What was the exact likelihood of obtaining a *t*-test value at least as extreme or as close to the one that was actually observed, assuming that the null hypothesis is true?

6. If the Levene's test for equality of variances was significant at $p \leq 0.05$, what SPSS output would the researcher need to report?

7. What does the numerator of the independent samples *t*-test represent?

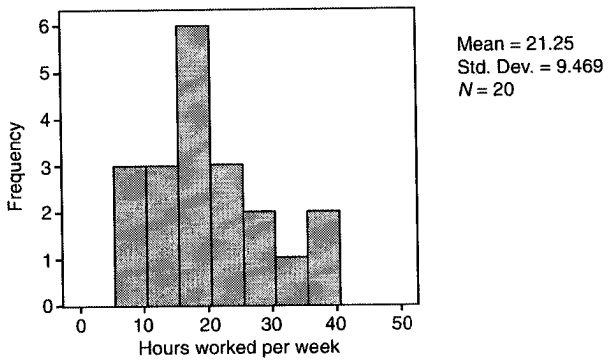
8. What does the denominator of the independent samples *t*-test represent?

9. What kind of design was used in the example?

10. Was the sample size adequate to detect differences between the two groups in this example? Provide a rationale for your answer.

Answers to Study Questions

1. The data appear slightly positively skewed as noted by a hand-drawn frequency distribution and using SPSS. See below for the SPSS frequency distribution.



2. The Shapiro-Wilk p value for hours worked was 0.142, indicating that the frequency distribution did not significantly deviate from normality.

3. Yes, the data meet criteria for "independent samples" because the dependent variable data were collected from two mutually exclusive groups of study participants. In addition, the study participants were randomly assigned to either the intervention or control group, which makes the groups independent.

4. The null hypothesis is: "There is no difference between the treatment and control groups in post-treatment weekly hours worked among veterans with spinal cord injuries."

5. The exact likelihood of obtaining a t-test value at least as extreme or as close to the one that was actually observed, assuming that the null hypothesis is true, was 0.1%. This value can be found in the "Independent Samples t-test" table in the SPSS output, where the exact p value is reported as 0.001.

6. If the Levene's test for equality of variances was significant at $p \leq 0.05$, the researcher would need to report the second row of SPSS output, containing the t-test value that has been adjusted for unequal variances.

7. The numerator represents the mean difference between the two groups (see formula for the independent samples t-test).

8. The denominator represents the extent to which there is dispersion among the values of the dependent variable.

9. The study design in the example was a randomized experimental design as evidenced by the

fact that the participants were randomly assigned to receiving the treatment or the control conditions.

10. The sample size was adequate to detect differences between the two groups, because a significant difference was found, $p = 0.001$. However, this sample is considered small, and as emphasized in Exercises 24 and 25, it is strongly recommended that a power analysis be conducted prior to the study beginning, in order to avoid the risk of Type II error.

DATA FOR ADDITIONAL COMPUTATIONAL PRACTICE FOR QUESTIONS TO BE GRADED

Using the same example from Ottomanelli and colleagues (2012), participants were randomized to receive the Supported Employment treatment or usual care ("Control"). A simulated subset was selected for this example so that the computations would be small and manageable. The independent variable in this example is the type of treatment received (Supported Employment) or the Control group and the dependent variable was the amount of weekly wages earned post-treatment. The null hypothesis is: "There is no difference between the treatment and control groups in post-treatment weekly wages earned among veterans with spinal cord injuries."

Compute the independent samples *t*-test on the data in Table 31-2 below.

TABLE 31-2 POST-TREATMENT WEEKLY WAGES EARNED BY TREATMENT GROUP

Participant # Treatment Group	Weekly Wages Earned Treatment Group	Participant # Control Group	Weekly Wages Earned Control Group
1	$200	11	$75
2	$225	12	$70
3	$157	13	$125
4	$165	14	$140
5	$330	15	$200
6	$210	16	$165
7	$330	17	$149
8	$290	18	$75
9	$250	19	$150
10	$170	20	$135

Questions to Be Graded

EXERCISE 31

Name: _____ Class: _____

Date: _____

Follow your instructor's directions to submit your answers to the following questions for grading. Your instructor may ask you to write your answers below and submit them as a hard copy for grading. Alternatively, your instructor may ask you to use the space below for notes and submit your answers online at http://evolve.elsevier.com/Grove/statistics/ under "Questions to Be Graded."

