CHAPTER 10

Effect Size, Confidence Intervals, and NHST: Two-Sample Designs

Summary

In a two-sample experiment, two populations are identified. The two populations might be reaction times after drinking caffeinated or decaffeinated coffee or test anxiety scores during the semester versus during finals week. In experimental research designs, two (or more) groups experience different ***treatments***, which are the levels of the independent variable. Groups are frequently identified as an ***experimental group***, which receives the treatment of interest, and a ***control group***, which is a comparison group that does not receive the treatment. A sample from each population is measured on the *dependent variable* and sample means are calculated. The goal is to determine if the mean of the populations that the groups came from are different, which would indicate that the treatment had some effect.

Paired-samples designs are two-group experiments in which the scores consist of pairs. In these designs, one member of the pair frequently serves in the experimental group and the other in the control group. The pairs might exist before the experiment begins (***natural pairs***) or might be formed by matching the participants on some variable related to the dependent variable (***matched pairs***). In some cases, participants experience a before-and-after design (***repeated measures***), in which case they are paired with themselves; *repeated measures* is the most common paired-samples design. The *t* distribution for a paired-samples design has N - 1 degrees of freedom, where N is the number of pairs.

Independent-samples designs occur when the scores are not paired in any logical way. Instead, ***random assignment*** is often used to assign individuals to the two groups. The *t* distribution for an independent-samples design has $N_1 + N_2 – 2$ degrees of freedom.

For both *paired-samples* and *independent-samples* designs, you can calculate effect size and confidence intervals. The *effect size* index, *d*, describes the size of the difference between the two populations. Although the formulas for *d* differ somewhat for *independent-sample* designs and *paired-sample* designs, values of 0.20, 0.50, and 0.80 indicate small, medium and large effects, respectively, for both designs.

For both *paired-samples* and *independent-samples* designs, you can calculate a *confidence interval* about the mean difference. A confidence interval consists of a lower and upper limit and is for a specified degree of confidence, such as 90, 95, or 99 percent. The two limits capture a range of values within which you can expect (with your specified degree of confidence) to find the mean difference that exists between the two populations from which the samples were drawn.

It is also possible to conduct null hypothesis significance testing (NHST) for two sample designs in order to make conclusions about the difference between *experimental* and *control* groups. The logic of NHST is to tentatively hypothesize that there will be no effect of the treatment (A). As such, the null hypothesis (H_0) will be $H_0 : \mu_{\text{A}} = \mu_{\text{no A}}$. The alternative hypothesis (H_1) will be that there is an effect; alternative hypotheses can offer a two-tailed alternative (H_1: $\mu_{\text{A}} \neq \mu_{\text{no A}}$) or one-tailed alternative (H_1: $\mu_{\text{A}} > \mu_{\text{no A}}$ or H_1: $\mu_{\text{A}} < \mu_{\text{no A}}$). Two-tailed tests, which are much more common, allow a conclusion that μ_{A} is larger *or* that it is smaller than $\mu_{\text{no A}}$. A one-tailed test places the entire rejection region in one tail of the sampling distribution; thus, if the relationship of the sample means is opposite that expected by the researcher, the null hypothesis cannot be rejected, no matter how different the means are.

This chapter also introduced four objections to NHST.

1. In practice, the null hypothesis will rarely ever be true. That is, the likelihood that two populations will be exactly equal is low, simply because of sampling error. Because of this, with a sufficiently large sample size, any amount of difference between two populations could reach statistical significance.

2. Making publishing decisions based on the $p < .05$ rule-of-thumb conflates *statistical significance* with *importance*. Any well-designed study, regardless of its *p*-values, can contribute important information to a field.
3. As reviewed in Chapter 9, *p*-values are often misunderstood and misinterpreted.
4. Finally, a sampling distribution *only* provides accurate probabilities when certain conditions are met. In practice, these conditions are rarely shown to be met.

Like any sampling distribution, the *t* distribution gives you accurate probabilities when certain conditions are met. For the independent-samples *t* test, the two conditions are:

1. the populations are normally distributed
2. the populations have variances that are equal

In addition to accurate probabilities, correct conclusions depend on control of extraneous variables. The most common way to control the extraneous variables is to *randomly assign* participants to levels of the independent variable.

