13 The Two-Group Design

The prototypical research study is the two-group design. Persons are in one of two groups. For example, one group receives an experimental treatment and another group serves as a control group. Examples of treatments are a new drug to cure cancer, an instructional program for disadvantaged children, a procedure to change attitudes, a pain relief strategy for childbirth, and an exercise program. The *control group* is a group of persons who are assumed to be identical to those in the treatment group except that individuals in the control group do not receive the treatment.

In this chapter a model for the analysis of the two-group design is presented. Also discussed are the design considerations concerning the assignment of persons to treatment groups and the formation of the two groups. Measures of the differences between the two groups are presented and statistical power considerations are discussed.

Model

The model for the two-group design is fairly simple. The model is

$$\begin{matrix}\text{dependent}\\ \text{variable}\end{matrix} = \text{constant} + \begin{matrix}\text{effect of the}\\ \text{independent}\\ \text{variable}\end{matrix} + \begin{matrix}\text{residual}\\ \text{variable}\end{matrix}$$

The restricted model is identical to the above model except that the independent variable has no effect on the dependent variable. Hence

$$\begin{matrix}\text{dependent}\\ \text{variable}\end{matrix} = \text{constant} + \begin{matrix}\text{residual}\\ \text{variable}\end{matrix}$$

The restricted model in this chapter is identical to the complete model discussed in the previous chapter.

Consider the terms in the complete model. The dependent variable is what changes or varies. It is the outcome that the treatment variable is designed to alter. The constant is the average score of persons in both groups. (The test of the constant was extensively discussed in Chapter 12.) The independent variable is a nominal variable with two levels—that is, a dichotomy. For instance, one level is the treatment group and the other is the control group. The independent variable is the variable that causes the dependent variable to change or vary. The residual variable represents variation in the dependent variable that is not explained by the independent variable. The residual variable is forced to have a mean of zero.

As an example, consider an experiment in which a researcher randomly assigns ten infants to one of two groups. All infants spend 20 minutes with a stranger. Then the infants are put into a situation with a number of fear-arousing stimuli. For five of the ten infants the stranger is present (present condition), and for the other five the stranger is absent (absent condition). The researcher measures the number of fear responses of the ten infants.

Present	*Absent*
6	12
4	6
3	8
7	10
4	7

The means are 4.8 for present and 8.6 for absent.

The central question with the two-group design is whether the independent variable affects the dependent variable. If the independent variable did not affect the dependent variable, as in the restricted model, then the *population* means for the two groups are equal. Because of sampling error, the two *sample* means are not exactly equal even if the restricted model is true. For the example it is not known whether the difference between the means of 4.8 and 8.6 can be explained by sampling error.

The two groups will be designated 1 and 2. The sample means will be designated $\bar{X}_1$ and $\bar{X}_2$, with sample sizes of n_1 and n_2, respectively. At issue is the amount of sampling error in the quantity $\bar{X}_1 - \bar{X}_2$ given that the population means are equal.

In Chapter 9 the idea that statistics vary was presented. The standard deviation of a statistic is called the standard error. As shown in Chapter 11, the standard error of the difference between two means randomly and independently sampled from the same population is

$$\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where σ is the population standard deviation of the observations and n_1 and n_2 are the sample sizes of the two sample means.

To estimate the standard error of the difference between two means an

estimate of σ is needed. In terms of the model, σ is the population standard deviation of the residual variable in the restricted model. The variance of the residual variable can be estimated by computing the variance of scores within each of the groups. Thus the variance is computed for each of the two groups. These variances are denoted as s_1^2 and s_2^2. Both of these are unbiased estimates of σ^2, the variance of the residual variable. Some way is needed to average or pool these variances to produce the most efficient estimate of the variance. When averaging variances the most *efficient* way to do so is to weight each variance by its denominator, $n - 1$. That is, weighting by $n - 1$ results in an estimate with the smallest standard error. The most efficient estimate of σ^2 is called s_p^2, given as follows:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

For the example, the variance for the present group is 2.7 and the variance for the absent group is 5.8. The pooled variance or s_p^2 is

$$\frac{(4)(2.7) + (4)(5.8)}{5 + 5 - 2} = 4.25$$

Now that there is an estimate of σ^2, the standard error of the difference between two means sampled from the same population can be estimated. That estimate is

$$\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$$

This formula states how variable the difference between means would be if the two sets of observations were drawn from the same population. Such an assumption is made in the restricted model. So to evaluate whether the independent variable causes the dependent variable (the complete model), a model in which the independent variable has no effect is tested. Given this restricted model, the population means of the two groups are equal. To test the restricted model and the null hypothesis of equal population means, the difference between sample means is compared to its standard error. For the example the standard error of the difference between two means is

$$\sqrt{4.25\left[\frac{1}{5} + \frac{1}{5}\right]} = 1.304$$

The difference between the means is $4.8 - 8.6 = -3.8$, and its standard error is 1.304. Their ratio is $-3.8/1.304 = -2.914$.

