CHAPTER 9

Hypothesis Testing, Effect Size and Confidence Intervals: Two-Sample Designs

Summary

For the simplest *experiment*, two populations are identified. The two populations might be reaction times after drinking coffee or decaf, test anxiety scores during the semester versus during finals week, or percent of recall after being awake or asleep. The question is whether the mean of one population, μ_1, is different from the mean of a second population, μ_2.

The two different populations are the two levels of the *independent variable.* A sample from each population is measured on the *dependent variable* and sample means are calculated. The means of the two samples will probably be different. Null hypothesis statistical testing (NHST) may permit you to decide the reason for the difference.

The *logic of null hypothesis statistical testing (NHST)* is to tentatively hypothesize that $\mu_1 = \mu_2$. This hypothesis is the null hypothesis and is symbolized H_0. If H_0 is true, then the difference in the two sample means is due to sampling fluctuation. A sampling distribution such as the *t* distribution shows the probability of differences between sample means *if H_0 is true.* By finding the probability of the observed difference in sample means, the researcher can draw a conclusion. If the probability is small (.05 or less, for example), H_0 can be rejected. (because such sample data are unlikely if H_0 is true). If the probability is larger (.051 or larger, for example), retain H_0. A *t* test gives you the probability of the observed data *if H_0 is true.*

The probability value that separates rejecting H_0 from retaining H_0 is called α. .05 is the largest commonly accepted α value. The α value is also referred to as the significance level. If the probability of the observed difference is less than α, the difference is *statistically significant.* Note that this logic is *identical* to the logic in the previous chapter. The difference is in the comparison. In this chapter, we compare *two sample means.* In Chapter 8, we compared the *mean of a sample to the mean of a population.*

Chapter 9

The last step in data analysis is a carefully worded statement about the effects of the independent variable on the dependent variable. Terms that describe the variables are used in this statement, not generic terms such as experimental group and dependent variable. Once the statistics are completed, write a clear explanation of what the data show.

In addition to deciding on an α level before the data are gathered, a researcher must also decide whether the *t* test will be one or two-tailed. *Two-tailed tests*, which are much more common, allow a conclusion that μ_1 is larger OR that it is smaller than μ_2. 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.

Paired-samples designs are two-group experiments in which the scores consist of pairs. The pairs might exist before the experiment begins (*natural pairs*); one member of the pair serves in the experimental group and the other in the control group. Pairs might be formed by matching the participants on some variable related to the dependent variable (*matched pairs*). Finally, participants might serve in a before-and-after design (*repeated measures*) which 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).

An independent-samples design occurs when the scores are not paired in any logical way. The *t* distribution for an independent-samples design has N_1+N_2-2 degrees of freedom.

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, and
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.

Chapter 9

Besides testing a null hypothesis about a difference between two levels of tl independent variable, 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 captur a range of values. Within this range, you can expect, with a certain degree of confidence, to find the difference that exists between the two population means tha the two samples are drawn from.

This chapter provides additional information about *effect size*. The *effect siz index, d*, describes the size of the difference between the two populations. Althougl 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.

The more powerful a statistical test is, the better its ability to detect a false null hypothesis. Power is equal to 1-β, where β is the probability of a Type II error. (For an excellent article on power, see Cohen [1992]). The factors that determine power are:

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

Multiple-Choice Questions

1. For a simple experiment, which of the following is true?
   (1) for the null hypothesis, assume that the two groups represent different populations;
   (2) apply the dependent variable to both groups and then measure the changes in the independent variable;
   (3) find the probability that the two samples are from different populations by using a *t* distribution;
   (4) treat both groups exactly alike except for one thing.

Chapter 9

2. The logic of hypothesis testing is to assume that two populations have
   (1) means that are equal and then see if sample data will permit you tc conclude that they are probably equal;
   (2) means that are equal and then see if sample data will permit you tc conclude that they are probably unequal;
   (3) means that are not equal and then see if sample data will permit you to conclude that they are probably unequal;
   (4) means that are not equal and then see if sample data will permit you to conclude that they are probably equal.