The more powerful a statistical test is, the better its ability to detect a false null hypothesis. ***Power*** is equal to $1 - \beta$, where β is the probability of a Type II error. The factors that determine power are:

1. the effect size (the size of the difference between populations)
2. the standard error of the difference, which is governed by *N* and sample variability
3. alpha (α)

For an excellent article on power, see Cohen (1992).

Multiple-Choice Questions

1. The logic of hypothesis testing is to assume that two populations have ______ and then see if sample data will permit you to conclude that they are ______.

a. means that are equal, probably equal
b. means that are equal, probably unequal
c. means that are not equal, probably unequal
d. means that are not equal, probably equal

2. In a two-tailed independent-samples design, the null hypothesis is that ______.

a. the population mean of one group is equal to that of a second group
b. the population mean of one group is larger or smaller than that of a second group
c. the sample mean of one group is equal to that of a second group
d. the sample mean of one group is larger or smaller than that of a second group

3. According to your text, the reason we conduct experiments is to be able to tell ______.

a. whether all extraneous variables were controlled
b. whether the samples were representative of the population
c. how dependent variable scores are affected by the independent variable
d. all of the above

4. An experimenter found one sample mean of 13 based on an *N* of 8. The second sample mean was 18 based on an *N* of 6. These data suggest a(n) ______ design.

a. paired-samples
b. independent-samples
c. either paired-samples or independent-samples
d. cannot be determined from the information given

5. A one-tailed test of significance produced a *t* equal to -2.30, significant at the .05 level. The design of this experiment ______.

a. was a paired-samples design
b. was an independent-samples design
c. neither a. nor b.
d. cannot be determined from the information given

6. The Montagues and Capulets blame each other for Romeo and Juliet falling in love. On a test of propensity to fall in love, the mean of 6 members of the Montague family was 54 and the mean of 10 members of the Capulet family was 64. When a statistician compared the families' scores with a *t* test, to determine if one family was more at fault, a *t* value of 2.13 was obtained. If you adopt an α level of .05 (two-tailed test), you should conclude that the Capulets are ______.

a. significantly more loving than the Montagues
b. significantly less loving than the Montagues
c. not significantly different from the Montagues
d. not yet comparable; additional information is needed

7. Which of the following variables affect the size of the standard error of a difference?

a. difference between sample means
b. sample size
c. both a. and b.
d. neither a. nor b.

8. $p < .05$ means that the difference between sample means ______.

a. fell outside the rejection region
b. should be attributed to chance rather than to the independent variable
c. should be declared "not significant"
d. none of the above

9. The power of a statistical test is defined as ______.

a. α
b. β
c. $1 - \alpha$
d. $1 - \beta$

10. The 95 percent confidence interval about a mean difference was -3.0 minutes to 6.5 minutes. The null hypothesis that the two population means are equal ______.

a. can be rejected at the .05 level
b. can be rejected at the .01 level
c. can be rejected at both the .05 and the .01 level
d. should be retained

11. For a normally distributed set of scores, if is often best to use the design that has the most power. Which of the following designs has the most power?

a. paired-sample
b. independent sample
c. neither paired nor independent samples
d. cannot be determined from the information given

12. Which of the following has an influence on the power of a statistical test?

a. sample size
b. alpha (α)
c. effect size
d. all of the above

13. Degrees of freedom are most closely related to ______.

a. sample size
b. alpha
c. the actual difference between population means
d. the choice of a one- or two-tailed test

14. Which of these phrases does **not** belong with the other three?

a. extraneous variable
b. independent variable
c. treatment
d. experimental group

15. An experiment allows a researcher to show that the null hypothesis is ______.

a. probably false
b. probably true
c. either probably false or probably true
d. none of the above

16. “The difference between the two _______ means is zero” is a statement of the null hypothesis.

a. sample
b. population
c. either sample or population
d. none of the above

17. If you find that there is a logical reason to pair the scores from the two groups in a two-group experiment, you know that you ______.

a. have an independent-samples design
b. have a paired-samples design
c. should use a two-tailed critical value
d. can use confidence intervals but not hypothesis testing

18. A student conducted a before-and-after study on college statistics students to see if the course improved ability to reason. In a later study, she used a mother-daughter sample to assess generational differences in attitudes toward social media. Her two designs were ______ and ______.

a. paired, paired
b. paired, independent
c. independent, independent
d. independent, paired

19. If the *t* distribution is to produce an accurate probability value, which of the following must be true?

a. sample size is the same for both samples
b. the populations compared have equal variances
c. the samples compared have equal means

d. all of the above

20. The size of the difference between the two populations that the samples are from is *most* closely associated with ______.

a. whether or not to use hypothesis testing
b. Type I and Type II errors
c. effect size
d. .05 or .01 α level

Short-Answer Questions

1. What are the factors that influence whether or not you reject the null hypothesis? Explain how each factor influences the final decision.