If it were known how the quantity

$$\frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

was distributed, it could be more precisely estimated how unusual the difference between the means is relative to its standard error. For the example, the question is how unusual is –2.914, the mean difference divided by its standard error. It happens that

$$\frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has a t distribution with $n_1 + n_2 - 2$ degrees of freedom, given a series of assumptions that are discussed in the following section.

As discussed in the previous two chapters, the t distribution closely resembles the Z or standard normal distribution except that it is less peaked and has fatter tails. The tails are fatter because the denominator of t is the statistic s, whereas the denominator of Z is the parameter σ. How fat the tails of t are depends on how precise the estimate of the variance is, and that precision depends on the degrees of freedom. There is then a family of t distributions, which vary by their degrees of freedom.

For the two-group study, the degrees of freedom are $n_1 + n_2 - 2$. Because $n_1 + n_2$ equals the number of persons in the study, the degrees of freedom are the total number of persons in the study less two. It is less two because the means for the two groups are estimated.

In the two-group study, to test the restricted model that the independent variable has no effect on the dependent variable, the test statistic is computed

$$\frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

The test statistic is then compared to the critical values, ignoring sign, in Appendix D for the appropriate degrees of freedom. As is explained in the previous chapter, if the exact degrees of freedom are not in the table, one rounds *down* to the nearest value and then determines whether the test statistic, ignoring sign, is larger than any critical value for the degrees of freedom. If it is, the null hypothesis of equal population means is rejected, and the test statistic is said to be statistically significant. The p value is determined by noting the largest value in the table that the test statistic exceeds. The p level is given by the column heading. If the test statistic, ignoring sign, is smaller than all values in the table, then the difference between means is not statistically significant and the null hypothesis of equal population means is retained.

For the example, the *df* are eight, and a –2.914 value is statistically significant at the .02 level of significance. If a computer is used to compute the test statistic, the exact p value is .0195. The null hypothesis of equal means is rejected.

What has just been described is a two-tailed test. The null hypothesis is rejected if t is either very positive or very negative. Instead, the researcher may wish to perform a one-tailed test, which requires that he or she specify a priori which mean should be larger. For the example, theory might say that fear responses should be lower when a familiar adult is present. If the sample means confirm the prediction, one proceeds as in a two-tailed test, but the p value is cut in half. As was explained in Chapter 12, one-tailed tests are not recommended because the researcher would probably still believe the result was statistically significant even if the result were not in the predicted direction. For instance, it could have happened that fear responses increased when the stranger was present.

Assumptions

There are three major assumptions for the two-group t test, all of which refer to the residual variable:

1. normal distribution,
2. homogeneous variance, and
3. independence of observations.

The score on the residual variable for a given person is estimated by taking each person's score and subtracting the group mean.

Normality

The residual term must have a normal distribution for

$$\frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

to have a t distribution under the restricted model. To test this assumption a histogram is constructed for the set of observations minus the group mean and determine whether their shape is normal. (The normality assumption refers to residual variable and not to the dependent variable itself.) If the distribution is skewed, then the one-stretch transformations discussed in Chapter 5 should be considered; or if it is bounded on both sides, a two-stretch transformations may be needed. When any transformation of the dependent variable is contemplated, it must be determined whether transformation will render the dependent variable uninterpretable.

In practice, the normality assumption is not usually examined. With small sample sizes, it is difficult to detect that the distribution is nonnormal. With

large samples, the effect of nonnormality does not disturb the t test very much. The reason for this is the central limit theorem: As the sample size increases, the distribution of $\bar{X}$ becomes more normal even though the distribution of the scores may not be normal. Given the central limit theorem, it is also true that the distribution of $\bar{X}_1 - \bar{X}_2$ approaches normality as n_1 and n_2 increase.