3. Which conclusion is *not* appropriate when using hypothesis testing?
   (1) The two sample means probably came from two different populations.
   (2) The two samples probably came from the same population.
   (3) Retain the hypothesis that the two sample means came from the same population.
   (4) All of the above.

4. In an independent-samples design, the null hypothesis is that
   (1) the population mean of one group is equal to that of a second group;
   (2) the population mean of one group is larger to smaller than that of a second group;
   (3) the sample mean of one group is equal to that of a second group;
   (4) the sample mean of one group is larger or smaller than that of a second group.

5. According to your text the reason we do experiments is to be able to tell
   (1) whether all extraneous variables were controlled;
   (2) whether the samples were representative of the population;
   (3) how dependent variable scores are affected by the independent variable;
   (4) all of the above.

6. " _____ depends on the number of observations minus the number of relations among the observations" is a statement about how to calculate
   (1) *df*;
   (2) the difference between population means;
   (3) $s_{\bar{X}}$;
   (4) none of the above.

7. 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. The design
   (1) was a paired-samples one;
   (2) was an independent-samples one;
   (3) could be either a paired- or an independent-samples one.

8. A one-tailed test of significance produced a *t* equal to –2.30, significant at th .05 level. The design of this experiment
   (1) was a paired-samples design;
   (2) was an independent-samples design;
   (3) could have been either a paired- or independent-samples design.

9. With an acknowledgment to Sesame Street, "Which of these things is not lik the others, which of these things doesn't belong?"
   (1) repeated measures;
   (2) natural pairs;
   (3) independent samples;
   (4) matched pairs.

10. There has been bad blood between the Montague family and the Capulet family for a good while. In this modern day, resolution can be achieved by using a psychological test. In the test of "propensity to fall in love," the mean of the 6 Montagues was 54 and the mean of the 10 Capulets was 64. (Italian norms show a national average of 100.) When a statistician compared the families with a *t* test, a value of 2.13 was obtained. If you adopt an α level of .05 (two-tailed test), you should conclude that the Capulets are
    (1) significantly more loving than the Montagues;
    (2) significantly less loving than the Montagues;
    (3) not significantly different from the Montagues;
    (4) not yet comparable; additional information is needed.

11. In an independent samples design the Hatfields had a mean score of 25; the mean score of the McCoys was 26. Low scores mean better performance. The researcher ran a two-tailed test with α at .05. A *t* value of 1.99 was found. Which of the following is true?
    (1) If $df = 40$, the Hatfields are significantly better than the McCoys;
    (2) If $df = 40$, the McCoys are significantly better than the Hatfields;
    (3) If $df = 120$, the Hatfields and the McCoys are not significantly different;
    (4) If $df = 120$, the Hatfields are significantly better than the McCoys;
    (5) If $df = 12$, the McCoys are significantly better than the Hatfields.

12. Which of the following variables affect the size of the standard error of a difference?
    (1) difference between sample means;
    (2) sample size;
    (3) both (1) and (2);
    (4) neither (1) nor (2).

13. $p < .05$ means that the difference between sample means
    (1) fell outside the rejection region;
    (2) should be attributed to chance rather than to the independent variable;
    (3) should be declared "not significant";
    (4) none of the above.

14. Which of the following answers has effect size indexes that are considered small?
    (1) 0.01 and 0.05;
    (2) 0.10 and 0.20;
    (3) both (1) and (2);
    (4) neither (1) nor (2).

15. The power of a statistical test is defined as
    (1) α;
    (2) β;
    (3) $1 - \alpha$;
    (4) $1 - \beta$.

16. The 95 percent confidence interval about a mean difference was –3.0 minute to 6.5 minutes. The null hypothesis that the two population means are equal
    (1) can be rejected at the .05 level;
    (2) can be rejected at the .01 level;
    (3) can be rejected at both the .05 and the .01 level;
    (4) should be retained.

