CHAPTER 8

Hypothesis Testing and Effect Size: One-Sample Designs

Summary

This chapter is the first of a two-chapter sequence that covers the basics of *null hypothesis statistical testing (NHST)*. NHST is a statistical tool that helps you make decisions about populations when you have sample data. This chapter introduces you to techniques that are appropriate when data are gathered from only *one sample*. Techniques that apply data from two or more samples are covered in later chapters.

There are several steps involved in NHST:

1. Hypothesize that the sample's statistic (for example, $\overline{X}$ or r) came from a particular population with a parameter (μ_0 or ρ). This is a hypothesis of equality and is called the null hypothesis (H_0).
2. Choose a sampling distribution that shows the probability of different sample statistics when samples are drawn from the population that has the particular parameter. In this chapter, the proper sampling distribution is the *t distribution*.
3. Using the proper sampling distribution, determine the probability of your sample mean, or one more extreme. This probability is correct, *if the null hypothesis is true.*
4. If the probability is small (equal to or less than .05), conclude that the null hypothesis is incorrect and that your sample data have likely come from some other population.
5. If the probability is large (greater than .05), conclude that the data are consistent with the null hypothesis and perhaps with other hypotheses as well. In statistical terms, you are now left with insufficient evidence to reject the null hypothesis.

Expanding on steps 4 and 5, the continuum of probability is divided into two regions. The *rejection region* is characterized by small probabilities and the decision to *reject the null hypothesis*. The other region is characterized by larger probabilities and the decision to *retain the null hypothesis,* which is also referred to as *failure to*

reject the null hypothesis. The dividing point on the continuum is α (alpha). The experimenter chooses the α level, which is sometimes referred to as the *level of significance*. Alpha is typically .05 or less.

For the two statistics in this chapter, $\bar{X}$ and *r*, the *t* distribution is the proper sampling distribution. A *t* test on one sample gives a *t* value. This *t* value is evaluated by comparing it to the values shown in the *t distribution* table, which shows selected *t* values called *critical values*. These critical values are those points on the distribution that correspond to commonly chosen α levels. To find the appropriate critical value in the *t* distribution table, you must know the *degrees of freedom* appropriate for the data. In testing a sample mean, $df = N\text{-}1$, where N is the number of scores. In testing a correlation coefficient, $df = N\text{-}2$, where *N* is the number of pairs of score. In later chapters that explain other designs, there are other formulas for *df*.

If the data-based *t* value is greater than the critical value associated with α in the table, *reject the null hypothesis*. If the data-based *t* value is less than the critical value, *retain the null hypothesis*. Results that lead to a rejected null hypothesis are said to be *statistically significant*. Not having sufficient evidence to reject of the null hypothesis does not mean that the experiment is a failure. There are several reasons that a test may fail to reject the null hypothesis.

In addition to the null hypothesis, NHST requires an alternative hypothesis (H_1). The most common alternative hypothesis is *two-tailed*, an alternative that places half the rejection region in each tail of the sampling distribution. A two-tailed test permits the rejection of the null hypothesis for means that are either larger or smaller than the null hypothesis mean and for correlation coefficients that are either positive or negative. One kind of one-tailed test permits rejection of the null hypothesis only if the sample mean is larger than the null hypothesis mean. Such a test cannot detect a sample mean that is smaller, no matter how small.

Like all decision-making aids, NHST can lead to wrong conclusions. If the null hypothesis is indeed true, and the data lead you to reject it, you have made a *Type I error*. The probability of a Type I error is never greater than α, which is set by the researcher.

Chapter 8

If the null hypothesis is actually false and the data lead you to retain it, you have made a *Type II error*. The probability of a Type II error is symbolized by β. Among the several factors that determine β are (1) α – the smaller α is, the larger β is, and vice versa, and (2) the actual difference between the population sampled from and the null hypothesis population – the larger the difference, the smaller β is.

The formula for the *effect size index, d*, tells you about the size of the difference between the mean of the population that the sample was drawn from and the mean of the null hypothesis population. *d* values of 0.20, 0.50, and 0.80 characterize small, medium and large effect sizes for a one-sample *t* test. An effect size estimate is an increasingly popular way to describe differences.

Multiple-Choice Questions

1. In NHST, the hypothesis that is tested is about a
   (1) sample;
   (2) population;
   (3) both (1) and (2);
   (4) neither (1) nor (2).