Equal Variances

The two-sample t test requires a pooling or an averaging of the two-sample variances, s_1^2 and s_2^2. The equal variance assumption requires that the population variances of both groups are equal to the same value. Although the means may differ, the variances are assumed not to. A procedure is needed for determining whether the sample variances are significantly different from one another. That is, a way is needed to determine whether the sample variances differ by more than the amount expected given sampling error. It turns out that the ratio of the two sample variances is distributed as F given the null hypothesis that their population variances are equal. The F test is presented in the next chapter.

If the variances differ significantly, there are a number of strategies available. First, one might consider transformations to promote equal variances. For instance, if the data are skewed, the one-stretch transformations described in Chapter 5 may make the variances in the groups more nearly equal. Second, if n_1 is nearly equal to n_2, the problem can be safely ignored, because the t test is only slightly affected by unequal variances. However, if the variances and sample sizes are unequal, caution must be exercised in interpreting p values. It must be determined which group has the larger variance. The t test results in too many Type I errors if the group with the larger variance also has the smaller n. The t test results in too few Type I errors if the group with the larger variance has the larger n.

Independence

The scores of persons on the residual variable are assumed to be uncorrelated. *Independence* requires that if one residual score is positive, the residual score of any other observation is no more likely to be positive or negative. There are a number of factors that aid in determining whether the observations are likely to be independent from each other. They concern (a) whether repeated observations are taken from the same person, (b) what the sampling unit is, and (c) whether there is social contact between the persons that generate the observations. Below is a consideration of each of these conditions.

First, whatever it is that generates the data is referred to as a *unit*. The unit may be a person, animal, or group of persons. For the two-group t test each observation must be from a different unit. So each unit, be it a person or nerve

cell, is measured only once. There must not be repeated measures from the same unit. For instance, assume that a person is measured before undergoing therapy and after and so each person is measured twice. These two observations are not likely to be independent. Also, if a behavior modification study is conducted using the same person, then the same person provides all the data and the scores are not likely to be independent. There are analysis procedures for these kinds of data structures, but they are different from the two-group *t* test.

Second, independence can be enhanced through the design of the study. The sampling unit of the study should be the unit that provides the observation. That is, each unit should enter the study singly. For instance, if married couples were in the study and both members provide data, then the independence assumption is likely to be violated because a husband is likely to be more similar to his wife than to someone else's wife. The observations must not come in pairs as in couples, friends, littermates, or twins. If they do, other statistical methods must be used.

Third, to achieve independent observations persons in the study must not influence others' responses. Once subjects enter the study, they should, if possible, be kept isolated so that they do not influence each other. They should not communicate with each other or know any other subject's response on the dependent variable. If they do communicate or observe each other, their observations are likely to be correlated because they may imitate or influence each other.

The effect of nonindependent observations is to bias the estimate of residual variance and, therefore, the standard error of the difference between means. Usually, though not always, the direction of bias in the two-group design is to make the estimate of the standard error too small, which makes researchers falsely confident that the means are significantly different. Unlike the normality and equal variance assumptions, even moderate violation of the independence assumption has very serious consequences. The failure to meet the independence assumption invalidates the *p* values.

One solution to the problem is to design the research so that observations are independent. If this is not possible, it may be possible to find a different way of analyzing the data to meet the assumption. There is one case in which observations are nonindependent, but data can be reanalyzed to meet the independence assumption. It is the case of paired observations, which is now discussed.

Paired t *Test*

Some two-group studies contain observations in which pairs of observations are linked. Each observation in one group is paired or linked to one other observation in the other group. Consider some examples:

1. Twenty-five persons enter a stop-smoking program. The number of cigarettes smoked before entering the program and six months after completion of the program are measured. There are two groups of observations: those before the treatment and at a six-month follow-up. The observations are paired, that is, each person provides two scores, one in the pretreatment group and another in the posttreatment group.
2. A researcher is interested in the different ways in which fathers and mothers treat their infant children. A total of 40 infants are observed, each with its father and mother. Again, the observations are paired. Each infant provides two data points, one in the mother group and one in the father group.
3. Pairs of rats from the same litter are used in an experiment on learning. One rat from the litter has an operation that is supposed to facilitate learning. The other rat does not have the operation. A total of 20 pairs are studied. Each litter provides two observations, one of which is in the operation group and the other in the nonoperation group.