2. A two-group experiment might have the phrase "p = .01." Explain what this means by finishing the sentence: "The probability is .01 that..."

3. List and explain the three factors identified in your text that influence whether or not you reach correct conclusions when you use a t test.

4. Explain what a powerful statistical test is. How can power of a statistical test be increased?

Problems

1. Identify the design and the degrees of freedom for each of the following experiments. If it is a paired samples design, identify the type of pairing.

a. The effect of soda on attention was measured by counting the number of lines of text participants could read in 5 minute. Fifty participants read, then drank a 12 oz. can of soda and took a short break, then read for another 5 minutes.
b. A group of 21 famous sociologists rated their attitude toward statistics and then identified their best student who had obtained a PhD in sociology. Attitudes toward statistics were then obtained from a former student of each sociologist.
c. A consumer group compared two detergents, Bold and Tide, to determine which was better. A sample of 24 white washcloths that had been soaked in mud for 10 hours were washed (12 cloths for each brand). Afterward the amount of light reflected from each cloth was measured with a photometer.
d. Researchers compared a classroom of 25 Native Americans to a classroom of 25 Hispanic Americans. For each student, attitude toward school was measured and mean attitudes compared.
e. Eight cancer patients rated their emotionality while sitting in a large blue waiting room. Later they rated their emotionality while sitting in a small yellow waiting room.

2. For each description, identify the independent variable, the dependent variable and whether a paired or an independent-samples design is described.

a. A 1957 article in the *Journal of Experimental Psychology* described the kind of dreams that are reported when you wake a person during REM sleep (rapid eye movement) or at other times during NREM sleep (non-REM). They found that you get vivid dreams reported from people woken from REM sleep and no dreams or vague dreams reported from people woken from NREM sleep.
b. Pat's handwriting was terrible. Pat converted this longstanding problem into an independent project for his Research Methods class. The participants in his experiment graded an essay on a

scale of 50-100. For 20 subjects, the essay was written in Pat's handwriting, and for another 23, the essay was beautifully written.

3. For the independent-samples data that follow, calculate a 95 percent confidence interval about the mean difference and write a sentence of interpretation.

Group 1 ($\bar{X}_1$)	Group 2 ($\bar{X}_2$)
11	6
8	4
5	3

4. For the paired-samples data that follow, calculate a 99 percent confidence interval about the mean difference and write a sentence of interpretation.

X	***Y***
14	8
9	4
7	6

5. Benjamin, Cavell, and Shallenberger (1984) tested the question of whether a student should change an answer on a multiple choice question when re-checking a test. Assume they found the following data. What is the appropriate statistical test for this experiment? Calculate a 95 percent confidence interval to determine if there was a difference in the performance of students who changed their answers and those who stayed with their initial answers. Then, calculate the effect size index and write a sentence of interpretation and make a conclusion about whether students should change their answers on multiple choice test when checking their answers.

Students who changed answers ($\bar{X}_1$)	Students who stayed with initial answers ($\bar{X}_2$)
85	74
73	75
69	88
99	70
87	71

6. For the paired-samples data that follow, conduct a *t* test. State the null hypothesis, set a two-tailed alternative hypothesis, perform a paired-samples *t* test with $\alpha = 0.01$, calculate the effect size index, and write a conclusion about the results.

X	*Y*
8	11
6	9
5	7
3	5
1	6

7. Yang and Urminsky (2018) studied how anticipated affective reactions, like smiling, can motivate different gift choices. In one part of their study, participants were asked to imagine the affective reactions that would be elicited by a personalized mug and an award-wining ergonomic mug. Participants were asked "How much of an affective reaction would receivers show in response to these gifts when receiving them?" and made a rating of each item. Below are summary data from Study 1 of Yang and Urminsky (2018;

retrieved from https://osf.io/fctr8/ on July 14, 2018) that will allow you to closely replicate their findings. Conduct a paired-samples *t* test to determine if there is a difference in the affective reaction people think would be elicited by the personalized or award-winning ergonomic mugs. State the null hypothesis, set a two-tailed alternative hypothesis, perform a *t* test, calculate the effect size index, and write a conclusion about the affective reaction participants expect to the personalized and ergonomic mugs. Set alpha at .05.

