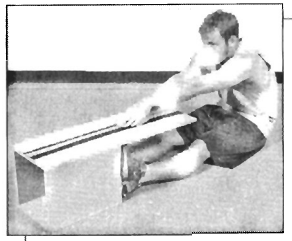

The gymnastics coach at Mountain View Community College wanted to know if female collegiate gymnasts at Mountain View had greater hip and low back flexibility than male gymnasts. To answer this question, she measured the flexibility of athletes on the men's and women's teams using a sit-and-reach box. At the end of the season, the average distance reached beyond the toes for men was 12 centimeters, while women averaged 15 centimeters. At this college, the women clearly were more flexible than the men. No further statistical analysis was needed.

Next, the coach wanted to compare all collegiate gymnasts in the entire state. This required her to measure all the athletes in the state, which was not feasible. Instead, she took random samples of male and female gymnasts from several colleges around the state. She then performed a t test to determine whether the means in the two samples accurately reflected the means of the populations from which they came. Predicting population parameters from sample statistics is called *inferential statistics*. In this chapter we learn the techniques for performing this type of analysis.

Recall from chapter 6 that a sample mean may be used as an estimate of a population mean. Remember also that we can determine the odds, or probability, that the population mean lies within certain numerical limits by using the sample mean as a predictor. This same technique can be used in reverse to determine if a given sample is likely to have been randomly selected from a specific population.

If the population mean is known or assumed to be a certain value, and if the sample mean is not close enough to the population mean to fall within the limits set by a selected level of confidence, then one of the following conclusions must be true: (a) The sample was not randomly drawn from the population, or (b) the sample was drawn from the population, but it has been modified so that it is no longer representative of the population from which it was originally drawn.

Using similar logic, we can make conclusions about two sets of data. If we draw two samples from the same population, and the means of these samples differ by amounts larger than would be expected based on normal distributions, one of the following conclusions must be true: (a) One or both of the samples were not randomly drawn from the population, or (b) some factor has affected one or both samples, causing them to deviate from the population from which they were originally drawn.

t Tests

When a sample is drawn from a population with a known or estimated mean (μ) and standard deviation (σ), the probability (or odds) that the mean of a randomly drawn sample ($\bar{X}$) will lie within certain limits of μ can be determined. To ascertain the probability that a given sample came from a certain population, the value of the standard error of the mean must be calculated by one of the following formulas. If the standard deviation of the population (σ) is known, the formula used is

$$\sigma_M = \frac{\sigma}{\sqrt{N}}, \tag{8.01}$$

where σ_M is the symbol for the actual standard error of the mean for a population with known mean and standard deviation. If σ is not known, the formula used is

$$SE_M = \frac{SD}{\sqrt{N}}, \tag{8.02}$$

where SE_M is the standard error of the mean estimated from the sample. Note that this is the same as equation 6.03.

Using either σ_M or SE_M, we can determine the odds that a sample is representative of the population from which it was drawn by doing a Z test, if the population is known,

$$Z = \frac{\bar{X} - \mu}{\sigma_M}, \tag{8.03}$$

or a t test,

$$t = \frac{\bar{X} - \mu}{SE_M}, \tag{8.04}$$

if the population is estimated.

Evaluating Z

When σ_M is known, Z indicates the significance of the difference between $\bar{X}$ and μ. Under these conditions, we can determine the significance of Z by comparing it to critical ratios of 1.65 at $p = .10$, 1.96 at $p = .05$, and 2.58 at $p = .01$. If Z exceeds one of these levels, we may conclude at the given level of confidence (LOC) that the sample was not randomly drawn from the population, or that it has been modified in some way so that it no longer represents the population from which it was drawn.

Evaluating t From a Single Sample

The approximation of σ_M by SE_M is not accurate in small samples (N less than 60). This was first demonstrated by an English statistician named William Sealy Gossett (1876–1937), who wrote under the pseudonym Student (Kotz & Johnson, 1982). He developed a series of approximations of the normal curve to account for the bias in the estimate of σ_M called *Student's t distribution.* Table A.3 in appendix A lists the values for Student's t distribution. If there were no error in the estimation,

table A.1 could be used in every case. But when samples are used to estimate population parameters, especially when the samples are small, the *t* distribution (table A.3) must be used to evaluate the *t* statistic. The values in table A.3 at the given *p* levels are called **critical ratios.** They represent the *t* ratio that must be reached to reject chance.

The *t* test for one sample produces the ratio of the **actual mean difference** between the sample and the population to the **expected mean difference,** (that amount of difference between $\bar{X}$ and μ that can be expected to occur by chance alone). The expected mean difference is estimated by equation 8.02 and is called the standard error of the mean. To interpret *t* for a single sample, we must first find the degrees of freedom, which can be calculated by the formula $df = N - 1$. The *t* ratio is compared to the values for a two-tailed test (one- and two-tailed tests will be explained later in this chapter) from the *t* distribution in table A.3 for the appropriate *df*.

When *t* exceeds the value in table A.3 for a given *p* level, we may conclude that $\bar{X}$ was not drawn from μ. When *t* is less than the critical ratio in table A.3, the null hypothesis (H_0) is accepted; there is no reliable difference between $\bar{X}$ and μ. When *t* exceeds the critical ratio, H_0 is rejected and H_1 is accepted; some factor other than chance is operating on the sample mean. Notice that in table A.3 when degrees of freedom are large ($df > 120$), the values for a two-tailed *t* test at a given *p* value are the same as the values read from table A.1 for a *Z* test (1.65, 1.96, and 2.58).

This technique is useful for determining if influences introduced by an experiment have an effect on the subjects. If we know or estimate the population parameters and then draw a random sample and treat it in a manner that is expected to alter its mean value, we can determine the odds that the treatment had an effect by using a *t* test. If *t* exceeds the critical ratios in table A.3, we can conclude that the treatment was effective because the odds are high that the sample is no longer representative of the population from which it was drawn. The treatment has caused the sample to change so that it does not match the characteristics of the parent population.

Assumptions for the *t* Test

Several assumptions must be met for the *t* test to be properly applied. If these assumptions are not met, the results may not be valid. When the investigator knows that one or more of these criteria are not met, a more conservative (i.e., $p = .01$ rather than $p = .05$) level should be selected to avoid errors. This allows us to be confident of the conclusions and helps to compensate for the fact that all assumptions were not met. The *t* test is quite **robust;** it produces reasonably reliable results, even if the assumptions are not met totally. The *t* test is based on the following assumptions:

- The population from which the samples are drawn is normally distributed. (See chapter 6 for methods of determining the amount of skewness and kurtosis in a data set.)

- The sample or samples are randomly selected from the population. If the samples are not randomly selected, a generalization from the sample to the population cannot be made.
- When two samples are drawn, the samples have approximately equal variance. The variance of one group should not be more than twice as large as the variance of the other. This is called **homogeneity of variance.**
- The data must be parametric, that is, based on an interval or ratio measurement scale (see chapter 1).

An Example From Physical Education

The faculty of a physical education department became concerned about the apparent lack of skill students developed in their 5-week volleyball units. Students are assigned to instructors in a random fashion alphabetically, based on last name. Typically the classes just play during the entire period and receive little or no practice on specific skills. A standardized test of volleyball serving ability (50 points, maximum) has been given to every student for many years and the data have been saved. The mean for more than 1,000 students (the population) on this test is 31 points (μ) with a standard deviation of 7.5 (σ).

One teacher decided to try a different approach. In one class (the sample), half the period was devoted to teaching skills, especially serving skills, and the students practiced under the teacher's direction for 20 minutes. Games were played only at the end of the period, and a tournament among the squads was held the last week of the volleyball block.