1. Do the example data meet the assumptions for the independent samples *t*-test? Provide a rationale for your answer.

2. If calculating by hand, draw the frequency distributions of the dependent variable, wages earned. What is the shape of the distribution? If using SPSS, what is the result of the Shapiro-Wilk test of normality for the dependent variable?

3. What are the means for two group's wages earned?

4. What is the independent samples *t*-test value?

5. Is the *t*-test significant at $a = 0.05$? Specify how you arrived at your answer.

6. If using SPSS, what is the exact likelihood of obtaining a *t*-test value at least as extreme or as close to the one that was actually observed, assuming that the null hypothesis is true?

7. Which group earned the most money post-treatment?

8. Write your interpretation of the results as you would in an APA-formatted journal.

9. What do the results indicate regarding the impact of the supported employment vocational rehabilitation on wages earned?

10. Was the sample size adequate to detect significant differences between the two groups in this example? Provide a rationale for your answer.

Calculating *t*-tests for Paired (Dependent) Samples

EXERCISE 32

A paired samples *t*-test (also referred to as a dependent samples *t*-test) is a statistical procedure that compares two sets of data from one group of people. The paired samples *t*-test was introduced in Exercise 17, which is focused on understanding these results in research reports. This exercise focuses on calculating and interpreting the results from paired samples *t*-tests. When samples are related, the formula used to calculate the *t* statistic is different from the formula for the independent samples *t*-test (see Exercise 31).

RESEARCH DESIGNS APPROPRIATE FOR THE PAIRED SAMPLES *t*-TEST

The term *paired samples* refers to a research design that repeatedly assesses the same group of people, an approach commonly referred to as **repeated measures**. Paired samples can also refer to naturally occurring pairs, such as siblings or spouses. The most common research design that may utilize a paired samples *t*-test is the one-group pretest-posttest design, wherein a single group of subjects is assessed at baseline and once again after receiving an intervention (Gliner, Morgan, & Leech, 2009). Another design that may utilize a paired samples *t*-test is where one group of participants is exposed to one level of an intervention and then those scores are compared with the same participants' responses to another level of the intervention, resulting in paired scores. This is called a **one-sample crossover design** (Gliner et al., 2009).

Example 1: A researcher conducts a one-sample pretest-posttest study wherein she assesses her sample for health status at baseline, and again post-treatment. Her research question is: "Is there a difference in health status from baseline to post-treatment?" The dependent variable is health status.

Null hypothesis: There is no difference between the baseline and post-treatment health status scores.

Example 2: A researcher conducts a crossover study wherein subjects receive a randomly generated order of two medications. One is a standard FDA-approved medication to reduce blood pressure, and the other is an experimental medication. The dependent variable is reduction in blood pressure (systolic and diastolic), and the independent variable is medication type. Her research question is: "Is there a difference between the experimental medication and the control medication in blood pressure reduction?"

Null hypothesis: There is no difference between the two medication trials in blood pressure reduction.

STATISTICAL FORMULA AND ASSUMPTIONS

Like the independent samples *t*-test, this *t*-test requires that differences between the paired scores be independent and normally or approximately normally distributed (see Exercise 17).

The formula for the paired samples *t*-test is:

$$t = \frac{\bar{D}}{s_{\bar{D}}}$$

where:

$\bar{D}$ = the mean difference of the paired data
$s_{\bar{D}}$ = the standard error of the difference

To compute the *t*-test, one must compute the denominator in the formula: the standard error of the difference:

$$s_{\bar{D}} = \frac{s_D}{\sqrt{n}}$$

where:

s_D = the standard deviation of the differences between the paired data
n = the number of subjects in the sample (or number of paired scores in the case of sibling or spousal data)

HAND CALCULATIONS

Using an example from a study examining the dental pocket depths among 10 adults receiving treatment for periodontitis, changes over time were investigated (Pope, Rossmann, Kerns, Beach, & Cipher, 2014). These data are presented in Table 32-1. A simulated subset was selected for this example so that the computations would be small and manageable. In actuality, studies involving paired samples *t*-tests need to be adequately powered (Aberson, 2010; Cohen, 1988). See Exercises 24 and 25 for more information regarding statistical power.