17. The *main* difference between a paired-sample and independent-sample *t* is
    (1) sample size;
    (2) *df*;
    (3) the organization of the data;
    (4) the analysis of the data.

18. 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?
    (1) paired-sample;
    (2) independent sample;
    (3) neither.

19. Which of the following has an influence on the *power* of a statistical test?
    (1) sample size;
    (2) alpha;
    (3) actual difference;
    (4) all of the above.

20. The *df* are most closely related to
    (1) sample size;
    (2) alpha;
    (3) actual difference between population means;
    (4) choice of a one- or two-tailed test.

Short-Answer Questions

1. Identify the design and the degrees of freedom for each of the following experiments.
   a. To determine which is the lowest form of humor, 14 sophomores rated a pun and then a limerick for lowness.
   b. To determine which is the lowest humor, a Greek physician found the amount of blood and the amount of lymph in the sole of each of 21 Greek philosophers.
   c. To determine which is the longest form of humerus, an anthropologist measured that bone in 15 men and 15 women and compared the sexes.
   d. To determine whether the psychologist or the philosopher had the lowest form of humus in his garden, 12 samples were taken from the garden of each. The amount of humus was determined for each sample.
   e. Every student knows, of course, that the very lowest form of humor is test humor. To determine if test humor has become even lower over time, the tests of 34 young professors were compared for lowness to the tests of each of their own teachers when they were young.

2. Identify the design and the degrees of freedom for each of the following experiments.
   a. The effect of cola on attention was measured by counting the number of "eye reversals" in videotapes of students reading the *Iliad*. Fifty participants were observed for 10 minutes. Each consumed 12 oz. of cola during 5 minutes. The 50 were then observed for 10 more minutes.
   b. 21 famous sociologists rate their attitude toward statistics and then identified their best student who had obtained a PhD in sociology. Attitudes toward statistics were then obtained from these 21 also.

c. A consumer group compared two detergents, Bold and Tide, to determine which was better. 24 white wash cloths 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. The mean IQ, reading level, and age of a classroom of 25 Native Americans was equal to that of 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.

3. List the factors that influence whether or not you reject the null hypothesis. Explain how each factor influences the final decision.

4. A two-group experiment might have the phrase "$p = .01$." Explain by finishing the sentence: The probability is .01 that...

5. Please list and explain the three factors your text identified that influence whether or not you reach correct conclusions when you use a *t* test.

6. Those interested in the nature of *Homo sapiens* have often asked, "Is experience necessary for this behavior, or will it develop without any experience?" One way of answering this question has been to study different cultures. The rationale is that, if the cultures are quite different but the behavior is the same, then experience is not necessary. In the case of walking, a behavior of fundamental importance, comparisons have been made between Native American cultures and Anglo American cultures. Some Native American babies spent most of the day bound to a board on their mothers' backs and had few opportunities to creep, crawl, and kick. The age (in months) at which children from the two cultures first walked was the dependent variable in this study.

   A dozen Native American children began to walk at a mean age of 12.3 months; for a dozen Anglo American children, the mean was 12.1 months. A *t* test on the means produced a value of 0.05. Identify the design as

independent samples or correlated samples, give the critical value for t at the .05 level, and write a conclusion about the effect of experience on walking behavior.

7. This is an experiment on set (previous experience) that is from the same tradition as the "two-string problem" in your text. This experiment is based on one by Luchins (1942) and is referred to as the "water-jar problem." Participants mentally used three jars to measure out a specific amount of water. For example, if the three jars held 12.4, and 3 units and the task was to obtain 5 units, you could fill the 12-unit jar and from it fill the 4- and 3-unit jars once, leaving 5 units in the larger jar. After giving the kind of explanation you have just received, Luchins gave participants the series of eight problems below. It will be worthwhile to work these problems in order yourself, noting in the margin the number of seconds it takes you to solve each problem. Work the eight problems before reading on.