The students in this class ($N = 30$) were also given the standardized serving test. Their average score ($\bar{X}$) was 35 points out of the 50 possible with a standard deviation (*SD*) of 8.3. Was the teacher effective in improving serving skills? In other words, if we assume the class of 30 students to be a random sample of the population of more than 1,000 students, is it likely that the average score for the class ($\bar{X} = 35$) is representative of the population mean ($\mu = 31$)? To test the hypothesis that the sample class represents the population, we calculate the standard error of mean for the population:

$$\sigma_M = \frac{\sigma}{\sqrt{N}} = \frac{7.5}{\sqrt{30}} = 1.37.$$

Then we conduct a *Z* test (because μ and σ are known) to determine the odds that the mean of a sample randomly drawn from the population would differ from the population mean by as much as 4 points:

$$Z = \frac{35 - 31}{1.37} = 2.92.$$

What are the odds that a *Z* score of 2.92 would be found if the sample were drawn from the population and not treated? Because *Z* is greater than 2.58, the probability that $\bar{X}$ did not come from μ is greater than 99 to 1, $p < .01$.

Two possibilities must be considered:

1. The class was not a random sample of the population and therefore does not represent the population. Perhaps by luck, or by design, these 30 students were better at the beginning of the 5-week block than the typical students assigned to the other classes.
2. The sample was random at the beginning of the 5-week block, but the treatment (instruction and practice) has changed the students in such a way that they are no longer representative of the population. In other words, the students have changed so that they now represent another population, one that has instruction and practice rather than free play.

If random assignment to instructors can be demonstrated, so that there is assurance that the class was representative at the start of the 5-week block, then only one conclusion is left. A difference of 4 points between the mean of the sample and the mean of the population would occur less than 1 time in 100 by chance alone. In other words, the odds that the instruction was effective are better than 99 to 1 (LOC = 99%). We reject H_0 and conclude that the instruction was effective at $p < .01$.

In the previous example, we used σ because we knew its value. But in most research, μ and σ are not known. If σ is not known, we estimate SE_M, the standard error of the mean, using equation 8.02 and the standard deviation of the sample. Then we use the t test rather than Z to determine the significance of the difference between the sample and the *estimated population mean* (using equation 8.02). First we calculate SE_M:

$$SE_M = \frac{8.3}{\sqrt{30}} = 1.52.$$

Then we use SE_M to determine t. If $\bar{X} = 35$, then

$$t = \frac{35 - 31}{1.52} = 2.63.$$

Notice that the answer for t is slightly smaller than the Z score (2.92). This demonstrates that the power to detect differences between samples and populations is greater when the population is known than when it is estimated. To find the level of confidence, we compare the t ratio (2.63) to the critical ratios of a two-tailed test from the t distribution in table A.3 for $df = 30 - 1 = 29$. The critical ratio at $df = 29$ in table A.3 for $p = .05$ is 2.045 and for $p = .01$ it is 2.756. Our t value falls between these values, so we accept the less significant value, $p < .05$. Note that we have a lower level of confidence (95%) for rejecting H_0 when the population is estimated than when the population parameters are known (99%). While Table A.3 is limited to three levels (.10, .05, and .01), computers will generate exact p values. Always use a computer to analyze real data.

Comparing Two Independent Samples (A Between Comparison)

The same concepts used to compare one sample mean to the population mean may be applied to compare two samples drawn from the same population. We will review the theory before considering the calculations. If the ratio exceeds the critical ratio from table A.3, the null hypothesis H_0 is rejected. H_0 states that both means represent (or were randomly drawn from) the same population. If the t ratio does not exceed the critical ratio, H_0 is accepted.

In this design, two independent samples are randomly drawn from a population. By **independent samples,** we mean that the subjects in one sample are different people and are not related, or correlated, in any way to the subjects in the other sample. One sample, the experimental group, is treated, and the other, the control group, is not. A t test is performed to compare the actual difference between the means of the two samples (the numerator of the t test) with the difference expected if only chance were operating (the denominator of the t test). This ratio is used to determine whether the two groups both represent the population from which they were drawn. If the calculated t ratio is larger than the critical ratio from the t distribution in table A.3, chance is rejected, and a probable cause is assumed. When H_0 is rejected for an experiment conducted under carefully controlled conditions, the treatment is identified as the probable cause. This design is represented in table 8.1 where two different groups of subjects each are compared before and after treatment.

Table 8.1 Research Design for Two Group Between Comparison

Group	Number	Pretest	Treatment	Posttest
Control	N_1	Yes	No	Yes
Experimental	N_2	Yes	Yes	Yes

If both groups are randomly drawn from the same population, a t test between the groups' means on the pretest should not be significant. If the treatment is effective, and all other variables are controlled, a t test between the groups on the posttest should be significant. This indicates that the experimental group is no longer representative of the original population from which it was drawn because the treatment has caused it to change.

In chapter 6 it was demonstrated that the means of a large number of random samples drawn from the same population will form a normal curve. The central limit theorem indicates this is true even if the population is skewed. **The**

differences between pairs of random samples drawn from a population will also form a normal curve. If enough pairs of sample means are randomly drawn, and the difference between each pair is consistently calculated by subtracting the second mean from the first, then the average difference of all the pairs will be zero and the distribution will be normal. In some cases, the first mean will be larger than the second, making a positive difference; in other cases, the second will be larger than the first, producing a negative difference. When added together, the positives should cancel out the negatives, and all the difference scores should sum to zero.

If we assume an infinite set of scores, each derived from the mean difference of a pair of randomly drawn samples, these differences will form a normal curve with a mean of zero. After calculating the standard deviation of the difference scores (called the *standard error of the difference,* SE_D), we could then compute t scores and look in table A.3 to determine significance. Note that in this case, SE_D becomes the denominator of the t test. It is the value that indicates the amount of difference between two randomly drawn sample means that can be expected (from the normal curve ± 1 *SD*) by chance alone.

When t is larger than a given critical ratio in table A.3, we may conclude that (a) one or both of the samples were not randomly drawn, or (b) some factor has intervened to cause one or both of the groups to alter their mean value.

Now let us review the process for conducting experiments with two sample groups. Two groups of subjects are randomly selected from a population (or they may be selected by categories; e.g., age, gender). One group, the control group, is controlled very closely in all respects; the other group, the experimental group, is controlled in all respects except one. That one factor, the independent variable, is allowed to influence only the experimental group.

After these conditions are met, measurements of the dependent variable are taken on both groups. If the t ratio between the actual mean difference ($\overline{X}_1 - \overline{X}_2$) and the expected mean difference (the standard error of the difference) is larger than the critical ratio for t (in table A.3) at a given p level, we conclude that the independent variable has had a significant effect, and we reject chance as a cause for the mean difference. The two groups are no longer representative of the same population from which they were drawn, because the experimental group has been changed by the independent variable.

Standard Error of the Difference

Equation 8.02, presented earlier, can be used to calculate the standard error of the mean (SE_M), or the amount that a single randomly drawn sample mean can be expected to deviate from a population mean by chance alone. A similar formula permits us to calculate the **standard error of the difference** (SE_D), the amount of difference between two randomly drawn sample means that may be attributed to chance alone.

When SE_M has been calculated for each of two sample means randomly drawn from the same population, we can use equation 8.05 to estimate the amount of the difference between the two means attributable to chance:

$$SE_D = \sqrt{(SE_{M1})^2 + (SE_{M2})^2}. \tag{8.05}$$

This formula (based on only two samples) estimates the size of the standard deviation of an infinitely large group of difference scores, each of which has been derived from randomly drawn pairs of sample means. Because it is a standard deviation, it may be interpreted in the same manner as any Z score on a normal curve.

Equation 8.06 presents the t test for independent samples:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{SE_D}. \tag{8.06}$$

This t test is a ratio of the actual difference between two means (the numerator of t) and the difference that would be expected due to chance for an infinite set of sample means (the denominator of t). If the t ratio exceeds the critical ratio at a given p level, then chance, or H_0, is rejected as a probable cause for the difference between the sample means, and it is concluded that another factor or factors caused the difference (H_1 is accepted). We may then say that the mean difference is significant.

The statistics in the t test do not identify the causative factor. Careful controls and proper experimental design are needed to identify the factor. If the control group is closely monitored to assure that only chance operates on it, and if only one independent variable is permitted to influence the experimental group, then the independent variable may be identified as the causative factor.