2. Using NHST, you can conclude that the null hypothesis is
   (1) probably true;
   (2) probably false;
   (3) both (1) and (2) are possible conclusions;
   (4) neither (1) nor (2) are possible conclusions.

3. The *t* distribution, as a sampling distribution, gives the probability of events when
   (1) the null hypothesis is true;
   (2) the null hypothesis is false;
   (3) both (1) and (2) are correct at times;
   (4) the alternative hypothesis is not identified.

4. Suppose the difference between a sample mean and the null hypothesis mean was in the rejection region of the sampling distribution. This means that the difference is
   (1) probably the result of mistakes;
   (2) unreasonable; it is either too large or too small;
   (3) probably due to chance;
   (4) probably not due to chance.

5. Which phrase goes with "in the rejection region"?
   (1) calculated probability is small;
   (2) reject the null hypothesis;
   (3) both (1) and (2);
   (4) neither (1) nor (2).

6. When an experimenter uses $\alpha = .05$, the rejection region is _____
   (1) 95 percent of the curve;
   (2) 5 percent of the curve;
   (3) 5 percent of a one-tailed test and 10 percent of a two-tailed test;
   (4) 10 percent of a one-tailed test and 5 percent of a two-tailed test.

7. When we reject the null hypothesis, we have evidence that the difference observed is
   (1) due to chance;
   (2) very small;
   (3) unlikely to be due to chance;
   (4) a Type I error.

8. An effect size index is most closely associated with which phrase below?
   (1) The α level chosen by the researcher;
   (2) The probability of a Type II error;
   (3) The size of the difference between the sample mean and the null hypothesis mean;
   (4) The size of the rejection region.

9. Suppose you obtained a sample from a population for which the null hypothesis was true. On the basis of a *t* test, you failed to reject the null hypothesis. You have made a
   (1) Type I error;
   (2) Type II error;
   (3) correct decision;
   (4) any of the above; more information is needed.

10. Suppose you obtained a sample from a population different from the one specified by the null hypothesis. On the basis of a *t* test, you failed to reject the null hypothesis. You have made a
   (1) Type I error;
   (2) Type II error;
   (3) correct decision;
   (4) any of the above; more information is needed.

11. "$p > .05$" means that
   (1) the null hypothesis should be rejected;
   (2) the difference between the statistic and the null hypothesis parameter is statistically significant;
   (3) both (1) and (2);
   (4) neither (1) nor (2).

12. We reject the null hypothesis when
   (1) $p > .05$;
   (2) $p < .05$;
   (3) not enough information to answer this question.

13. Which of the following shows a correct match-up of an alternative hypothesis and its one- or two-tailed test?
   (1) H_1: $\mu_0 = \mu_1$; one-tailed;
   (2) H_1: $\mu_0 < \mu_1$; two-tailed;
   (3) both (1) and (2);
   (4) neither (1) nor (2).

Chapter 8

14. The choice of an alternative hypothesis has an effect on
    (1) conclusions you may draw;
    (2) α level;
    (3) which null hypothesis you are testing;
    (4) all of the above.

15. Which answer below belongs with the concept of a two-tailed test of significance?
    (1) $H_1: \mu_1 > \mu_0$;
    (2) Type II error;
    (3) both (1) and (2);
    (4) a divided rejection region.

16. A one-tailed test is proper when
    (1) you do not have enough data for a two-tailed test;
    (2) you have only one sample, not two;
    (3) you are interested in finding out only that the effect of a treatment is to increase the scores;
    (4) you want to make the standard error of the mean as small as possible.

17. When the *t* distribution is used to determine the significance of a correlation coefficient, the null hypothesis is that the population correlation coefficient is
    (1) –1.00;
    (2) 0.00;
    (3) 1.00;
    (4) the coefficient obtained from the sample.

18. Your text concluded that the Frito-Lay corporation, which claims to put 269.3 grams of Doritos tortilla chips in their packages, actually puts in _____ that amount.
    (1) about;
    (2) exactly;
    (3) significantly more than;
    (4) significantly less than.

19. The custom of using an α level of .05 got its start in the field of
    (1) astronomy;
    (2) agriculture;
    (3) physics;
    (4) government.

20. If the *t* test value is less than the critical value on the table, _____ the null hypothesis even though you could be making a _____ error.
    (1) reject, Type I;
    (2) retain, Type II;
    (3) reject, Type II;
    (4) retain, Type I.