TABLE 32-1 PROBING DEPTHS AT BASELINE AND POST-TREATMENT

Participant #	Baseline Probing Depths (mm)	Post-treatment Probing Depths (mm)	Difference Scores
1	5	4	1
2	4	2	2
3	5	3	2
4	6	3	3
5	5	3	2
6	4	2	2
7	5	4	1
8	6	4	2
9	7	5	2
10	5	3	2

The independent variable in this example is treatment over time, meaning that all 10 participants received treatment for periodontitis. The dependent variable was probing depth, defined as a tooth's dental pocket depths (in mm), with higher scores representing more undesirable depths. The null hypothesis is: "There is no reduction in probing depth from baseline to post-treatment for patients with periodontitis."

The computations for the *t*-test are as follows:

Step 1: Compute the difference between each subject's pair of data (see last column of Table 32-1).

Step 2: Compute the mean of the difference scores, which becomes the numerator of the *t*-test:

$$\bar{D} = 1.90$$

Step 3: Compute the standard error of the difference.

a. Compute the standard deviation of the difference scores

$$s = 0.57$$

b. Plug into the standard error of the difference formula

$$s_{\bar{D}} = \frac{s_D}{\sqrt{n}}$$

$$s_{\bar{D}} = \frac{0.57}{\sqrt{10}}$$

$$\mathbf{s_{\bar{D}} = 0.18}$$

Step 4: Compute *t* value:

$$t = \frac{\bar{D}}{s_{\bar{D}}}$$

$$t = \frac{1.90}{0.18}$$

$$\mathbf{t = 10.56}$$

Step 5: Compute the degrees of freedom:

$$df = n - 1$$

$$df = 10 - 1$$

$$df = 9$$

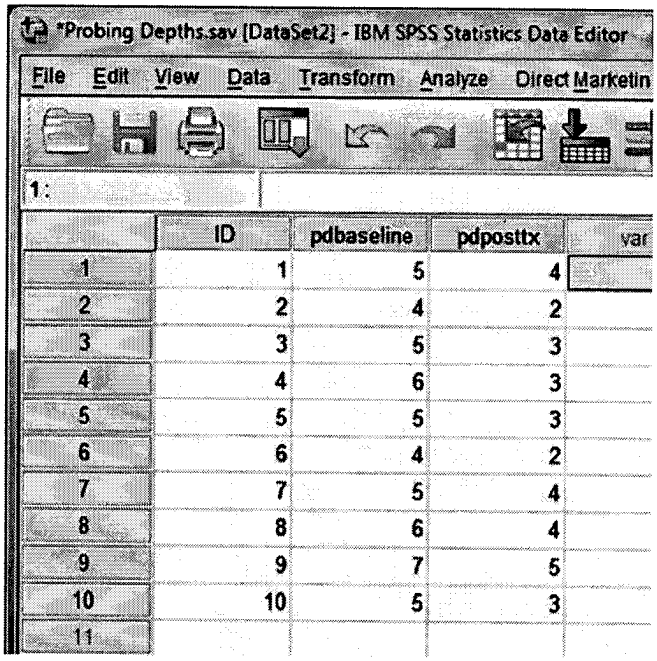
Step 6: Locate the critical *t* value on the *t* distribution table and compare it to the obtained *t*.

The critical *t* value for 9 degrees of freedom at alpha (α) = 0.05 is 2.26 for a two-tailed test. Our obtained *t* is 10.59, exceeding the critical value (Appendix A) , which means our

t-test is statistically significant and represents a real difference between the two groups. Therefore, we can reject our null hypothesis. This means that if we viewed the *t* distribution for $df = 9$, the middle 95% of the distribution would be marked by −2.26 and 2.26. It should be noted that if the obtained *t* was −10.56, it would also be considered statistically significant, because the absolute value of the obtained *t* is compared to the critical *t* value.