To use table A.3 in appendix A for a two-group, independent t test, we must first calculate the degrees of freedom, *df*:

$$df = (N_1 - 1) + (N_2 - 1). \tag{8.07}$$

One degree of freedom is lost for each group. Then we compare the calculated t ratio with the critical ratios in table A.3 for the appropriate degrees of freedom. If the calculated t ratio exceeds the critical ratio for t, the mean difference is declared to be significant at the indicated p level, and H_0 is rejected.

An Example From Pedagogy

A pedagogy researcher wanted to know if a 10-minute verbal lesson in basketball shooting techniques would have any effect on the shooting ability of high school physical education students. The population for this study was all students at a given high school. The investigator randomly selected 100 students from a list of all students enrolled in the school. The subjects were then contacted

by phone, and appointments were made with each one to meet in the gym dressed for activity.

The odd-numbered subjects (the control group) listened to a 10-minute verbal lesson on zone and one-on-one defense, warmed up for 5 minutes, and then made 20 free throws at the basket. The even-numbered subjects (experimental group) listened to a 10-minute verbal lesson on basketball shooting techniques, warmed up for 5 minutes, and then made 20 free throws at the basket.

The lesson on defenses for the control group equalized the time spent with the experimental group but was irrelevant to the shooting task. In both cases, the lesson was verbal, with no audiovisual equipment, and identical for each subject in the group. The following data on the number of successful shots out of 20 attempts were collected:

Control Group Defense lesson	Experimental Group Shooting technique lesson
$N_1 = 50$	$N_2 = 50$
$\bar{X}_1 = 9.5$	$\bar{X}_2 = 10.3$
$SD_1 = 2.7$	$SD_2 = 3.2$

Did the shooting lesson have an effect? The data presented previously show that there is a difference in favor of the experimental group, but it could be the result of random selection. If the experiment were to be repeated, but with the defense lesson given to both groups, a difference this large may still be obtained. A *t* test is needed to determine if the observed difference is random or significant.

To perform the *t* test, we need to calculate the standard error of the mean for each group and the standard error of the difference between the groups.

The standard error of the mean for the control group is

$$SE_{M1} = \frac{2.7}{\sqrt{50}} = .38.$$

And the standard error of the mean for the experimental group is

$$SE_{M2} = \frac{3.2}{\sqrt{50}} = .45.$$

Using equation 8.05, we calculate the standard error of the difference to be

$$SE_D = \sqrt{.38^2 + .45^2} = .59.$$

Remember, the standard error of the difference is the standard deviation of an infinite set of mean differences that form a normal curve. So its value ($SE_D = .59$) represents the value of the mean difference that would be expected 68% of the time between two randomly drawn samples influenced only by chance. This value

is compared with the actual difference observed between the control and experimental means by a *t* test:

$$t = \frac{9.5 - 10.3}{.59} = -1.36.$$

The *t* value is negative because a larger value was subtracted from a smaller value in the numerator. The sign of the *t* ratio is not important in the interpretation of *t*, because it may be positive or negative depending on which group is listed first in the numerator. Only the absolute value of *t* is considered in determining significance.

The obtained *t* ratio (1.36) is compared to the critical ratios for a two-tailed test (one-tailed and two-tailed tests are explained later in this chapter) in table A.3 for $df = (50 - 1) + (50 - 1) = 98$. Table A.3 does not list a *df* value of 98, so we use the next lowest value, 60. For $df = 60$, the critical ratio at the most liberal *p* level, $p = .10$, is 1.671. Our obtained value of 1.36 does not reach this level. Because *t* is not significant at $p = .10$, we do not need to compare it at higher levels. We conclude that the differences could have happened by chance alone, so we accept the null hypothesis, H_0. The verbal shooting lesson had no significant effect.

An Example From Leisure Studies/Recreation

The *t* test can be applied to problems in many disciplines. A recreation director wanted to know if daily positive verbal reinforcement affected a person's ability to bowl. So the director conducted a 10-week experiment with two recreational bowling classes. In one class, the director praised the bowlers on their daily game scores, recorded the scores in a special book, and posted the scores on the wall.

In a second class, the director was neutral in relationships with the bowlers and made no comments, either positive or negative, on performance. The two groups were assumed to be equal in bowling ability at the start. To find out whether the praise had an effect on the average bowling scores at the end of the 10-week period, a *t* test was conducted on the following data:

Class 1 (praise)	Class 2 (neutral)
$N = 30$	$N = 30$
$\bar{X}_1 = 145$	$\bar{X}_2 = 135$
$SD_1 = 18.1$	$SD_2 = 14.9$

The standard error of the mean for the experimental group is

$$SE_{M1} = \frac{18.1}{\sqrt{30}} = 3.3.$$

For the control group,

$$SE_{M2} = \frac{14.9}{\sqrt{30}} = 2.7.$$

The standard error of the difference is

$$SE_D = \sqrt{(3.3)^2 + (2.7)^2} = 4.3.$$

This results in a t value of 2.33:

$$t = \frac{145 - 135}{4.3} = 2.33$$

$$df = (30 - 1) + (30 - 1) = 58$$

Table A.3 indicates that the t ratio of 2.33 reaches the $p < .05$ level for $df = 40$ (no value is listed for $df = 58$). In other words, the odds of finding a mean difference as large as 10 points by chance alone are less than 5 in 100. The director concluded that praise did make a difference with a level of confidence better than 95% (H_0 was rejected and H_1 was accepted). The results of a t test are often reported in tabular form as shown in table 8.2.

Table 8.2 Effects of Praise on Bowling Scores

Group	Mean	*SD*	SE_M	SE_D	*t*	*p*
Praise	145	18.1	3.3	4.3	2.33	< .05
Neutral	135	14.9	2.7			

The t Test With Unequal Values of N

The examples presented thus far have involved two groups with equal values of N. In practical research, this is almost never the case. Subjects often drop out of experiments (this phenomenon is interestingly referred to as subject mortality), and the two groups usually do not have equal numbers of subjects. The formula for standard error of the difference must be modified to account for the differences in N.

The standard error of the difference formula (equation 8.05) sums the two standard errors of the mean values based on the assumption that both values contribute equally to the standard error of the difference. This is true if $N_1 = N_2$. But if N_1 is twice as large as N_2, N_1 should contribute two-thirds of the total value of standard error of the difference. Yet equation 8.05 permits it to contribute only half. For this reason, the following alternate formula for standard error of the difference is used when values of N are unequal:

$$SE_D \sqrt{\left[\frac{(N_1 - 1)(SD_1)^2 + (N_2 - 1)(SD_2)^2}{N_1 + N_2 - 2}\right]\left[\frac{1}{N_1} + \frac{1}{N_2}\right]}. \quad (8.08)$$

This formula does not require the calculation of standard error of the mean for each group. Standard error of the difference is calculated directly from the standard deviations and N_1 and N_2. This saves one step in the process. Equation 8.08 produces the same answer as equation 8.05 when $N_1 = N_2$, so it could be used in every case.

When the values of N are large and only slightly unequal, the error introduced by using the simpler equation 8.05 to compute SE_D is probably not critical. But when the values of N_1 and N_2 are small, and approach a ratio of 2:1, the error introduced by equation 8.05 is considerable. If there is any doubt about which equation is appropriate, equation 8.08 should be used so that maximum confidence can be placed in the result.