$N = 213$
Average affective reaction for personalized mug ($\bar{X}$) = 6
Average affective reaction for ergonomic mug ($\bar{Y}$) = 5
$\Sigma D = 257$
$\Sigma D^2 = 1021$

8. Yang and Urminsky (2018) also examined how we might focus differently on anticipated affective reactions when making choices for others versus ourselves. In this part of the study, half of the participants imagined *giving* a personalized mug and the other half imagined *receiving* the personalized mug. After this, participants were asked how much they would personally prefer the mug. Below are summary data from Study 1 of Yang and Urminsky (2018; retrieved from https://osf.io/fctr8/ on July 14, 2018) that will allow you to closely replicate their findings. Conduct an independent-samples *t* test to determine if givers or receivers more strongly preferred the personalized mugs. State the null hypothesis, set a two-tailed alternative hypothesis, perform a *t* test, calculate the effect size index, and write a conclusion about the affective reaction participants expect to the personalized and ergonomic mugs. Set alpha at .05.

Receiver Preference (X_1)	Giver Preference (X_2)
$N_1 = 106$	$N_2 = 106$
$\Sigma X_1 = 508$	$\Sigma X_2 = 569$
$\Sigma X_1^2 = 2816$	$\Sigma X_2^2 = 3453$
$\bar{X}_1 = 4.79$	$\bar{X}_2 = 5.37$

ANSWERS

Multiple-Choice Questions

1. b. means that are equal, probably unequal
   Explanation: A NHST always starts with the null hypothesis, which assumes that the population means from which the samples were taken are equal. We are testing whether we should keep or reject this assumption, so we can conclude that they are probably unequal, but we never prove that they are equal with this type of test.
2. a. the population mean of one group is equal to that of a second group
   Explanation: Null hypotheses are never about sample means. And, for two-tailed tests, only alternative hypotheses can focus on identifying one group as having a larger or smaller mean.
3. c. how dependent variable scores are affected by the independent variable
   Explanation: Although we do want to control extraneous variable and we hope that samples are representative of populations, those are not the reasons we do experiments. Experiments are primarily about trying to identify cause and effect relationships and we study this by manipulating the independent variable and measuring changes in the dependent variable.
4. b. independent-samples
   Explanation: Because there are two different sample sizes, we can rule out that data are pairs.
5. d. cannot be determined from the information given
   Explanation: Because we can conduct a one-tailed NHST with a paired- or independent-samples design, there is no way to tell the design just from knowing the t value.
6. c. not significantly different from the Montagues
   Explanation: Because the *N*s are different, we have an independent-samples design. The $df = N_1 + N_2 – 2 = 10 + 6 - 2 = 14$. The critical value for a two-tailed test is $t_{.05}$ (14) = ±2.145. Because the obtained value of *t* is smaller than the critical value, we cannot reject the H_0, so we would say that the two groups are not significantly different.
7. b. sample size
   Explanation: Difference between sample means and sample size both affect power, but only sample size affects standard error of a difference.
8. d. none of the above
   Explanation: In this problem, *p* refers to the probability of getting the difference between sample means you did, if the two groups actually do come from the same population (or populations with equal means). If the probability of getting the difference you did is smaller than .05, it means that you found a statistically significant difference. Your *t* value would fall in the rejection region, is unlikely due to chance, and is significant.
9. d. 1 – β
   Explanation: Beta (β) represents the probability of making a Type II error (incorrect retention of the null hypothesis). Thus, $1 – \beta$ represents the probability of *not* making a Type II error (correctly rejecting the null hypothesis), which is the definition of statistical power.
10. d. should be retained
    Explanation: Because zero is in the confidence interval, we cannot reject the hypothesis that the difference between the two population means is zero. In other words, we are 95 percent confident that the difference between the two population means *could be* zero.
11. a. paired-sample
    Explanation: The standard error of the difference is one factor that affects power and sample variability is one factor that affects the standard error of the difference. The act of pairing scores in a meaningful way reduces sample variability, relative to independent-samples designs. Thus, the standard error of the difference is smaller in paired-samples designs, which serves to increase power.
12. d. all of the above
13. a. sample size
14. a. extraneous variable
15. a. probably false
    Explanation: A NHST is only testing whether we can reject the null hypothesis, so NHST cannot prove the null hypothesis true. As a result, a NHST really only allows us to say that the H_0 is

probably false. It is also correct that we use the word “probably” – because if we set our alpha level at .05, we are saying there is still a 5% likelihood we would see the difference we see, just due to chance.