These three examples illustrate the key element of the paired design. Each observation is linked to one and only one other observation in the other group. Thus, each of n observations in one treatment group is linked to one of n observations in a second group. The degree to which the observations are linked can be measured by a correlation coefficient.

When observations are linked in this way, the independence assumption is violated because the linked observations are likely to be correlated. This lack of independence makes the two-group analysis that has been described in this chapter no longer valid because normally the t test will yield more Type I errors than it should. It happens that the one-group t test described in the previous chapter can be applied to the paired two-group design.

The key idea is to compute a difference score, always subtracting the scores in the same way. For example, the pretreatment score is always subtracted from the posttreatment score. The test that the mean of the difference score equals zero is equivalent to the hypothesis that the two groups have the same mean. The use of the one-sample t test with difference scores is called a *paired t test.*

In a paired t test, each of the n pairs of scores is differenced. The mean of the differences, $\bar{X}_D$, and the standard deviation of the differences, s_D, are computed. Then, the quantity

$$\frac{\bar{X}_D}{s_D/\sqrt{n}}$$

has a t distribution with $n - 1$ degrees of freedom, given the restricted model. Recall that n is the number of pairs and not the number of scores. If the t is statistically significant, the restricted model that the independent variable has no effect on the dependent variable is rejected.

Computational Formulas

Earlier s_p^2 was defined as the pooled or average variance across the two groups. Its formula is

$$\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Given the definition of s_p^2 the above formula can be rewritten as

$$\frac{\sum X_1^2 - (\sum X_1)^2/n_1 + \sum X_2^2 - (\sum X_2)^2/n_2}{n_1 + n_2 - 2}$$

This is the formula generally used to compute s_p^2. So for the example, the formula for s_p^2 is

$$\frac{126 - (24)^2/5 + 393 - (43)^2/5}{5 + 5 - 2} = 4.25$$

The formula for $1/n_1 + 1/n_2$ can be more simply computed by

$$\frac{n_1 + n_2}{n_1 n_2}$$

These computational formulas can be entered into the formula for t resulting in the following formula.

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{n_1 + n_2}{n_1 n_2}\right)\frac{\sum X_1^2 - (\sum X_1)^2/n_1 + \sum X_2^2 - (\sum X_2)^2/n_2}{n_1 + n_2 - 2}}}$$

Ordinarily t is computed to three decimal places.

The computational formula for the paired t test is

$$\frac{\bar{X}_D}{\sqrt{\frac{\sum D^2 - (\sum D)^2/n}{n(n-1)}}}$$

where D is the difference between linked scores and n the number of linked scores.

Effect Size and Power

Even if the restricted model is rejected, it is not known how large the treatment effect is. Statistical significance cannot be equated with scientific significance because statistical significance depends on theoretically unimportant factors such as sample size. For instance, consider two studies that attempt to reduce cigarette smoking. It is possible for the t statistic for one

study to be 8.433, yet the treatment reduces cigarette smoking by two cigarettes. Whereas in a second study the t statistic could be only 2.108, yet the program reduces the level of the smoking by 20 cigarettes. This could happen if the first study has 16,000 subjects, the second only 10 subjects, and the pooled standard deviation is 15.

Effect Size

The most commonly used measure of how much the treatment affects the dependent variable is a measure called *effect size* or *Cohen's d*. The quantity d is defined as

$$\frac{\mu_1 - \mu_2}{\sigma}$$

The numerator is the difference between the population means. The denominator is the standard deviation of the residual variable. The size of d can range from negative to positive infinity, but values larger than two are quite rare. Most values of d vary from zero to one.

Cohen's d is like a Z score in that its denominator is a standard deviation. It measures how different the means of the two groups are relative to the standard deviation within groups. Cohen (1977) describes three different effect sizes. They are

small d = .2
medium d = .5
large d = .8

A small effect is so small that to detect it one needs a statistical analysis. An example of an effect size of this magnitude is the difference in height between 15- and 16-year-old girls (Cohen, 1977). A medium effect is one that is large enough to see without doing statistical analysis. It is reflected by the difference in height between 14- and 18-year-old girls. A large effect is so large that statistics are hardly even necessary. It is reflected by the size of the difference in height between 13- and 18-year-old girls.