An Example From Biomechanics

The following example applies equation 8.08 to the mean values obtained in a laboratory test comparing hip and low-back flexion of randomly selected males and females. The following measurements in centimeters were obtained using the sit-and-reach test:

Males	Females
$\bar{X}_1 = 22.5$	$\bar{X}_2 = 25.6$
$SD_1 = 2.5$	$SD_2 = 3.0$
$N_1 = 10$	$N_2 = 8$

Using equation 8.08, we calculate that $SE_D = 1.29$:

$$SE_D \sqrt{\left[\frac{(10 - 1)(2.5)^2 + (8 - 1)(3.0)^2}{10 + 8 - 2}\right]\left[\frac{1}{10} + \frac{1}{8}\right]} = 1.29.$$

Once SE_D is known, t can be calculated:

$$t = \frac{22.5 - 25.6}{1.29} = -2.40.$$

The degrees of freedom are determined using equation 8.07. In this example $df = (10 - 1) + (8 - 1) = 16$. Table A.3 indicates that for $df = 16$, a t ratio of 2.40 is

significant at $p < .05$. So H_0 is rejected, and H_1 is accepted. The researcher concludes with better than 95% confidence that females are more flexible than males in the hip and low-back joints as measured by the sit-and-reach test.

Repeated Measures Design (A Within Comparison)

The standard formulas for calculating t assume no correlation between the groups. Both groups must be randomly selected from the population and independent of each other. But when a researcher tests a group of subjects twice, such as in a pre-post comparison, the groups are no longer independent. **Dependent samples** assume that there is a relationship, or correlation, between the scores and that a person's score on the posttest is partially dependent on his or her pretest score.

This is always the case when the same subjects are measured twice. A group of subjects is given a pretest, subsequently treated in some way, and then given a posttest. The difference between the pretest and posttest means is computed to determine the effects of the treatment. This arrangement is often referred to as a repeated measures design or within comparison.

Because both sets of scores are made up of the same subjects, there is a relationship, or correlation, between the scores of each subject on the pre- and posttests. The differences between the pre- and posttest scores are usually smaller than they would be if we were testing two different groups of people. Two test scores of a single person are more likely to be similar than are the scores of two different people.

If there is a positive correlation between the two groups, a high pretest score is associated with a high posttest score. The same is true of low scores. Consequently, the difference between the two means will tend to be smaller with single group pre-post comparisons than with independent two-group comparisons. This may result in a false conclusion that there is no significant difference between the pretest and posttest means.

This same argument holds true for studies using matched pairs, pairs of subjects who are intentionally chosen because they have similar characteristics on the variable of interest. These matched pairs—sometimes called research twins—are then divided between two groups so that the means of the groups on the pretest are essentially equal. One group is treated, and the other group acts as control; then the posttest means are compared.

We expect the matched group means to have smaller differences than if the two groups were not matched on the pretest. In effect, we have forced the groups to be equal on the pretest so that posttest comparisons may be made with more clarity. The matched twins in each group may be considered to be the same person, and the correlation between them on the dependent variable can be calculated.

To accomplish this matching process, all the subjects are given a pretest and then ranked according to score. Using a technique sometimes referred to as the ABBA assignment procedure, the researcher places the first (highest scoring) subject in group A, the second and third subjects into group B, the fourth and fifth into group A, the sixth and seventh into group B, and so forth until all subjects have been assigned. The alternation of subjects into groups ensures that for each pair of subjects (1 and 2, 3 and 4, 5 and 6, etc.) one group does not always get the higher score of the pair.

This technique usually results in a correlation between the groups on the dependent variable and in smaller mean differences on the posttest. But because the two groups start with almost equal means on the pretest, it is easier to identify the independent variable as the cause of posttest differences.

If correlated samples are used (either the same subjects or matched pairs) and if no correction is made, the researcher may falsely conclude that there is no difference between the means, when in fact a real, or significant, difference does exist but is smaller than expected.

Correction for Correlated Samples

The correction for correlated samples is made in the formula for standard error of the difference (SE_D) because this value indicates the difference to be expected by chance alone. By adjusting the SE_D formula, we can regulate t to more correctly reflect any real difference that may exist in dependent samples.

The correction is made by factoring out, or subtracting, the effects of the correlation between the two samples in the formula for standard error of the difference:

$$SE_D = \sqrt{(SE_{M1})^2 + (SE_{M2})^2 - 2\, r\, (SE_{M1})\, (SE_{M2})}. \qquad (8.09)$$

A positive value for r reduces the SE_D, increases t, and provides a greater chance of finding significance. When r is negative (which is very unlikely with matched pairs or repeated measures), t becomes smaller.

When r is zero, the term $2r\,(SE_{M1})\,(SE_{M2})$ becomes zero, and the formula reverts to its original form. Actually, equation 8.09 is the more generalized form of equation 8.05; however, $2r(SE_{M1})(SE_{M2})$ is usually not included as a component when the groups are independent, because r is assumed to be zero.

Note that all subjects must have data points on both variables (usually a pre-post companion). If any subject is missing data on either variable, they must be eliminated from the study.

A researcher compared 30 high school students on vertical jump in inches before and after 3 weeks of leg strength development. The following example shows how an incorrect conclusion that no differences exist could be made on correlated samples of $N = 30$, if $r = .60$, and the correction in SE_D is not made. The degrees of freedom for a dependent t test are $N_{pairs} - 1$.

Suppose the following values resulted from the study:

Vertical jump [pretest]	Vertical jump [posttest]
$\bar{X}_1 = 15$	$\bar{X}_2 = 17.5$
$SE_{M1} = .9$	$SE_{M2} = 1.5$

If we do not correct for correlated samples, $SE_D = 1.75$:

$$SE_D = \sqrt{(.9)^2 + (1.5)^2} = 1.75.$$

And $t = -1.43$ (which is not significant at 29 *df*):

$$t = \frac{15 - 17.5}{1.75} = -1.43, \; N.S.$$

When we apply the correction,

$$SE_D = \sqrt{(.9)^2 + (1.5)^2 - 2\,(.6)\,(.9)\,(1.5)} = 1.20,$$

and

$$t = \frac{15 - 17.5}{1.20} = -2.08, \; p < .05.$$

At $df = 29$, the uncorrected t (–1.43) is not significant, but the corrected t (–2.08) is significant at $p < .05$. Because the subjects in both tests are the same people, a serious error would be made without the correction factor in the formula for SE_D. The problem of correcting for unequal N never arises with matched pairs or repeated measure designs, because the pairs are matched and the Ns are always equal.

Another method to compute t for correlated samples, which does not require the calculation of r, is called the direct difference method. This method is sometimes preferred because it is easier to calculate by hand. The formula for the direct difference method is

$$t = \frac{\Sigma D}{\sqrt{[N\Sigma D^2 - (\Sigma D)^2]/(N-1)}}, \tag{8.10}$$

where D = the difference between the pre- and posttest scores for each subject and N = the number of pairs of scores. The results are identical with both methods.

An Example From Leisure Studies/Recreation

A graduate student in leisure studies wanted to know the short-term effect of a 4-day bicycle tour on the self-esteem of the participants. To measure self-esteem, the student administered the Cooper-Smith self-esteem survey to 45 bicycle riders immediately before and after the 4-day trip. The results are shown in table 8.3.

The correlation between the pre- and posttests is quite high (.92). Because of this high correlation, the standard error of the difference is very small (.37). This low error value permits the researcher to find significant differences between the two mean values ($df = 44$, $p < .01$). Based on this analysis, the graduate student rejected H_0 and concluded that the 4-day bicycle tour did have an effect on self-esteem. However, note that the mean difference was only 2.3 points (40.7 – 38.4). While this difference did not happen by chance, is it large enough to be meaningful? The next section will provide a method to answer this question.

Table 8.3 Effects of a 4-Day Bicycle Tour on Self-Esteem

Variable	Mean	*SD*	SE_M	SE_D	*r*	*t*
Pretest	38.4	6.1	.90	.37	.92	–6.35
Posttest	40.7	6.3	.94			

The Magnitude of the Difference (Size of Effect)

It is common to report the probability of error (p value) reached by the t ratio. Declaring t to be significant at $p = .05$ or some similar level only indicates the odds that the differences are real and that they did not occur by chance. This is often termed statistical significance. But we must also consider practical significance. If the values of N are large enough, if standard deviations are small enough, and especially if the design is repeated measures, statistically significant differences may be found between means that are quite close together in value. This small but statistically significant difference may not be large enough to be of much use in a practical application. How important is the size of the mean difference?