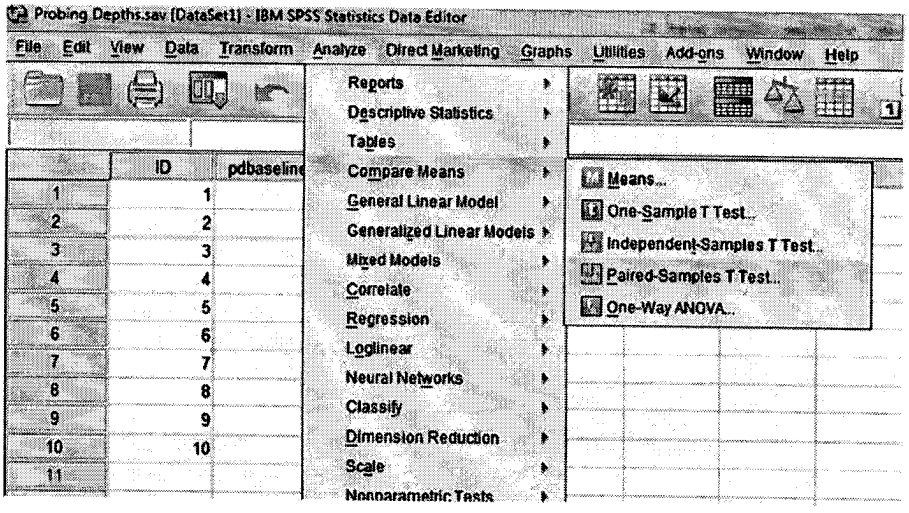
SPSS COMPUTATIONS

This is how our dataset looks in SPSS.

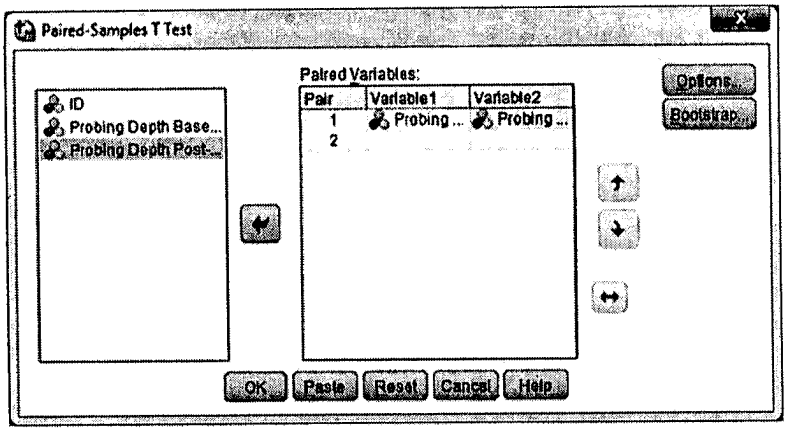


	ID	pdbaseline	pdposttx	var
1	1	5	4	
2	2	4	2	
3	3	5	3	
4	4	6	3	
5	5	5	3	
6	6	4	2	
7	7	5	4	
8	8	6	4	
9	9	7	5	
10	10	5	3	
11				

Step 1: From the Analyze menu, choose "Compare Means" and "Paired-Samples T Test".



Step 2: Move both variables over to the right, as in the window below. Click "OK".



INTERPRETATION OF SPSS OUTPUT

The following tables are generated from SPSS. The first table contains descriptive statistics for the two variables. The second table contains the Pearson product-moment correlation between the two variables. The last table contains the *t*-test results.

T-Test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Probing Depth Baseline	5.20	10	.919	.291
	Probing Depth Post-Treatment	3.30	10	.949	.300

Note the means for the two variables

The first table displays descriptive statistics that allow us to note the means at baseline and post-treatment. This table is important because we can observe that the mean probing depth at baseline was 5.20, and the mean probing depth post-treatment was 3.30, indicating a decrease in depths over time (a desirable trend).

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Probing Depth Baseline & Probing Depth Post-Treatment	10	.816	.004

Note that the two variables are correlated

The second table displays the Pearson product-moment correlation coefficient that was computed between the two variables. It is common that the two variables are significantly correlated, because the sample is being assessed twice, and, therefore, it is logical that a person's post-treatment value is affected by his or her baseline value in a repeated measures design. Although this table is a standard part of the SPSS output for a paired samples *t*-test, the contents are not reported in the results of published studies.

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Probing Depth Baseline - Probing Depth Post-Treatment	1.900	.568	.180	1.494	2.306	10.585	9	.000

Note that the *t* value is 10.585

The exact *p* value is <0.0001

The last table contains the actual *t*-test value, along with the values that compose the *t*-test formula. Note that the first value in the table, 1.90, was the mean difference that we computed in Step 2 of our hand calculations. The next two values in the table, 0.568 and 0.18, were the values we computed in Steps 3a and 3b of our hand calculations. The *t*-test value of 10.585 is slightly higher than what we obtained in our hand calculations. This is because we rounded to the hundredth decimal place in our hand calculations, when the actual standard error value is actually 0.1795, which yields a *t*-test value of 1.90 ÷ 0.1795 = 10.585. The last value in the table is the *p* value, otherwise known as the probability of obtaining a statistical value at least as extreme or as close to the one that was actually observed, assuming that the null hypothesis is true. Even though SPSS has printed a value of ".000," it is not possible to have an absolute zero *p* value. Rather, SPSS has rounded the *p* value to three decimal places and is more accurately reported as "$p < 0.001$."