To better understand the meaning of the d measure of effect size, imagine that you are considering which of two movies to see one night. Assume that you have access to a survey that was done that measured the extent to which college students enjoyed each of the two movies. If there was sufficient information in the survey you could measure the d for the two movies. The value of d would indicate the degree to which one movie was enjoyed by more college students than the other. If d was small, say .2, that would indicate that if you saw both movies, the probability that you would prefer the one others found to be enjoyable would be .56. If d was .5, the probability that you would prefer the more popular movie would be .64 and if d was .8, the probability would be .71. (The probabilities of .56 for small, .64 for moderate, and .71 for large are determined from the standard normal distribution.)

In research areas where empirical data are lacking, one must make an intelligent guess of the value of d in order to estimate power. If previous studies have been conducted, d can be estimated by

$$\frac{\bar{X}_1 - \bar{X}_2}{s_p}$$

or the mathematically equivalent formula of

$$t\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

When the sample sizes are equal ($n_1 = n_2 = n$), d equals

$$t\sqrt{\frac{2}{n}}$$

So for the example, the effect size equals $-2.914\sqrt{2/5}$, or -1.84. If the paired t test is used, the estimate of d is

$$t\sqrt{\frac{2(1 - r)}{n}}$$

where t is the paired t, n the number of pairs, and r the correlation between the paired scores.

Power

One reason for determining the value of d is that d must be known to ascertain the power of the two-sample t test. In the previous chapter, power is defined as the probability of rejecting the restricted model when it is false. It also equals one minus the probability of making a Type II error. The power of the two-group or two-sample t test depends on three factors: the difference between means, the residual variance, and the sample sizes. The difference between means can be increased by choosing more extreme treatments. Instead of comparing one week of psychotherapy versus none, one year could be compared to none. Although power can be enhanced in this way, generalizibility may suffer because extreme groups may be atypical of everyday treatments.

The residual variance can be reduced by choosing to study persons who are relatively similar. Animal researchers minimize variability by choosing organisms from the same strain. Variability can also be reduced by carefully measuring the dependent variable. A third way to reduce the residual variance is to use a paired design. The residual variance of the paired design is reduced to the degree that there is a correlation between paired observations. A paired design tends to have more power than an unpaired design.

Increasing sample size enhances power in two ways. First, it increases degrees of freedom of the t test, so the difference between means need not be as large to be significant. Second, it reduces the standard error of the mean because $1/n_1 + 1/n_2$ is part of the formula. If the total sample size is fixed, the way to minimize the standard error of the mean difference is by having n_1 equal to n_2.

For a given value of Cohen's d, a given n, and a given alpha, power can be determined. In Table 13.1 is the power for the two-sample t test for small, medium, and large effect sizes. They are given for the .05 level of significance. The n in the table is the sample size in each of the two groups. So, the total sample size of the study is $2n$. The entry in the table is the power multiplied by 100. So if a researcher contemplates doing a study with 20 persons in each group and the effect size is moderate, from Table 13.1 the chances of rejecting the null hypothesis is .33. This means that for every three times that the experiment is done, the null hypothesis is rejected once.

For a given d, alpha, and level of power desired, the n that is needed for that power can be determined. These sample sizes are given in Table 13.2. For instance, if d is .5 with an alpha of .05 and power of .80, a researcher would need 64 subjects in each of the two conditions.

Adjustments need to be made to d if a paired t test is planned and Tables 13.1 or 13.2 are employed. In this case the new d' value is equal to $d/\sqrt{(1-r)}$, where r is the degree of correlation between the paired observations. Also, if the sample sizes are unequal, the n in the tables must be adjusted. The new n, denoted n', equals $2n_1n_2/(n_1 + n_2)$.

Design Considerations

Before the results of a two-group experiment can be interpreted, various design issues must be considered. Two important questions are, first, the rule

TABLE 13.1 **Power Tables for the Two-Sample t Test,[a] with Alpha = .05 and $n = n_1 = n_2$**

	Effect Size (Cohen's d)		
n	.2	.5	.8
10	7	18	39
20	9	33	69
40	14	60	94
80	24	88	99
100	29	94	99
200	51	99	99

[a]Taken from Cohen (1977).
NOTE: Entry in the table is the probability of rejecting the null hypothesis times 100 for a given effect size and sample size.

TABLE 13.2 **Sample Size Required for a Two-Sample *t* Test[a] to Achieve a Given Level of Power for a Given Effect Size and Alpha of .05**

	Effect Size (Cohen's *d*)		
Power	.2	.5	.8
.25	84	14	6
.50	193	32	13
.60	246	40	16
.70	310	50	20
.80	393	64	26
.90	526	85	34
.95	651	105	42
.99	920	148	58

[a]Taken from Cohen (1977).
NOTE: Entry in the table is the sample size for each of the two groups.

by which persons are assigned to groups and, second, the manner in which the two groups are formed.