Thomas and Nelson (2001, p. 139) suggest the use of omega squared (ω^2) to determine the importance, or usefulness, of the mean difference. Omega squared is an estimate of the percentage of the total variance (the difference between the means) that can be explained by the influence of the independent variable (the treatment). For a t test, the formula for omega squared is

$$\omega^2 = \frac{t^2 - 1}{t^2 + N_1 + N_2 - 1}. \tag{8.11}$$

Applying equation 8.11 to the data from the earlier problem comparing male and female hip and low-back flexibility yields

$$\omega^2 = \frac{(-2.4)^2 - 1}{(-2.4)^2 + 10 + 8 - 1} = .21.$$

In this case, 21% of the differences between males and females in hip and low-back flexibility can be attributed to gender. The remaining 79% of the variance is due to individual differences among subjects, other unidentified factors, and errors of measurement.

How large must omega squared be before it is considered important? The answer to that question is not statistically based. Each investigator or consumer of the research must determine the importance of omega squared. In this example, it is meaningful to know that 21% of the variance can be explained. But gender clearly is not the only variable that affects hip and low-back flexibility. Other factors, unidentified in this study, are at work.

Another method of determining the importance of the mean difference is the **effect size** (*ES*), which may be estimated by the ratio of the mean difference over the standard deviation of the control group, or the pooled variance of the treatment groups if there is no control group:

$$ES = \frac{\bar{X}_1 - \bar{X}_2}{SD_{\text{Control}}}. \tag{8.12}$$

The control group is normally used as an estimate of the variance because it has not been contaminated by the treatment effect. In the example of the impact of praise on learning to bowl, the effect size is

$$ES = \frac{145 - 135}{14.9} = .67.$$

Jacob Cohen (1988, p. 21), as quoted in the work of Winer, et al. (1991, p. 122), proposes that *ES* values of .2 represent small differences; .5, moderate differences; and .8+, large differences. Winer, et al. (1991), also suggests that *ES* may be interpreted as a *Z* (standardized) score of mean differences. In the bowling example, an *ES* of .67 indicates that the effect of praise on learning to bowl was moderate.

The amount of improvement from the pretest to the posttest in repeated measures designs can be determined by assessing the percent of change. The following formula will determine the percent of change (improvement) between two repeated measures:

$$\text{Percent improvement} = \left(\frac{\bar{X}_2 - \bar{X}_1}{\bar{X}_1} \right) \times 100, \tag{8.13}$$

where $\bar{X}_1$ and $\bar{X}_2$ represent the pre- and posttest mean values.

In the example of the effects of participation in a 4-day bicycle tour on self-esteem (table 8.3), the pre-post improvement is small—only 6% ([40.7 – 38.4]/38.4 × 100 = 6%). Although the paired *t* value (–6.35) is significant at $p < .01$, the improvement analysis indicates that the change was minimal at best.

Omega squared, effect size, and percent improvement are important attributes to report when mean differences are studied. They provide additional information to the consumers of the research to assist them in determining the usefulness of the conclusions. These values may be more meaningful than the *p* value, especially if *p* just misses being significant (i.e., $p = .06$).

Type I and Type II Errors

Two kinds of errors can be made in accepting and rejecting hypotheses. A **type I error** is committed when the null hypothesis is true but is erroneously rejected (differences are found that in reality do not exist). A **type II error** is made when the null hypothesis is not true but is incorrectly accepted (the research fails to detect differences that really do exist).

Neophyte researchers are sometimes accused of making type I errors in their zeal to find significant differences. But failure to find a difference does not render the research worthless. It is just as important to know that differences do not exist as it is to know that differences do exist. The experienced and competent researcher is honestly seeking the truth, and a correct conclusion from a quality research project is valuable even if the research hypothesis is rejected.

Table 8.4 demonstrates the conditions under which type I and type II errors may be made. The dilemma facing the researcher is that one can never absolutely know which, if either, type of error is being made. The experimental design provides the means to determine the odds of each type of error, but complete assurance is never possible.

The researcher must decide which type of error is the most costly and then protect against it. If concluding that a difference exists when it does not (a type I error) is

Table 8.4 Type I and Type II Errors

	Reality	
	Null hypothesis is true	Null hypothesis is false
Accept null	No error; conclusion is correct.	Type II error; conclusion is incorrect.
Reject null	Type I error; conclusion is incorrect.	No error; conclusion is correct.

likely to risk human life or commit large amounts of resources to a false conclusion, then this is an expensive error and should be avoided. But if differences do exist, and the study fails to find them (a type II error), consumers of the research will never be able to take advantage of knowledge that may be helpful in solving problems.

Setting an appropriate p level to protect against either of the possible errors is critical. When using the null hypothesis, if we set a p level too low ($p = .10$ rather than $p = .05$, for example) and found the resulting t value to be significant at just barely $p = .10$, then we would reject the null hypothesis and conclude that a real difference exists. But in this case the odds are 10 out of 100 that the difference is not real. If there really is no difference between the means, and the resultant t just happens to be that 1-in-10 event that occurs by chance alone, we have committed a type I error.

It is also possible to err in the other direction. For example, if we use the null hypothesis with $p = .01$ and the t value is not high enough to reach the critical ratio in table A.3 (perhaps due to a small N or to measurement or other experimental errors), then we accept the null hypothesis. But if in reality the means do differ, we have committed a type II error.

We can never know absolutely when we have made either of these kinds of errors. Statistical techniques only permit us to make probability statements about the truth. To reduce the probability of type I errors, use a more conservative p value ($p = .01$ instead of $p = .05$). To guard against type II errors, set a more liberal p value. Researchers must make a trade-off decision: Protecting against one type of error increases the probability of committing the other type.

The critical factor in this decision is the consequence of being wrong. The confidence level should be set to protect against the most costly error. We must ask, Which is worse: to accept the null hypothesis when it is really false or to reject it when it is really true?

Possible Causes of Error

Type I	Type II
1. Measurement error	1. Measurement error
2. Lack of random sample	2. Lack of sufficient power (N too small)
3. p value too liberal ($p = .10$)	3. p value too conservative ($p = .01$)
4. Investigator bias	4. Treatment effect not properly applied
5. Improper use of one-tailed test	

Two- and One-Tailed Tests

Most research is done because the results of the experiment are not known beforehand. If the researcher can answer the research question through logical reasoning or a review of related literature, the experiment is not necessary. When review of

all prior research does not yield an answer, the researcher proposes the null hypothesis, H_0, and conducts an experiment to test that hypothesis.

One way to state the null hypothesis is to predict that the difference between two population means is zero ($\mu_1 - \mu_2 = 0$) and that small differences in either direction (plus or minus) on the sample means are considered to be chance occurrences. The direction, or sign, of the difference is not important, because we do not know before we collect data which mean will be larger. We are simply looking for differences in either direction.

Two-Tailed Test

The null hypothesis is tested with a **two-tailed test.** If t does not reach the critical ratio for $p = .05$, then we know that the sample mean difference falls within the area of the normal curve that includes 95% of all possible differences, and we can accept H_0 with only a 5% chance of being wrong. When N values are large ($df > 120$), this corresponds to a Z score of ±1.96 (see figure 8.1). In this case the 5% rejection area (alpha) is divided between the two tails of the curve; each tail includes 2.5% of the area under the curve ($\alpha/2$).

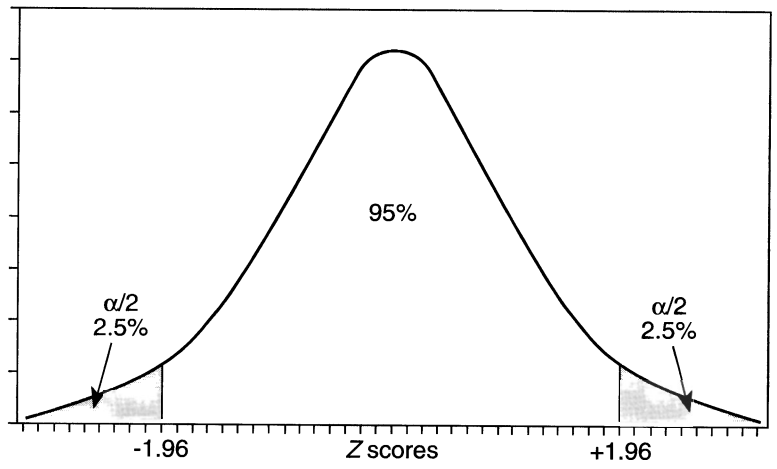

Figure 8.1 Distribution of alpha rejection area for a two-tailed test.