FINAL INTERPRETATION IN AMERICAN PSYCHOLOGICAL ASSOCIATION (APA) FORMAT

The following interpretation is written as it might appear in a research article, formatted according to APA guidelines (APA, 2010). A paired samples *t*-test computed on probing depths revealed that the patients undergoing treatment for periodontitis had significantly lower probing depths from baseline to post-treatment, $t(9) = 10.59$, $p < 0.001$; $\bar{X} = 5.20$ versus 3.30, respectively. Thus, the particular type of treatment approach implemented to improve the health of patients' gums was effective as represented by decreases in probing depths.

STUDY QUESTIONS

1. Are the example data normally distributed? Provide a rationale for your answer.

2. If you have access to SPSS, compute the Shapiro-Wilk tests of normality for the two variables (as demonstrated in Exercise 26). If you do not have access to SPSS, plot the frequency distributions by hand. What do the results indicate?

3. Do the data meet criteria for "paired samples"? Provide a rationale for your answer.

4. What is the null hypothesis in the Pope et al. (2014) study?

5. What was the exact likelihood of obtaining a *t*-test value at least as extreme or as close to the one that was actually observed, assuming that the null hypothesis is true?

6. Why do you think the baseline and post-treatment variables were correlated?

7. What does the numerator of the paired samples *t*-test represent?

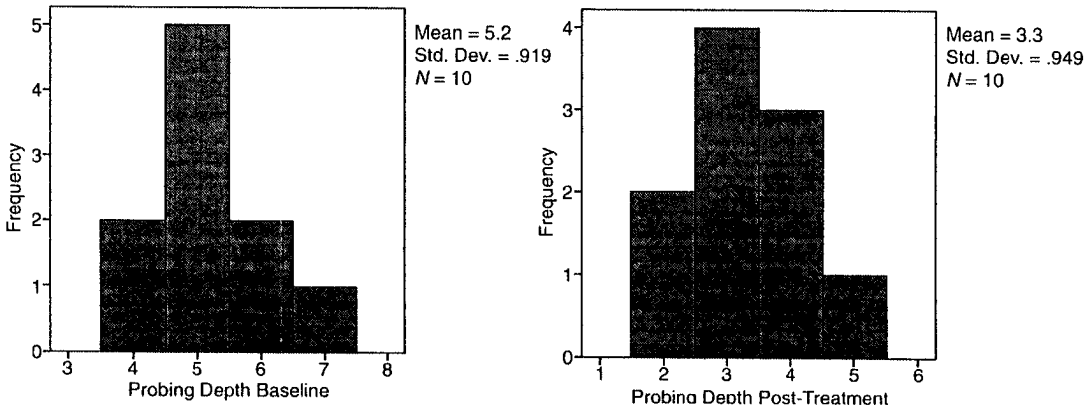
8. What does the denominator of the paired samples *t*-test represent?

9. Why would a one-sample crossover design also be suitable for a paired samples *t*-test?

10. The researchers concluded that the periodontal treatment program resulted in a significant decrease in probing depths. What are some alternative explanations for these improvements? (These alternative scientific explanations would apply to any single-sample repeated measures design.)

Answers to Study Questions

1. Yes, the data are approximately normally distributed as noted by the two SPSS frequency distributions below.



2. The Shapiro-Wilk *p* values for baseline and post-treatment probing depths were 0.15 and 0.29, respectively, indicating that the two frequency distributions did not significantly deviate from normality. Moreover, visual inspection of the frequency distributions indicates that the variables are approximately normally distributed.

3. Yes, the data meet criteria for "paired samples" because the two variables were collected from the same single group of study participants before and after the treatment.

4. The null hypothesis is: "There is no reduction in probing depth from baseline to post-treatment for patients with periodontitis."

5. The exact likelihood of obtaining a *t*-test value at least as extreme as or as close to the one that was actually observed, assuming that the null hypothesis is true, was less than 0.001%.

6. The baseline and post-treatment variables were correlated because the sample of patients was assessed twice, and therefore it is logical that a person's post-treatment value is affected by his or her baseline value in a repeated measures design.

7. The numerator represents the mean difference between the two variables (see formula for the paired samples *t*-test).

8. The denominator represents the extent to which there is dispersion among the entire dataset's values.

9. A one-sample crossover design also would be suitable for a paired samples *t*-test because one group of participants is exposed to one level of an intervention and then those scores are