There are two basic ways in which persons are assigned to groups. They can be assigned randomly or on the basis of some variable.

Random assignment requires that each person has the same probability of being assigned to a given group. Random assignment can be accomplished by coin flip, dice roll, or a random number table. With random assignment, each person has an equal probability of being assigned to a given group. In the absence of treatment effects, the difference between the means is totally explained by sampling error. However, if the means differ by a statistically significant amount, that difference can be attributed to the independent variable. The advantage of a random assignment rule is that it is known that the treatment means differ either due to sampling error or due the independent variable.

A nonrandom rule is one in which persons are assigned to groups on the basis of some variable. For instance, persons are assigned to a surgical procedure on the basis of some clinical test. To analyze the design correctly with a nonrandom assignment rule, that variable must be controlled in analysis. One way in which this can be accomplished is through multiple regression, which is described in advanced statistical texts. Most of the time when assignment is nonrandom, however, it is not known exactly which variable made the groups different and so it is not known which variable to control in the analysis. If the variable that determines assignment to levels of the independent variable cannot be controlled, then when the means differ it is not known whether the treatment made them different or whether the variable that assigned persons to groups made the groups different. A random assignment rule is preferable to a nonrandom rule in order to establish the causal connection between the independent variable and the dependent variable.

It is important to distinguish random *assignment* from random *selection.* Random selection refers to the entry into the study, whereas random assignment refers to the entry into levels of the independent variable. Random selection of persons means that the sample is representative of the population from which it is sampled. Random assignment yields strong causal inference.

The second major design consideration is the formation of the two groups. There is more than one way to study the effects of the independent variable. For instance, consider a study of the effects of jogging: Two groups of persons would be formed, a jogging group and the other a control (that is, no jogging) group. There are many ways to form the two groups:

1. Marathon runners are compared to persons who engage in no physical exercise.
2. Persons who jog ten miles a week are compared to those who swim four times a week.
3. Rats who run mazes for two hours a day are compared to rats who are confined to a cage all day.

The advantage of plan 1 is that the maximum effect of jogging could be estimated, but the disadvantage is that it does not estimate the potential benefit of jogging to most persons. Plan 2 would test the effect of jogging over an alternative exercise plan, but it probably would have very low power. Plan 3 would allow for randomization and exactly measure the effect of exercise, but it would have dubious generality to humans. No one plan is best for all purposes, and each has serious drawbacks. So, when a two-group experiment is undertaken, its interpretation depends on how subjects are assigned to levels of the independent variable and how the two groups are formed.

Illustrations

In this section four different examples are considered. These examples illustrate the computation required for the two-group design.

Example 1

One group consists of ten persons in a smoking cessation program, and the other group contains ten persons who were put on a waiting list. The two groups were formed randomly. The dependent variable is the number of cigarettes smoked per day two weeks after the program is completed. The scores of the treatment group are

0, 15, 12, 9, 10, 0, 0, 25, 5, 3

and the control group

18, 23, 15, 10, 8, 16, 13, 10, 20, 16

The mean for the treatment group is 7.9, and for the control group the mean is 14.9. The pooled variance is

$$\frac{1209 - 79^2/10 + 2423 - 149^2/10}{10 + 10 - 2} = 43.767$$

The standard error of the difference between means is

$$\sqrt{43.767\left[\frac{1}{10} + \frac{1}{10}\right]} = 2.959$$

The test of no effect of the treatment is

$$t(18) = \frac{7.9 - 14.9}{2.959} = -2.366$$

which with 18 degrees of freedom is statistically significant at the .05 level of significance. Thus, the program lowered the level of cigarette smoking to an extent that cannot be explained by sampling error. Because groups were formed randomly, the difference can be attributed to the program and not to any other variable. The value of Cohen's d, using the formula $t\sqrt{2/n}$, is $-2.366\sqrt{2/10} = -1.06$.