Use the null hypothesis and a two-tailed test when prior research or logical reasoning does not clearly indicate that a significant difference between the mean values should be expected. Use the columns for a two-tailed test in table A.3 in appendix A to determine the t value needed to reject the null hypothesis at the predetermined p value.

One-Tailed Test

Sometimes the review of literature and/or reasoned logic strongly suggests that a difference does exist between two mean values. The researcher is confident that the direction of the difference is well established but is not sure of the size of the difference. In this case, the researcher may test the research hypothesis (H_1), but such a situation is rare. The evidence suggesting the direction of the mean difference must be strong to justify testing H_1. The opinion of the investigator alone is not sufficient.

The researcher predicts that two population means are not equal. By convention, the mean expected to be larger is designated as μ_1. Because the first mean is predicted to be greater than the second, the direction of the difference is established as positive.

Because the difference to be tested is always positive ($\mu_1 > \mu_2$), we are interested in only the positive side of the normal curve. If an observational comparison of sample means shows $\bar{X}_2$ to be larger (even by the slightest amount) or equal to $\bar{X}_1$, the hypothesis that $\mu_1 > \mu_2$ is proven false and must be rejected.

When the direction of the difference is well established before data collection, use the research hypothesis and the one-tailed columns in table A.3 to determine the critical *t* ratio.

The **one-tailed test** places the full 5% of the alpha area representing error at one end of the curve (see figure 8.2). The *Z* score that represents this point (1.65) is lower than the *Z* score for a two-tailed test (1.96).

Table A.3 (we are assuming $df > 120$ for this discussion) shows that the *t* value is 1.65 for $p = .05$ in a one-tailed test (45% of the area under the curve). For $p = .01$ (49% of the area under the curve), the one-tailed *t* value is 2.33. Note that the *t*

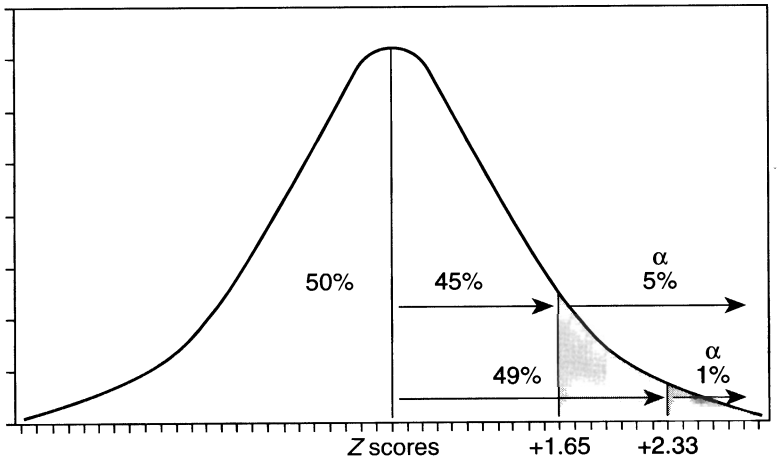

Figure 8.2 Distribution of alpha rejection area for a one-tailed test.

score does not need to be as large in a one-tailed test as in a two-tailed test (the $p = .01$ two-tailed *t* value is 2.58). Therefore, it is easier to find significant differences when a one-tailed test is used. For this reason, the one-tailed test is more powerful, or more likely to find significant differences, than the two-tailed test. But, it also presents a greater risk for a type I error. The next section discusses the power of a test and how it is calculated.

Determining Power and Sample Size

Power is the ability of a test to detect a real effect in a population based on a sample taken from that population. In other words, power is the probability of correctly rejecting the null hypothesis when it is false. Generally speaking, as sample size increases, power will increase. Using the concepts and formulas presented in this section, it is possible to calculate how large *N* must be in a sample to reach a given level of power (i.e. 80% power, or 80% probability that a real effect in the population will be detected). See Tran in *Measurement in Physical Education and Exercise Science*, vol. 1, no.1, 1997, p. 89, for an excellent discussion of the importance of calculating power. Because the critical *t* values are lower for a one-tailed test than for a two-tailed test (see table A.3), the one-tailed test is considered more powerful at a given *p* value. Also, a value of $p = .05$ is more powerful than $p = .01$ because the *t* value does not need to be as large to reach the critical ratio.

In figure 8.3, the range of the control group represents all the possible values for the mean of the population from which a random sample was taken. The most likely value is at the center of the curve, and the least likely values are at the extremes. The range of the experimental group represents all possible values of the population from which it was taken after treatment has been applied. The alpha point (Z_α) on the control curve is the point at which a null hypothesis is rejected for a given mean value in the experimental group.

Any value for the mean of the experimental group that lies to the right of Z_α ($1 - \beta$ area) will be judged significantly different from the mean of control at $p < .05$ (i.e., not taken from the same population). If H_0 is really true, this represents a type I error. Conversely, any value for the mean of the experimental group that lies to the left of Z_α (the beta area) will be judged to be not significantly different from the mean of control group. If H_0 is false, this represents a type II error.

It then follows that the area of $1 - \beta$ in the experimental group is the area of power, the area where a false null hypothesis will be correctly rejected. This area represents all of the possible values for the mean of the experimental population that fall beyond the Z_α level of the control population. Power is calculated by determining Z_β, converting it to a percentile using table A.1 in appendix A, and adding this percent of area to the 50% of the curve to the right of the experimental mean.

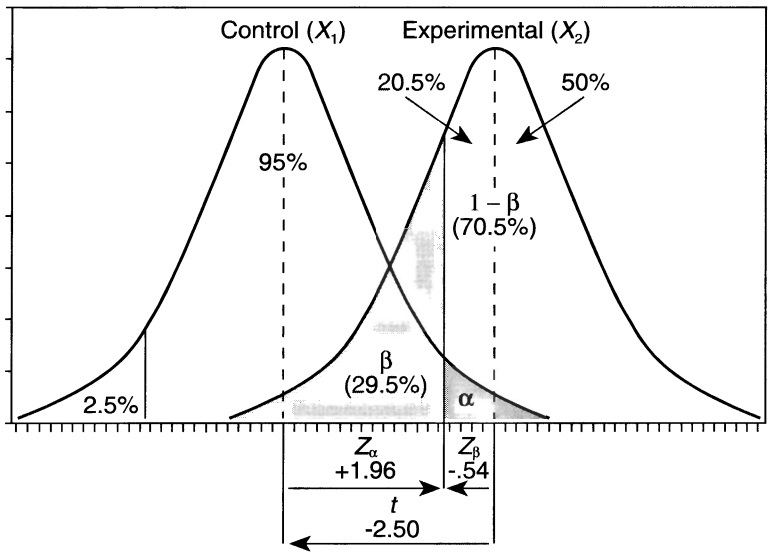

Figure 8.3 Calculation of power for an independent *t* test.

In figure 8.3, 1 – β represents 70.5% of all the possible mean values for the experimental group. So there is a 70.5% chance that a false null hypothesis will be rejected; power = 70.5%. Let's consider how power is calculated.

As figure 8.3 shows, power is dependent on four factors:

1. The Z_α level set by the researcher (the level set to protect against type I errors; $p = .05$, $p = .01$, etc.). It is represented by a Z_α score from the normal distribution table [(A.1), (Z_α (.10) = 1.65, Z_α (.05) = 1.96, Z_α (.01) = 2.58)].
2. The difference, Δ (Δ = $\bar{X}_1 - \bar{X}_2$), between the two mean values being compared (where $\bar{X}_1$ is the mean of the control group and $\bar{X}_2$ is the mean of the experimental group).
3. The standard deviations of the two groups (*SD*), which determine the spread of the curves.
4. The sample size, *N*, of each of the two groups.