16. b. population
17. b. have a paired-samples design
18. a. paired, paired
19. b. the populations compared have equal variances
20. c. effect size
    Explanation: The size of the difference between two populations is the definition of effect size. It is also the case, though, that effect size is related to power, which is the probability of *not* making a Type II error.

Short-Answer Questions

1. There are three factors that influence the decision to reject H_0. First, the size of the difference between the populations (effect size). The greater the difference, the more likely we are to reject the null hypothesis. Second, the size of the standard error of the difference. The smaller it is, the more likely we are to reject the null hypothesis. You can reduce the standard error by increasing sample size or reducing sample variability. Third, the alpha level. The larger the alpha is, the more likely we are to reject the null hypothesis.
2. If the two populations are exactly equal, the probability is .01 that you would obtain the observed difference just due to chance. You could also say that the probability is .01 of obtaining a difference as large or larger as the one observed if there really is no difference between the two populations.
3. a. The populations from which the samples are drawn are normally distributed.
   b. The populations from which the samples are drawn have variances that are equal.
   c. Extraneous variables are eliminated, likely by random assignment of participants to conditions.

4. A powerful statistical test is one that is likely to detect an actual difference between populations. Increasing sample size, reducing sample variability, and increasing the alpha level can all increase the power of a statistical test. Power is also affected by the size of the difference between the populations (effect size).

Problems

1. a. paired samples (repeated measures); df = 49
   b. paired samples (natural pairs); df = 20
   c. independent samples; df = 22
   d. independent samples; df = 48
   e. paired samples (repeated measures); df = 7

2. a. independent variable: stage of sleep when woken up (REM or NREM); dependent variable: vividness of dreams; paired-samples design
   b. independent variable: handwriting quality; dependent variable: essay grade; independent-samples design

3. We can be 95% confident that the true population mean difference is between -1.72 and 9.06. Because this interval contains zero, we cannot call these groups significantly different.

Group 1 ($\bar{X}_1$)	${\bar{X}_1}^2$	Group 2 ($\bar{X}_2$)	${\bar{X}_2}^2$
11	121	6	36
8	64	4	16
5	25	3	9
$\sum \bar{X}_1$ = 24	$\sum {\bar{X}_1}^2$ = 210	$\sum \bar{X}_2$ = 13	$\sum {\bar{X}_2}^2$ = 61

$$\bar{X}_1 = \frac{\Sigma X}{N} = \frac{24}{3} = 8$$

$$\bar{X}_2 = \frac{\Sigma X}{N} = \frac{13}{3} = 4.33$$

$$s_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sum X_1{}^2 - \frac{(\sum X_1)^2}{N_1} + \sum X_2{}^2 - \frac{(\sum X_2)^2}{N_2}}{N_1(N_2-1)}}$$

$$= \sqrt{\frac{210 - \frac{24^2}{3} + 61 - \frac{13^2}{3}}{3(3-1)}} = \sqrt{\frac{210 - 192 + 61 - 56.33}{6}}$$

$$= \sqrt{\frac{18 + 4.67}{6}} = \sqrt{\frac{22.67}{6}} = \sqrt{3.78} = 1.94$$

$df = N_1 + N_2 - 2 = 3 + 3 - 2 = 4$

Critical value: $t_{95}(4) = \pm 2.776$

$$LL = (\bar{X}_1 - \bar{X}_2) - t_\alpha\left(s_{\bar{X}_1-\bar{X}_2}\right) = (8 - 4.33) - (2.776)(1.94)$$

$$= 3.67 - 5.39 = -1.72$$

$$UL = (\bar{X}_1 - \bar{X}_2) + t_\alpha\left(s_{\bar{X}_1-\bar{X}_2}\right) = (8 - 4.33) + (2.776)(1.94)$$

$$= 3.67 + 5.39 = 9.06$$

4. We can be 99% confident that the true population mean difference is between -11.19 and 19.19. Because this interval contains zero, we cannot call these groups significantly different.