Example 2

Five persons undergo a drug treatment to reduce blood pressure and five others receive an inert drug. There are two groups: a drug and a placebo group. Their changes in blood pressure are

Drug: –15, –17, –14, –6, 4

Placebo: 0, –6, 8, 9, –7

The means are –9.6 and .8 for the drug and placebo groups, respectively. The sums of squared scores are 762 and 230 in the drug and treatment groups, respectively. The pooled variance is

$$\frac{762 - (-48)^2/5 + 230 - 4^2/5}{5 + 5 - 2} = 66.000$$

The t test value is

$$t(8) = \frac{-9.6 - .8}{\sqrt{66.000\left[\frac{1}{5} + \frac{1}{5}\right]}} = -2.024$$

The t value of –2.024 is not significant at the .05 level. It is, however, significant at the .10 level and some researchers refer to this level of significance as *marginal significance*. There is, then, not very compelling evidence from this study that the drug reduces blood pressure. The effect size equals $-2.024\sqrt{2/5}$, which is –1.28. Even though the effect size is –1.28, the sample size makes the power so low that the result is not statistically significant.

Example 3

Of 28 people involved in a study on attitude change, 13 received a message from a high-status source and 15 from a low-status source. The resulting attitude changes for the two groups are

High-status source: 5, 6, 9, 3, 0, 4, 10, 6, 9, 5, 6, 5, 7

Low-status source: –1, 0, 3, –4, –6, –2, –1, 0, 3, 6, –3, –2, –1, –2, 1

A positive change indicates change in the direction consistent with the message, whereas a negative change indicates the reverse. The means for the high- and low-status groups are 5.77 and –.60. The sums of the squared scores for the two groups are 519 and 131. The pooled variance is

$$\frac{519 - 75^2/13 + 131 - (-9)^2/15}{13 + 15 - 2} = 8.150$$

The t test value is

$$t(26) = \frac{5.77 - (-.60)}{\sqrt{8.150\left[\frac{1}{13} + \frac{1}{15}\right]}} = 5.888$$

This result is statistically significant at the .001 level. There was more attitude change that was consistent with the message in the high-status than in the low-status group. The value of Cohen's d is

$$5.888\sqrt{\frac{1}{13} + \frac{1}{15}} = 2.23$$

Example 4

Each of ten six-year-old children interact with a different four-year-old child. Measured is the degree of social responsiveness by each person in the conversation. The hypothesis is that six-year-olds are more socially responsive than four-year-olds. The data are paired because two persons of different ages interact. The scores are

Pair	*Six-Year-Old*	*Four-Year-Old*
1	6	5
2	5	4
3	4	5
4	7	6
5	6	3
6	7	5
7	3	4
8	6	3
9	8	6
10	5	4

The correlation between scores is .45. The differences between each four-year-old and each six-year-old are 1, 1, –1, 1, 3, 2, –1, 3, 2, and 1, and the mean is 1.2. The variance of the different scores is

$$\frac{32 - 12^2/10}{9} = 1.956$$

The t test value is

$$t(9) = \frac{1.2}{\sqrt{1.956/10}} = 2.713$$

This value of t is statistically significant at the .05 level. Thus six-year-olds are more socially responsive than four-year-olds. The value of d is

$$2.713\sqrt{\frac{2(1 - .45)}{10}} = .90$$

(The correlation between the two scores is equal to .45.)

Summary

The complete model for the two-group design involves a dichotomous independent variable that causes the dependent variable. In the restricted model the independent variable has no effect on the dependent variable. This model is evaluated by computing the difference between the means divided by the standard error of the difference between means. This standard error equals the pooled standard deviation of the two groups times the square root of $1/n_1 + 1/n_2$. When the restricted model is true, the difference between means divided by its standard error has a t distribution, with $n_1 + n_2 - 2$ degrees of freedom. The test of the restricted model presumes that the residual variable has a normal distribution, that the variances in the two groups are equal, and that the observations are independent.

When observations are paired, differences are computed and the mean of the differences evaluates the equality of the group means. The size of the

treatment effect is measured by *Cohen's d*, which is called a measure of *effect size*. With the sample size, alpha, and the effect size, the power of the t test can be determined. The interpretation of a significant t test depends upon design considerations. If the units are randomly assigned to levels of the independent variable, then significant differences on the dependent variable can be attributed to the independent variable.

In the next chapter the independent variable may take on more than two levels.

Problems

1. Determine the minimum value of t needed to achieve the given significance levels with the corresponding degrees of freedom.

	Alpha	*df*
a.	.05	26
b.	.01	6
c.	.10	44
d.	.02	62
e.	.001	132
f	.05	77

2. The following scores are taken from a study that compared two different methods of increasing vocabulary. The scores of ten persons, five under each method, are:

   A: 16, 19, 20, 18, 24

   B: 12, 15, 16, 15, 14

   Is there any evidence that one method is superior to the other?