Only *N* and Z_α are under the control of the researcher, and Z_α usually cannot be radically manipulated because of the need to protect against type I errors. Therefore, the researcher can control power primarily by manipulating the size of *N*.

Calculating Power

The following process is used to calculate power—that is, to determine the 1 – β area in figure 8.3.

The researcher sets Z_α, the *p* value set to reject the null hypothesis. The values for the means, standard deviations, *N* for each group, the standard error of the difference, and *t* are calculated. Figure 8.3 demonstrates that *t* is the sum of Z_α and Z_β. To determine the power area of the experimental curve, we must find the value of Z_β, which is the percent of the area on the experimental curve between $\bar{X}_2$ and Z_β.

In figure 8.3, Z_α is a positive value; it proceeds to the right of the control mean ($\bar{X}_1$). The *t* value is negative because $\bar{X}_1 < \bar{X}_2$. The Z_β value is also negative; it proceeds to the left of the experimental mean X_2. If the analysis were made with $\bar{X}_1 > \bar{X}_2$ then *t* and Z_β would be positive values, and Z_α would be negative. In order for the following formulas to be applied toward either tail of the curve, the values of *t*, Z_α, and Z_β will all be considered absolute. Then *t* is equal to the sum of Z_α and Z_β:

$$t = Z_\alpha + Z_\beta. \tag{8.14}$$

Conversely,

$$Z_\beta = t - Z_\alpha. \tag{8.15}$$

Let us assume the following data apply to figure 8.3:

$\bar{X}_1 = 30$, $\bar{X}_2 = 32.5$ (Δ = 2.5)
$SD = 5$ for each group
$N = 50$ for each group
$SE_M = .71$ for each group, ($5 / \sqrt{50} = .71$)
$SE_D = 1.00$, ($\sqrt{.71^2 + .71^2} = 1.00$)
$t = 2.50$, (2.5 / 1.0 = 2.5)
$Z_\alpha = 1.96$ ($p = .05$)

Equation 8.15 may now be used to determine Z_β:

$$Z_\beta = 2.50 - 1.96 = .54.$$

We convert the Z_β value of .54 to a percentile from table A.1 ($Z_\beta = .54 = 20.5\%$ of the area under the normal curve) and compute the area of 1 – β (20.5% + 50% = 70.5%). Therefore 70.5% of all possible values of the experimental population mean lie to the right of Z_α. In other words, there is a 70.5% chance of rejecting the null hypothesis if the values given in the data section above are true; power = 70.5%.

Calculating Sample Size

The only factor in these equations that is easily manipulated by the researcher is *N*. We could increase our power by increasing *N*, but how large does *N* need to be to produce a given power level?

Equation 8.14 may be solved for N as follows. The formula for t is

$$t = Z_\alpha + Z_\beta.$$

Because $t = \Delta / SE_D$, we may substitute Δ / SE_D for t.

$$\frac{\Delta}{SE_D} = Z_\alpha + Z_\beta.$$

Recall that for $N_1 = N_2$ (equation 8.05),

$$SE_D = \sqrt{(SE_{M1})^2 + (SE_{M2})^2}$$

and (equation 8.02)

$$SE_M = \frac{SD}{\sqrt{N}}$$

When $N_1 = N_2$ and $SD_1 = SD_2$:

$$SE_D = \sqrt{\left(\frac{SD_1}{\sqrt{N_1}}\right)^2 + \left(\frac{SD_2}{\sqrt{N_2}}\right)^2} = \sqrt{\frac{2SD^2}{N}}.$$

Substituting this value for SE_D, we obtain the following:

$$\frac{\Delta}{\sqrt{\frac{2SD^2}{N}}} = Z_\alpha + Z_\beta$$

$$\Delta = \left(\sqrt{\frac{2SD^2}{N}}\right)(Z_\alpha + Z_\beta)$$

$$\Delta^2 = \left(\frac{2SD^2}{N}\right)(Z_\alpha + Z_\beta)^2$$

$$N\Delta^2 = 2SD^2 (Z_\alpha + Z_\beta)^2$$

Therefore,

$$N = \frac{2SD^2 (Z_\alpha + Z_\beta)^2}{\Delta^2}. \qquad (8.16)$$

With equation 8.16, we can determine the N needed for a given power level if we know the other values. Suppose we want power = .80 at $p = .05$. Then the area of Z_β must be 30% and $1 - \beta$ is 80%. We look up 30% in table A.1 to find that $Z_\beta = .84$. If $\Delta = 5$, and for each group $SD = 6$, and we set Z_α at 1.96, then

$$N = \frac{2(6)^2 (1.96 + .84)^2}{5^2} = 22.6.$$

We conclude that to achieve 80% power under these conditions, we must use a sample size of approximately 23 in each group.

The calculation of power is a major factor in experimental design. It is important to know what the odds are that real differences between group means may be detected before we conduct expensive and time-consuming research. Research performed with insufficient power (i.e., N is too small) may result in a type II error (failure to reject a false null hypothesis) or may waste valuable resources on a study that has little chance of rejecting the null.

In a power calculation, the values for the means and standard deviations are not usually known beforehand. To calculate power before the data are collected, these values must be estimated from pilot data or from prior research on similar subjects.

The previous power calculation example is applicable only to a t test of independent means, with equal values of both N and SD. This is the most simple application of the concept of power. Similar calculations may be made for unequal values of N or for dependent tests.

A software program titled *PC Size* is described by Dallal (1986). It can be used to perform power calculations for simple research designs. Additional discussions of power may be found in Kachigan (1986, p. 185), and Thomas and Nelson (2001, p. 108-110).

The *t* Test for Proportions

The techniques for estimating error in sample means and determining the significance of the difference between two means may be modified to apply to proportions. The following example illustrates this procedure.

An Example From Administration

A teacher surveyed 150 girls in a large school to determine their favorite subject and found that 60% chose physical education. The principal doubted these findings and claimed that the true population value of those favoring physical education couldn't be more than 50%. He challenged the 60% figure and asked the teacher to prove it.

If we assume that the teacher's survey was properly conducted and that subjects were randomly drawn from the population, what are the odds that another random survey from the same population could result in a value of 50%? To answer such a question, we need to know the error that can be expected in a proportion. This error can be estimated with the following formula for the standard error of a proportion,

$$SE_p = \sqrt{\frac{pq}{N}}, \tag{8.17}$$

where p is the obtained proportion, $q = 1 - p$, and N = sample size. Applying this formula to the problem at hand yields

$$SE_p = \sqrt{\frac{.60\ (1-.60)}{150}} = .04.$$

The standard error of a proportion (SE_p) may be interpreted as a Z score. Therefore, the odds that the true proportion lies between .56 and .64 ($.60 \pm .04$) are 68 to 32. Multiplying SE_p by 1.65 ($p = .10$), 1.96 ($p = .05$), or 2.58 ($p = .01$) will produce the limits of the population at a given level of confidence. For example, at $p = .05$, $.60 \pm (1.96).04 = .60 \pm .078$ indicating that the odds that the true population mean lies between .522 and .678 are 95 to 5.

Based on this analysis, the principal agreed that the true proportion for the population was probably not 50%.

This concept may be applied to a t test between two proportions (Bruning & Kintz, 1977, p. 222). The formula for a t test between proportions, t_p, is

$$t_p = \frac{P_1 - P_2}{\sqrt{\frac{p(1-p)}{N_1} + \frac{p(1-p)}{N_2}}}, \tag{8.18}$$

where P_1 and P_2 are the proportions to be compared, and p under the radical is

$$p = \frac{N_1 P_1 + N_2 P_2}{N_1 + N_2}. \tag{8.19}$$

The t test for proportions should not be used when either p or q times N is less than 5. Under these conditions, use nonparametric statistics (see Witte, 1985, p. 155).