X	Y	D	D^2
14	8	6	36
9	4	5	25
7	6	1	1
$\sum X = 30$	$\sum Y = 18$	$\sum D = 12$	$\sum D^2 = 62$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{30}{3} = 10.0$$

$$\bar{Y} = \frac{\Sigma X}{N} = \frac{18}{3} = 6.0$$

$$\hat{s}_D = \sqrt{\frac{\sum D^2 - \frac{(\sum D)^2}{N}}{N-1}} = \sqrt{\frac{62 - \frac{(12)^2}{3}}{2}}$$

$$= \sqrt{\frac{62 - 48}{2}} = \sqrt{\frac{14}{2}} = \sqrt{7.0} = 2.65$$

$$s_{\bar{D}} = \hat{s}_D \Big/ \sqrt{N} = \frac{2.65}{\sqrt{3}} = \frac{2.65}{1.73} = 1.53$$

$df = N - 1 = 3 - 1 = 2$

Critical value: $t_{99}(2) = \pm 9.925$

$$LL = (\bar{X} - \bar{Y}) - t_\alpha(s_{\bar{D}}) = (10.0 - 6.0) - (9.925)(1.53) = 4.0 - 15.19 = -11.19$$

$$UL = (\bar{X} - \bar{Y}) + t_\alpha(s_{\bar{D}}) = (10.0 - 6.0) + (9.925)(1.53) = 4.0 + 15.19 = 19.19$$

5. **The appropriate test is an independent-samples *t* test, because there is no indication of any matching, natural pairs, or repeated-measures designs.**

   We can be 95% confident that the true population mean difference is between -7.41 and 21.41. Because this interval contains zero, we cannot call these groups significantly different. Students who changed their answers ($\bar{X}_1$ = 82.6) did not perform significantly differently from those who stayed with their initial answers ($\bar{X}_2$ = 75.6). The effect size index, *d* = 0.71, was medium.

Students who changed answers ($\bar{X}_1$)	$\bar{X}_1^{\,2}$	Students who stayed with initial answers ($\bar{X}_2$)	$\bar{X}_2^{\,2}$
85	7225	74	5476
73	5329	75	5625
69	4761	88	7744
99	9801	70	4900
87	7569	71	5041
$\sum \bar{X}_1 = 413$	$\sum \bar{X}_1^{\,2} = 34{,}685$	$\sum \bar{X}_2 = 378$	$\sum \bar{X}_2^{\,2} = 28{,}786$

$$\bar{X}_1 = \frac{\Sigma X}{N} = \frac{413}{5} = 82.6$$

$$\bar{X}_2 = \frac{\Sigma X}{N} = \frac{378}{5} = 75.6$$

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sum X_1^{\,2} - \frac{(\sum X_1)^2}{N_1} + \sum X_2^{\,2} - \frac{(\sum X_2)^2}{N_2}}{N_1(N_2 - 1)}}$$

$$= \sqrt{\frac{34{,}685 - \frac{413^2}{5} + 28{,}786 - \frac{378^2}{5}}{5(5-1)}}$$

$$= \sqrt{\frac{34{,}685 - 34{,}113.8 + 28{,}786 - 28{,}576.8}{20}}$$

$$= \sqrt{\frac{571.2 + 209.2}{20}} = \sqrt{\frac{780.4}{20}} = \sqrt{39.02} = 6.25$$

$df = N_1 + N_2 - 2 = 5 + 5 - 2 = 8$

Critical value: $t_{.05}(8) = \pm\ 2.306$

$$LL = (\bar{X}_1 - \bar{X}_2) - t_\alpha\left(s_{\bar{X}_1 - \bar{X}_2}\right) = (82.6 - 75.6) - (2.306)(6.25)$$

$$= 7.0 - 14.413 = -7.41$$

$$UL = (\bar{X}_1 - \bar{X}_2) + t_\alpha\left(s_{\bar{X}_1 - \bar{X}_2}\right) = (82.6 - 75.6) + (2.306)(6.25)$$

$$= 7.0 + 14.413 = 21.41$$

$$\hat{s}_1 = \sqrt{\frac{\sum {X_1}^2 - \frac{(\sum X_1)^2}{N}}{N - 1}} = \sqrt{\frac{34{,}685 - \frac{413^2}{5}}{4}}$$