3. Compute a paired t test to evaluate the effectiveness of a weight loss program.

Person	*Before*	*After*
1	163	150
2	149	143
3	236	240
4	189	180
5	176	160
6	216	205

4. For the following t values compute d.

   a. $t(20) = 1.380$, $n_1 = 11$, $n_2 = 11$
   b. $t(98) = 2.110$, $n_1 = 50$, $n_2 = 50$

c. $t(10) = 1.530$, $n_1 = 8$, $n_2 = 4$
d. $t(54) = -.470$, $n_1 = 30$, $n_2 = 26$

5. Determine the power of the following tests.

a. $n_1 = n_2 = 20$ and $d = .5$
b. $n_1 = n_2 = 100$ and $d = .2$
c. $n_1 = n_2 = 80$ and $d = .8$
d. $n_1 = n_2 = 10$ and $d = .8$
e. paired design; $d = .22$, $r = .8$, and $n = 20$
f. $n_1 = 11$, $n_2 = 100$, and $d = .8$

6. A program is developed to improve the intelligence scores (IQ) of preschool children. Two groups of children are randomly formed. Test whether the program affects IQ:

Treated group: 109, 123, 141, 119, 133, 117, 118, 120

Control group: 106, 103, 114, 120, 116, 107, 98

7. Twenty persons are randomly assigned to one of two treatments. In the treatment group, ten persons are taught a series of strategies to improve their memory. The control group learned none of the strategies. The scores on a memory test are

Memory group: 88, 76, 83, 75, 64, 80, 76, 73, 84, 78

Control group: 84, 73, 84, 78, 68, 78, 71, 70, 80, 79

Are the two groups significantly different?

8. Describe the advantages and disadvantages of using the control groups in a study to evaluate the effect of group therapy to reduce cigarette smoking.

a. individual therapy
b. hypnosis condition
c. a film that encourages quitting

9. A psychologist studies the degree of happiness of people at various stages in life. His measure of general happiness varies from 0 to 60. In one study he compared the happiness of married and single men aged 25. Is there a significant difference between the two groups?

Married	*Single*
58	57
45	44
50	59
54	44
49	39
39	60
50	44
51	

10. Nine persons were asked to rate the taste of cola A and cola B on a scale from one to ten. Is one drink significantly preferred to the other?

Person	*Cola A*	*Cola B*
1	7	7
2	8	9
3	8	7
4	9	5
5	10	9
6	9	7
7	8	6
8	8	10
9	7	8

11. For the following studies estimate Cohen's d.

    a. $t = 2.910$, $n_1 = 10$, $n_2 = 12$
    b. $t = -.410$, $n_1 = 5$, $n_2 = 5$
    c. a paired design in which $t = 5.910$, there are 8 pairs, and $r = .8$
    d. $t = -.970$, $n_1 = n_2 = 80$

12. A researcher wishes to test whether eight-grade girls outscore eighth-grade boys in vocabulary. She tested 30 boys and 42 girls and found means of 64.53 for boys and 66.42 for girls. The standard deviations are 12.34 for boys and 12.59 for girls. Compute Cohen's d for this study and interpret it. Evaluate whether the sex difference is statistically significant.

13. The data for Example 1 in the chapter are repeated here: The scores of the treatment group are

    0, 15, 12, 9, 10, 0, 0, 25, 5, 3

    and for the control group are

    18, 23, 15, 10, 8, 16, 13, 10, 20, 16

    Compute the standard deviations for group and evaluate in words the assumption of equal variances and its effect on the p value.

14. Diehl, Kluender, and Parker (1985) tested for the recognition of auditory stimuli on two tasks. Each of 13 subjects received a score on each task, the maximum being 40. Does performance on the tasks significantly vary?

	Task			*Task*	
Subject	*A*	*B*	*Subject*	*A*	*B*
DS	20	19	LD	30	24
MM	21	18	RL	23	18
JH	28	24	TW	29	19
JS	17	6	TA	30	16
MC	15	13	VS	34	29
CM	20	13	CJ	21	20
LG	28	22			

15. For the following effect sizes and designated power, state the necessary sample size needed in each group.

	Effect Size	*Power*
a.	.5	.50
b.	.8	.80
c.	.5	.95
d.	.2	.25