An Example Comparing Two Proportions

If 60% of 150 girls and 70% of 125 boys chose physical education as their favorite subject, is there a significant difference between the girls and the boys at $p = .05$? To answer this question, we calculate

$$p = \frac{(150)(.60) + (125)(.70)}{150 + 125} = .65$$

and

$$t_p = \frac{.60 - .70}{\sqrt{\frac{.65\,(1-.65)}{150} + \frac{.65\,(1-.65)}{125}}} = 1.74.$$

Then we look in table A.3 to interpret t_p for $df = N_1 + N_2 - 2$. The t value of 1.74 is not large enough to reject chance at $p = .05$ because it does not reach the critical ratio of 1.96 for $df = 273$. There is no significant difference between the proportion chosen by the girls and the proportion chosen by the boys.

Summary

It is very unlikely that means of two random samples from the same population will be identical. Differences will almost always be observed. This is not unexpected; people do not always perform exactly the same, and even if they did, the measurement of their performance is not perfect. Because of these random errors, we always expect mean values to differ. The question is, how much can they differ before we suspect that the difference is caused by something other than chance?

As we discussed earlier in this chapter, the purpose of a t test is to determine whether the difference between two sample mean values is large enough to reject chance as a probable cause. Using the concepts of the normal curve, we can determine the amount of difference between any two sample means that can be attributed to chance alone. If the observed difference is larger than this estimated difference, then we reject chance as a cause and look for another reason to explain the mean difference.

The t test is the technique by which we perform this analysis. The t value is simply the ratio of the observed difference (the numerator) to the expected difference (the denominator). If the observed difference is larger than the expected difference, the t ratio can be compared to the critical ratios in table A.3 to determine the probability that the observed difference occurred by chance. The table values, or critical ratios, are the values of t for selected sample sizes, or degrees of freedom, that would be expected to occur by pure chance. When our obtained t exceeds these values, we reject the null hypothesis (H_0), accept H_1 and declare the differences to be significant (i.e., not caused by chance).

The t test is useful for conducting experimental research. If we want to know the effect of some treatment, we compare a group that has had the treatment with one that has not. If the treatment is ineffective, we expect only chance differences between the groups. If the treatment is effective, the differences will exceed the expected difference. The t test may be modified to make comparisons between observed proportions.

Following is a list of essential steps that need to be completed to properly determine the significance of the difference between two population means using randomly selected sample groups.

1. Define the population of interest.
2. State the problem.
3. Review the literature. Determine if the problem is solved by prior research.
4. If the problem is not solved, state an hypothesis (H_0 = null, H_1 = directional).
5. Select a level of confidence (consider consequences of error).
6. Select a power level and determine appropriate sample size.
7. Randomly select two samples from the population. Treat one sample with the independent variable. Provide appropriate controls for the other sample.
8. Compute mean, standard deviation ($N - 1$), and standard error of the mean for each sample.
9. Compute standard error of the difference and t.
10. Determine degrees of freedom.
11. Compare obtained t to critical values in a table of t or read p value from a computer.
12. Make a conclusion by accepting or rejecting the hypothesis at the given level of confidence.
13. Determine the practical importance of the conclusion by calculating the size of the effect.

Problems to Solve

1. Calculate the standard error of the difference for each of the following sets of independent data.

	$\bar{X}_1$	$\bar{X}_2$	N_1	N_2	SD_1	SD_2
A.	172	175	50	50	20	18
B.	9.7	7.0	10	15	2.5	3.1

2. What are the t values for problems 1A and 1B? Is either significant? At what level of confidence?
3. Give a verbal definition of the meaning of the standard error of the difference.
4. In a study of absolute errors in active versus passive arm positioning, an investigator collected data in centimeters on 20 college-age subjects (from the motor learning laboratory, California State University Northridge, courtesy of Tami Abourezk). Is there a significant difference in the errors made by the active group ($N_1 = 10$) versus the passive group ($N_2 = 10$) on arm positioning?

Subject	Active	Subject	Passive
1	2.65	11	3.30
2	2.42	12	2.00
3	3.30	13	0.09
4	0.19	14	0.04
5	1.25	15	4.56
6	2.00	16	3.33
7	3.34	17	1.02
8	4.08	18	0.89
9	0.70	19	2.78
10	2.89	20	1.65

A. Compute the t value and the p value by hand, then accept or reject the null hypothesis.

B. Confirm your results on a computer. How does the computer output compare with your hand calculated results?
Hint: In the computer database, create two columns. In column 1 (the grouping variable) enter the number 1 in rows 1 to 10, and the number 2 in rows 11 to 20. In column 2 (the score column), enter the absolute error values for subjects 1–10 in rows 1–10, and the values for subjects 11–20 in rows 11–20.

5. A graduate student in biomechanics was interested in stride length of cross-country skiers. Stride length is an important factor in the development of speed for racing. Data were collected on 20 athletes from the cross-country ski team (the experimental group). The data were compared to those of a second group of 17 students (the control group) who were not varsity athletes but did participate in recreational skiing. The researcher assumed that no significant differences would be found. Is there a significant difference between the athletes and the nonathletes in stride length? What is the level of confidence? What is ω^2, and what is the effect size? (Adapted from Duoos, 1984, data fabricated).

Athletes	Nonathletes
$\bar{X}_1 = .90$ meters	$\bar{X}_2 = .70$ meters
$N_1 = 20$	$N_2 = 17$
$SD_1 = .17$	$SD_2 = .23$

6. Find the t value for the following data on dominant versus nondominant grip strength. Twenty subjects were measured twice, once with the dominant hand and once again with the nondominant hand. Which hypothesis, H_0 or H_1, might be appropriate in this study? Is the difference significant, and if so, at what level of confidence?

Dominant		Nondominant
$\bar{X}_1 = 40$ pounds		$\bar{X}_2 = 35$ pounds
$SD_1 = 12.7$	$r = .83$	$SD_2 = 14.2$

7. A researcher in exercise physiology wanted to know if body composition differs among prepubescent males and females. To test the null hypothesis, she measured skinfolds in millimeters on 5 males and 6 females with the following results.

Males	Females
21	22
25	19
19	18
17	24
18	21
	23

   A. Is the difference in the means significant? If so, at what level of confidence?
   B. Do you accept or reject the null hypothesis?
   C. Check your hand calculations on a computer. See hint in problem #4.

8. A sport psychologist wondered if motivation could improve aerobic capacity. To answer the question, he measured 10 subjects on $\dot{V}O_2$max (ml/kg/min) on a treadmill. Students were instructed verbally to "do your very best." One week later, the subjects were measured again but this time they were told they would receive $100 if their $\dot{V}O_2$max was higher than the first time. Following are the data.

First	Second
45	54
33	50
59	58
32	38
30	42
27	35
29	38
59	66
44	48
40	49

   A. Did the money cause them to significantly raise their $\dot{V}O_2$max values?
   B. If so, what is the probability of error in your conclusion? Should you accept or reject the null hypothesis?
   C. Check your hand calculations on a computer.
      Hint: In the computer database, enter values for each subject on the same row. Use column 1 for the first test, and column 2 for the second test.

9. An aerobic dance teacher wanted to know if two workouts per week of 30 minutes each were enough to increase $\dot{V}O_2$max in sedentary middle-aged women. The teacher proposed to compare $\dot{V}O_2$max in some women who had been in the aerobic dance classes for 6 months to another group of sedentary women. Based on related literature, the teacher estimated that $\dot{V}O_2$max would be about 7 milliliters per kilogram per minute higher in the aerobic dancers, with a standard deviation of 4.5. If this is a fair estimate of the data to be expected, how many women must she test in each group to produce a power coefficient of .90 at $p = .05$?

See appendix C for answers to problems.

Key Words

Actual mean difference
Critical ratio
Dependent sample
Effect size
Expected mean difference
Homogenity of variance
Independent sample
One-tailed test
Power
Robust
Standard error of the difference
Two-tailed test
Type I error
Type II error