$$= \sqrt{\frac{571.2}{4}} = \sqrt{142.8} = 11.95$$

$$\hat{s}_2 = \sqrt{\frac{\sum {X_2}^2 - \frac{(\sum X_2)^2}{N}}{N - 1}} = \sqrt{\frac{28{,}786 - \frac{378^2}{5}}{4}}$$

$$= \sqrt{\frac{209.2}{4}} = \sqrt{52.3} = 7.23$$

$$\hat{s}_p = \sqrt{\frac{(N_1 - 1){\hat{s}_1}^2 + (N_2 - 1){\hat{s}_2}^2}{N_1 + N_2 - 2}} = \sqrt{\frac{(5 - 1)11.95^2 + (5 - 1)7.23^2}{5 + 5 - 2}}$$

$$= \sqrt{\frac{(4)(142.80) + (4)(52.27)}{8}} = \sqrt{\frac{571.2 + 209.09}{8}} = \sqrt{\frac{780.29}{8}}$$

$$= \sqrt{97.54} = 9.88$$

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\hat{s}_p} = \frac{82.6 - 75.6}{9.88} = \frac{7.0}{9.88} = 0.71$$

6. We can be 95% confident that the true population mean difference is between -5.53 and -0.47. Because this interval does not contain zero, we can call these groups significantly different. The average scores on X ($\bar{X}$ = 4.60) were significantly different than average scores on Y ($\bar{Y}$ = 7.60). The effect size index, d = 2.46, was large.

X	Y	D	D^2
8	11	-3	9
6	9	-3	9
5	7	-2	4
3	5	-2	4
1	6	-5	25
$\sum X$ = 23	$\sum Y$ = 38	$\sum D$ = -15	$\sum D^2$ = 51

$$\bar{X} = \frac{\Sigma X}{N} = \frac{23}{5} = 4.6$$

$$\bar{Y} = \frac{\Sigma X}{N} = \frac{38}{5} = 7.6$$

$$\hat{s}_D = \sqrt{\frac{\sum D^2 - \frac{(\sum D)^2}{N}}{N-1}} = \sqrt{\frac{51 - \frac{(-15)^2}{5}}{4}}$$

$$= \sqrt{\frac{51-45}{4}} = \sqrt{\frac{6}{4}} = \sqrt{1.5} = 1.22$$

$$s_{\bar{D}} = {}^{\hat{s}_D}\!\Big/_{\sqrt{N}} = \frac{1.22}{\sqrt{5}} = \frac{1.22}{2.24} = 0.55$$

$df = N - 1 = 5 - 1 = 4$

Critical value: $t_{.01}(4) = \pm\ 4.604$

$$LL = (\bar{X} - \bar{Y}) - t_\alpha(s_{\bar{D}}) = (4.6 - 7.6) - (4.604)(0.55)$$

$$= -3.0 - 2.53 = -5.53$$

$$UL = (\bar{X} - \bar{Y}) + t_\alpha(s_{\bar{D}}) = (4.6 - 7.6) + (4.604)(0.55)$$

$$= -3.0 + 2.53 = -.47$$

$$d = \frac{\bar{X} - \bar{Y}}{\hat{s}_D} = \frac{4.6 - 7.6}{1.22} = \frac{-3}{1.22} = |-2.46|$$

7. H$_0$: $\mu_{\text{personalized}} = \mu_{\text{ergonomic}}$
H$_1$: $\mu_{\text{personalized}} \neq \mu_{\text{ergonomic}}$

$$\hat{s}_D = \sqrt{\frac{\sum D^2 - \frac{(\sum D)^2}{N}}{N-1}} = \sqrt{\frac{1021 - \frac{(257)^2}{213}}{213-1}}$$

$$\hat{s}_D = \sqrt{\frac{1021 - 310.09}{212}} = \sqrt{\frac{710.91}{212}} = \sqrt{3.35} = 1.83$$

$$s_{\bar{D}} = {}^{\hat{s}_D}\!\Big/_{\sqrt{N}} = \frac{1.83}{\sqrt{213}} = \frac{1.83}{14.59} = 0.13$$

$$t = \frac{\bar{X} - \bar{Y}}{s_{\bar{D}}} = \frac{6-5}{0.13} = \frac{1}{0.13} = 7.69$$

$$d = \frac{\bar{X} - \bar{Y}}{\hat{s}_D} = \frac{6-5}{1.83} = \frac{1}{1.83} = |0.55|$$