Problem	Jars Contain			Obtain
1	21	127	3	100
2	14	163	25	99
3	18	43	10	5
4	9	42	6	21
5	20	59	4	31
6	23	49	3	20
7	15	39	3	18
8	28	76	3	25

Some participants worked the problems in the order that you followed, and some started with Problem No. 8. The dependent variable was the time necessary to solve Problem No. 8. The independent variable was whether the participant had received the "set" generated by working Problems 1 through 7. (If you worked the problems as suggested, you established such a set.)

Those who worked Problem No. 8 first required significantly less time to find an answer than those who worked the problem last. Identify the design as independent- or repeated-samples, and write a conclusion about this experiment on set.

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

9. What is the best type of statistic to use if you are comparing a pre-test post-test experiment?

10. List two reasons to run a paired-sample *t* when it is possible.

Problems

1. In an early study of the effects of frustration on feelings of hostility, Miller and Bugelski (1948) had a group of boys at a camp rate their attitudes toward two minority groups (Mexicans and Japanese). The campers then participated in a long, difficult testing session that kept them away from their weekly movie. Finally the boys again rated their attitudes toward the minority groups. The scores below are similar to those of Miller and Bugelski; they represent the number of unfavorable traits attributed to minorities. Analyze them with a *t* test and an effect size index, and explain your analysis.

Participant	Before Testing	After Testing
1	5	6
2	4	4
3	3	5
4	3	4
5	2	4
6	2	3
7	1	3
8	0	2

2. R. S. Lazarus (1964) had two groups of participants watch a film that showed accidents occurring in a workshop. The accidents were gruesome events such as fingers being cut off and a plank being thrown through a man's midsection by a circular saw. One group was instructed to remain detached from the events. The other group was instructed to become involved. Heart rate was monitored and increases noted. The data that follow are similar to those obtained by Lazarus. Analyze the data with a *t* test and an effect size index

and comment on human ability to control emotions (as measured by heart rate increase).

Detached	Involved
23	31
21	27
19	24
15	23
14	21
12	14
10	

3. As a result of research before 1900, E. L. Thorndike concluded that animals were incapable of learning by imitation. In 1901, however, L. L. Hobhouse reported that cats, dogs, otters, elephants, monkeys, and chimpanzees could learn by imitation. Suppose the following study was conducted to study the question. One group of hungry cats was shown food being obtained from under a vase. Another group was not, although there was food under the vase. Shortly afterward, the time (in seconds) required to upset the vase and find the food was recorded for each animal. Results are given below. Calculate a 95% confidence interval. Which of the two theorists does your analysis support?

Shown	Not Shown
18	25
15	22
15	21
12	19
11	16
9	15

4. A number of studies have used animals to examine the relationship between neuroticism and alcoholism. Here is a typical study. From each of seven litters two cats were randomly selected and assigned to one of two groups. One group was subjected to a procedure that induced

temporary neurosis. Then all cats were offered milk spiked with 5% alcohol. The amount consumed in three minutes was measured in milliliters. Decide if this is a correlated-samples design or an independent-samples design. Analyze the data with a 99% confidence interval. Comment on the relationship between neuroticism and alcohol consumption.

Littermates	No Experimental Neurosis	Experimental Neurosis
1	63	88
2	59	90
3	52	74
4	51	78
5	46	78
6	44	61
7	38	54

5. A 1984 study by Benjamin, Cavell, and Shallenberger tested the question of whether a student should change an answer on a multiple choice question when re-checking a test. They found that changing answers on multiple-choice exams *increased* scores on exams. Assume they found the following data.

Scores on exams for students who changed answers	Scores on exams for students who stuck with initial response
85	74
73	75
69	88
99	70
87	71

What is the appropriate statistical test for this experiment? Analyze the data and write a conclusion.