Short-Answer Questions _____

1. Distinguish between Type I and Type II errors.

2. Distinguish between $\alpha = .05$ and $p = .05$.

3. Distinguish between rejecting and retaining the null hypothesis.

4. Write an interpretation of each of the following situations.
    a. Sometimes natural events produce experiences that cannot be duplicated in the laboratory. The abduction, confinement, and release of hostages is an example. Does this experience have any effect on the personality characteristic of dominance? The California Personality Inventory has a scale for Dominance (D_o); high scores indicate confidence, assertiveness, and task orientation, and low scores indicate an unassuming and unforceful person. A score of 50 is the mean for the D_o scale; the standard deviation is 10. Suppose that the mean D_o score for a group of 8 released hostages was 42. A *t* test produced a value of 2.25. The calculated value of *d* was 0.80. Write the null hypothesis and a conclusion about the effects of confinement.
    b. Are teachers accurate at assessing the honesty of their students? Is there any relationship between teacher ratings and test scores on a test of honesty? This last question can be answered with a correlation coefficient. Murphy and Davidshofer (1991, p. 121) report a Pearson *r*

of .62 between teacher ratings of honesty and a test designed to assess honesty. If the data were based on 15 students, what conclusion can you draw?

5. Write an interpretation of each of the following situations.
   a. Stanley Milgram found in the early 1960s that a cross section of Americans was willing to administer an average of 285 volts to other participants in an experiment (Milgram, 1963). Have times changed? Suppose that the study was replicated today with 20 participants who were willing to administer an average of only 240 volts. A *t* test produced a value of 2.00. Write the null hypothesis and a conclusion about the difference between today and the early 1960s in people's willingness to administer shock to others.
   b. Matsumoto, Kasri, and Kooken (1999) found that there was a strong correlation between people of different cultures and facial expressions. That is, there was strong agreement of what a facial expression represented across cultures. If the correlation was .44, and was based on a sample of 25, what conclusion would be appropriate?

6. Write a paragraph or two explaining NHST.

Problems

1. Remember the Personal Control Scores in Chapter 2? The mean PC score for all college students is 51. What about students who are in academic difficulty? How do they feel about the control of their personal lives? The hypothetical data in this problem are PC scores for students in academic difficulty. Perform a *t* test, calculate an effect size index, and write a conclusion about feelings of personal control among students in academic difficulty.

48 44 53 35 58 42 55 37 50 46 40 32

2. Every year thousands of college-bound American high school seniors take the Scholastic Aptitude Test (SAT). During one recent year the mean score was 896 (math plus verbal). Summary statistics for one small high school follow. How do the students compare to the national norm? State the null hypothesis, choose an alternative hypothesis, perform a *t* test, calculate the effect size index, and write a conclusion.

$\Sigma X = 24{,}206$ $\Sigma X^2 = 22{,}716{,}411$

$N = 26$

3. Miller (1956), in a classic study, demonstrated that participants can store 7 bits of information in short term memory. Assume you believe you can train students to do better by having them memorize 10 digit phone numbers. You train your participants for 10 days, test them and obtain the information below. Analyze the data with a *t* test and an effect size index. Write an interpretation.

$\Sigma X = 235$ $\Sigma X^2 = 1931$

$N = 30$

Raw Scores:

9	5	8	7	9	8	6	6	7	5
8	6	9	8	7	11	10	8	6	9
11	8	8	9	6	9	6	7	7	12

4. The Rathus Assertiveness Schedule is a 30-item questionnaire; the mean score for men is 11. Suppose that the student development office at your college conducted a 6-week assertiveness course. At the end of the course, the male participants had the following scores. Perform a *t* test, calculate the effect size index, and write a conclusion about the assertiveness course.

25 36 -1 28 24 53 -3 46 26

Chapter 8

5. As you may recall from problem 30 in the textbook, when people choose a number between 1 and 10, the mean is 6. To investigate the effects of blatant, direct stimuli on behavior, a researcher embedded many low numbers (1 to 3) in a video clip. Each one was very obvious. Afterward the participants were asked to "choose a number between 1 and 10." Analyze the data with a *t* test and an effect size index. Write an interpretation of these data. In addition, using your analysis of Problem 30, write an interpretation based on both data sets.

$\Sigma X = 200$ $\Sigma X^2 = 1206$ $N = 40$

Raw Scores:

5	1	3	2	3	3	4	4	7	5
7	6	6	7	3	8	4	9	5	6	4
3	6	1	2	8	7	5	4	7	6	3
7	8	2	8	1	9	7	4			