$df = N - 1 = 213 - 1 = 212$

Critical value: $t_{.05}(120) = \pm\ 1.980$

According to the data, we reject the null hypothesis. The average expected affective reaction for a personalized mug ($\bar{X}$ = 6.00) was significantly greater than the expected affective reaction for an ergonomic mug ($\bar{Y}$ = 5.00). The effect size index, d = 0.55, was medium. Thus, participants expected a stronger positive affective reaction after receiving a personalized gift.

APA format: The average expected affective reaction to a personalized mug (M = 6.00) is significantly higher than the expected affective reaction to a personalized mug (M = 5.00), t(212) = 7.69, p < .05. The size of the difference was medium, d = 0.55.

8. H_0: $\mu_{\text{receiver}} = \mu_{\text{giver}}$
H_1: $\mu_{\text{receiver}} \neq \mu_{\text{giver}}$

$$s_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sum X_1{}^2 - \frac{(\sum X_1)^2}{N_1} + \sum X_2{}^2 - \frac{(\sum X_2)^2}{N_2}}{N_1(N_2-1)}}$$

$$= \sqrt{\frac{2816 - \frac{508^2}{106} + 3453 - \frac{569^2}{106}}{106(106-1)}} = \sqrt{\frac{2816 - 2434.57 + 3453 - 3054.35}{11130}}$$

$$= \sqrt{\frac{381.43 + 398.65}{11130}} = \sqrt{\frac{780.08}{11130}} = \sqrt{.07} = .26$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1-\bar{X}_2}} = \frac{4.79 - 5.37}{0.26} = \frac{-0.58}{0.26} = -2.23$$

$df = N_1 + N_2 - 2 = 106 + 106 - 2 = 210$

Critical value: $t_{.05}(120) = \pm\ 1.980$

$$\hat{s}_1 = \sqrt{\frac{\sum X_1{}^2 - \frac{(\sum X_1)^2}{N}}{N-1}} = \sqrt{\frac{2816 - \frac{508^2}{106}}{105}}$$

$$= \sqrt{\frac{381.43}{105}} = \sqrt{3.63} = 1.91$$

$$\hat{s}_2 = \sqrt{\frac{\sum X_2{}^2 - \frac{(\sum X_2)^2}{N}}{N-1}} = \sqrt{\frac{3453 - \frac{569^2}{106}}{105}}$$

$$= \sqrt{\frac{398.65}{105}} = \sqrt{3.80} = 1.95$$

$$\hat{s}_p = \sqrt{\frac{(N_1-1)\hat{s}_1{}^2 + (N_2-1)\hat{s}_2{}^2}{N_1+N_2-2}} = \sqrt{\frac{(106-1)1.91^2 + (106-1)1.95^2}{106+106-2}}$$

$$= \sqrt{\frac{(105)(3.65) + (105)(3.80)}{210}} = \sqrt{\frac{383.25 + 399}{210}} = \sqrt{\frac{782.25}{210}}$$

$$= \sqrt{3.73} = 1.93$$

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\hat{s}_p} = \frac{4.79 - 5.37}{1.93} = \frac{-0.58}{1.93} = |-0.30|$$

According to the data, we reject the null hypothesis. Participants who imagined receiving a personalized mug ($\bar{X}_1$ = 4.79) were significantly less likely to prefer the mug than those who imagined giving the personalized mug ($\bar{X}_2$ = 5.37). The effect size index, *d* = 0.30, was small.

APA format: Participants who imagined giving a personalized mug were significantly more likely to prefer the mug (*M* = 5.37) than those who imagined receiving the personalized mug (*M* = 4.79), *t*(210) = -2.23, *p* < .05. The size of the difference was small, *d* = 0.30.

References

Benjamin, L. T., Cavell, T. A., & Shallenberger, W. R. (1984). Staying with initial answers on objective tests: Is it a myth? *Teaching of Psychology*, *11*(3), 133-141.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, *1*, 98-101.

Yang, A. X., & Urminsky, O. (2018). The smile-seeking hypothesis: How immediate affective reactions motivate and reward gift giving. *Psychological Science*. Advance online publication. doi: 10.1177/0956797618761373